DOI: 10.1534/genetics.104.027607

The Origin of Subfunctions and Modular Gene Regulation

Allan Force,*[,1] William A. Cresko,[†,‡] F. Bryan Pickett,[§] Steven R. Proulx,[‡] Chris Amemiya* and Michael Lynch**

**Benaroya Research Institute at Virginia Mason, Seattle, Washington 98101, [†]Institute of Neuroscience, University of Oregon, Eugene, Oregon 97403, [§]Department of Biology, Loyola University, Chicago, Illinois 60626, [‡]Center for Ecology and Evolutionary Biology, University of Oregon, Eugene, Oregon 97403 and **Department of Biology, Indiana University, Bloomington, Indiana 47405*

Manuscript received February 12, 2004
Accepted for publication February 8, 2005

ABSTRACT

Evolutionary explanations for the origin of modularity in genetic and developmental pathways generally assume that modularity confers a selective advantage. However, our results suggest that even in the absence of any direct selective advantage, genotypic modularity may increase through the formation of new subfunctions under near-neutral processes. Two subfunctions may be formed from a single ancestral subfunction by the process of fission. Subfunction fission occurs when multiple functions under unified genetic control become subdivided into more restricted functions under independent genetic control. Provided that population size is sufficiently small, random genetic drift and mutation can conspire to produce changes in the number of subfunctions in the genome of a species without necessarily altering the phenotype. Extensive genotypic modularity may then accrue in a near-neutral fashion in permissive population-genetic environments, potentially opening novel pathways to morphological evolution. Many aspects of gene complexity in multicellular eukaryotes may have arisen passively as population size reductions accompanied increases in organism size, with the adaptive exploitation of such complexity occurring secondarily.

EUKARYOTIC gene regulation is a remarkably complex process, with each gene displaying multiple functions in discrete tissues and times during development. To accomplish such tasks, the noncoding DNA of individual genes often harbors numerous small *cis*-acting elements that cooperatively interact with multiple *trans*-acting factors to tune levels of transcription (Davidson 2001). Mutations in these regulatory regions may influence many aspects of phenotypic evolution by imposing the loss or gain of gene expression (Raff 1996; Gerhart and Kirschner 1997; Force *et al.* 1999; Carroll 2001; Carroll *et al.* 2001). Over evolutionary time, an increase in the particulate nature of gene regulation seems to correlate with the subdivision and specialization of body plans of multicellular organisms, leading to organisms in which traits are capable of following independent evolutionary trajectories (Wagner 1996; Wagner and Altenberg 1996; Raff and Sly 2000). Increases in the particulate nature of gene regulation that affect the structure of developmental networks may be thought of as increases in genotypic modularity, while the subdivision and specialization of body regions at the phenotypic level may be thought of as increases in phenotypic modularity. Although it is tempting to view regulatory-region complexity as a prerequisite for the adaptive origin of morphological complexity, the causal link between genotypic and phenotypic modularity remains unclear, and a formal theoretical framework for the evolutionary origin of regulatory gene structure remains to be developed.

Renewed interest in the evolutionary fates of duplicate genes (Piatgorsky and Wistow 1991; Clark 1994; Hughes 1994; Walsh 1995; Sidow 1996; Nowak *et al.* 1997; Wagner 1994, 1998; Force *et al.* 1999; Stoltzfus 1999; Lynch and Force 2000; Lynch *et al.* 2001; Wagner 2001; Rodin and Riggs 2003) has resulted in the development of models that explicitly incorporate the complex, multifunctional organization of eukaryotic genes. For example, under the duplication-degeneration-complementation (DDC) model (Force *et al.* 1999; Lynch and Force 2000; Lynch *et al.* 2001), genes are posited to contain independently mutable subfunctions that can be partitioned among descendant copies following a gene-duplication event. A *gene subfunction* has been defined as an independently mutable function of a gene that falls into a distinct complementation class (Force *et al.* 1999). The defining characteristic of a subfunction is not the number or types of its DNA components, but their integrated operation in performing a task that is mutationally independent of other suites of functionally integrated elements acting at the same locus. A subfunction component may correspond to regulatory elements (*e.g.*, transcription-factor binding

[1]*Corresponding author:* Benaroya Research Institute at Virginia Mason, 1201 Ninth Ave., Seattle, WA 98101.
E-mail: force@vmresearch.org

sites), splice junctions, mRNA stability elements, and/or coding regions (*e.g.*, functional motifs), among other possibilities (Force *et al.* 1999, 2004). Under the general DDC model, a variety of population-level mechanisms may lead to duplicate-gene preservation and the partitioning of gene subfunctions (Piatgorsky and Wistow 1991; Hughes 1994; Force *et al.* 1999; Lynch and Force 2000; Lynch *et al.* 2001; Adams *et al.* 2003; Rodin and Riggs 2003). However, although these mechanisms might explain the evolutionary fates of a large fraction of duplicate genes in metazoans and vascular plants, they also beg the question—How do new gene subfunctions arise in the first place? In this article we present a model for the origin of new regulatory subfunctions by a near-neutral process via cycles of information accretion and loss at individual loci. Our intention is to show how mutation, duplication, and genetic drift can drive the evolution of genotypic modularity.

In this article, we restrict our attention to the evolution of new regulatory subfunctions. Several studies on the structure of genetic networks and regulatory regions have provided clues as to the proper structure of models for the evolution of modularity at the level of gene regulation. Small regulatory elements can arise by *de novo* mutation or by transpositional insertion, providing many potential degrees of freedom for altering the number and type of transcription-factor binding sites (Arnosti *et al.* 1996; Wray 1998; Yuh *et al.* 1998; Brosius 1999; von Dassow and Monro 1999; Edelman *et al.* 2000; Stone and Wray 2001; MacArthur and Brookfield 2004). If various permutations of such elements provide for functionally equivalent outputs of a gene (see Edelman *et al.* 2000), then it follows that the stochastic turnover of control elements by nearly neutral processes may play a critical role in the evolution of regulatory regions (Bonneton *et al.* 1997; Ludwig *et al.* 1998; Hancock *et al.* 1999; Ludwig *et al.* 2000; McGregor *et al.* 2001; Shaw *et al.* 2001). Although some regulatory-region structures may endow their associated alleles with higher fitness than others, the range of effectively equivalent states will necessarily increase in populations with smaller size where the efficiency of selection is diminished and the intrusion of new transcription-factor binding sites into the system will proceed passively.

A CONCEPTUAL MODEL FOR SUBFUNCTION FISSION

There are two broad views of how new regulatory elements may be incorporated into genes to form new regulatory subfunctions: *subfunction cooption* and *subfunction fission* (see Raff 1996; Carroll *et al.* 2001; Davidson 2001; Force *et al.* 2004). Subfunction cooption involves the evolution of a new function not carried out by the ancestral gene, whereas subfunction fission involves subdivision of a function already present in the ancestral gene. The effects of subfunction cooption may range from the altered expression of a single structural gene to the dramatic activation of a whole developmental pathway in a new location. Although cooption of *new* gene functions and expression domains is widely discussed in the literature (see Raff 1996; Carroll *et al.* 2001; Davidson 2001) and has undoubtedly played an important role in the evolution of new morphological structures, cooption may not be the most common pathway for the origin of new regulatory subfunctions.

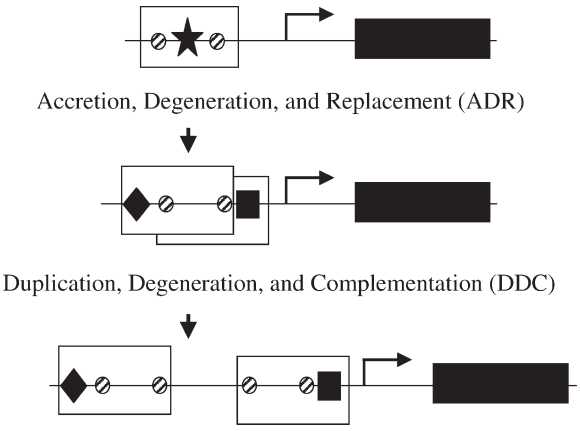

Figure 1.—General outline of the subfunction fission model. During phase 1, accretion of two new regulatory elements (diamond and square) occurs, with each redundantly driving a portion of the ancestral expression domain, which was previously under the control of a single positive ancestral element (star). Subsequently, the ancestral shared element (star) degenerates, resulting in the replacement of the ancestral shared regulation with semi-independent regulation. In phase 2, duplication, degeneration, and complementation of the regulatory region lead to two independent regulatory modules, each driving independent expression domains and functions. Hatched circles represent shared positive and negative regulatory sites that are required for all functions.

In contrast to subfunction cooption, subfunction fission occurs when *multiple* functions under shared genetic control evolve to be under *independent* genetic control. The pattern of tissue-specific gene expression remains conserved during subfunction fission while the underlying molecular mechanisms for achieving that pattern undergo evolutionary modification. Two general phases may contribute to subfunction fission: (1) the replacement of completely shared regulatory binding sites with independent binding sites to form a semi-independent enhancer and (2) duplication of the semi-independent enhancers, followed by the formation of two entirely independent regions, each critical to a single regulatory subfunction, by complementary degenerative mutations (Figure 1). Phase 1 involves accretion, degeneration, and the replacement (ADR) of ancestral transcription-factor binding sites and phase 2 involves the DDC of binding sites within enhancers. The series of events involved in phase 2 are essentially the same as those underlying the subfunctionalization of gene duplicates, except in this case the duplicated region comprises just the enhancers within a gene.

