Multiple Routes to Subfunctionalization and Gene Duplicate Specialization

Stephen R. Proulx¹

Ecology, Evolution, and Marine Biology Department, University of California, Santa Barbara, California 93106-9620

ABSTRACT Gene duplication is arguably the most significant source of new functional genetic material. A better understanding of the processes that lead to the stable incorporation of gene duplications into the genome is important both because it relates to interspecific differences in genome composition and because it can shed light on why some classes of gene are more prone to duplication than others. Typically, models of gene duplication consider the periods before duplication, during the spread and fixation of a new duplicate, and following duplication as distinct phases without a common underlying selective environment. I consider a scenario where a gene that is initially expressed in multiple contexts can undergo mutations that alter its expression profile or its functional coding sequence. The selective regime that acts on the functional output of the allele copies carried by an individual is constant. If there is a potential selective benefit to having different coding sequences expressed in each context, then, regardless of the constraints on functional variation at the single-locus gene, the waiting time until a gene duplication is incorporated goes down as population size increases.

GENE duplication has long been viewed as a mechanism that promotes diversification of functional genes in the genome (Ohno 1970; Taylor and Raes 2004; Conant and Wolfe 2008). Simply stated, this view holds that once a gene locus has been duplicated, the pair of loci can go their own way without decreasing fitness. Whether the newly formed and subsequently diverged loci are then maintained over longer evolutionary periods obviously depends on the fitness costs of deleting one of the loci. While models of duplication universally agree on this point (Innan and Kondrashov 2010), they differ in their view of how the duplication itself originally spreads and how the loci then diverge. Here I develop models of the gene duplication process that consider the functional effects of mutations in coding or regulatory regions with consistent selection acting on variation before, during, and following duplication. By doing this I am able to compare the total waiting time for a gene to go from a single copy to a pair of stably maintained duplicates. I find that the waiting time until duplication depends on whether the net effect of selection on coding variants at the single-

copy gene is stabilizing, the magnitude of the potential fitness gain from having diverged duplicate genes, and the population size. Regardless of the specific assumptions on the fitness effects of coding and regulatory mutations, I find that there typically are multiple routes to duplication that are driven by selection and therefore speed up as population size increases.

Previous models apply selection inconsistently

Three main modes for the adaptive maintenance of duplications have been proposed: neofunctionalization, subfunctionalization, and the divergence of multifunctional genes. The theoretical framework proposed for each of these modes suffers from a deep logical flaw: the effect of selection is applied when it is convenient to support the conclusion of the proposed mechanism.

Under the neofunctionalization model, the duplication becomes fixed by drift and one of the duplicate loci takes on a completely new function while the other maintains the old one (Force *et al.* 1999). This view assumes that somehow evolution is frozen before the random fixation of a duplication. A mutation that appeared in one of the alleles at the single-copy ancestor is just as likely to provide some incremental ability to perform a “new function” as a mutation in one of the alleles of a duplicate pair of loci. More generally, the allelic state (or states) of a single-copy gene is subject to

Copyright © 2012 by the Genetics Society of America

doi: 10.1534/genetics.111.135590

Manuscript received October 7, 2011; accepted for publication November 29, 2011

Supporting information is available online at <http://www.genetics.org/content/suppl/2011/12/05/genetics.111.135590.DC1>.

¹Address for correspondence: Ecology, Evolution, and Marine Biology Department, University of California, Santa Barbara, CA 93106-9610. E-mail: proulx@lifesci.ucsb.edu

the same selection pressures that operate at a pair of duplicate loci. If a change in the environmental circumstances of the species is posited, it would be very unlikely that this shift would happen at precisely the same time as the fixation of a duplication by drift. The set of mutationally accessible alleles determines the opportunity for neofunctionalization; it is not the fixation of a duplication that creates opportunities for beneficial mutations. Models of neofunctionalization artificially restrict the question of when mutation and selection are able to discover new functions to the postduplication phase (but see Walsh 2003).

Under the divergence of multifunctional genes model, a single locus is fixed for an allele that has multiple functions and duplication provides an opportunity for each locus to optimize a single function (Hughes 2005). The main proponents of this mode of duplication (Hughes 2005; Des Marais and Rausher 2008) have not relied on formal mathematical models (Innan and Kondrashov 2010), making it more difficult to draw conclusions about the generality and tempo of this process. This framework involves the assumption that the single-copy locus is fixed for an allele coding for a multifunctional protein that does not perform either of its functions optimally. If a duplication becomes fixed, then the two loci may diverge so that each locus is fixed for an allele that codes for a protein optimized on just one function. This framework again assumes that during the preduplication phase, evolution is frozen and no genetic variation is possible. In effect, it argues that duplications diverge because of multifunctional proteins but kicks back the question of how multifunctional genes arise. The problem with this view is that the evolution of multifunctionality and the divergence of multifunctional duplicate genes both depend on the relationship between the multiallelic genotype and fitness (e.g., Proulx and Phillips 2006). In particular, the fitness effects of allelic variation at a single-copy gene determine whether multifunctionality will evolve, and a subset of the conditions that promote multifunctionality also promotes the divergence of genes following duplication. The critical finding of Proulx and Phillips (2006) was that most parameter values that promote divergence of duplicate loci actually promote allelic divergence and the evolution of heterozygote advantage at the single-locus gene (*i.e.*, before a duplication becomes fixed in the population).

Under the subfunctionalization model [and the duplication-degeneration-complementation (DDC) model in particular], duplications are fixed by drift followed by the stochastic fixation of loss of subfunction mutations in each of the duplicate loci (Force *et al.* 1999, 2005; Lynch and Force 2000; Lynch *et al.* 2001; Walsh 2003). The starting point for these models is that existing loci have multiple functions, that duplication itself does not alter fitness, and that these functions can be partitioned into distinct subfunctions without any loss (or gain) of fitness. This assumption is often operationalized by considering genes with multiple *cis*-regulatory binding sites and assuming that the number of alleles expressed in a specific regulatory context has no effect on function (and therefore

fitness). The particularity of this assumption is rarely discussed. Under these assumptions, alleles that have mutationally lost a subfunction can drift to fixation at either of the duplicate loci. The time for this to occur is quite sensitive to the assumptions, in particular that subfunctions are completely separable. If mutations that destroy one subfunction are even slightly disadvantageous as homozygotes, say because they also cause a change in the speed or variability of transcription initiation in the other context, then the waiting time until such mutations fix increases rapidly with population size. This mutation rate asymmetry can increase the waiting times well above those previously noted.

Further, the subfunctionalization framework assumes that multifunctionality arose in the distant past and that the forces selecting for multifunctionality have little to do with the postduplication process. Preduplication, however, selection must be acting on both the regulatory and coding regions of alleles. Whatever the physiological, developmental, or genetic factors are that determine preduplication evolution will determine how postduplication mutations in coding and regulatory alleles affect fitness. Even though the DDC model posits that duplications become stable because of a series of neutral substitutions, the postduplication evolution of the two loci will still be affected by both coding and further regulatory mutations. The subfunctionalization framework ignores the evolutionary process that makes subfunctionalization possible, but the preduplication process sets the conditions that allow (or do not allow) subfunctionalization to proceed. The subfunctionalization model can be taken seriously only if it is shown that the evolutionary processes that precede duplication tend to produce alleles whose mutational neighborhood fits the assumptions of the DDC model. The point is that evolution preceding duplication will determine whether mutations that delete transcription factor (TF) binding sites behave neutrally or do not.

Developing a consistent model for the entire duplication process

The overarching theme of this article is that evolution both before and after duplications arise is governed by the same biochemical, physiological, and developmental effects of changes in genotype. The same mechanisms that cause fitness to depend on the two alleles an individual carries at a single locus also act on individuals that carry four alleles. If changes in dosage affect fitness, then so too must changes in expression at a single locus. If a mutation altering one allele in a set of four affects fitness, so too will a mutation altering one allele in a set of two. Taking the simplifying view that the process of duplication can be broken up into independent phases is not just a benign approximation because it generates predictions that are qualitatively different from those made by a consistent model.

Our previous work considered changes in allele function related by a trade-off between performance in two distinct contexts. We showed that the conditions that favor the evolution of multifunctional genes, a necessary precursor to

the multifunctional gene model, can lead to divergence of loci following duplication. Our results also showed that the same conditions that allow multifunctional genes to diverge postduplication are likely to cause divergence of alleles preduplication (termed allelic divergence) followed by the selectively driven spread of a duplication (Proulx and Phillips 2006). These results also demonstrate that allelic divergence and duplication can happen on much shorter timescales relative to the timing of subfunctionalization for all but the smallest population sizes. In this article, I extend this analysis to a scenario where both *cis*-regulatory and coding mutations occur.

Previous works by Force, Lynch, and co-workers (Force *et al.* 1999, 2005; Lynch and Force 2000; Lynch *et al.* 2001) have typically considered mutations that knock out regulatory regions and lead to loss of expression in one or more situations. These studies have followed the duplication and subfunctionalization process and allowed for future change in coding regions that could specialize the new duplications that have had their expression patterns subdivided. One argument for this rationale is that mutations removing binding sites for TFs are expected to be more common than mutations that cause conditionally advantageous change in coding sequence. However, subfunctionalization models work because duplications are able to fix, essentially by drift, in smaller populations. This process can require enormous amounts of time because it will be limited either by the waiting time to the appearance of a duplication that is destined to become fixed ($1/\text{duplication rate}$) or by the time that it takes for a duplication to spread when it is destined to fix ($\sim 4N_e$ generations in diploids). If either of these waiting times is large, then the total time waiting for a duplication to fix via drift will be large.

What happens to populations during such long waiting times? Even if the rate at which adaptive mutations arise is relatively low, their spread through populations will be much quicker than fixation via drift. In this article I explore the waiting times for a set of alternative pathways that eventually lead to a population fixed for duplicate genes that have diverged both in regulatory and in coding sequence.

In this model there are two contexts and the focal gene can have promoter sites that induce expression in each context. The coding region of the gene may also experience mutation that can improve performance in one context while reducing performance in the other context. The structure of the fitness landscape determines whether mutations that alter the coding region can spread in the ancestral population, leading to the evolution of heterozygote advantage followed by the spread and fixation of gene duplicates. When the effect of altering the coding region alone is a net reduction in fitness, this direct pathway to duplication is selectively disfavored. However, an alternative pathway involving first the loss of expression in one context and then a mutation in the coding region is possible through a form of stochastic tunneling. Stochastic tunneling occurs

when a segregating mutation gives rise to a beneficial secondary mutation that then fixes (Iwasa *et al.* 2004; Weissman *et al.* 2009; Proulx 2011). In addition, duplication events that result in alleles missing some fraction of the *cis*-regulatory region are circum-neutral and can therefore drift to high frequencies [genotypes are considered circum-neutral when all differences in their population genetic dynamics can be attributed to their genetic context rather than to direct effects on reproductive output (Proulx and Adler 2010)]. Coding mutations are then directly advantageous and rapidly spread in the population. I then compare the waiting times for the different possible pathways toward duplication to determine how the fitness landscape, mutational parameters, and population size determine the rate at which duplications are incorporated into genomes.

Model Framework

Available mutations

Many eukaryotic genes are regulated to be expressed in multiple contexts. In this article I consider two kinds of mutations, regulatory and coding. Mutations that alter the *cis*-regulatory sequence can cause the allele to be expressed only in one context, often called subfunctionalization (Force *et al.* 1999). I refer to these as subfunctionalizing mutations and use the subscripts s_i to denote an allele that is expressed only in context i .

Mutations in the protein-coding sequence that improve the function of the protein in context i are indicated by c_i (note that I explain below why such mutations are expected to degrade the function of the protein in the other context). We can also refer to alleles with two mutations in a similar way, such as $s_i c_j$. An allele that is expressed only in context i and has a coding mutation that improves function in context i is indicated with the shorthand sc_i (in other words, sc_i is the same as $s_1 c_1$). I refer to such alleles as *subspecialized*. Haplotypes with a duplicate copy of the gene are indicated with a separator|. For example, a haplotype that has one copy of the ancestral allele and one copy of a subspecialized allele would be $A|sc_1$. Alleles that are expressed only in the context in which they are deleterious are expected to be rapidly lost from the population and I do not track their frequencies in the analytical analysis (but they are included in the stochastic simulations). Figure 1 shows a schematic diagram of the mutational network, limited to alleles that are not unconditionally deleterious.