Under subfunction fission, the new gene architecture diverges beneath a constant phenotype. Despite this initial invariance of expression patterns, subfunction fission may open up previously inaccessible evolutionary pathways by eliminating some pleiotropic constraints associated with shared regulatory regions while creating others. Therefore, the evolutionary potential of such alterations may not be realized until a new selective environment and/or appropriate mix of mutations is encountered, at which point modularity at the level of gene architecture may promote the evolution of phenotypic modularity. However, the model that we present highlights the logical distinction between the causal nonadaptive forces that may lead to the restructuring of genomic architecture and the secondary consequences of such change for phenotypic evolution. We now formalize the theory for the two phases contributing to subfunction fission.

Phase 1—accretion, degeneration, and replacement: We start by considering the process by which an allele with two overlapping regulatory subfunctions arises from an allele with one regulatory subfunction (Figure 2). We assume a starting point where several transcription factors are present, some of which are general and some tissue specific. For an initial shared regulatory state in which the same positive transcription factor (TF), TFA, drives the gene's expression in two tissues via the same binding site (A), the ancestral allele has only one subfunction because degenerative mutations in binding site A reduce expression in both domains similarly. We further assume the presence of tissue-specific transcription factors (TFB and TFC), expressed in a complementary manner (one in each tissue) with respect to the ancestral expression of TFA. The existence of these tissue-specific transcription factors provides the essential setting in which an allele using TFA as an activator of expression can give rise to a semi-independently regulated derivative requiring both TFB and TFC (Figure 2). Such a transition requires the addition of binding sites for TFB and TFC, followed by the loss of binding sites for TFA. Although positive Darwinian selection may directly promote subfunction formation, we restrict our attention in this article to a near-neutral process.

If we assume that binding sites for all three types of transcription factor are subject to mutational accretion and degeneration, then there is no permanent allelic state under this model, as the alternative classes of shared and semi-independently regulated alleles are free to drift in frequency. However, for heuristic purposes, we first focus on the case in which the population is initially fixed for the shared regulated allele and evaluate the expected time to a transient state of fixation by the semi-independently regulated allele. Even for the simple model outlined above, there are nine alternative allelic states (Figure 3), eight representing all possible permutations of the three types of transcription-factor binding sites (A, B, and C) and an additional coding-region null allele. We assume that each transcription-factor binding site is added to an allele at rate μ_a and deleted at rate μ_d. We denote the presence of a site with an uppercase letter and the absence of a site with a lowercase letter. Thus, for example, the *Abc* allele containing only binding site A becomes either an *ABc* allele at rate μ_a or an allele without any sites *abc* (but still having an intact coding region) at rate μ_d. In addition, each expressed allele mutates to the coding null class, denoted by *xxx*, at rate μ_c. The *abc* and *xxx* alleles are functionally equivalent but differ in their ability to mutate back into a viable state.

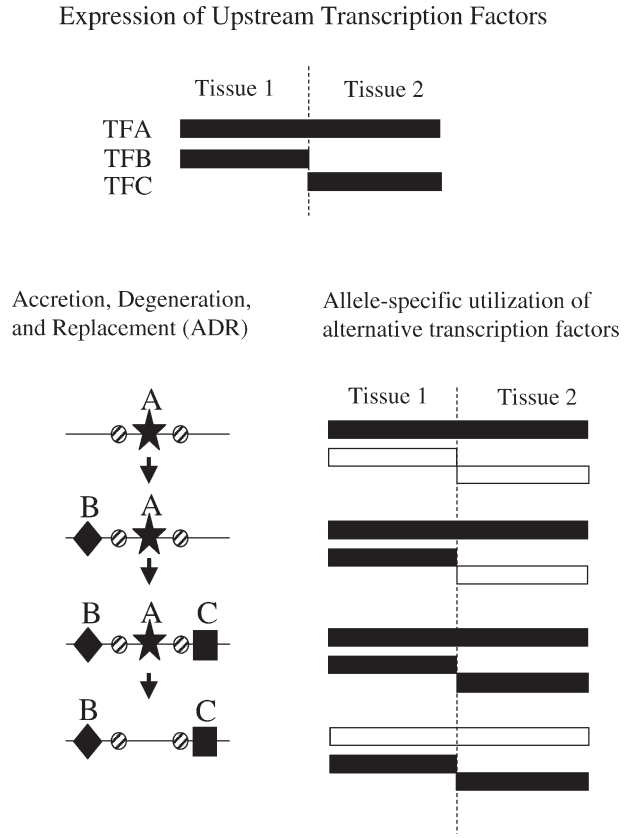

FIGURE 2.—Subfunction fission phase 1: accretion, degeneration, and replacement (ADR). (Top) Positive transcription factors are expressed in both tissues 1 and 2 (TFA) or only in tissue 1 (TFB) or tissue 2 (TFC). (Bottom) The ancestral enhancer (left) drives expression (right) in both tissues via a regulatory element (A). Consecutive regulatory element accretion events involving B and C sites produce an enhancer that redundantly drives expression in both tissues. Degenerative mutations lead to the loss of the ancestral positive regulatory element (A), resulting in an enhancer with semi-independent regulation and two regulatory subfunctions (B and C). The star, diamond, and square represent the A, B, and C binding sites as in Figure 1. Hatched circles represent shared positive and negative regulatory sites that are required for all functions.

To simplify the following small-population-size approximation, we exclude all nonfixable classes of alleles. Only alleles containing minimally an A site or both B and C sites may go to fixation, so this reduces our consideration to just the *Abc*, *ABc*, *AbC*, *ABC*, and *aBC* alleles. It is convenient to further group these alleles

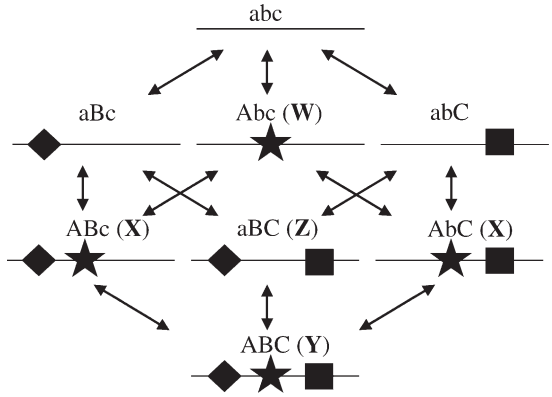

FIGURE 3.—The alleles and mutational pathways for ADR. The eight alleles containing functional binding sites and their possible transitions are shown. The star, diamond, and square represent the A, B, and C binding sites as in Figure 1. We denote the presence of a site with an uppercase letter and the absence of a site with a lowercase letter. The shortest path for the ADR process involves the transition of alleles from the W → X → Y → Z classes.

into four classes: the W class containing the *Abc* allele, the X class containing the *ABc* and *AbC* alleles, the Y class containing the *ABC* allele, and the Z class containing the *aBC* allele. The shortest path from an ancestral state fixed for the shared regulated *Abc* allele to a fixed state involving the semi-independently regulated *aBC* allele involves just three sequential transitions: W to X, X to Y, and finally Y to Z. However, many other longer paths can lead to the same end result. Assuming a sufficiently small population size, all transitions between alternative allelic states will proceed independently in an effectively neutral fashion, and a transition matrix can be used to obtain the entire distribution of transition times between these (or any other) two states.

To clarify the definition of the various transition probabilities, we first consider the elements necessary for the shortest (W → X → Y → Z) route. First, because the *Abc* allele mutates to either the *ABc* or the *AbC* allele at rate μ_a, and because the expected time to neutral fixation is $4N$ generations (where N is the effective population size), the expected transition time from state W to the adjacent state X is $[1/(2\mu_a) + 4N]$, and the approximate per-generation probability of transition between these two states is the reciprocal of this quantity. Second, although the transition to the X class may involve either an *ABc* or an *AbC* allele, in both cases, the formation of the Y class results from the addition of a single binding site, so the approximate per-generation probability of this transition is $[(1/\mu_a) + 4N]^{-1}$. Finally, the transition from the Y to the Z class involves the loss of the A site, so the approximate transition probability is $[(1/\mu_d) + 4N]^{-1}$. Conditional on taking the direct route W → X → Y → Z, the mean time for the transition from the *Abc* to the *aBC* state is the sum of the three stepwise transition times. However, because the W ↔ X and X ↔ Y transitions are reversible, the average time to conversion to the semiregulated state is necessarily larger than that obtained by the shortest path.

To account for all potential paths, we use the transition-matrix **P**, with the element in the ith row and jth column, P_{ij}, denoting the probability of transition from state j to i. For our particular application, the four rows and columns are the classes W, X, Y, and Z. Following from the logic developed in the preceding paragraph, the nonzero off-diagonal elements are $P_{WX} = [(1/\mu_d) + 4N]^{-1}$, $P_{XW} = [(1/(2\mu_a)) + 4N]^{-1}$, $P_{XY} = [(1/(2\mu_d)) + 4N]^{-1}$, $P_{YX} = [(1/\mu_a) + 4N]^{-1}$, and $P_{ZY} = [(1/\mu_d) + 4N]^{-1}$. All remaining off-diagonal entries are equal to zero, and columns of elements must sum to one, so $P_{WW} = 1 - P_{XW}$, $P_{XX} = 1 - P_{WX} - P_{YX}$, and $P_{YY} = 1 - P_{XY} - P_{ZY}$. Note that because our interest here is simply in the time to first arrival at state Z, we treat this final state as an absorbing boundary, even though class Z can mutate back to Y, so $P_{ZZ} = 1$. The mean time to move from the fixed *Abc* state to the fixed *aBC* state is obtained by recursively multiplying the transition matrix by the column vector $[p_W, p_X, p_Y, p_Z]^T$, where the elements denote the probability that the population is in each state, starting with the vector $[1, 0, 0, 0]^T$. The time to first arrival at the fixed *aBC* state is then

$$\bar{t}_F = \sum_{t=1}^{\infty} (p_{Z,t} - p_{Z,t-1})\, t. \qquad (1)$$

This is a first-order approximation, as it assumes instantaneous transitions between monomorphic states, ignoring the complexities of the polymorphic transition between fixed states, which can become important at large population sizes (see below).

To examine the validity of this small-population-size approximation (Figures 4 and 5), we compared the analytical results with those obtained by stochastic simulations. In this article we assume an ideal random-mating population where N is equal to the effective population size. The simulations kept track of genotype frequencies after random sampling of gametes in diploid, sexual populations. We assumed that homozygous recessive genotypes lacking expression in either tissue had a fitness of zero, whereas all other genotypes were assigned a fitness of one. The simulations were started with the *Abc* allele at a frequency of one and stopped when the sum of frequencies of all A-site-bearing alleles (*Abc*, *ABc*, *AbC*, and *ABC*) equaled zero. Provided that $4N\mu_a \ll 1$ and $4N\mu_d \ll 1$ (*i.e.*, the power of random genetic drift is well in excess of the mutation rates), the purely neutral theory provides a very good approximation to the time to first arrival at the fixed state of the semi-independently regulated *aBC* allele (Figure 4). At any given mutation rates in sufficiently small populations, the transition time is essentially independent of population size, because the rates of movement between alternative states are primarily determined by the waiting times for muta-

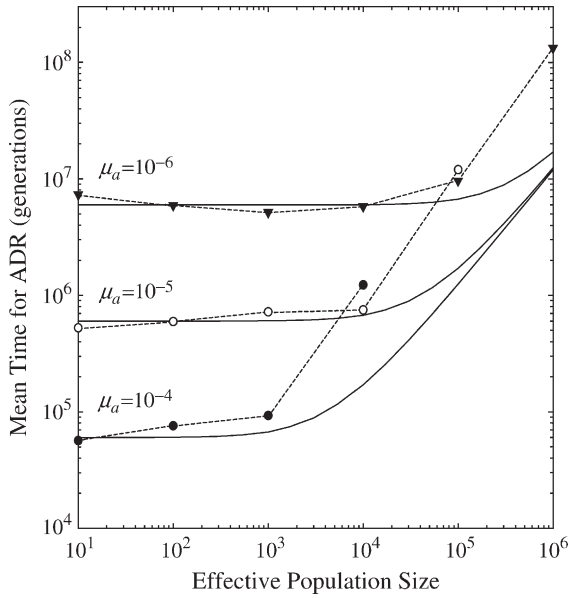

FIGURE 4.—Time to transient fixation of a semi-independently regulated *aBC* allele, starting from a state of fixation for the shared regulated *Abc* allele during ADR. Solid lines denote the solution of Equation 1 in the text, whereas the data points connected by dotted lines were obtained by simulation, as described in the text. Here the ratio of forward and reverse mutation rates was held constant ($\mu_a/\mu_d = 1$), and the mutation rate to coding-region null alleles was $\mu_c = 0.00001$. For most data points, ≥100 replicates were run.

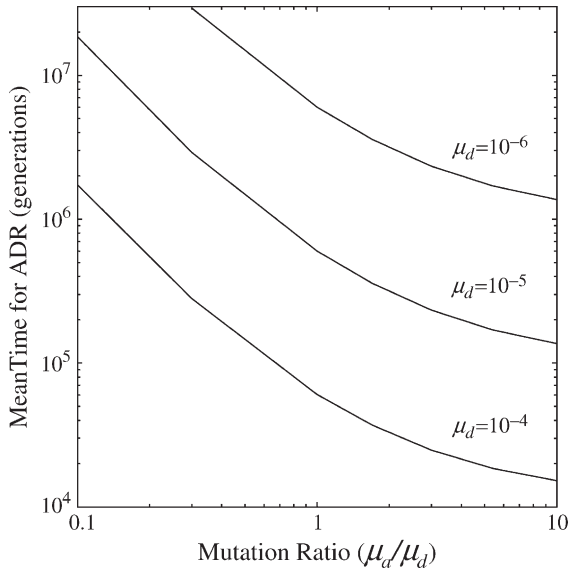

FIGURE 5.—Time to transient fixation of a semi-independently regulated *aBC* allele, starting from a state of fixation for the shared regulated *Abc* allele in a small population during ADR [small = $N(\mu_a + \mu_d) \leq 0.1$). Results are given for various values of μ_a for three fixed values of μ_d.

tions, rather than by the time for such mutations to drift to fixation. In general, the time to subfunction fission declines as the ratio of μ_a to μ_d increases, eventually reaching an asymptote at $1/\mu_d$ generations at high μ_a/μ_d, as the final degenerative step involving the Y → Z transition becomes the limiting factor (Figure 5).

The conditions $4N\mu_a \ll 1$ and $4N\mu_d \ll 1$ may be frequently met in eukaryotes. We know that the average value of $4N\mu$, where μ is the substitutional mutation rate per nucleotide, is $\tilde{0}.002$ for vertebrates, $\tilde{0}.010$ for invertebrates and land plants, and $\tilde{0}.1$ for eukaryotic microbes (LYNCH and CONERY 2003). Thus, given that transcription-factor binding sites typically contain 4–20 nucleotides, we can expect $4N\mu_d$ to be on the order of 4–20 times $4N\mu$ and hence $\ll 1$ for most multicellular species. Although new sites may arise in regions surrounding the existing enhancer, μ_a is generally unlikely to greatly exceed μ_d; therefore $4N\mu_a$ is also expected to be $\ll 1$. These rough approximations suggest that the population-genetic environments of most multicellular organisms enable alternative alleles like those in Figure 3 to drift freely back and forth to transient states of fixation in an effectively neutral fashion.

For larger N, the neutral theory progressively underestimates the transition time to fixed states (Figure 4). There appear to be two reasons for this behavior. First, because a semi-independently regulated allele has an additional transcription-factor binding site relative to the ancestral shared regulated allele and mutates at twice the rate to a nonfixable allele, the former is at a weak selective disadvantage of order μ_d. Second, because of the reversibility of mutations and the extended time to fixation with increasing N, arrival at the stopping criterion for our simulations of a completely monomorphic state becomes increasingly unlikely, and focusing on such an extreme state as an indicator of the availability of *aBC* alleles becomes increasingly misleading.

Because large populations will typically harbor polymorphisms involving the full spectrum of alleles in Figure 3, an alternative way to consider the potential for a locus to undergo an ADR transition is to consider the equilibrium distribution of average allele frequencies. To accomplish this, we ran simulations of single replicate populations and averaged the frequencies of the alleles over a large number of generations (Figure 6, top, bottom left). The clear result is that the average frequency of the semi-independently regulated *aBC* allele decreases with increasing population size, irrespective of the specific mutation rates μ_a and μ_d. Figure 6, bottom right, shows the infinite population size equilibrium frequencies of the four allele classes (see the APPENDIX for the analytical solution). Such behavior results from the increased efficiency of mutationally induced selection against *aBC* alleles at large N relative to the less mutationally sensitive alleles *ABC*, *ABc*, *AbC*, and *Abc*. When the rate of accretionary mutation equals or exceeds the rate of degenerative mutation, the redundantly regulated *ABC* allele dominates, as a consequence of the mutational pressure toward gain of bind-