I consider scenarios where several coding mutations and regulatory mutations are accessible from the ancestral allele. Of course, the ancestral allele is just the most recently fixed allele in the population. Levins' notion of a fitness set is particularly useful for describing the series of substitutions that can lead to our ancestral allele. On the basis of arguments developed in supporting information, [File S1](#), I assume that mutations from the A allele exhibit antagonistic pleiotropy, because mutations that increase fitness in one

Figure 1 Schematic diagram of the mutational network. The ancestral genotype is in the middle and labeled A. The ancestral allele can be mutated by losing a TF binding site (alleles s_1 and s_2) or by a change in coding region that causes the protein to be more favorable in one context (alleles c_1 and c_2). These mutants can further mutate to produce alleles that are expressed in a single context and specialized to that context (alleles sc_1 and sc_2). Not shown are mutations that cause complete loss of expression and mutations that produce a mismatch between expression and coding sequence. Any allele can be duplicated, and this happens with probability μ_d . There are 144 alleles that involve duplications so they are not all shown. The fading arrows indicate linkages to the portions of the mutational network that are not drawn.

context decrease fitness in the other context (see File S1, section 1).

The parameter space can be divided into two regions on the basis of the fitness effects of mutations that affect the coding sequence:

1. *Allelic divergence*: Specialized alleles can invade when rare and reach a deterministic equilibrium frequency. Even though there is still antagonistic pleiotropy, mutations near A are at a net advantage when heterozygous. Coding mutations near A are then maintained in the population and can create direct selection favoring duplications. This scenario additionally leads to selection to alter the regulatory region to create subspecialized alleles.
2. *Net stabilizing coding selection*: There is antagonistic pleiotropy that causes mutants that increase fitness in one context to be at a net disadvantage both as heterozygotes and as homozygotes. In this scenario, mutations that silence expression in one context act as recessive lethal (or recessive sick) mutations and can be stochastically maintained at appreciable frequencies. Secondary mutations produce alleles that are expressed only in the context to which their coding region is adapted. These alleles are actively maintained by selection and open the door to complementary mutations that specialize in the other contexts. Duplications are then directly advantageous and can spread due to selection.

Fitness model

In this section, I define a mechanistic model of fitness that allows dominance and epistasis to emerge without adding a large number of parameters. I follow the assumption that only the relative amount of expressed protein determines fitness (Proulx and Phillips 2006).

Context-specific fitness is assumed to be a function of the number and type of proteins expressed in each context. If only the ancestral protein is expressed, then context-specific fitness is assigned to be 1. Instead of assigning pairwise and three-way dominance, I assume that the ancestral protein provides an impulse to keep tissue-specific fitness at 1 that is scaled by a coefficient h (similar to dominance). Each specialized allele that is expressed in a given context provides either a positive or a negative impulse on fitness. This results in a model that describes interactions among nine allele states using only five parameters. The parameters describe the context-specific fitness of each protein-coding state (two coding states in two contexts giving four parameters) and the degree of dominance of the ancestral coding state. For simplicity I assume that fitness is 0 if no protein is expressed in either context and that there is no epistasis between contexts.

Using this framework, context-specific fitness is given by

$$\Phi_{\kappa} = 1 + \left(\frac{\sum_{i=1}^2 w_{i,\kappa} E_{i,\kappa}}{\sum_{j=1}^2 E_{j,\kappa}} \right) 2(1-h) \left(\frac{\sum_{j=1}^2 E_{j,\kappa}}{\sum_{j=0}^2 E_{j,\kappa}} \right), \quad (1)$$

where $i = 0$ represents the ancestral coding sequence, $w_{i,\kappa}$ represents the fitness component for protein i in context κ , $E_{i,\kappa}$ represents the number of expressed alleles that code for protein i in context κ , and h relates to the dominance of the ancestral protein state. If $h = 1$, then the ancestral sequence is fully dominant, but if $h = \frac{1}{2}$, then the ancestral coding sequence is codominant. This formulation is fairly flexible and can smoothly move between the conditions assumed in the standard DDC model to conditions where selection acts on coding changes. Because there is no epistasis, total fitness is simply $\Phi_1 \Phi_2$. I write total fitness, W , as a function of the set of alleles that an individual carries. For example, $W(A, c_1)$ represents the fitness of an individual with one ancestral allele and one coding mutant allele.

Calculating approximate waiting times

I assume that ancestral populations are fixed for the *A* allele that is expressed in both contexts. The evolutionary process allows for mutations to both the coding and the regulatory regions, as well as knockout mutations that irrevocably silence the allele. For simplicity, I assume that each allele has the same knockout mutation rate.

Throughout this article I write the total number of haploid genomes as N and assume that $N_e \approx N$. When $N\mu \ll 1$ [the weak mutation assumption of Gillespie's strong selection–weak mutation model (Gillespie 1991)], then the population is well described by the nonstochastic population genetic equilibria most of the time but occasionally transitions between states following the successful introduction of a new mutation. That is to say, without frequency-dependent selection we expect most populations to be monomorphic and with frequency dependence we expect the population to be near the frequency-dependent equilibrium. The population can change state if a mutation arises, is not lost when rare, and is deterministically maintained in the population. However, stochastic fluctuations in allele frequency are considered during the invasion of a new haplotype. This modeling framework is related to Gillespie's strong selection–weak mutation formalism (Gillespie 1991) but makes allowances for situations with weak or frequency-dependent selection. Much inspiration was drawn from Hammerstein's (1996) streetcar approach.

The steps that go into calculating the waiting time for each evolutionary transition are presented in more detail in [File S1](#). Under the assumption that $N\mu < \frac{1}{2}$ the waiting time for a mutation that is favored when rare is simply

$$T \approx \frac{1}{N\mu} \frac{1}{2s},$$

where s is the difference between the relative fitness of the mutant and 1. I ignore the time required to approach population genetic equilibrium for alleles under selection because it is usually orders of magnitude smaller than the waiting time for the appearance of a successful mutation.

Simulation framework

I simulated the full evolutionary process to observe evolutionary trajectories and to compare the waiting times until duplications become resident. The simulation was performed using Mathematica (code available, see [File S2](#)). I assumed constant population size where regulation occurred by exact culling of juveniles so that the number of adults is constant. The order of events was mating → selection → recombination → mutation → culling. The simulation was streamlined by tracking counts of haplotypes in the gamete stage and by calculating the total probability that each adult in the next generation would have a particular genotype. The distribution of haplotypes that contribute to the next generation is a composite of selection, mutation, and recombination and is expected to be multinomial dis-

tributed (Proulx 2000). By first calculating the multinomial coefficients the number of random variables drawn could be kept low so that simulations of large populations could still be performed in reasonable amounts of time.

Evolutionary Trajectories of Duplication

I analyze four different scenarios on the basis of fitness landscapes and the types of duplicating mutations considered. For each, I calculate the expected waiting time until a duplication is stably maintained and compare the results to stochastic simulations.

No coding selection

This scenario reflects the classic DDC assumption that there is no genetic variation for context-specific adaptation of the coding region. The double-recessive model commonly assumed in models of subfunctionalization is assumed.

By considering transitions between populations that are effectively monomorphic the waiting time for the DDC process to reach completion can be calculated (Lynch and Force 2000; Lynch *et al.* 2001; Walsh 2003; Force *et al.* 2005). To go from an ancestral state with a single locus expressed in two contexts to a population fixed for a pair of duplicate genes, each expressed in a single context, three population states must be visited. First a duplication must spread to fixation. Then, a mutation knocking out expression in one context must spread to fixation at one of the duplicate loci. If the duplication is lost before this second step, then the process must start over again. Once one gene copy has lost expression in one context, the locus that is expressed in both contexts can no longer be lost by drift. However, the gene copy that is expressed only in a single context may still be lost by drift, returning the population to be fixed for the *A* allele. Finally, a mutation knocking out expression in the alternative context must spread to fixation at the other gene copy. At this point the pair of gene copies is under strong selection to maintain function and the duplication is expected to be preserved. Because knockout mutations and drift can remove a duplicate gene just as easily as they can result in the fixation of a new duplication, most instances of this process will require many false starts to reach completion.

The transitions between the four possible states of the population can be described by a Markov transition matrix (see Force *et al.* 2005 for a similar approach). The population states are indexed on the basis of the haplotype fixed in the population: *A* (the ancestral allele present in a single copy), *A*/*A* (a haplotype carrying duplicate copies of the ancestral allele), *s*₁/*A* (a haplotype with one copy of the ancestral allele and one copy of a subfunctionalized allele), and *s*₁/*s*₂ (a haplotype with complementary subfunctionalized alleles). For convenience I label the first subfunctional mutant to arise as *s*₁, regardless of which context expression is lost in. Because each transition is a neutral substitution, the per generation probability that a new mutant destined

for fixation arises is simply the rate of each type of mutation. The transition $A|A \rightarrow s_1|A$ can happen by loss of one regulatory element at either locus (with probability $2\mu_s$), while the transition $s_1|A \rightarrow s_1|s_2$ requires the loss of a specific regulatory element at a specific gene copy (with probability $\mu_s/2$):

$$\mathbf{M} = \begin{pmatrix} A & A|A & s_1|A & s_1|s_2 \\ 1 - \mu_d & \mu_d & 0 & 0 \\ 2\mu_k & -2\mu_k - 2\mu_s + 1 & -2\mu_s & 0 \\ \mu_k & 0 & -\mu_k - \frac{\mu_s}{2} + 1 & \frac{\mu_s}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix}. \quad (2)$$

The number of haplotypes in the population is defined as N and assumed for simplicity to be approximately equal to the effective number of haplotypes in the population. Using the fact that each neutral fixation takes an average of $2N$ generations, first-step analysis can be used to calculate the average waiting time until the DDC process is complete (Taylor and Karlin 1984) (see [File S1](#), section 2.1 for the details of the calculation of waiting time). Assuming that $\mu = \mu_d = \mu_s$ and $\gamma = \mu_k/\mu$, then

$$\bar{T}_{\text{DDC}} \approx 2N((2\gamma + 1)(2\gamma + 3)) + \frac{4\gamma^2 + 8\gamma + 7}{2\mu}, \quad (3)$$

where the $2N$ term represents the time spent during drift of mutations destined to fix and the second term represents the time waiting for mutations. For instance, if $\gamma = 1$, then the DDC process requires 15 neutral fixation events. The number of fixation events increases quadratically as γ increases. The waiting time under the pure DDC process is plotted for some sample mutation rates in Figure 2. Differences in the rate of silencing mutations can have just as large an effect on waiting time as differences in population size.

I simulated this process simply by setting the coding mutation rate to 0 in the full model ($\mu_c = 0$). Figure 2 shows the predicted and observed mean waiting times until a stable duplication (i.e., $s_1|s_2$ or $s_2|s_1$) is maintained. The variance in waiting times is large, on the order of the square of the waiting time. When the mutation rates are low, the assumptions of the approximation are met and the fit is quite good.

However, the approximation breaks down as the mutation rate becomes large ($N\mu \gg 1$) and overestimates the waiting times. To make these calculations, I have ignored the possibility that multiple mutations occur before the population becomes effectively fixed for a substitution involving only a single mutation. For instance, while the $A|A$ haplotype is segregating one of the A copies could become subfunctionalized and then drift to fixation. This is a form of stochastic tunneling, but in this case it involves two mutations that are neutral. Weissman *et al.* (2009) developed techniques to determine when the stochastic tunneling regime can be applied and when deterministic models are better descriptors. In the DDC case each potential substitution is neutral, which can violate the assumptions of the tunneling models when $N\mu \gg 1$. Unfortunately, neither the deterministic approximation nor the stochastic tunneling approximation applies in this

regime and accurate estimates of the waiting times are not available. However, for biologically reasonable parameters the prediction of this model holds.

Allelic divergence

Proulx and Phillips (2006) showed that selection acting on function in two contexts can lead to the maintenance of alternate diverged alleles at a single-copy gene. This then creates selection for the spread of gene duplicates. While this process can be described by deterministic dynamics, there is still a stochastic component that will play a role in finite populations simply because of variance in the waiting times for mutations to appear and because adaptive mutations can be lost through drift when rare. Claessen *et al.* (2007) showed that evolutionary branching can have significant time lags before alternative genotypes are maintained. The total waiting time can be calculated as the average of the path-dependent waiting times weighted by the probability that each path is taken. However, the probability of taking a path is generally correlated with the waiting time, so that pathways involving shorter waiting times are much more likely to be taken. For each of the three main pathways for duplication under divergent coding selection, the waiting time decreases with increases in population size, mutation rate, and selection coefficient.

The waiting times for the pathways shown in Figure 3 are calculated in detail in [File S1](#), section 2.2. For the three pathways they are

$$T_{P_1} \approx \frac{3}{N\mu_c} \frac{1}{2s_c} + \frac{1}{N\mu_d} \frac{1}{2s_{c|c}} + \frac{1}{N\mu_s} \frac{1}{2s_{sc_1}} + \frac{1}{N\mu_s/2} \frac{1}{2s_{sc_2}} \quad (4)$$

$$T_{P_2} \approx \frac{3}{N\mu_c} \frac{1}{2s_c} + \frac{6}{N\mu_s} \frac{1}{2s_{sc}} + \frac{1}{N\mu_d} \frac{1}{2\sqrt{r/2s_{sc_1|sc_2}}} \quad (5)$$

$$T_{P_3} \approx \frac{3}{N\mu_c} \frac{1}{2s_c} + \frac{6}{N\mu_s} \frac{1}{2s_{sc}} + \frac{1}{Np_{sc}\mu_d} \frac{1}{2s_{sc_1|sc_1}} + \frac{1}{Np_{sc_1|sc_1}(p_{sc_1|c_2}p_{sc_2})r} \frac{1}{2s_{sc_1|sc_2}}, \quad (6)$$

where p_x refers to the population genetic equilibrium frequency of haplotype x at the previous population state and s_x refers to the selection coefficient for the rare mutant of type x . Note that the selection coefficients are also context dependent and may incorporate multiple genetic backgrounds that the focal haplotype may be found in.