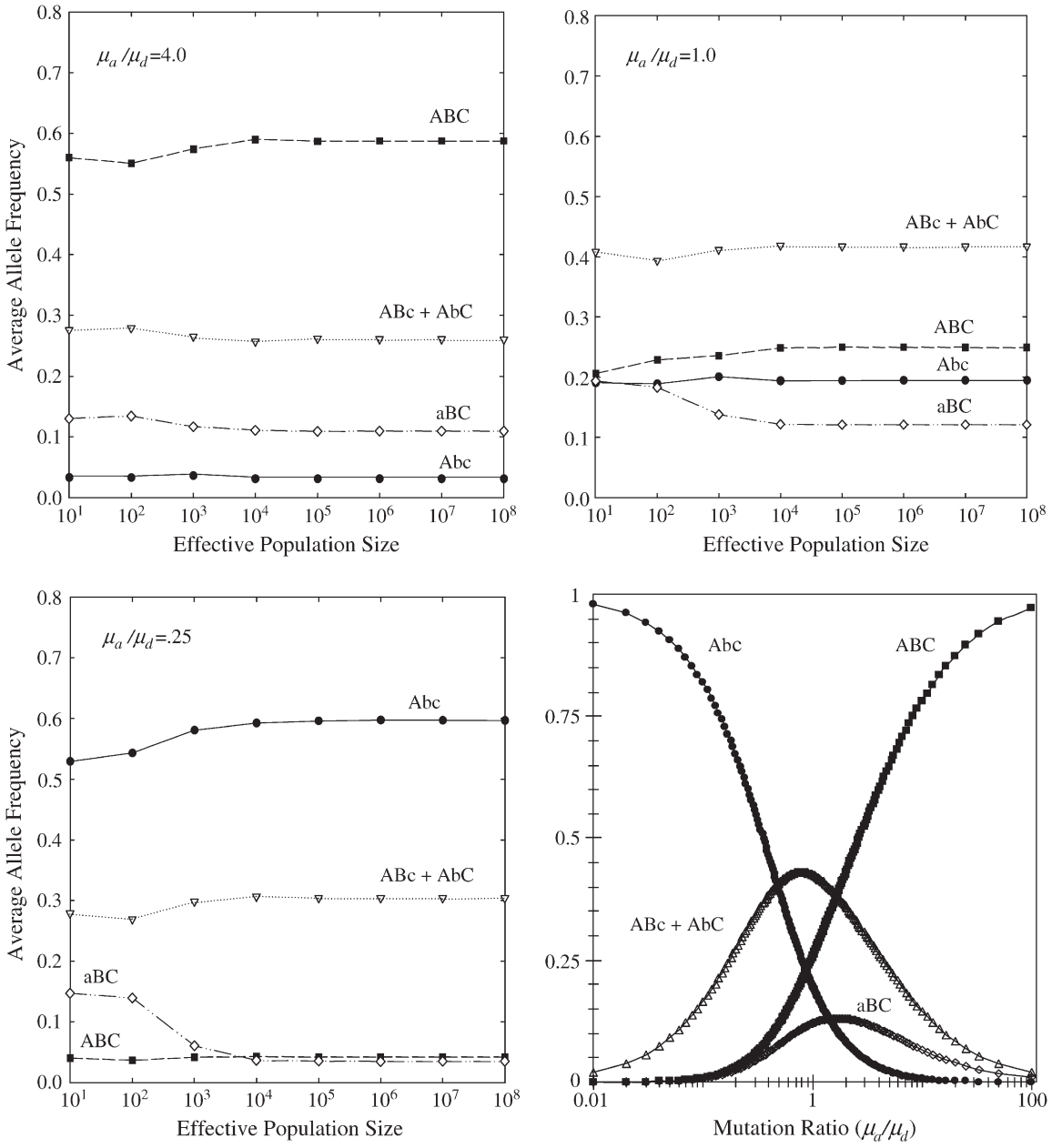

FIGURE 6.—Average frequencies for the four classes of fully functional alleles, as a function of effective population size for three mutation ratios (all with $\mu_a + \mu_d = 0.0005$ and $\mu_c = 0.0001$) and as a function of the mutation ratio for infinite populations (see APPENDIX for derivation). In the latter case, the equilibrium frequencies depend only on the ratio of mutation rates, not on their absolute values (bottom right).

ing sites. When $\mu_a/\mu_d < 1$, the shared regulated *Abc* allele dominates for the opposite reason. For all conditions examined, the semi-independently regulated *aBC* allele maintains equilibrium frequencies of at least 0.04, so these results indicate that most large populations are potentially poised to make the transition to a semiregulated state.

One criticism that may be raised against our simple model is that it involves only three binding sites (A, B, and C), whereas most enhancer regions contain multiple copies of different transcription-factor binding sites. To address this concern we ran simulations for a six-site model, in which an allele may contain up to two each of the A, B, and C sites, making a model comprising 65 alleles. Genotypes carrying at least one functional A site or genotypes carrying at least one each of functional B and C sites with zero A sites were assumed to have a fitness of 1. Simulations were initiated with the locus fixed for the *AAbbcc* allele and ended when the frequency of all *A*-bearing alleles had gone to zero. Results for the six-site model show that the behavior is qualitatively the same as that of the three-site model, except that at small population sizes the time for ADR of *A*-bearing alleles is ~60% shorter (Figure 7, top). While

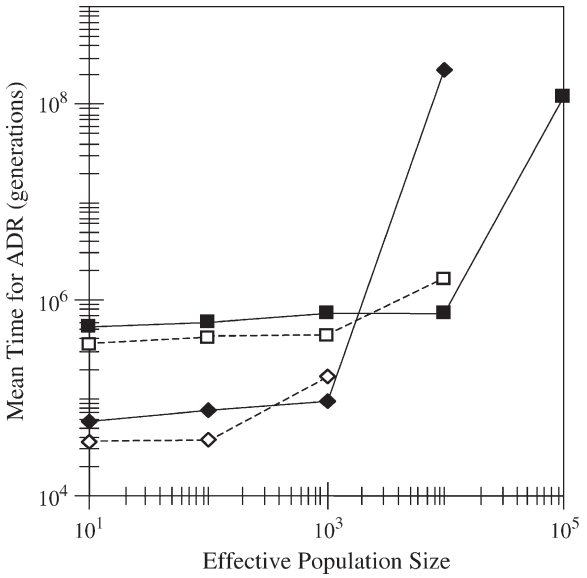

FIGURE 7.—Time to transient fixation of a semi-independently regulated *aaBXCX* allele for the six-site model starting from a state of fixation for the shared regulated *AAbbcc* allele during ADR. We denote the presence of a site with an uppercase letter, the absence of a site with a lowercase letter, and either presence or absence with an *X*. The dotted lines connect simulation data for the six-site model and the solid lines connect simulation data for the three-site model for comparison. For the three-site and six-site models $\mu_a/\mu_d = 1.0$ and $\mu_c = 0.00001$. Squares and diamonds represent $\mu_a = 0.00001$ and $\mu_a = 0.0001$, respectively.

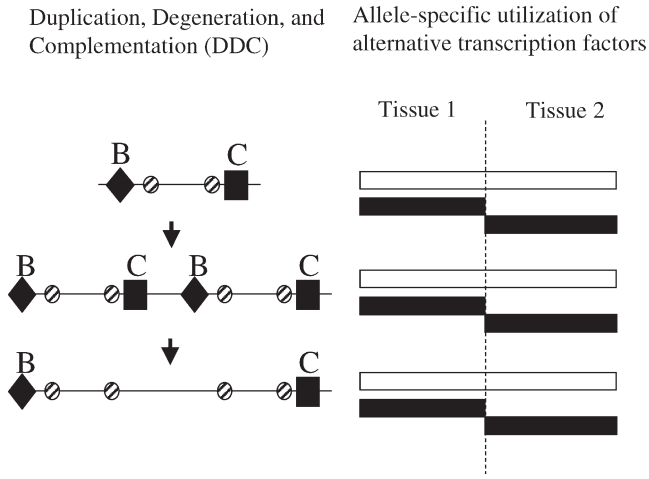

FIGURE 8.—Subfunction fission phase 2: enhancer subfunctionalization via duplication, degeneration, and complementation. The ancestral semi-independent enhancer (left) drives expression (right) in both tissues via regulatory elements B and C. The semiindependently regulated enhancer is duplicated within a gene and the duplicated enhancers may be preserved by subfunctionalization through complementary degenerative mutations. The structure of the enhancers has been expanded and given greater independence (left) while the expression patterns have remained conserved (right). The sites undergoing complementation are the independent B (diamond) and C (square) sites. Hatched circles represent shared positive and negative regulatory sites that are required for all functions.

there are two ancestral A sites to lose, four (two B and two C) sites may be gained. The additional viable combinations decrease the time to fission by doubling the rate of addition of B and C sites. In addition, the simulation results show that ~95% of the alleles at equilibrium exhibit at least partial redundancy with respect to the ancestral A function and ~12% are of the semi-independently regulated class similar to the results of the three-site model (see Figure 6, top right, and data not shown). This analysis may suggest that enhancers exhibiting at least partial redundancy should be common and that semi-independently regulated alleles may have appreciable frequencies when the rates of addition and deletion are nearly equal.

Phase 2—the formation of independent modular regulatory regions via enhancer subfunctionalization: The ADR phase presented above can lead to the formation of two subfunctions with semi-independent regulation. Subsequent ADR events of both positive and negative regulatory elements may act to reinforce the initial fission event. Another type of reinforcing event may involve enhancer duplication. The rate of duplication of entire genes is known to be on the order of 1% per gene per million years (LYNCH and CONERY 2003), and small duplications of <1000 bp are far more frequent than whole-gene duplication (KATJU and LYNCH 2003), so internal duplication has the potential to contribute significantly to regulatory-region evolution. Local duplication of a regulatory region may facilitate the formation and preservation of two fully independent regulatory elements by a process that we call *enhancer subfunctionalization*. The end product of this internal duplication event is a state in which each of the duplicate regulatory regions becomes restricted to driving expression in a single tissue.

For simplicity, we again consider a single continuous stretch of DNA, with tissue-specific transcription factors binding to unique B and C sites (see Figures 1 and 8) with a fraction of the DNA-binding sites in the enhancer being required by both overlapping subfunctions. If such an enhancer duplicates to a local site that is completely linked to the ancestral site, subfunctionalization may eventually preserve the duplicate enhancers through complementary degenerative mutations under genetic drift. This process has previously been investigated through theory and simulations for gene duplicates (FORCE *et al.* 1999; LYNCH and FORCE 2000; LYNCH *et al.* 2001), and we follow the previous terminology for enhancer subfunctionalization.

Degenerative mutations in the shared component lead to loss of both subfunctions at rate μ_c, while degenerative mutations in the unique component B and C sites lead to the loss of each subfunction at rate μ_r. Therefore, the total rate of mutation for the overlapping enhancer corresponding to a subfunction pair is $2\mu_r + \mu_c$, and the total rate for duplicate enhancer pairs is $4\mu_r + 2\mu_c$. Assuming that the duplicated regions are entirely functionally redundant, such that alleles car-

rying the duplicates have identical fitness to those with the ancestral state, the probability of initial fixation of a duplicate enhancer allele is simply its initial frequency, *i.e.*, the neutral expectation, $1/(2N)$, where N is the size of the diploid population. Subfunctionalization requires that the first degenerative mutation to become fixed falls in a B or C site, the probability of which is $2\mu_r/(2\mu_r + \mu_c)$. Finally, the last mutation to fix must fall in the nonshared unique site remaining in the still intact enhancer, and this will occur with probability $\mu_r/(2\mu_r + \mu_c)$ in populations of sufficiently small N. The product of the three terms is the probability that a newly arisen allele with a duplicate enhancer will be converted to a state of a nonoverlapping pair of enhancers by degenerative mutation,

$$\Pr(\text{Sub}) = 2\left(\frac{1}{2N}\right)\left(\frac{\mu_r}{2\mu_r + \mu_c}\right)^2 = \frac{(\alpha)^2}{N}, \qquad (2)$$

where $\alpha = \mu_r/(2\mu_r + \mu_c)$.

As noted previously, this simple approach fails when $\mu_c N$ begins to exceed ~0.1 (Lynch and Force 2000; Lynch *et al.* 2001). Subfunctional alleles mutate to null alleles at a higher rate, μ_c, than nonfunctional alleles carrying an intact overlapping enhancer and one dead overlapping enhancer. The mutation rate difference is small and begins to significantly affect the dynamics when the effective population size is large enough. In the appendix, we derive two estimates for the probability of enhancer subfunctionalization that account for this change in behavior in large populations. The behavior of the approximations was verified by individual-based computer simulations incorporating the sequential processes of mutation, selection (against lethal null homozygotes), and random gamete sampling using the same procedures outlined in our previous work on gene duplication (Figure 9; Lynch *et al.* 2001). For any given set of parameters (N, μ_r, and μ_c), 10^5 independent simulations were evaluated to obtain a precise estimate of Pr(Sub).

Because of the inverse scaling of Pr(Sub) with N, it is convenient to focus on the scaled probability of subfunctionalization, θ_{Sub}, which is the ratio of the actual probability of subfunctionalization and the neutral probability of fixation $1/(2N)$. Provided $N\mu_c < 0.1$, θ_{Sub} is very close to the prediction of the small-population theory, α^2/N (Figure 9). θ_{Sub} then slowly increases with N with a maximum slightly greater than the small-population prediction occurring in the vicinity of $N\mu_c \simeq 1$. However, with further increases in N, θ_{Sub} drops very rapidly, with $\theta_{Sub} \simeq 0$ for $N\mu_c > 10$.

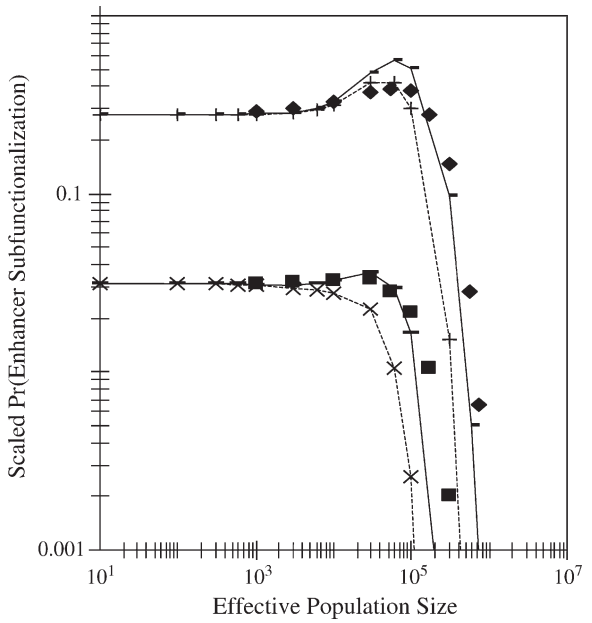

Figure 9.—The probability of enhancer subfunctionalization as a function of effective population size. The ratio of nonfunctionalizing mutations to all mutations in the top curve (diamonds) is $\mu_c/(2\mu_r + \mu_c) = 0.25$ and in the bottom curve (squares) is $\mu_c/(2\mu_r + \mu_c) = 0.75$ and the total rate of mutation per enhancer is $(2\mu_r + \mu_c) = 0.00001$. Solid lines (slightly deleterious) and dotted lines (neutral) represent approximations to enhancer subfunctionalization derived in the appendix.

SUBFUNCTION FORMATION AND RESOLUTION LEAD TO THE MODULAR RESTRUCTURING OF GENE NETWORKS

The results reported in this communication and in our previous work may have more global significance when considered in their totality (Force *et al.* 1999, 2004; Lynch and Force 2000; Lynch *et al.* 2001). Over evolutionary time there has been a general increase in the functional and structural specialization of body regions in multicellular eukaryotes, such as the long-term increase in the numbers and types of arthropod limb morphologies (Carroll *et al.* 2001). The trend in morphological specialization and developmental regionalization may be due to changes in the underlying circuitry of the developmental gene networks. The predominant form of morphological evolution at the phenotypic level among metazoans may involve parcellation of existing developmental modules at the genotypic level (Wagner 1996; Wagner and Altenberg 1996; Force *et al.* 2004). The formation of new subfunctions and the association of subfunctions with different genes following gene duplication events will change the circuitry of the underlying developmental genetic networks.

We refer to the subfunction formation and resolution processes, as they impact both genes and networks, as *modular restructuring* (Figure 10). First, new subfunctions are formed within a gene by subfunction fission or cooption processes (formation). Second, subfunctions are partitioned among different gene copies following gene duplication by DDC mechanisms (resolution). The modular restructuring process might have immediate effects on the phenotype. More importantly, however, under