The total waiting time until a stable duplication is maintained depends on how likely it is that each pathway will be taken. The difference between pathway P_1 and P_3 is the time at which the gene duplicates. If a duplication happens to occur before the subspecialized alleles arise, then we expect the process to proceed down P_1 and otherwise move toward branch point B_2 . The route at branch point B_2 depends on the fitness parameters. Pathway P_2 is likely to occur only if the fitness of the heterozygote carrying alternate

Figure 2 (A) The pathways that lead to duplication under the DDC model. (B and C) Plots of the expected waiting times until subfunctionalization is complete. For B and C, $\mu_s = 10^{-6}$ and $\mu_d = 10^{-8}$. In B, $\mu_k = 10^{-6}$ and the x-axis is N , while in C, $N = 10^5$ and the x-axis is μ_k . The waiting time is largely insensitive to population size when $N < \mu_d$. When $\mu_k > \mu_s$, the waiting times can be quite large. (D) shows that the waiting time decreases as the mutation rates increase. The population size was held at 5000 and the recombination rate was 10^{-3} . The simulation was stopped when the total number of s_1s_2 and s_2s_1 haplotypes reached 80% of the total population size. The mutation rates were held equal to each other, $\mu_s = \mu_d = \mu_k$. The gray curve is the prediction from Equation 3 and the black dots show the mean of the simulation runs with the 95% confidence interval.

subspecialized alleles is high. Generally, P_1 and P_3 have similar waiting times because they depend on the same events but in different orders. Figure 3 shows the expected waiting time for a sample set of parameters. Because this process is largely driven by selection, the waiting time goes down as population size and the selection coefficients increase.

Simulations were used to check the accuracy of the waiting time calculations for small population sizes. Figure 3 shows the simulated waiting times when the value of μ_s was set to be much lower than μ_c . This decreases the likelihood that subfunctionalized mutants would appear first. Higher levels of μ_s result in waiting times that are shorter than predicted because they use paths that are considered in the next section. The calculations for pathways P_1 , P_2 , and P_3 are upper bounds for the waiting time.

Net stabilizing coding selection

While the DDC process is characterized by neutral fixations and allelic divergence involves a series of events driven by selection, the process when there is net stabilizing coding selection combines elements of stochastic population genetics and selection-driven change. Starting from the ancestral population state where the A allele is fixed, mutations that alter either the coding region or the regulatory region are not favored when rare. Because coding changes are actively selected against (even when heterozygous), we expect them to remain at a low fluctuating frequency that depends on the mutation rate. Thus, eventual fixation of gene duplicates is unlikely to proceed through an intermediate stage of coding allele divergence.

Figure 3 Alternative pathways to duplication following allelic divergence. The ovals represent distinct population states where each haplotype in the oval is maintained at a deterministic population genetic equilibrium. The initial population state is at the top where a single allele is fixed. The population state can change because of sequential fixation of alleles (gray arrows) or through simultaneous acquisition of two symmetric mutations (dashed arrows). The composition of each population state is labeled with haplotypes separated by commas. Many more paths are possible, in particular the symmetric paths where mutational change alters the performance/expression in the red context first. Two branch points (B_1 and B_2) lead down three complete pathways (P_1 , P_2 , and P_3) to the stable maintenance of diverged gene duplicates. (B) and (C) show expected waiting times until a stable pair of duplicate genes is maintained. In B $\mu_s = 10^{-6}$, $\mu_d = 10^{-9}$, and $\mu_c = 10^{-8}$. The fitness parameters were set so that the increase in relative fitness for each further refinement of the genotype is proportional to a coefficient s . The scheme is $W(A, c_1) = 1 + s$, $W(c_1, c_2) = 1 + 2s$, $W(c_1, sc_2) = 1 + 4s$, $W(c_1, sc_2, sc_2) = 1 + 5s$, $W(sc_1, sc_2) = 1 + 7s$, and $W(c_1, c_2) = 1 - s/4$. For pathways P_1 and P_3 waiting time to recombination of the stable duplicate is ignored. B shows the waiting time for pathways P_1 and P_3 in black (lines overlap) and P_2 with $r = 10^{-3}$ in blue and $r = 10^{-8}$ in red. For comparison, the waiting time for the DDC model is shown in green. Selection is assumed to be weak with $s = 10^{-3}$. The waiting time decreases as population size increases in a similar way for each pathway. (C) shows the effect of selection with $N = 10^5$ and $r = 10^{-3}$ with pathways P_1 and P_3 in black (lines overlap) and P_2 in blue ($\mu_s = 10^{-6}$, $\mu_d = 10^{-8}$, and $\mu_c = 10^{-7}$). For comparison, the waiting time for the DDC model is shown in green. The waiting time for pathway P_2 shows a nonlinear response to the strength of selection because the waiting time for a duplicate to fix via stochastic tunneling does not change and eventually dominates the waiting time along that pathway. (D) shows simulation results. The parameters were $r = 10^{-3}$, $\mu_s = 10^{-7}$, $\mu_c = 10^{-5}$, $\mu_d = 10^{-5}$, and $\mu_k = 10^{-5}$. The gray curve is the prediction from Equation 4 and the black dots show the mean of the simulation runs with the 95% confidence interval.

Losses of context-specific expression, in contrast, behave as recessive lethal mutations. Such mutations are characterized by stochastic population genetic dynamics where their mean frequency increases with the square root of the mutation rate and with population size (Nei 1968; Crow and Kimura 1970; Robertson and Narain 1971). In effect, such mutations behave neutrally when rare but interfere with themselves when they become more common. This interference is stochastically exacerbated in small populations.

Because the square root of the mutation rate is much larger than the mutation rate itself, these recessive lethal mutants occur in large enough numbers to offer a significant opportunity for secondary mutations to arise and fix [*i.e.*, through stochastic tunneling (Iwasa *et al.* 2004; Weissman *et al.* 2009; Proulx 2011)]. Here, this means that secondary mutations that alter the coding region arise from stochastically segregating loss of expression alleles and create subspecialized alleles. Subspecialized alleles are always beneficial when rare (*i.e.*, as heterozygotes) but are assumed to be lethal as homozygotes (Figure 4).

Once subspecialized alleles arise, they are maintained at frequency-dependent equilibria (see Figure 5 for a sample simulation showing the sequence of substitutions). Once the subspecialized alleles are maintained, duplications of the subspecialized alleles are directly favored. They do not spread to fixation but reach a population genetic equilibrium. Recombination between subspecialized duplicate haplotypes and either the ancestral allele or the other subspecialized allele creates a haplotype that deterministically spreads to fixation. Each successive step in the sequence takes a smaller amount of time because the frequency of the haplotype that participates in the next step continues to increase, creating greater and greater opportunity for further adaptive mutations to arise and spread.

I consider three pathways to duplication under net stabilizing coding selection (Figure 4). In the first case, a subspecialized allele arises, duplicates, and recombines to create a stably maintained duplication. In the second and third cases, both subspecialized alleles become resident before either one is duplicated. The details of the calculation of the waiting times are presented in File S1, section 2.3. The total waiting time when all three pathways are considered is

$$T_{P_{1|2|3}} = \frac{1}{2} \left(\frac{1}{\bar{p}_s N \mu_c 2s_{sc}} + \frac{1}{p_{sc} N \mu_d 2s_{sc_1|sc_1}} + \frac{1}{(\bar{p}_s N \mu_c 2s_{sc}) + (p_{sc} N \mu_d 2s_{sc_1|sc_1})} \right) + \frac{1}{p_{sc_1|sc_1} (1 - p_{sc_1|sc_1}) N r 2s_{A|sc_1}}. \quad (7)$$

The total waiting time goes down as both population size and the selection coefficients increase and agrees well with simulations (Figure 4).

Duplication with loss of regulatory regions

The molecular mechanisms responsible for gene duplication can result in a duplicate locus that does not include the full regulatory sequence and in some cases does not even include

all exons. This process has been termed “partial duplication” and has been shown to be common in *Caenorhabditis elegans* (Katju and Lynch 2003). This means that a single mutational event sometimes creates a gene copy with altered expression. This is particularly interesting for the net stabilizing coding selection scenario because it opens up another pathway to the stable maintenance of a gene duplication.

The first step of this pathway involves the production of haplotypes carrying one ancestral allele and one subfunctionalized allele (*i.e.*, $A|s_1$, see Figure 6). This haplotype has the same direct fitness as the ancestral allele haplotype but does not behave neutrally because of its position in the mutational network [*i.e.*, it is circum-neutral (Proulx and Adler 2010)]. Thus, a lineage founded by an $A|s_1$ mutant can produce a significant probability of producing a secondary mutant before going extinct. This is known as stochastic tunneling, and the general expression for the probability of stochastic tunneling in a Wright–Fisher model was derived in Proulx (2011). The probability that a lineage of $A|s_1$ mutants gives rise to an $A|sc_1$ mutant that then is not lost is

$$T_{(A) \rightarrow (A|sc_1)} \approx \frac{1}{N \mu_{ds} 2 \sqrt{s_{sc_1} \mu_c / 2}},$$

where μ_{ds} is probability that allele A mutates into allele $A|s_1$ or $A|s_2$ and s_{sc_1} is the invasion selection coefficient for $A|sc_1$ haplotypes in a population of all A alleles. Note that only half of the possible coding mutations result in subspecialized alleles. As a point of comparison, $T_{(A) \rightarrow (A|sc_1)}$ will be shorter than the waiting time for allele $A|sc_1$ to drift to fixation ($1/(\mu_{ds})$) so long as $N^2 > 2/(s_{sc_1} \mu_c)$. This does not pose a particularly stringent condition, even though we already require $N \mu_c < 1$ for each type of mutation we consider.

Once a subspecialized duplication has been established in the population, mutations are favored that cause the ancestral allele to lose expression in the context that the subspecialized allele is expressed. Such mutations decrease the amount of interference that the subspecialized allele faces but do not reduce function in the other context. These can be followed by specialization of the coding sequence, giving the total waiting time of

$$T_{P_1} = T_{(A) \rightarrow (A|sc_1)} + T_{(A|sc_1) \rightarrow (sc_1|s_2)} + T_{(sc_1|s_2) \rightarrow (sc_1|sc_2)} \\ = \frac{1}{N \mu_{ds} 2 \sqrt{s_{sc_1} \mu_c / 2}} + \frac{1}{N \mu_s / 2} \frac{1}{2s_{sc_1|s_2}} + \frac{1}{N \mu_c / 2} \frac{1}{2s_{sc_1|sc_2}}. \quad (8)$$

Figure 6 shows how the expected waiting times decrease with increasing population size and selection coefficient and the agreement with simulations.

Discussion

The goal of this article is to understand how alternative pathways toward gene duplication relate to each other and

Figure 4 Pathways and waiting times until duplication under net stabilizing coding selection. (A) The ovals represent distinct population states where each haplotype in the oval is maintained at a deterministic population genetic equilibrium. The composition of each population state is labeled with haplotypes separated by commas. The population state can change because of sequential fixation of alleles (gray arrows) or through stochastic tunneling where a first mutation can give rise to a second mutation that is then maintained. The dashed arrows represent the stochastic production of recessive lethal mutations that arise, sojourn, and go extinct. Two branch points (B_1 and B_2) lead down three complete pathways (P_1 , P_2 , and P_3) to the stable maintenance of diverged gene duplicates. (B) and (C) compare the expected waiting time under the DDC model and under net stabilizing coding selection. The parameters are $r = 10^{-3}$, $\mu_s = 10^{-6}$, $\mu_c = 10^{-7}$, $\mu_d = 10^{-7}$. In (B), the selective advantage of a subspecialized allele as a heterozygote was set at 0.01 following Equation 1. In (C), $N = 10^6$ and the selection coefficient is varied. The red curve shows the waiting time for the net stabilizing selection pathways, the blue curve shows the waiting time for the DDC model when $\mu_k = 0$, and the green curve shows the waiting time for the DDC model when $\mu_k = 10^{-5}$. (D) compares simulation results with analytical predictions. The parameters were $r = 10^{-3}$, $N = 5000$, $\mu_s = 10^{-4}$, $\mu_c = 10^{-4}$, and $\mu_k = 10^{-4}$. The gray curve is the prediction from Equation 7 and the black dots show the mean of the simulation runs with the 95% confidence interval.