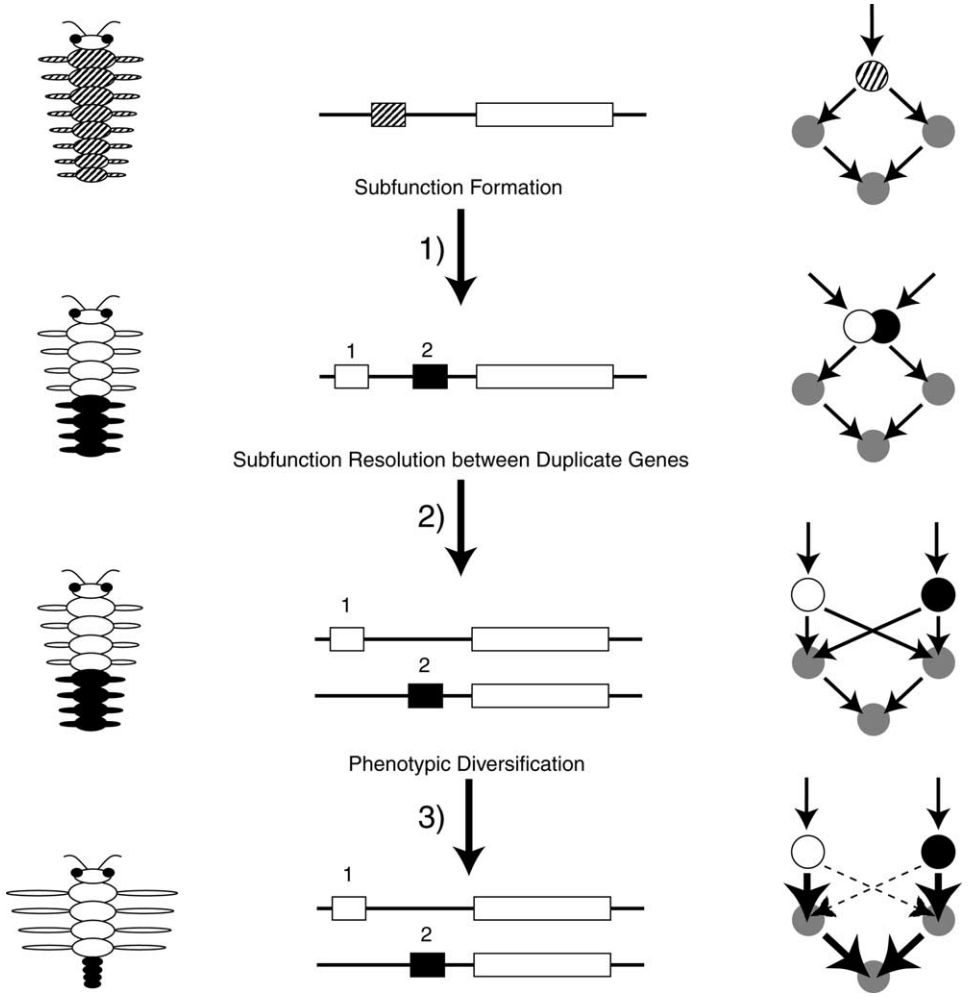

FIGURE 10.—Modular restructuring by subfunction formation and gene duplication. Near-neutral processes in small populations may change genotypic modularity passively. Subfunction formation and the resolution of subfunctions between gene duplicates lead to changes in the underlying genotypic-phenotypic map (columns 2 and 3) without affecting the phenotype (first column), first three rows. Column 2 illustrates changes at the level of a gene and column 3 illustrates changes at the gene network level. In the fourth row, the effects of modular restructuring accrued passively permit subsequent adaptive changes at the phenotypic level. Hatched circles and squares represent an ancestral subfunction that undergoes fission into two subfunctions represented by open and solid circles and squares.

relatively constant environmental conditions the phenotype may not be affected in any discernible way. Third, the new underlying genetic circuitry created by modular restructuring may open up new pathways for rapid morphological change. For instance, mutations may have unique phenotypic effects on the new genetic architecture that are now beneficial due to the removal of ancestral pleiotropic effects. Modular restructured genetic architectures may provide the evolutionary potential for rapid responses to novel environmental conditions. Therefore, modular restructuring at the genomic level may in part provide a population-level mechanism for the frequent observation of relatively rapid bursts of evolution and long periods of stasis observed in the fossil record (GOULD and ELDREDGE 1977, 1993, for example).

DISCUSSION

We have investigated the origin of new regulatory subfunctions via fission where multiple expression domains under unified genetic control become subdivided into more restricted expression domains under independent genetic control. Subfunction fission may proceed by the replacement of ancestral transcription-factor binding sites with new sites that drive more restricted patterns of gene expression. Duplication of overlapping enhancers may then lead to the formation of two entirely independent and modular regulatory regions through enhancer subfunctionalization.

Population size plays a key role in this process. Provided the effective population sizes are sufficiently small where $N(\mu_a + \mu_d) \leq 0.1$, the time for ADR closely reflects the behavior of the small population theory derived here. The time for ADR is extended in large populations because of a small mutationally induced selective advantage of the most redundant *ABC* allele. In the case of the *ABC* allele, if any of the three sites is deleted, the resulting allele is viable when homozygous and may be fixed in the population. In contrast, deletion of any site in the *aBC* allele results in a nonviable allele when homozygous, which is unlikely to go to fixation in very large populations. Therefore, while the fixation of the *aBC* allele is inhibited in very large populations, its immediate precursor, the *ABC* allele, is increased by differential mutation pressure. However, mutation pressure begins to strongly affect the allele frequencies only when $N(\mu_a + \mu_d) > 0.1$. Given that the rates for

μ_a and μ_d are not likely to be $>10^{-6}$ and could be orders of magnitude less, the minimum effective population size where the distribution of allele frequencies begins to be affected by mutation-induced selection pressure is ~100,000. Therefore, mutationally induced selection drives enhancers and genes toward overlapping function and nonmodularity in large populations. However, in small populations, ADR and DDC processes allow systems to diffuse toward modularity, meaning populations are likely to harbor a distribution of nonmodular, quasi-modular, and modular enhancer structures.

Previously, Stone and Wray simulated the evolution of new transcription-factor binding sites by point mutation to estimate the time to formation of new sites and then estimated their time to fixation using neutral theory (Stone and Wray 2001). While their calculations for the time to fixation were incorrect (MacArthur and Brookfield 2004), their treatment of the time for origin of new sites by mutation is consistent with our results. If we assume $\mu_a \cong \mu_d \cong 10^{-7}$, the expected number of generations for ADR is ~50,000,000 at an effective population size of 100,000 (see Figure 4). This is a slow rate for the evolution of new subfunctions by fission, but given that the number of subfunctions in the genome is greater than the number of genes, it may be a significant process over long-term evolution. Our model provides a baseline for the time to formation of a new regulatory subfunction, which may be reduced significantly by selection. For instance, the time for accretion to form the *ABC* allele under the above conditions is reduced ~1000-fold to about ~50,000 generations when the partially redundant *AbC* and *ABc* alleles each have a selective advantage of $s = 0.01$ and their effects are additive (data not shown and also see MacArthur and Brookfield 2004).

In contrast to our near-neutral fission process, models of compensatory evolution suggest enhancers may evolve through pairs of individually deleterious mutations that are beneficial in combination (Carter and Wagner 2002). Under this model, evolution of functionally conserved enhancers may occur by a two-step process with the first step involving a deleterious intermediate and the second step involving a beneficial compensatory mutation. This model makes the prediction that enhancers will evolve faster in very large populations where the double-mutant allele arises from segregating single-mutant deleterious alleles in the population. For conserved enhancers to evolve faster than under neutrality in large populations requires the assumption that the new double-mutant enhancer allele has a higher fitness than the ancestral enhancer allele. It is not clear how frequently this would be the case, unless the process were cyclical where slightly deleterious mutations would become fixed by drift in small populations and then a new deleterious allele could act as an intermediate in the formation of a compensatory beneficial allele. Interestingly, long-term cyclical population size would also aid the near-neutral fission process. If populations go through prolonged periods of large effective size followed by prolonged periods of small effective size, a buildup of the redundant precursor *ABC* allele during the former period and the fission *aBC* allele during the latter period is possible. It is unlikely the slightly deleterious *aBc* and *abC* alleles would contribute significantly to the evolution of the *aBC* alleles under our assumption of near neutrality. However, if the *aBC* allele was strictly beneficial, then these indirect pathways would be expected to contribute significantly.

Our results suggest why a population-genetic perspective, incorporating random genetic drift, is central to understanding the evolution of genotypic and potentially phenotypic modularity. While many (Wagner 1995; Cheverud 1996; Wagner and Altenberg 1996; Gerhart and Kirschner 1997; Raff and Raff 2000; Raff and Sly 2000) have argued that modularity may be directly selected for at the individual level, or indirectly selected for at the population level as an enhancer of evolvability, our work suggests that new subfunctions and genetic modularity can evolve under certain population-genetic scenarios via a nearly neutral process, without any selection promoting modularity itself. In addition, an increase in the number of regulatory subfunctions corresponds to an increase in the complexity of developmental genetic networks, which may be the foundation for phenotypic complexity. Finally, the results of this article support the hypothesis that many of the complex features of eukaryotic genomes may arise as simple by-products of random genetic drift in populations with small effective sizes (small being potentially as large as 10^7) (Force *et al.* 1999, 2004; Lynch and Force 2000; Lynch *et al.* 2001; Lynch and Conery 2003).

The authors thank Thomas Hansen and Ashley Carter for their comments on an early draft of the manuscript. Our work has been funded, in part, by grants from the National Institutes of Health (NIH) (RR14085, HG02526-01, 5F32GM020892), and the National Science Foundation (IBN-0321461, IBN-023639). Work in the Pickett laboratory is supported by grants NIH 2R15GM061620-02 and U.S. Department of Agriculture 2003-35304-13252.

LITERATURE CITED

Adams, K. L., R. Cronn, R. Percifield and J. F. Wendel, 2003 Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. Proc. Natl. Acad. Sci. USA **100** (8): 4649–4654.

Arnosti, D. N., S. Barolo, M. Levine and S. Small, 1996 The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development **122** (1): 205–214.

Bonneton, F., P. J. Shaw, C. Fazakerley, M. Shi and G. A. Dover, 1997 Comparison of bicoid-dependent regulation of hunchback between *Musca domestica* and *Drosophila melanogaster*. Mech. Dev. **66:** 143–156.

Brosius, J., 1999 Genomes were forged by massive bombardments with retroelements and retrosequences. Genetica **107:** 209–238.

Carroll, S. B., 2001 Chance and necessity: the evolution of morphological complexity and diversity. Nature **409:** 1102–1109.

Carroll, S. B., J. K. Grenier and S. D. Weatherbee, 2001 *From DNA to Diversity.* Blackwell Science, Malden, MA.
Carter, A. J. R., and G. P. Wagner, 2002 Evolution of functionally conserved enhancers can be accelerated in large populations: a population genetic model. Proc. R. Soc. Biol. **169:** 953–960.
Cheverud, J. M., 1996 Developmental integration and the evolution of pleiotropy. Am. Zool. **36:** 44–50.
Clark, A. G., 1994 Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA **91:** 2950–2954.
Davidson, E. H., 2001 *Genomic Regulatory Systems: Development and Evolution.* Academic Press, San Diego.
Edelman, G. M., R. Meech, G. C. Owens and F. S. Jones, 2000 Synthetic promoter elements obtained by nucleotide sequence variation and selection for activity. Proc. Natl. Acad. Sci. USA **97:** 3038–3043.
Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531–1545.
Force, A. G., W. A. Cresko and F. B. Pickett, 2004 Informational accretion, gene duplication, and the mechanisms of genetic module parcellation, pp. 315–337 in *Modularity in Evolution and Development,* edited by G. Schlosser and G. P. Wagner. University of Chicago Press, Chicago.
Gerhart, J., and M. Kirschner, 1997 *Cells, Embryos and Evolution.* Blackwell Science, Malden, MA.
Gould, S., and N. Eldredge, 1977 Punctuated equilibria: the tempo and mode of evolution reconsidered. Palaeobiology **3:** 115–151.
Gould, S., and N. Eldredge, 1993 Punctuated equilibrium comes of age. Nature **366:** 223–227.
Hancock, J. M., P. J. Shaw, F. Bonneton and G. A. Dover, 1999 High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. Mol. Biol. Evol. **16:** 253–265.
Hughes, A. L., 1994 The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. Ser. B **256:** 119–124.
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.
Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics **47:** 713–719.
Ludwig, M. Z., N. H. Patel and M. Kreitman, 1998 Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development **125** (5): 949–958.
Ludwig, M. Z., C. Bergman, N. H. Patel and M. Kreitman, 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403** (6769): 564–567.
Lynch, M., and J. S. Conery, 2003 The origins of genome complexity. Science **302:** 1401–1404.
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.
MacArthur, S., and J. F. Y. Brookfield, 2004 Expected rates and modes of evolution of enhancer sequences. Mol. Biol. Evol. **21** (6): 1064–1073.
McGregor, A. P., P. J. Shaw, J. M. Hancock, D. Bopp, M. Hediger *et al.*, 2001 Rapid restructuring of bicoid-dependent hunchback promoters within and between dipteran species: implications for molecular coevolution. Evol. Dev. **3:** 397–407.
Nowak, M. A., M. C. Boerlijst, J. Cooke and J. M. Smith, 1997 Evolution of genetic redundancy. Nature **388:** 167–171.
Piatgorsky, J., and G. Wistow, 1991 The recruitment of crystallins: new functions precede gene duplications. Science **252:** 1078–1079.
Raff, E. C., and R. A. Raff, 2000 Dissociability, modularity, evolvability. Evol. Dev. **2:** 235–237.
Raff, R., 1996 *The Shape of Life.* University of Chicago Press, Chicago.
Raff, R. A., and B. J. Sly, 2000 Modularity and dissociation in the evolution of gene expression territories in development. Evol. Dev. **2:** 102–113.
Rodin, S. N., and A. D. Riggs, 2003 Epigenetic silencing may aid evolution by gene duplication. J. Mol. Evol. **56:** 718–729.
Shaw, P. J., A. Salameh, A. P. McGregor, S. Bala and G. A. Dover, 2001 Divergent structure and function of the bicoid gene in muscoidea fly species. Evol. Dev. **3:** 251–262.
Sidow, A., 1996 Gen(om)e duplications in the evolution of early vertebrates. Curr. Opin. Genet. Dev. **6:** 715–722.
Stoltzfus, A., 1999 On the possibility of constructive neutral evolution. J. Mol. Evol. **49** (2): 169–181.
Stone, J. R., and G. A. Wray, 2001 Rapid evolution of cis-regulatory sequences via local point mutations. Mol. Biol. Evol. **18** (9): 1764–1770.
von Dassow, G., and E. Monro, 1999 Modularity in animal development and evolution: elements of a conceptual framework for evodevo. J. Exp. Zool. **285:** 307–325.
Wagner, A., 1994 Evolution of gene networks by gene duplications: a mathematical model and its implications on genome organization. Proc. Natl. Acad. Sci. USA **91** (10): 4387–4391.
Wagner, A., 1998 The fate of duplicated genes: loss or new function? BioEssays **20** (10): 785–788.
Wagner, A., 2001 Birth and death of duplicated genes in completely sequenced eukaryotes. Trends Genet. **17** (5): 237–239.
Wagner, G. P., 1995 Adaptation and the modular design of organisms. Adv. Artif. Life **929:** 317–328.
Wagner, G. P., 1996 Homologues, natural kinds and the evolution of modularity. Am. Zool. **36:** 36–43.
Wagner, G. P., and L. Altenberg, 1996 Complex adaptations and the evolution of evolvability. Evolution **50:** 967–976.
Walsh, J. B., 1995 How often do duplicated genes evolve new functions? Genetics **110:** 345–364.
Wray, G. A., 1998 Promoter logic. Science **279** (5358): 1871–1872.
Yuh, C. H., H. Bolouri and E. H. Davidson, 1998 Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science **279:** 1896–1902.

Communicating editor: D. M. Rand

APPENDIX

Equilibrium frequencies of fission alleles in an infinite population: Here we compute the distribution of allele frequencies in an infinite population using a matrix-modeling approach. The probability that a binding site is added is μ_a and the probability that a binding site is lost through mutation is μ_d. All individuals that have a genotype where both expression domains are covered produce the same expected number of offspring, while individuals that do not produce zero offspring.

We can define a matrix with elements $\mathbf{A}_{ij}$ that describe the number of genotype i offspring (row i) produced by a single adult of genotype j (column j). Because the genotypes *AbC* and *ABc* have the same mutational properties (each goes to *Abc*, *ABC*, and either *aBc* or *abC* with the same probability) we can lump them into a single class. We index the genotypes as

1	ABC
2	AbC or ABc
3	Abc
4	aBC

We can compute the matrix entries by considering the probability an individual in class j produces an individual in class i. This depends on the probability of each binding site being mutated. For example, $\mathbf{A}_{11} = (1 - \mu_d)^3$ because an individual with all three functional binding sites will produce an individual with three functional binding sites only if none of those sites are lost to mutation. Thus, the matrix **A** is given by

$$\begin{pmatrix} (1-\mu_d)^3 & \mu_a(1-\mu_d)^2 & \mu_a^2(1-\mu_d) & \mu_a(1-\mu_d)^2 \\ 2(1-\mu_d)^2\mu_d & (1-\mu_a)(1-\mu_d)^2 + \mu_a(1-\mu_d)\mu_d & 2(1-\mu_a)\mu_a(1-\mu_d) & 2\mu_a(1-\mu_d)\mu_d \\ (1-\mu_d)\mu_d^2 & (1-\mu_a)(1-\mu_d)\mu_d & (1-\mu_a)^2(1-\mu_d) & \mu_a\mu_d^2 \\ (1-\mu_d)^2\mu_d & \mu_a(1-\mu_d)\mu_d & \mu_a^2\mu_d & (1-\mu_a)(1-\mu_d)^2 \end{pmatrix}. \quad \text{(A1)}$$

The long-term frequency of genotype class i can be found by computing the dominant right eigenvector of **A**. This represents the stable distribution of genotype densities, excluding the genotype classes that have no reproductive success. This vector can be normalized to produce frequencies by dividing by the sum of the vector elements. Note that because class 2 contains two genotypes, the frequency of genotypes *AbC* and *ABc* is half of the frequency of class 2.

The probability of enhancer subfunctionalization: Here we derive an approximation to the probability of enhancer subfunctionalization. We begin by considering a pair of linked overlapping duplicate enhancers within a gene. Each of two independent regulatory elements, *B* and *C*, within a single overlapping enhancer is knocked out at rate μ_r and the entire overlapping enhancer with shared regulatory elements is knocked out at rate μ_c (see Figure 8). We use the following shorthand: A single intact enhancer is denoted by *BC* for the two independently mutable components of the overlapping enhancer structure. If either site is functionally deleted it is replaced with an *, and if both independent components or the shared components are functionally deleted the dead enhancer is denoted by **. Four allele classes are formed during the process of enhancer subfunctionalization that are viable when fixed in populations. These include the duplicate enhancer alleles *BC*|*BC* with frequency *p*, the partial subfunctional alleles **C*|*BC*, *B**|*BC*, *BC*|**C*, and *BC*|*B** with frequency *x*, the subfunctional alleles **C*|*B** and *B**|**C* with frequency *y*, and the single nonfunctional enhancer class alleles ****|*BC* and *BC*|** with frequency *q*. The initial nonduplicate enhancer *BC* alleles, with frequency q', are identical in state to the single enhancer class alleles but are kept track of separately because they are not descendants of the original duplicate *BC*|*BC* allele. Furthermore, we refer to an allele class by its respective uppercase letter, *P*, *X*, *Y*, *Q*, and Q'.