Figure 5 Simulation of the process when there is net stabilizing selection on coding sequence. In this simulation the parameters were $r = 10^{-3}$, $N = 10^5$, $\mu_s = 10^{-5}$, $\mu_c = 10^{-5}$, $\mu_d = 10^{-5}$, and $\mu_k = 10^{-5}$. The population is initialized with all individuals homozygous for the ancestral allele (A). During the first 15,500 generations, subfunctionalized mutations occur but do not reach high frequencies. By generation 15,500 a subspecialized mutation sc_1 has reached appreciable frequency and is unlikely to go extinct because of drift. This allele fluctuates in frequency around a deterministic equilibrium value of 0.05. At about generation 16,500 a duplication occurs that creates an $sc_1|sc_1$ haplotype. This haplotype spreads in the population until it nears the deterministic equilibrium. Soon after this a recombination event creates the $Alsc_1$ haplotype that spreads in the population. This haplotype could become fixed, but another mutation happens before it does, creating the $sc_1|sc_2$ haplotype that rapidly spreads. Finally, around generation 17,000, another mutation event creates the $sc_1|sc_2$ haplotype that spreads to fixation.

determine the total rate at which stable gene duplications are incorporated into the genome. My framework relies on a consistent view of how changes in gene expression and coding sequence determine the phenotypic output and organismal fitness. This is not to say that the fitness effects of mutational substitutions are expected to remain constant, only that such changes are not viewed as occurring only after a gene duplication has already become fixed.

The models considered here can be categorized on the basis of the type of selection that acts on changes in the coding sequence in the absence of related regulatory changes. I have shown that regardless of whether selection on the coding sequence is net stabilizing or leads to allelic divergence, increased population size and increased selection for context-specific alleles speed up the incorporation of duplications into the genome. This is because there are always routes toward gene duplication that are, at least in part, driven by selection. Many pathways can lead from an ancestral genotype to a maintained duplicate, and some of these pathways involve selection and therefore accelerate as both population size and the selection coefficients increase. Even if many possible pathways are unlikely to occur, because they either are selected against or require many fortuitous events, the presence of even a single adaptive pathway to duplication has a large impact on reducing the total waiting time.

Selection for multifunctional proteins can lead to allelic divergence followed by duplication, and most conditions that promote the origin of multifunctional proteins also

create an adaptive pathway to gene duplication (Proulx and Phillips 2006). In the context of the current study, allelic divergence can lead to incorporation of gene duplications in relatively short periods of time (see Figure 3). Even for fairly weak selection, very low adaptive coding mutation rates, and moderately large population size (10^5) the adaptive duplication pathway is much faster than the DDC pathway. These pathways need not operate exclusively, however. If a duplication does drift to fixation or high frequency, subsequent coding mutations will be under positive selection and lead to the stable maintenance of the gene duplication.

The pattern is similar for net stabilizing coding selection but the selection coefficient and population size must be larger to achieve the same waiting time (see Figure 4). The main difference between the allelic divergence and net stabilizing coding selection regimes is that in the stabilizing regime the first adaptive step involves a form of evolutionary tunneling (Iwasa *et al.* 2004; Weissman *et al.* 2009; Proulx 2011). This step depends on a term involving the product of the mean frequency of subfunctionalized alleles under mutation–selection balance and the coding mutation rate. When population size is small, the DDC pathway is expected to dominate, but as population size increases the adaptive duplication pathways dominate. The overall pattern is expected to follow the minimum of these waiting times, so that the overall pattern is for waiting time to be flat for small populations but then drop off as population size becomes larger.

The rate of duplicate retention is dependent on the silencing rate for the DDC pathway, but not for adaptive pathways. If the silencing mutation rate (μ_k) is much larger than the other mutation rates, then the DDC waiting time can increase by orders of magnitude. This is a likely scenario when many possible coding and regulatory mutations knock out or completely disable gene function and this rate is expected to depend on both gene length and the structure of the gene in terms of intron number and UTR length (Lynch 2007). Under the DDC model, variation in gene structure can lead to substantial variation in the rate of duplicate retention that is equal to or larger than the variance in duplicate retention due to changes in population size alone. This effect can greatly increase the waiting time for stable maintenance of duplications compared with previous calculations of the waiting times for the DDC process.

Tandem gene duplications involve the replication of a chromosomal segment. A duplication that copies only part of the coding sequence is likely to produce a nonfunctional gene that will have a very low probability of ever mutating into a functional gene copy. On the other hand, a duplication or retrotransposition that copies only part of the regulatory region may create a functional gene that is expressed only in certain contexts. This can open up another adaptive path toward duplication where a coding mutation hits a segregating duplicate haplotype carrying $A|s_1$. This occurs at a rate that involves the product of the duplication rate and the square root of the coding mutation rate. This tends to be faster than the pathway that first involves the acquisition of

Figure 6 Pathways and duplication times under simultaneous duplication and subfunctionalization. (A) The ovals represent distinct population states based on the haplotypes present in the population. The solid arrows represent transitions toward population states that have deterministic population genetic equilibria. The dashed arrows represent transitions to population states that are characterized by stochastic dynamics and are not expected to be fixed states (i.e., streetcar stops). The composition of each population state is labeled with haplotypes separated by commas. (B) and (C) show the expected waiting times based on the analytical predictions. For B and C, $\mu_s = 10^{-6}$, $\mu_d = 10^{-7}$, $\mu_c = 10^{-8}$, and $r = 10^{-3}$. The fitness parameters were set so that the increase in relative fitness for each further refinement of the genotype is proportional to a coefficient s . The scheme is $W(A, A, sc_1) = 1 + s$, $W(A, A, sc_1, sc_1) = 1 + 2s$, $W(A, sc_1, sc_1, s_2) = 1 + 3s$, $W(sc_1, sc_1, s_2, s_2) = 1 + 4s$, and $W(sc_1, sc_1, s_2, sc_2) = 1 + 5s$. (B) shows that the waiting time decreases as population size increases. (C) shows the effect of changing the selection coefficient when $N = 10^6$. (D) compares the simulation results with the analytical predictions. The parameters were $r = 10^{-3}$, $N = 5000$, $h = \frac{1}{2}$. All of the mutation rates were set equal to each other. The gray curve is the prediction from Equation 8 and the black dots show the mean of the simulation runs with the 95% confidence interval.

the subspecialized double mutant for two reasons. First, subfunctional alleles act as recessive lethals at single-copy genes. It has long been known that their frequency can be large compared with dominant deleterious alleles and that this effect depends on population size. In particular, in large populations their mean frequency approaches the square root of the subfunctionalization mutation rate. This is quite similar to

the tunneling pathway involving duplicate haplotypes carrying a subfunctionalized allele, where the rate of tunneling is related to the square root of the coding mutation rate. However, in smaller populations the frequency of recessive lethals is significantly lower, so that even when coding and regulatory mutation rates are equal, the pathway starting with a duplication producing the $A|s_1$ haplotype is faster.

Figure 7 The reduction in the expected waiting time when duplications include subfunctionalization. The parameters are $N = 10^5$, $r = 10^{-3}$, $\mu_s = 10^{-7}$, $\mu_c = 10^{-8}$, $\mu_d = 10^{-7}$, and $\mu_k = 10^{-7}$. The selective advantage of a subspecialized allele as a heterozygote was set at 0.01 following Equation 1. The proportion of duplications that result in haplotype Als_1 , labeled “proportion subfunctionalized”, was varied from 0 to 0.9. The red curve shows the expected waiting time following the pathway described by Equation 7 while the orange curve shows the pathway that involved duplicate tunneling. For this plot I used a more accurate expression for the tunneling probability that involves solving a transcendental equation similar to Equation 8 (see Proulx 2011, for more details). The green curve shows the waiting time under subfunctionalization but allowing for duplications that directly produce Als_1 haplotypes. As the proportion subfunctionalized increases, both the orange and the green curves go down, but the effect is much larger on the orange curve.

Second, in both pathways a type of double mutant arises and the rate of the next step depends on the equilibrium frequency of the double mutant. This again falls out in favor of the pathway starting from $A|s_1$ because the $A|sc_1$ haplotype can spread to fixation, whereas the sc_1 haplotype is under negative frequency-dependent selection and tends to maintain a low equilibrium frequency. Putting these together, pathways starting with the $A|s_1$ haplotype can greatly accelerate the rate of duplicate incorporation even when the rate of duplications that also involve subfunctionalization is low (Figure 7).

Overall, the picture painted by this study is that adaptive processes are likely to be a component of most successful duplication events. When knockout mutations are included in models of the DDC process, I find that the waiting time until duplicate retention increases by orders of magnitude, calling into question the conclusion that typical multicellular eukaryote lineages experience population sizes amenable to the DDC process. Only in the exact scenario posited by the DDC model, where the potential for specialization of the gene toward specific tissues is absent or associated with very small selection coefficients, do we predict that adaptive routes toward duplication are unavailable. Because adaptive routes to duplication are present even under net stabilizing selection on coding regions, we expect duplication rates to increase with population size and selection strength. This creates an apparent paradox in that lineages with small effective population size have higher rates of gene duplication and lineages with enormous population size have lower rates of gene duplication.

This apparent paradox can be immediately resolved by noting that all known transitions to multicellularity produce

a correlation between N_e , mating system, and internal tissue complexity. The dynamics of lethal alleles are critically related to the mating system. In species that have high rates of selfing, recessive lethal alleles are selected against even when rare because one-quarter of the offspring of individuals carrying a lethal allele will be homozygous for the lethal allele. In haploid asexual species, there is not even the possibility of the spread of lethal alleles, so subfunctional mutants (*i.e.*, s_1 and s_2) are immediately selected against. Organisms that have multiple tissues that exhibit polyphenism or experience multiple distinct environments during a single life span will also have more opportunity for multifunctional proteins to evolve simply because there are more contexts that genes can become specialized for. This suggests that in addition to shifts in population size between basal eukaryotes and multicellular eukaryotes, changes in mating system and organismal complexity may have increased the rate of duplicate retention.

Acknowledgments

F. R. Adler is specially thanked for pointing out the ansatz for the Wright–Fisher tunneling problem and A. Yanchukov is specially thanked for a careful reading of a draft of this article. The comments of two anonymous reviewers contributed to both conceptual clarity and presentation of this work. This work was supported by National Science Foundation grant EF-0742582 (to S.R.P.).

Literature Cited

- Claessen, D., J. Andersson, L. Persson, and A. M. de Roos, 2007 Delayed evolutionary branching in small populations. *Evol. Ecol. Res.* 9(1): 51–69.
- Conant, G. C., and K. H. Wolfe, 2008 Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.* 9: 938–950.
- Connallon, T., and A. Clark, 2011 The resolution of sexual antagonism by gene duplication. *Genetics* 187(3): 919–937.
- Crow, J. F., and M. Kimura, 1970 *An Introduction to Population Genetics Theory*. Harper & Row, New York.
- Des Marais, D. L., and M. D. Rausher, 2008 Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* 454(7205): 762–765.
- Force, A., M. Lynch, F. Pickett, A. Amores, Y. Yan *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151(4): 1531–1545.
- Force, A., W. A. Cresko, F. B. Pickett, S. R. Proulx, C. Amemiya *et al.*, 2005 The origin of subfunctions and modular gene regulation. *Genetics* 170(1): 433–446.
- Gillespie, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford.
- Hammerstein, P., 1996 Darwinian adaptation, population genetics and the streetcar theory of evolution. *J. Math. Biol.* 34(5–6): 511–532.
- Hughes, A. L., 2005 Gene duplication and the origin of novel proteins. *Proc. Natl. Acad. Sci. USA* 102(25): 8791–8792.
- Innan, H., and F. Kondrashov, 2010 The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* 11(2): 97–108.
- Iwasa, Y., F. Michor, and M. A. Nowak, 2004 Stochastic tunnels in evolutionary dynamics. *Genetics* 166(3): 1571–1579.

Katju, V., and M. Lynch, 2003 The structure and early evolution of recently arisen gene duplicates in the *Caenorhabditis elegans* genome. *Genetics* 165: 1793–1803.

Lynch, M., 2007 *The Origins of Genome Architecture*. Sinauer Associates, Sunderland, MA.

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

Lynch, M., M. O'Hely, B. Walsh, and A. Force, 2001 The probability of preservation of a newly arisen gene duplicate. *Genetics* 159: 1789–1804.

Nei, M., 1968 The frequency distribution of lethal chromosomes in finite populations. *Proc. Natl. Acad. Sci. USA* 60(2): 517–524.

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

Otto, S. P., and P. Yong, 2002 The evolution of gene duplicates. *Adv. Genet.* 46: 451–483.

Proulx, S., and P. Phillips, 2006 Allelic divergence precedes and promotes gene duplication. *Evolution* 60(5): 881–892.

Proulx, S. R., 2000 The ESS under spatial variation with applications to sex allocation. *Theor. Popul. Biol.* 58(1): 33–47.

Proulx, S. R., 2011 The rate of multi-step evolution in Moran and Wright-Fisher populations. *Theor. Popul. Biol.* 80(3): 197–207.

Proulx, S. R., and F. R. Adler, 2010 The standard of neutrality: Still flapping in the breeze? *J. Evol. Biol.* 23(7): 1339–1350.

Robertson, A., and P. Narain, 1971 The survival of recessive lethals in finite populations. *Theor. Popul. Biol.* 2(1): 24–50.

Ross, S., 1988 *A First Course in Probability*. Macmillan, New York.

Taylor, H. M., and S. Karlin, 1984 *An Introduction to Stochastic Modeling*. Academic Press, Orlando, FL.

Taylor, J. S., and J. Raes, 2004 Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* 38: 615–643.

Walsh, B., 2003 Population-genetic models of the fates of duplicate genes. *Genetica* 118(2–3): 279–294.

Weissman, D. B., M. M. Desai, D. S. Fisher, and M. W. Feldman, 2009 The rate at which asexual populations cross fitness valleys. *Theor. Popul. Biol.* 75(4): 286–300.

Communicating editor: L. M. Wahl

Appendix: Probability of Loss of an In-Phase Duplication

When a duplicate haplotype first arises via a tandem duplication, it may consist of two copies of the same allele, termed an in-phase duplicate haplotype. The new duplicate haplotype can recombine with alleles at the original locus to create out-of-phase haplotypes. The relative fitness of an individual carrying the in-phase haplotype may be greater than or less than 1, while the relative fitness of an individual carrying the out-of-phase haplotype is >1 (fitness is relative to the mean fitness of the population in which the duplication arises). Define ω_1 as the relative fitness of an individual carrying three copies of the same specialized allele [i.e., $(c_1, c_1|c_1)$] and ω_2 as the relative fitness of an individual carrying the out-of-phase haplotype [either $(c_1, c_2|c_1)$ or $(c_2, c_2|c_1)$, which are assumed by symmetry to be equal].

In many studies of the dynamics of gene duplicate evolution, the eigenvalue for the spread of the duplicate is derived and used as a measure of selection or fixation (Otto and Yong 2002; Proulx and Phillips 2006; Connallon and Clark 2011). Consider a population of haplotypes carrying the c_1 and c_2 alleles in which a duplication occurs creating a $c_1|c_1$ haplotype. The spread of the duplicate haplotype involves both $c_1|c_1$ (in-phase) haplotypes and $c_2|c_1$ (out-of-phase) haplotypes. This two-state transition matrix is

$$\mathbf{M} = \begin{matrix} c_1|c_1 & c_2|c_1 \\ c_2|c_1 & c_1|c_1 \end{matrix} \begin{pmatrix} p\omega_1 + (1-p)\omega_2(1-r) & (1-p)\omega_2r \\ p\omega_2r & p\omega_2(1-r) + (1-p)\omega_2 \end{pmatrix}, \quad (\text{A1})$$

where p is the equilibrium frequency of the c_1 allele in the absence of the duplicate haplotype. The dominant eigenvalue can be found by standard techniques and is

$$\lambda_{\cdot|c_1} = \frac{1}{2} (\omega_1 p + \omega_2 (2 - p - r) + \sqrt{(\omega_1 - \omega_2)^2 p^2 - 2(\omega_1 - \omega_2)\omega_2 p(1 - 2p)r + \omega_2^2 r^2}). \quad (\text{A2})$$

The effective selection coefficient for the rare duplicate haplotype is $s_{\cdot|c_1} = \lambda_{\cdot|c_1} - 1$. In the case where $\omega_1 = \omega_2$, this system reduces to a standard selection problem with the well-known result that the probability of nonloss of the invading haplotype is $\sim 2s_{\cdot|c_1}$.

A contrasting approach is to directly calculate the probability of nonloss of the duplicate haplotype undergoing selection and recombination. This can be done using first-step analysis for the multitype branching process (Ross 1988). Assume that diploid adults produce a Poisson-distributed number of offspring and that the number that undergoes recombination is binomially distributed. Let D_1 be the probability that a haplotype lineage starting with 1 copy of the in-phase duplication eventually goes extinct (that is, no haplotypes carrying either the in-phase or the out-of-phase duplication are left in the population). Likewise, let D_2 be the probability that a haplotype lineage starting with 1 copy of the out-of-phase duplication eventually goes extinct. Because this is a branching process, the probability of eventual extinction of a set of duplicate haplotypes is simply the probability that the lineage produced by each individual goes extinct (see Proulx 2011 for a rigorous limit for Wright-Fisher populations). This gives

$$D_1 = p \left(\sum_{i=0}^{\infty} \frac{e^{-\omega_1} \omega_1^i}{i!} D_1^i \right) + (1-p) \left(\sum_{i=0}^{\infty} \frac{e^{-\omega_2} \omega_2^i}{i!} \sum_{j=0}^i \binom{i}{j} r^j (1-r)^{i-j} D_1^j D_2^{i-j} \right)$$

$$D_2 = p \left(\sum_{i=0}^{\infty} \frac{e^{-\omega_2} \omega_2^i}{i!} \sum_{j=0}^i \binom{i}{r} r^j (1-r)^{i-j} D_1^j D_2^{i-j} \right) + (1-p) \left(\sum_{i=0}^{\infty} \frac{e^{-\omega_2} \omega_2^i}{i!} D_1^i \right).$$

This can be simplified to give

$$D_1 = pe^{-\omega_1(1-D_1)} + (1-p)e^{-\omega_2(1-D_1-D_2r)} \quad (\text{A3})$$

$$D_2 = pe^{-\omega_2(1-D_1r-D_2(1-r))} + (1-p)e^{-\omega_2(1-D_2)}. \quad (\text{A4})$$

The classic result for fixation probability can be recovered if $\omega = \omega_1 = \omega_2$ (which also implies that $D = D_1 = D_2$), giving an implicit formula for D as

$$D = e^{(-\omega(1-D))}.$$

This transcendental equation cannot be further simplified, but can be approximated when ω is slightly larger than 1 (Proulx 2011) to give $D \approx 2(\omega - 1)$.

The joint solution to Equations A3 and A4 can be found numerically for specific values of ω_1 , ω_2 , r , and p . For the remainder of this discussion I assume that $p = \frac{1}{2}$. Figure A1 shows the probabilities of nonloss of the two duplicate haplotypes and the probability of loss estimated from the eigenvalue.

In all cases, the probability of nonloss is $< 2(\lambda - 1)$, where λ is the eigenvalue. When $r \approx 0$, the probability of loss is determined by the behavior of the in-phase duplicate. At $r = \frac{1}{2}$ the difference from the eigenvalue expectation is due to the probability that the initial in-phase duplication goes extinct immediately before any recombination is possible. Otherwise the eigenvector is immediately reached and the eigenvalue approximation for the probability of nonloss applies. Therefore, the probabilities of nonloss must interpolate between twice the mean fitness of rare in-phase haplotypes and $2(\lambda - 1)$.

The behavior of the system can be understood by considering three qualitatively different scenarios. In the first case, the mean fitness of the in-phase duplication is > 1 [because $(\omega_1 + \omega_2)/2 > 1$]. Because we will be considering only cases where $\omega_2 > 1$ the eigenvalue is also always > 1 . In this case, when $r = 0$, then the probability of nonloss is simply $2((\omega_1 + \omega_2)/2 - 1)$. As r increases, the probability of nonloss monotonically increases toward $2(\lambda - 1)$ (Figure A1A). Interestingly, the eigenvalue approach is most deceiving when r is small. This case also applies when the in-phase duplication has mean relative fitness of 1.

In the second case the mean fitness of the in-phase duplication is < 1 but the eigenvalue for $r = \frac{1}{2}$ is > 1 . Near $r = 0$ the probability of loss is close to 1, in contradiction to the eigenvalue result, which argues that the spread of the duplicate is fastest when r is small. However, the probability of nonloss rapidly increases in r and does reach a maximum value for intermediate r (Figure A1B). The probability of nonloss is always $< 2(\lambda - 1)$.

In the third case, both the mean fitness of the in-phase duplication is < 1 and the eigenvalue at $r = \frac{1}{2}$ is < 1 . In this case, for large enough r , the probability that a rare duplicate

Figure A1 The probability of nonloss of duplicate haplotypes. The probability of nonloss of the in-phase (blue curves) and out-of-phase (green curves) haplotypes is shown as a function of the recombination rate r . Also plotted is $2(\lambda - 1)$, twice the difference of the eigenvalue and 1. In each case, the value for the out-of-phase duplications is much closer to the eigenvalue curve. For A–C, $P = \frac{1}{2}$, $\omega_2 = 1.002$. The values of ω_1 are 0.999 in (A), 0.995 in (B), and 0.9935 in (C).

haplotype is lost approaches 1. The probability of loss for a single copy of the out-of-phase duplicate and $2(\lambda - 1)$ are virtually identical. The probability of loss of the in-phase haplotype is 0 for small r , increases to a maximum value for intermediate r , and then decreases until it becomes 0 when the eigenvalue reaches 1.

GENETICS

Supporting Information

<http://www.genetics.org/content/suppl/2011/12/05/genetics.111.135590.DC1>

Multiple Routes to Subfunctionalization and Gene Duplicate Specialization

Stephen R. Proulx

File S1

Supplementary information and equations

1 The position of the ancestral allele

So long as heterozygotes have context-specific fitness that is intermediate between the two homozygotes, the population will become fixed for new alleles that increase total fitness. The ancestral allele S is defined as the allele in the set that has maximum fitness as a homozygote. The population remains monomorphic when it is away from state A because homozygotes that are more fit can both invade and become fixed. Once the state A is reached, however, no other homozygous genotypes have higher total fitness (See figure 1). Mutations near A can be characterized as showing antagonistic pleiotropy because they improve function in one context but degrade function in the other context. Thus, given that the ancestral allele we take as our starting point is the product of previous evolution, the nearby mutants that increase fitness in one context necessarily decrease fitness in the other context. I refer to allele c_1 to describe a mutation that, as a homozygote, increases fitness in context 1, decreases fitness in context 2, and decreases total fitness. While this immediately implies that no mutations can replace A , it does not tell us whether or not mutations that are phenotypically similar to A can become established.

A coding mutant that is phenotypically near A can become established if heterozygotes carrying one copy of the mutant and one copy of A have higher fitness than individuals homozygous for A . While the mutant c_i must increase fitness in one context while decreasing fitness in the other, the net effect on fitness depends on the relative dominance of the mutant in the two contexts. A trivial example can illustrate this phenomena: If the mutant increases fitness in context 1 and is partially dominant in that context but reduces fitness in context 2 and is completely recessive in context 2, then it will surely increase in frequency. For more moderate scenarios where the mutant allele is partially dominant in each context, the weighted fitness gains in one context must outweigh the weighted fitness losses in the other for the mutant to increase in frequency when rare. As it increases in frequency, however, the probability that it will be found in a homozygous state goes up, and so the marginal fitness of the mutant allele decreases. Thus, while the fitness of each genotype is constant, the expected fitness of gametes bearing the mutant allele is frequency dependent.

Figure 1: Fitness-set description of allelic substitutions leading to the ancestral allele. The area between the axes and the purple curve represents the fitness-set; it is the biologically feasible set of homozygous allele effects. Total fitness goes up as fitness in each context increases, but because total fitness is the product of fitness in the two contexts the contours are curved. The circles represent populations that are monomorphic for a single allele and are displayed for a series of substitution transitions shown by green arrows. The boxes with negative numbers represent allele states prior to the A allele. The A allele is located on the edge of the fitness set where it is just tangent to the fitness contour. Near the A allele, c_1 and c_2 may be able to invade as a pair and replace A .

We can write down the fitness of the genotypes containing alleles A and c_1 or c_2 . For simplicity I will illustrate this for c_1 (see table 1). In this illustration I define the phenotype as the log fitness in each context. The A allele has phenotype $\ln(w_i)$ in context i . For any c_1 we expect that fitness in context 1 is larger than for the A allele so it has phenotype $\ln(w_1) + \ln(x)$ where $x > 1$ is the change in context 1 fitness. Likewise phenotype in context 2 is $\ln(w_2) + \ln(y)$ where $y < 1$. For A to be at the point where the fitness contour is tangent to the fitness set $\ln(x) + \ln(y)$ must be negative.

The context specific fitness of the heterozygote depends on dominance, as shown in

table 1. The c_1 allele can invade if $x^{h_1}y^{h_2} > 1$. Because $x > 1$ and $y < 1$, this can only be true if $h_2 < h_1$. In other words the mutant allele must have greater dominance in the context on which it specializes (see PROULX and PHILLIPS, 2006, for a similar approach based on derivatives of the fitness function)).