We divide the problem into a series of fixation events of interval length $4N$ generations, because on average this is the coalescence time of a single neutral allele in a diploid population. We can then estimate the probability of enhancer subfunctionalization Pr(Sub) as a summation of the individual probabilities of fixation of subfunctional alleles during each interval. Thus,

$$\Pr(\text{Sub}) = \sum_{i=1}^{\infty} \Pr(\bar{y}_i), \quad \text{(A2)}$$

where $\Pr(\bar{y}_i)$ is the probability of fixation of the subfunctional *Y* alleles in interval i and $\bar{y}_i$ is the mean frequency of subfunctional *Y* alleles in interval i.

The original duplicate enhancer *P* allele begins as a single copy in the population. We make the assumption that during each interval one of the viable alleles goes to fixation. This assumption is clearly valid when the population size is small, as the probability of homozygosity is close to one. In large population sizes the full spectrum of alleles will be present at the time of fixation and will deviate slightly from this assumption. Following the first fixation event where either a duplicate *P* allele or partial subfunctional *X* allele is fixed, subfunctional *Y* alleles can be derived from either ancestor. Therefore, we can rewrite Pr(Sub) as

$$\Pr(\text{Sub}) = \sum_{i=1}^{\infty} (\Pr(\bar{y}_{i,P}) + \Pr(\bar{y}_{i,X})), \quad \text{(A3)}$$

where $\Pr(\bar{y}_{i,P})$ and $\Pr(\bar{y}_{i,X})$ are the probabilities of fixation of the subfunctional alleles at the ith fixation event derived from a fixed duplicate *P* allele or a fixed partial subfunctional *X* allele at the end of interval $i - 1$.

For the frequencies of alleles at the end of each interval, we use a subscript referencing the interval number and/or allele from which it was derived. Thus, $\bar{p}_1$ is the frequency of *P* after the first interval and $\bar{p}_2$ is the frequency of *P* after the second and subsequent intervals derived from a *P* ancestor. Similarly, $\bar{x}_1$ is the frequency of *X* after the first interval, $\bar{x}_{2,P}$ is the frequency of *X* after the second and subsequent intervals derived from a *P* ancestor, and $\bar{x}_{2,X}$ is the frequency of *X* after the second and subsequent intervals derived from a *X* ancestor. The probability of

fixation of Y for the first interval is $\Pr(\bar{y}_1)$. The probabilities of fixation of Y for the second interval and subsequent intervals are $\Pr(\bar{y}_{2,P})$ for Y derived from P and $\Pr(\bar{y}_{2,X})$ for Y derived from X. The probabilities of fixation of Y within the second and subsequent intervals remain the same but are weighted by the decaying cumulative frequencies of P and X. Using the geometric series, it can be shown the probability of enhancer subfunctionalization is

$$\Pr(\text{Sub}) = \Pr(\bar{y}_1) + \Pr(\bar{y}_{2,P})\left(\frac{\bar{p}_1}{1-\bar{p}_2}\right) + \Pr(\bar{y}_{2,X})\left[\frac{\bar{x}_1}{1-\bar{x}_{2,X}} + \left(\frac{\bar{x}_{2,X}}{1-\bar{x}_{2,X}}\right)\left(\frac{\bar{p}_1\bar{p}_2}{1-\bar{p}_2}\right)\right] \tag{A4}$$

and the scaled probability of subfunctionalization is

$$\theta_{\text{Sub}} = 2N\Pr(\text{Sub}). \tag{A5}$$

Next, we determine expressions for the frequencies of the alleles after each interval. For the first interval and subsequent intervals the mean frequencies of the alleles derived from duplicate P alleles can be approximated in the following manner. The duplicate enhancer P alleles with arrangement $BC|BC$ mutate at a total rate of $(4\mu_r + 2\mu_c)$ into X, Y, and Q alleles. The frequency of the duplicate P alleles after t generations is described by the decay equation,

$$p_t = p_0 e^{(-(4\mu_r+2\mu_c)t)}. \tag{A6}$$

The partial subfunctional X alleles, with four arrangements $*C|BC$, $B*|BC$, $BC|*C$, and $BC|B*$, originate from the P alleles. The formation of X alleles requires a single independent regulatory element knockout at rate μ_r and requires that no other mutations occur in the remaining three independent regulatory sites (at rate $3\mu_r$) and the two shared regulatory regions (at rate $2\mu_c$). Thus, the frequency of X derived from a duplicate P ancestor is

$$x_t = p_0 4 e^{(-(3\mu_r+2\mu_c)t)}(1 - e^{(-\mu_r t)}). \tag{A7}$$

In a similar fashion, equations for the frequencies of Y, Q, and Q' alleles can be obtained

$$y_t = p_0 2 e^{(-(2\mu_r+2\mu_c)t)}(1 - e^{(-\mu_r t)})^2, \tag{A8}$$

$$\begin{aligned} q_t = 2p_0[&e^{(-(4\mu_r+\mu_c)t)}(1 - e^{(-\mu_c t)}) + e^{(-(2\mu_r+2\mu_c)t)}(1 - e^{(-\mu_r t)})^2 + e^{(-(2\mu_r+\mu_c)t)}(1 - e^{(-\mu_r t)})^2(1 - e^{(-\mu_c t)}) \\ &+ 2e^{(-(3\mu_r+\mu_c)t)}(1 - e^{(-\mu_c t)})(1 - e^{(-\mu_r t)}), \end{aligned} \tag{A9}$$

$$q_t' = q_0 e^{(-(2\mu_r+\mu_c)t)}. \tag{A10}$$

After normalization the expected frequencies of the alleles across all possible populations at the time of fixation are $\bar{p}_t = p_t/_F$, $\bar{x}_t = x_t/_F$, $\bar{y}_t = y_t/_F$, $\bar{q}_t = q_t/_F$, and $\bar{q}_t' = q_t/_F$, where $F = p_t + x_t + y_t + q_t + q_t'$.

To obtain the frequencies ($\bar{p}_1$, $\bar{x}_1$, $\bar{y}_1$, $\bar{q}_1$, and $\bar{q}_1'$) for the first interval we set $p_0 = 1/2N$, and $q_0 = 1 - 1/2N$. If following the first fixation event a duplicate P allele is fixed, the above equations (A6)–(A10) are used to determine the allele frequencies ($\bar{p}_{2,P}$, $\bar{x}_{2,P}$, $\bar{y}_{2,P}$, and $\bar{q}_{2,P}$) after $4N$ generations for the second and subsequent intervals where $p_0 = 1$. If following the first fixation event a partial subfunctional X allele is fixed, we can obtain the allele frequencies ($\bar{x}_{2,X}$, $\bar{y}_{2,X}$, and $\bar{y}_{2,X}$), in a similar manner to those above with x_0 set equal to 1:

$$x_t = x_0 e^{(-(3\mu_r+2\mu_c)t)}, \tag{A11}$$

$$y_t = x_0 e^{(-(2\mu_r+2\mu_c)t)}(1 - e^{(-\mu_r t)}), \tag{A12}$$

$$q_t = x e_0^{(-t(2\mu_r+2\mu_c))}(1 - e^{(-t\mu_r)}) + e^{(-t(3\mu_r+2\mu_c))}(1 - e^{(-t\mu_c)}) + e^{(-t(2\mu_r+\mu_c))}(1 - e^{(-t\mu_c)})(1 - e^{(-t\mu_r)}). \tag{A13}$$

The expected frequencies of the alleles across all possible populations at the time of fixation are then $\bar{x}_t = x_t/_F$, $\bar{y}_t = y_t/_F$, $\bar{q}_t = q_t/_F$, and $\bar{q}_t' = \bar{q}_t'/_F$, where $F = x_t + y_t + q_t$.

The three probabilities of fixation, $\Pr(\bar{y}_{2,P})$, $\Pr(\bar{y}_{2,X})$, and $\Pr(\bar{y}_1)$, are functions of the subfunctional Y allele frequency during each interval. In Figure 9, we plot two approximations, the first where we treat the fixation of Y alleles as a neutral process and the second where we treat the fixation of Y alleles as the fixation of a slightly deleterious allele. In the first case, the fixation probabilities are equal to the frequencies of Y at each interval i. In the second case, we use the diffusion approximation for the probability of fixation of a beneficial allele in a diploid population

$$P(\bar{y}_i) = \frac{1 - e^{-(4\bar{y}_i sN)}}{1 - e^{-(4sN)}}$$

(Kimura 1962), where $\bar{y}_i$ is the frequency of the slightly deleterious subfunctional alleles, s is the selection coefficient, and N is the population size. Consider the following two alleles, the Y allele $B^*|^*C$ and the Q allele $BC|^{**}$. The rate of mutation of the Y alleles to a nonfixable allele state is $2\mu_c + 2\mu_r$ and the rate of mutation of the Q alleles to a nonfixable allele state is $\mu_c + 2\mu_r$. Therefore, the subfunctional alleles die at a rate μ_c relative to the single nonfunctional enhancer Q alleles, suggesting the selection coefficient s for the Y alleles is $-\mu_c$.