Thus we can consider two types of fitness relationship:

1. *Allelic Divergence*: Specialized alleles can invade when rare when $x^{h_1}y^{h_2} > 1$. Even though there is still antagonistic pleiotropy, mutations near A are at a net advantage when heterozygous. Coding mutations near A are then maintained in the population and can create direct selection favoring duplications. This scenario additionally leads to selection to alter the regulatory region created subspecialized alleles.
2. *Net Stabilizing Coding Selection*: Specialized alleles cannot invade when rare when $x^{h_1}y^{h_2} < 1$. There is antagonistic pleiotropy which causes mutants that increase fitness in one context to be at a net disadvantage both as heterozygotes and as homozygotes. In this scenario, mutations that silence expression in one context act as recessive lethal (or recessive sick) mutations and can be stochastically maintained at appreciable frequencies. Secondary mutations can produce alleles that are expressed only in the context to which their coding region is adapted. These alleles are actively maintained by selection and open the door to complementary mutations that specialize on the other contexts. Duplications are then directly advantageous and can spread due to selection.

Genotype	Phenotype in context 1	Phenotype in context 2	Total Fitness
A, A	$\ln(w_1)$	$\ln(w_2)$	$w_1 w_2$
c_1, c_1	$\ln(w_1) + \ln(x)$	$\ln(w_2) + \ln(y)$	$w_1 w_2 x y$
A, c_1	$\ln(w_1) + h_1 \ln(x)$	$\ln(w_2) + h_2 \ln(y)$	$w_1 w_2 x^{h_1} y^{h_2}$

Table 1: Context specific fitness

2 Calculating approximate waiting times

I assume that ancestral populations are fixed for the allele that has the multifunctional coding sequence and both promoters. In this ancestral population, each individual expresses the multifunctional allele in both contexts. The evolutionary process allows for mutations in both the coding and regulatory region, as well as knockout mutations that irrevocably silence the allele. For simplicity, I assume that each allele has the same knockout mutation rate.

Throughout this paper I write the total number of haploid genomes as N and assume that $N_e \approx N$. When $N\mu \ll 1$ then the population is well described by the non-stochastic

population genetic equilibria most of the time but occasionally transitions between states following the successful introduction of a new mutation. That is to say, without frequency dependent selection we expect most populations to be monomorphic and with frequency dependence we expect the population to be near the frequency dependent equilibrium. The population can change state if a mutation arises, is not lost when rare, and is deterministically maintained in the population. However, stochastic fluctuations in allele frequency are considered during the invasion of a new haplotype. This modeling framework fits the streetcar approach put forward by HAMMERSTEIN (1996).

I use a stochastic processes approach to calculate the mean and variance for waiting times until a transition in population state. In particular, I focus on waiting times to go from the ancestral population state to one in which a gene duplication is selectively maintained. This final state can be viewed as an absorbing state of a Markov chain where states reflect the allelic composition of a population that is in population genetic equilibrium (the final stop of the streetcar). The population genetic equilibria are calculated assuming no stochasticity and no mutation. I calculate waiting times for both the neutral and adaptive process and use these both to compare the relative probability of each path and also to calculate the total expected waiting time along any path (See e.g. WEINREICH and CHAO, 2005). To the extent that neutral duplications can be produced and spread by drift regardless of how far the adaptive process has proceeded these represent independent trajectories and can be directly compared.

The waiting time until adaptive duplication can be calculated by following the series of events that lead up to it. The general approach is to consider each possible transition that follows a trajectory that ultimately ends in the stable maintenance of diverged duplicate genes. These steps may include deleterious or neutral intermediates. In such cases I use the stochastic tunneling paradigm and I calculate the waiting time until a secondary mutation will arise and be deterministically maintained in the population (IWASA *et al.*, 2004; WEISSMAN *et al.*, 2009; PROULX, 2011). In some cases, multiple similar steps can occur that result in symmetrical movement along the trajectory and I calculate the waiting time until the first of these occurs.

The basic approach for calculating the waiting time between population states is most straightforward for mutations that are under positive selection when rare. Each generation a mutation may occur and that rare mutation's fate will be described by the probability that it starts a lineage that either goes extinct when rare or increases in frequency until stochastic effects become small. If it escapes rarity, then further changes in gene frequency will be well described by the deterministic population genetic dynamics.

For a mutation that occurs with probability μ in a population of size N , the number of new mutations each generation is binomially distributed with parameters N and μ . For reasonably large population size (say above 100) and $N\mu$ small the probability that no new mutations appear in each generation is well described by the zero term of the Poisson distribution with mean $N\mu$. Under these assumptions, the waiting time between the appearance of new mutations is simply $1/(1 - e^{-N\mu})$. An even simpler approximation

holds when $N\mu \ll 1$ where the waiting time is $1/(N\mu)$. So long as $N\mu < 1/2$ this approximation works well for realistic mutation rates. Once $N\mu > 1/2$ the waiting time for mutations to arise is only a few generations and can be safely ignored.

Most mutations that are under positive selection still go extinct while rare. Because we are interested in processes where some mutations are favored when rare but then are under frequency dependent selection, we consider the probability that they are lost when rare rather than the probability of fixation. So long as population size is not too small (say less than 100) and selection coefficients are not too large, the probability that a mutant arising as a single copy is not lost is well approximated by $2s$. Under these assumptions the waiting time for a mutation that is favored when rare is

$$T \approx \frac{1}{N\mu} \frac{1}{2s},$$

where s is the difference between the relative fitness of the mutant and 1. Again, this approximation applies when $N\mu$ is not too large; if it exceeds about $1/2$ then we must return to the Poisson formula for the waiting time. I ignore the time required to approach population genetic equilibrium for alleles under selection because it is usually orders of magnitude smaller than the waiting time for the appearance of a successful mutation.

2.1 No coding selection

The transition matrix, $\mathbf{A}$, for the DDC process is given in equation 2 in the main text. The expected waiting time until a complementary pair of subfunctional alleles is fixed can be computed using first step analysis (ROSS, 1988). Define the expected waiting time until a duplication haplotype with a complementary pair of alleles is fixed for a population in state i as D_i . Assume that the census population size is approximately the effective population size and that N is the census count of haplotypes (in contrast to the common formalism where the number of haplotypes is set to $2N$). Transitions between states take on average $2N$ generations, while a population that stays in the same state only adds 1 generation to the waiting time. We have

$$D_i = A_{ij}(1 + D_i) + \sum_{j \neq i} M_{ij}(2N + D_j), \quad (1)$$

with the boundary condition that $D_{\text{Complementary}} = 0$. Solving this system and defining $\gamma = \mu_k/\mu_s$ gives

$$D_{\text{Single}} = 2N((2\gamma + 1)(2\gamma + 3)) + \left(\frac{(\gamma + 1)(2\gamma + 1)}{\mu_d} + \frac{2\gamma + 5}{2\mu_s} - (2\gamma + 1)(2\gamma + 3) \right). \quad (2)$$

The term multiplying $2N_e$ accounts for the waiting time for drift to fixation in each of the transitions, while the large term in parentheses accounts for the time waiting for mutations.

Because N_e , $1/\mu_s$, and $1/\mu_d$ are typically much larger than γ^2 we can approximate D as

$$\bar{T}_{\text{DDC}} = D_{\text{Single}} \approx 2N((2\gamma + 1)(2\gamma + 3)) + \left(\frac{(\gamma + 1)(2\gamma + 1)}{\mu_d} + \frac{2\gamma + 5}{2\mu_s} \right). \quad (3)$$

The number of extra transitions goes up with the square of γ while the waiting time for mutations goes up both as the μ_s and μ_d go down and as γ goes up.

2.2 Allelic Divergence

The waiting time until a duplicate is maintained can be found by calculating the waiting times for each step of the pathway and by calculating the probability of traversing each branch in the pathway. Figure 3 in the main text shows some of the pathways that are possible for this process. The number of alternative pathways grows quite large because there are many possible orders for mutational change. The figure shows the paths where the blue context is affected first, but the symmetric path involving changes in the red context first contribute to the calculations in this section. I consider only pathways where allelic divergence happens first because this process is expected to be quicker and have a higher probability of occurring. Inclusion of the other pathways will only decrease the expected waiting time until the final state of maintained duplicates is reached. Most of the transitions in figure 3 in the main text involve the spread of a new allele via selection. However, duplication events result in an in-phase haplotype carrying two copies of the same allele. In some scenarios this in-phase haplotype still spread via selection (for example an $sc_1|sc_1$ haplotype), but in others the in-phase duplicate is not selected for and may be weakly selected against. In such situations the spread of the duplicate depends on recombination which can be modeled either as a continuous or stochastic process. If the recombination rate is high enough, the eigenvalue of the spread of the duplication can be calculated and used as a measure of the invasion selection coefficient. If recombination rate is low, however, the process can be understood by tracking the stochastic dynamics of the lineage of duplicate haplotypes descending from the initial mutant duplicate and calculating the probability that a recombination event will occur and result in a recombinant that itself is not destined to be stochastically lost before the lineage of deleterious in-phase duplicate haplotype goes extinct. The stochastic tunneling framework describes just such situations (IWASA *et al.*, 2004; WEISSMAN *et al.*, 2009; PROULX, 2011).

When a process can follow two distinct paths and the probability of starting down each path is independent then the total waiting time is a function of the waiting time to start down each path. In practice, unless the parameters are particularly balanced, one pathway will be more likely than the others and the total waiting time will be well approximately the waiting time following the quicker pathway. In addition, the true total waiting time will typically be less than the waiting time down any specific path. Thus, a reasonable estimate of the upper bound for the waiting time can be found by calculating the waiting time down any path of our choice.

The first transition in figure 3 in the main text involves waiting until one of the adaptive coding mutations becomes resident. The waiting time for this to occur is

$$T_{(A) \rightarrow (A, c_1)} = \frac{1}{N\mu_c} \frac{1}{2s_c}$$

where the total coding mutation rate is μ_c and s_c is the selection coefficient for a coding mutation that specializes towards a single context. Note that a mutation could affect either context first, but because of symmetry I simply use the label c_1 . If the population is reasonably large, a complementary coding mutation may become resident while the first coding mutation is still rare. Note that while the probability of a coding mutation is μ_c , the probability of a coding mutation improving performance in context i is $\mu_c/2$. For two independent processes with the same exponential parameter $\lambda/2$, the expected waiting time until both events have occurred is simply $\frac{3}{\lambda}$. Putting these facts together, a good approximation for the waiting time until both coding mutations are resident is

$$T_{(A) \rightarrow (c_1, c_2)} = \frac{3}{N\mu_c} \frac{1}{2s_c}.$$

Once both coding mutants have been established they will deterministically replace the ancestral allele.

Alternatively, the process could involve a coding change followed by a change in expression of the specialized coding allele (not shown in figure 3 in the main text). The probability that this would precede both coding changes becoming resident depends on the relative mutation rates for coding regions and loss of expression and on the population genetic equilibrium frequency of the specialized coding allele, p_{c_1} . If p_{c_1} is low, then mutations are much more likely to hit the high-frequency ancestral alleles, but if p_{c_1} is high then the next mutation may be one that creates a subspecialized allele. The waiting times for the appearance of a subspecialized allele is

$$T_{(A, c_1) \rightarrow (A, s_{c_1})} = \frac{1}{Np_{c_1}\mu_s/2} \frac{1}{2s_{s_{c_1}}}$$

where $s_{s_{c_1}} = (p_{c_1}W(c_1, s_{c_1}) + (1 - p_{c_1})W_{A, s_{c_1}})/\bar{W} - 1$ and $\bar{W} = p_{c_1}^2W(c_1, c_1) + 2p_{c_1}(1 - p_{c_1})W(A, c_1) + (1 - p_{c_1})^2W(A, A)$. Because the ancestral allele is expected to have evolved to balance benefits in the two contexts, alleles that have specialized coding function (c_1) are expected to be maintained only at low frequencies until a counterbalancing allele (c_2) arises. This means that p_{c_1} is generally expected to be low. Because of this, I do not include this pathway in estimates of the upper bound on the time to duplication.

Returning to the path involving the joint acquisition of two complementary coding mutations (paths **P₁**, **P₂** and **P₃** in figure 3 in the main text), the next step could involve either a duplication event or the invasion of subspecialized alleles. The spread of a duplicate haplotype carrying specialized alleles can be described by the eigenvalue of the increase

in frequency for the recombining duplicate haplotype (OTTO and YONG, 2002; PROULX and PHILLIPS, 2006). These conditions have been provided by PROULX and PHILLIPS (2006) where it was shown that such duplications always have eigenvalues greater than 1. Under these conditions and so long as r is not too small, the probability of non-loss of the duplicate haplotype are approximately equal to $2(\lambda - 1)$ where λ is the eigenvalue for the spread of the duplicate haplotype (see Appendix). The probability of non-loss of the duplicate haplotypes can also be calculated directly for using first step analysis (see Appendix). The approximate per generation probability that an out-of-phase duplication (a duplicate haplotype with two different alleles) arises and is destined to become fixed

$$T_{(c_1, c_2) \rightarrow (c_1 | c_2)} = \frac{1}{N\mu_d} \frac{1}{2s_{c|c}},$$

where $s_{c|c}$ is the eigenvalue of the spread of the duplicate haplotype.

Following the fixation of an out-of-phase duplicate (state $(c_1 | c_2)$), the following steps in pathway **P₁** are quite straightforward because each successive substitution is expected to reach fixation. The total waiting time for these substitutions to occur is then

$$\begin{aligned} T_{P_1} &= T_{(A) \rightarrow (c_1, c_2)} + T_{(c_1, c_2) \rightarrow (c_1 | c_2)} + T_{(c_1 | c_2) \rightarrow (sc_1 | c_2)} + T_{(sc_1 | c_2) \rightarrow (sc_1, sc_2)} \\ &= \frac{3}{N\mu_c} \frac{1}{2s_c} + \frac{1}{N\mu_d} \frac{1}{2s_{c|c}} + \frac{1}{N\mu_s} \frac{1}{2s_{sc_1}} + \frac{1}{N\mu_s/2} \frac{1}{2s_{sc_2}}. \end{aligned}$$

Under the assumption of symmetric fitness effects recall that $W(i, i) = W(i)$ and that $W(i, j) = W(j, i)$. The values for the invasion fitnesses are

$$\begin{aligned} s_c &= W(A, c_1) - 1 \\ s_{c|c} &= \frac{1}{2} \left(\omega_1 p + \omega_2 (2 - p - r) + \sqrt{(\omega_1 - \omega_2)^2 p^2 - 2(\omega_1 - \omega_2) \omega_2 p (1 - 2p)r + \omega_2^2 r^2} \right) - 1 \\ \omega_1 &= \left(\frac{W(c_1)}{1/2W(c_1) + 1/2W(c_1, c_2)} \right) \\ \omega_2 &= \left(\frac{W(c_1, c_1, c_2)}{1/2W(c_1) + 1/2W(c_1, c_2)} \right) \\ sc_1 &= \frac{W(sc_1, c_2)}{W(c_1, c_2)} - 1 \\ s_{sc_2} &= \frac{W(sc_1, sc_2)}{W(sc_1, c_2)} - 1 \end{aligned}$$

Pathway **P₂** and **P₃** both involve the acquisition of the subspecialized alleles followed by a tandem duplication of one of the subspecialized alleles. These two pathways differ in terms of the population genetic equilibrium that is approached when both specialized and

subspecialized alleles are segregating. In pathway **P₂** the specialized alleles are completely replaced by the subspecialized alleles, even though they are homozygous lethal. In both cases the waiting time for both subspecialized alleles to become resident from the starting population state of (c_1, c_2) is

$$T_{(c_1, c_2) \rightarrow (c_1, c_2, sc_1, sc_2)} = \frac{6}{N\mu_s} \frac{1}{2s_{sc}},$$

where the invasion selection coefficient for the sc alleles is measured in a population composed of c_1 and c_2 alleles. The actual waiting time will be less than this because the presence of one subspecialized allele actually increases the invasion fitness of the other.

For pathway **P₂** the transition to (sc_1, sc_2) is based solely on the deterministic population genetics and will usually be short enough to be ignored. Once this point is reached, in-phase duplications of either subspecialized allele will have the same marginal fitness as the single-copy haplotypes. The out-of-phase duplication can reach fixation by stochastic tunneling when the a neutral duplicate sojourns and recombines before being lost (IWASA *et al.*, 2004; WEISSMAN *et al.*, 2009; PROULX, 2011). The waiting time for this is

$$T_{(sc_1, sc_2) \rightarrow (sc_1|sc_2)} \approx \frac{1}{N\mu_d} \frac{1}{2\sqrt{r/2s_{sc_1|sc_2}}},$$

where the probability that a rare $sc_1|sc_1$ haplotype recombines to create a $sc_1|sc_2$ haplotype is $r/2$ if the subspecialized alleles are each at frequency of $1/2$ (PROULX, 2011).

For pathway **P₂** the population state (c_1, c_2, sc_1, sc_2) is transient because at population genetic equilibrium the frequencies of c_1 and c_2 go to zero. The the waiting time is

$$\begin{aligned} T_{P_2} &= T_{(A) \rightarrow (c_1, c_2)} + T_{(c_1, c_2) \rightarrow (sc_1, sc_2)} + T_{(sc_1, sc_2) \rightarrow (sc_1|sc_2)} \\ &= \frac{3}{N\mu_c} \frac{1}{2s_c} + \frac{6}{N\mu_s} \frac{1}{2s_{sc}} + \frac{1}{N\mu_d} \frac{1}{2\sqrt{r/2s_{sc_1|sc_2}}}. \end{aligned}$$

Under the assumption of symmetric fitness effects, the values for the invasion fitnesses are

$$\begin{aligned} s_c &= W(A, c_1) - 1 \\ s_{sc} &= \frac{W(sc_1, c_1)/2 + W(sc_1, c_2)/2}{1/2W(c_1) + 1/2W(c_1, c_2)} - 1 \\ s_{sc_1|sc_2} &= \frac{W(sc_1, sc_1, sc_2)/2 + W(sc_1, sc_2, sc_2)/2}{W(sc_1, sc_1)/2 + W(sc_1, sc_2)/2}. \end{aligned}$$

Now consider pathway **P₃** at the point where the population composition is (c_1, c_2, sc_1, sc_2) . When the subspecialized alleles are maintained at equilibrium in the population, duplications of the subspecialized alleles will have mean fitness greater than 1 even when they are in-phase. This is because they have higher marginal fitness than the subspecialized allele

from which they are derived. This can be seen by noting that the duplicate haplotype has increased dosage of the specialized allele in the context in which it is favored, but does not alter the relative dosage of alleles in the other context. Regardless of the haplotype it is paired with this produces higher fitness than the haplotype with one copy of the subspecialized allele. Because the in-phase duplications are favored, we can calculate the sequential probability that the duplication is not lost and reaches its population genetic equilibrium. The waiting time until duplication is given by

$$T_{(c_1, c_2, sc_1, sc_2) \rightarrow (c_1, c_2, sc_2, sc_1 | sc_1)} = \frac{1}{N p_{sc} \mu_d} \frac{1}{2 s_{sc_1 | sc_1}},$$

where p_{sc} represents the combined frequency of the subspecialized alleles and $s_{sc_1 | sc_1}$ is the invasion selection coefficient of a haplotype carrying two copies of the same subspecialized allele. Recombination with either a specialized or subspecialized allele on the other context can now create a stable duplication that can become fixed in the population.

$$T_{(c_1, c_2, sc_2, sc_1 | sc_1) \rightarrow (sc_1 | sc_2)} = \frac{1}{N p_{sc_1 | sc_1} (p_{sc_1 | c_2} + p_{sc_2}) r} \frac{1}{2 s_{sc_1 | sc_2}},$$

where the frequencies of the haplotypes are calculated at population genetic equilibrium. The waiting time for the recombination step can be ignored if the rate of recombination is much larger than the mutation rates. For pathway **P₃** the waiting time is

$$\begin{aligned} T_{P_3} &= T_{(A) \rightarrow (c_1, c_2)} + T_{(c_1, c_2) \rightarrow (c_1, c_2, sc_1, sc_2)} + T_{(c_1, c_2, sc_1, sc_2) \rightarrow (c_1, c_2, sc_2, sc_1 | sc_1)} \\ &\quad + T_{(c_1, c_2, sc_2, sc_1 | sc_1) \rightarrow (sc_1 | sc_2)} \\ &= \frac{3}{N \mu_c} \frac{1}{2 s_c} + \frac{6}{N \mu_s} \frac{1}{2 s_{sc}} + \frac{1}{N p_{sc} \mu_d} \frac{1}{2 s_{sc_1 | sc_1}} + \frac{1}{N p_{sc_1 | sc_1} (p_{sc_1 | c_2} + p_{sc_2}) r} \frac{1}{2 s_{sc_1 | sc_2}}. \end{aligned}$$

Under the assumption of symmetric fitness effects, the values for the invasion fitnesses are

$$\begin{aligned} s_c &= W(A, c_1) - 1 \\ s_{sc} &= \frac{W(sc_1, c_1)/2 + W(sc_1, c_2)/2}{1/2 W(c_1) + 1/2 W(c_1, c_2)} - 1 \\ p_{sc} &= \frac{W(c_1, c_1) + W(c_1, c_2) - W(c_1, sc_1) - W(c_1, sc_2)}{W(c_1, c_1) + W(c_1 + c_2) - 2W(c_1, sc_1) - 2W(c_1, sc_2) + W(sc_1, sc_1) + W(sc_1, sc_2)} \\ s_{sc_1 | sc_1} &= \frac{p_{sc}(W(sc_1) + W(sc_1, sc_2))/2 + (1 - p_{sc})(W(c_1, sc_1, sc_1) + W(c_2, sc_1, sc_1))/2}{\bar{W}}. \end{aligned}$$

Values for $p_{sc_1 | sc_1}$, $p_{sc_1 | c_2}$, p_{sc_2} , and $s_{sc_1 | sc_2}$ can be numerically determined.

2.3 Net Stabilizing Coding Selection

In this scenario, regardless of precisely which path is taken, one step involves the transient presence of subfunctionalized alleles that behave as recessive lethals. Many aspects of

the stochastic dynamics and stationary frequency distribution for recessive lethals have been worked out (NEI, 1968; ROBERTSON and NARAIN, 1971; CROW and KIMURA, 1970). In large populations, the mean frequency of the recessive lethal allele approaches $\sqrt{2\mu}$ (CROW and KIMURA, 1970), but in smaller populations the mean frequency is lower (NEI, 1968). The stationary distribution for the frequency of the recessive lethal mutations is also known from a diffusion approximation approach (NEI, 1968). The number of mutant progeny produced by a single initial mutant before extinction of the mutant lineage can be calculated for finite populations using a Markov chain approach, but there are no known approximations for this distribution (ROBERTSON and NARAIN, 1971).

To find the waiting time for a population fixed for the ancestral allele to gain a sub-specialized allele we could use the stationary distribution for the number of subfunctional alleles present in a given generation and then calculate the per generation probability that no successful subspecialized alleles arise. However, the tunneling approximation for alleles that are selected against shows that the probability of tunneling can be approximated based on the mean frequency of those alleles (PROULX, 2011) and this approach appears to be valid for recessive lethals as well. In particular, the stationary distribution of subfunctional alleles has a mode at 0 if $2N\mu_s < 1$ and is otherwise distributed around a mean value of

$$\bar{p}_s = \frac{\Gamma(N\frac{\mu_s}{2} + \frac{1}{2})}{\sqrt{N}\Gamma(N\frac{\mu_s}{2})}, \quad (4)$$

where the relevant mutation rate is $\mu_s/2$, because that is the probability for each class of subfunctional mutation (s_1 and s_2). So long as $\mu_c 2s_c \ll 1$ and $N\bar{p}_s$ is not too large, the probability that a successful subspecialized allele will arise is approximately linear in the number of subfunctional alleles present and therefore is

$$N \cdot 2\bar{p}_s \cdot \mu_c/2,$$

where the total frequency of subfunctionalized alleles is $2\bar{p}_s$ and the probability of an appropriate coding mutation is $\mu_c/2$.

The waiting time until a secondary mutation produces either the sc_1 or sc_2 allele and that such an allele would then not be lost by drift is

$$T_{(A) \rightarrow (sc_1)} = \frac{1}{\bar{p}_s N \mu_c} \frac{1}{2s_{sc}},$$

where s_{sc} is the invasion fitness of a subspecialized allele. Once the subspecialized allele arises it will be maintained at frequency dependent equilibrium. The next event would be the duplication of the subspecialized allele, which has a waiting time given by

$$T_{(sc_1) \rightarrow (sc_1|sc_1)} = \frac{1}{p_{sc_1} N \mu_d} \frac{1}{2s_{sc_1|sc_1}}.$$

Here, p_{sc_1} is the population genetic equilibrium frequency of the subspecialized allele and is given by $p_{sc_1} = \frac{w(A, sc_1) - 1}{1 + 2(w(A, sc_1) - 1)}$ where I have arbitrarily labeled the first subspecialized allele to arise as sc_1 . The selection coefficient is calculated at the population genetic equilibrium for the subspecialized allele and refers to selection for the haplotype containing two copies of the same subspecialized allele. Note that $s_{sc_1|sc_1}$ is always positive because this haplotype only increases the relative dosage of the coding mutation in the context to which it is adapted. The subspecialized alleles make no contribution to the other context, so their duplication is unconditionally beneficial.

As in the diversifying coding scenario, the fate of the duplication will depend on the recombination rate. If the recombination rate is high then we can calculate the eigenvalue for the spread of new duplicates, taking into account the fact that the spreading duplicate will experience multiple haplotype backgrounds. If the rate of recombination is low then the duplication may spread to high frequency before a successful recombination event. If r is small then the waiting time for recombination to create a stable duplicate haplotype is

$$T_{(sc_1|sc_1) \rightarrow (A|sc_1)} = \frac{1}{p_{sc_1|sc_1}(1 - p_{sc_1|sc_1})Nr} \frac{1}{2s_{A|sc_1}},$$

where $p_{sc_1|sc_1}$ is the population genetic equilibrium for the subspecialized duplicate haplotype, r is the recombination rate, and $s_{A|sc_1}$ refers to the selection coefficient for the out of phase duplicate haplotype. Once a duplicate haplotype is formed that contains a subspecialized allele and an allele that is expressed in the other tissue the duplication is considered stable. The total time is then

$$T_{P_1} = T_{(A) \rightarrow (sc_1)} + T_{(sc_1) \rightarrow (sc_1|sc_1)} + T_{(sc_1|sc_1) \rightarrow (A|sc_1)} \quad (5)$$

$$= \frac{1}{\bar{p}_s N \mu_c} \frac{1}{2s_{sc}} + \frac{1}{p_{sc_1} N \mu_d} \frac{1}{2s_{sc_1|sc_1}} + \frac{1}{p_{sc_1|sc_1}(1 - p_{sc_1|sc_1})Nr} \frac{1}{2s_{A|sc_1}}. \quad (6)$$

The subspecialized allele that becomes resident first is arbitrarily labeled sc_1 . Under the assumptions that $W(A) = 1$ and symmetric fitness effects in the two contexts the interme-

diate calculations are

$$\begin{aligned}
\bar{p}_s &= \frac{\Gamma(N\frac{\mu_s}{2} + \frac{1}{2})}{\sqrt{N}\Gamma(N\frac{\mu_s}{2})} \\
s_{sc_1} &= W(A, sc_1) - 1 \\
p_{sc_1} &= \frac{W(A, sc_1) - 1}{1 + 2(W(A, sc_1) - 1)} \\
s_{sc_1|sc_1} &= (1 - p_{sc_1})W(A, sc_1, sc_1) - \left(\frac{W(A, sc_1)^2}{2W(A, sc_1) - 1} \right) \\
p_{sc_1|sc_1} &= \frac{W(A, sc_1, sc_1) - 1}{1 + 2(W(A, sc_1, sc_1) - 1)} \\
s_{A|sc_1} &= p_{sc_1|sc_1}W(A, sc_1, sc_1) + (1 - p_{sc_1|sc_1})W(A, A, sc_1) - \left(\frac{W(A, sc_1, sc_1)^2}{2W(A, sc_1, sc_1) - 1} \right).
\end{aligned}$$

Pathways **P₂** and **P₃** involve the acquisition of both subspecialized alleles before a duplication becomes resident (figure 4 in the main text). The probability of joint acquisition is just 3/2 times the probability of acquiring just one subspecialized allele, so that

$$T_{(A) \rightarrow (sc_1, sc_2)} = \frac{3}{2\bar{p}_s N \mu_c} \frac{1}{2s_{sc}}.$$

Following the branching point B₂, the pathways are symmetrical so we can calculate the waiting time down either of them. Because two subspecialized alleles are resident, the probability of either one duplicating must be taken into account. The mean frequency of the subspecialized alleles is slightly higher when both are present, so the rate of successful duplication is more than twice as large as in pathway **P₁**. This gives

$$T_{(sc_1, sc_2) \rightarrow (sc_1|sc_1)} = \frac{1}{(p_{sc_1} + p_{sc_2})N\mu_d} \frac{1}{2s_{sc_1|sc_1}},$$

where without loss of generalization the first allele to duplicate is labeled sc_1 .

Once the duplication $sc_1|sc_1$ is resident in the population it will achieve population genetic equilibrium where the frequency of non-duplicate sc_1 alleles goes to zero but A and sc_2 alleles are deterministically maintained. At this point a recombination event could create either $A|sc_1$ or $sc_1|sc_2$ haplotypes and both have positive invasion selection coefficients. Because the frequency of the A alleles is typically much higher than the frequency of the sc_2 allele and because the invasion fitness of the $sc_1|sc_2$ haplotype is larger than the invasion fitness of the $A|sc_1$ haplotype we can use the waiting time for a $A|sc_1$ allele to become resident as an upper bound on the waiting time for maintenance of a stable duplicate haplotype. This is

$$T_{(sc_1|sc_1) \rightarrow (A|sc_1)} = \frac{1}{p_{sc_1|sc_1}(1 - p_{sc_1|sc_1})Nr} \frac{1}{2s_{A|sc_1}}.$$

The total waiting time to go down either pathway **P₂** or **P₃** is

$$T_{P_{2|3}} = T_{(A) \rightarrow (sc_1, sc_2)} + T_{(sc_1, sc_2) \rightarrow (sc_1|sc_1)} + T_{(sc_1|sc_1) \rightarrow (A|sc_1)} \quad (7)$$

$$= \frac{3}{2\bar{p}_s N \mu_c} \frac{1}{2s_{sc.}} + \frac{1}{2p_{sc.} N \mu_d} \frac{1}{2s_{sc_1|sc_1}} + \frac{1}{p_{sc_1|sc_1} (1 - p_{sc_1|sc_1}) N r} \frac{1}{2s_{A|sc_1}}, \quad (8)$$

where $p_{sc.} = p_{sc_1} + p_{sc_2}$ represents the total frequency of subspecialized alleles and $s_{sc.}$ represents the selection coefficient for either subspecialized allele. Under the assumptions that $W(A) = 1$ and symmetric fitness effects in the two contexts we have

$$\begin{aligned} \bar{p}_s &= \frac{\Gamma(N \frac{\mu_s}{2} + \frac{1}{2})}{\sqrt{N} \Gamma(N \frac{\mu_s}{2})} \\ s_{sc.} &= W(A, sc.) - 1 \\ p_{sc.} &= \frac{W(A, sc_1) - 1}{W(sc_1, sc_2) + 2(W(A, sc_1) - 1)} \\ s_{sc_1|sc_1} &= \frac{(1 - 2p_{sc.})W(A, sc_1, sc_1) + p_{sc.}W(A, sc_1, sc_1, sc_1) + p_{sc.}W(A, sc_1, sc_1, sc_2)}{(1 - 2p_{sc.})W(A, sc_1) + p_{sc.}W(sc_1, sc_2)} - 1 \\ p_{sc_1|sc_1} &= \frac{W(A, sc_i, sc_i) - 1}{1 + 2(W(A, sc_i, sc_i) - 1)} \\ s_{A|sc_1} &= p_{sc_1|sc_1}W(A, sc_i, sc_i) + (1 - p_{sc_1|sc_1})W(A, A, sc_i) - \left(\frac{W(A, sc_i, sc_i)^2}{2W(A, sc_i, sc_i) - 1} \right). \end{aligned}$$

The pathways for the waiting times for divergence under stabilizing coding selection can also involve transitions down both parallel and serial pathways. To calculate the total waiting time until a stable duplication is fixed we need to construct a formula for this situation. The process can proceed by either sc_1 or sc_2 becoming resident first. Following this, either the other subspecialized allele or a duplication of the resident subspecialized allele may occur. If both subspecialized alleles become resident then the probability of a duplication is approximately twice as large because there are now two alleles each maintained at frequency dependent selection balance. The presence of both subspecialized alleles has a slight positive effect on the frequency of each subspecialized allele, but this is generally small enough to be ignored.

Call the per generation probability of sc_1 becoming resident λ_1 , the per generation probability of sc_2 becoming resident λ_2 , and the per generation probability of $sc_i|sc_i$ becoming resident λ_3 (given that sc_i is already resident). To calculate the total waiting time we must integrate the probability of each path taking an amount of time x multiplied by the time it took over all waiting times and each pathway. Consider first the contribution to the integral when sc_1 becomes resident first and duplication precedes the maintenance

of sc_2 .

$$\begin{aligned} T(sc_1, sc_1|sc_1) &= \int_0^\infty \int_{x_1}^\infty x_3 e^{-x_1 \lambda_2} \lambda_1 e^{-x_1 \lambda_1} e^{-(x_3-x_1) \lambda_2} \lambda_3 e^{-(x_3-x_1) \lambda_3} dx_3 dx_1 \\ &= \frac{\lambda_1 \lambda_3 (\lambda_1 + 2\lambda_2 + \lambda_3)}{(\lambda_1 + \lambda_2)^2 (\lambda_2 + \lambda_3)^2} \end{aligned}$$

Now consider the contribution to the integral when sc_1 becomes resident first, sc_2 becomes resident next and duplication follows. In this case, the probability of duplication increases to $2\lambda_3$ once both sc_1 and sc_2 are maintained.

$$\begin{aligned} T(sc_1, sc_2, sc_i|sc_i) &= \int_0^\infty \int_{x_1}^\infty \int_{x_2}^\infty x_3 e^{-x_1 \lambda_2} \lambda_1 e^{-x_1 \lambda_1} \lambda_2 e^{-(x_2-x_1) \lambda_2} \\ &\quad e^{-(x_2-x_1) \lambda_3} 2\lambda_3 e^{-(x_3-x_2) 2\lambda_3} dx_3 dx_2 dx_1 \\ &= \frac{\lambda_1 \lambda_2 (\lambda_2^2 + 5\lambda_3 \lambda_2 + 2\lambda_3^2 + \lambda_1 (\lambda_2 + 3\lambda_3))}{2(\lambda_1 + \lambda_2)^2 \lambda_3 (\lambda_2 + \lambda_3)^2}. \end{aligned}$$

By symmetry $\lambda_2 = \lambda_1$. The total waiting time is

$$\begin{aligned} T(sc_1, sc_1|sc_1) + T(sc_1, sc_2, sc_i|sc_i) + T(sc_2, sc_2|sc_2) + T(sc_2, sc_1, sc_i|sc_i) \\ = \frac{1}{2} \left(\frac{1}{\lambda_1} + \frac{1}{\lambda_3} + \frac{1}{\lambda_1 + \lambda_3} \right). \end{aligned}$$

A similar calculation can be done to find the variance in waiting time.

Using this approach we find that the total time is given by

$$T_{P_1|2|3} = T_{(A) \rightarrow (sc_1|sc_1)} + T_{(sc_1|sc_1) \rightarrow (A|sc_1)} \quad (9)$$

$$\begin{aligned} &= \frac{1}{2} \left(\frac{1}{\bar{p}_s N \mu_c} \frac{1}{2s_{sc.}} + \frac{1}{p_{sc.} N \mu_d} \frac{1}{2s_{sc_1|sc_1}} + \frac{1}{(\bar{p}_s N \mu_c 2s_{sc.}) + (p_{sc.} N \mu_d 2s_{sc_1|sc_1})} \right) \\ &\quad + \frac{1}{p_{sc_1|sc_1} (1 - p_{sc_1|sc_1}) N r} \frac{1}{2s_{A|sc_1}}. \end{aligned} \quad (10)$$

For simplicity (and as an upper bound on the waiting time) the equilibrium frequencies and invasion fitnesses from pathway $\mathbf{P}_1$ are used for all pathways.

References

- CROW, J. F. and M. KIMURA, 1970 *An introduction to population genetics theory*. New York: Harper & Row.
- HAMMERSTEIN, P., 1996 Darwinian adaptation, population genetics and the streetcar theory of evolution. *J. Math. Biol.* **34**(5-6): 511–532.

- IWASA, Y., F. MICHOR, and M. A. NOWAK, 2004 Stochastic tunnels in evolutionary dynamics. *Genetics* *166*(3): 1571–9.
- NEI, M., 1968 The frequency distribution of lethal chromosomes in finite populations. *P Natl Acad Sci Usa* *60*(2): 517–24.
- OTTO, S. P. and P. YONG, 2002 The evolution of gene duplicates. *Adv. Genet.* **46**: 451–483.
- PROULX, S. and P. PHILLIPS, 2006, Jan) Allelic divergence precedes and promotes gene duplication. *Evolution* *60*(5): 881–892.
- PROULX, S. R., 2011 The rate of multi-step evolution in Moran and Wright-Fisher populations. *Theor. Pop. Biol.*
- ROBERTSON, A. and P. NARAIN, 1971 The survival of recessive lethals in finite populations. *Theoretical population biology* *2*(1): 24–50.
- ROSS, S., 1988 *A first course in probability*. New York: Macmillan.
- WEINREICH, D. and L. CHAO, 2005, Jan) Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution* *59*(6): 1175–1182.
- WEISSMAN, D. B., M. M. DESAI, D. S. FISHER, and M. W. FELDMAN, 2009 The rate at which asexual populations cross fitness valleys. *Theoretical population biology* *75*(4): 286–300.

File S2

Mathematica File

File S2 is available for download as a compressed folder at

<http://www.genetics.org/content/suppl/2011/12/05/genetics.111.135590.DC1>