# The Origins of Eukaryotic Gene Structure

*Michael Lynch*

Department of Biology, Indiana University, Bloomington

Most of the phenotypic diversity that we perceive in the natural world is directly attributable to the peculiar structure of the eukaryotic gene, which harbors numerous embellishments relative to the situation in prokaryotes. The most profound changes include introns that must be spliced out of precursor mRNAs, transcribed but untranslated leader and trailer sequences (untranslated regions), modular regulatory elements that drive patterns of gene expression, and expansive intergenic regions that harbor additional diffuse control mechanisms. Explaining the origins of these features is difficult because they each impose an intrinsic disadvantage by increasing the genic mutation rate to defective alleles. To address these issues, a general hypothesis for the emergence of eukaryotic gene structure is provided here. Extensive information on absolute population sizes, recombination rates, and mutation rates strongly supports the view that eukaryotes have reduced genetic effective population sizes relative to prokaryotes, with especially extreme reductions being the rule in multicellular lineages. The resultant increase in the power of random genetic drift appears to be sufficient to overwhelm the weak mutational disadvantages associated with most novel aspects of the eukaryotic gene, supporting the idea that most such changes are simple outcomes of semi-neutral processes rather than direct products of natural selection. However, by establishing an essentially permanent change in the population-genetic environment permissive to the genome-wide repatterning of gene structure, the eukaryotic condition also promoted a reliable resource from which natural selection could secondarily build novel forms of organismal complexity. Under this hypothesis, arguments based on molecular, cellular, and/or physiological constraints are insufficient to explain the disparities in gene, genomic, and phenotypic complexity between prokaryotes and eukaryotes.

## Introduction

Although full-genome sequencing has revealed numerous patterns of variation in genomic architecture among major taxonomic groups, a formidable, remaining challenge is to transform the descriptive field of comparative genomics into a more mechanistic theory of evolutionary genomics. Such an enterprise does not have to start from scratch. Nearly a century of mathematical derivation has resulted in a formal theory for evolution based on the expected dynamics of gene-frequency changes. Initially dubbed the Modern Synthesis by Huxley in 1942 and having experienced further enhancements since then, this theory has survived so much empirical scrutiny that the credibility of any proposed scenario for genome evolution must remain in doubt until shown to be consistent with basic population-genetic principles. In turn, if a mechanistic understanding of genome evolution is to be achieved, population-genetic theory will need to go beyond its reliance on algebraic formulations involving selection, mutation, recombination, and random genetic drift to incorporate the DNA-level constraints that are now known to define the evolutionary playing field.

Ever since Darwin, the vast majority of biologists have invoked natural selection as the primary, and in many cases the only, explanation for observed patterns of variation at most levels of organization. This greatly oversimplifies the evolutionary process. For example, Kimura, Ohta, and several contemporaries showed why numerous aspects of DNA sequence evolution cannot be explained entirely in terms of adaptive processes (reviewed in Kimura 1983; Ohta 1997). The neutral (or nearly neutral) theory that emerged from this work still enjoys a central place in the field of molecular evolution and has been applied to some aspects of evolutionary genomics (Force et al. 1999, 2005; Lynch et al. 2001, Lynch 2002; Lynch and Conery 2003; Lynch, Scofield, and Hong 2005*b*). The goal of this paper is to expand on these previous results to demonstrate the plausibility of the hypothesis that many of the unique complexities of the eukaryotic gene arose by semi-neutral processes with little, if any, direct involvement of positive selection.

Although eukaryotes share many basic aspects of transcription, translation, and replication with their prokaryotic ancestors, there are profound differences at the level of gene architecture. Prokaryotic genes are often organized into operons that are transcribed into polycistronic units, whereas with few exceptions, eukaryotic genes are transcribed as single-gene units. Unlike prokaryotic genes, eukaryotic genes often have complex regulatory regions, and in multicellular species such regions often have a modular structure that helps facilitate tissue-specific expression. Eukaryotic protein-coding genes also often contain introns, whereas prokaryotic genes do not, and eukaryotic transcripts generally contain longer untranslated leader and terminal sequences (untranslated regions [UTRs]) than do those of prokaryotes.

Three general observations have encouraged the view that that these kinds of increases in gene complexity were necessary prerequisites to the origin of organisms with multiple cell types: (i) most aspects of gene architecture are much more elaborate in multicellular than unicellular eukaryotes; (ii) similar forms of genomic architecture are found in the two major and independently evolved multicellular lineages, animals and land plants; and (iii) more complex genes can often carry out more complex sets of tasks (Raff 1996; Gerhart and Kirschner 1997; Davidson 2001; Carroll, Grenier, and Weatherbee 2001). However, despite these clear associations, the direction of causality in the link between genome and organismal complexity is far from certain. There is no direct evidence that

Key words: complexity, gene networks, gene regulation, gene structure, genome evolution, genetic draft, introns, modularity, mutation, natural selection, population size, pleiotropy, random genetic drift, recombination, subfunctionalization, transcription factors, UTR.

E-mail: milynch@indiana.edu.

*Mol. Biol. Evol.* 23(2):450–468. 2006
doi:10.1093/molbev/msj050
Advance Access publication November 9, 2005

multicellularity itself was promoted by adaptive processes, and the fact that many prokaryotes are capable of cell differentiation reminds us that the evolution of multicellularity need not have awaited the emergence of eukaryotes.

The key point to be made below is that the types of genomic evolution that can occur within a species are not so much dependent on aspects of cell biology as on the constraints imposed by population-level processes, most notably by population size itself. The arguments underlying this hypothesis will be laid out in three sections. First, I will review the role that chance plays in evolution and why this depends on population size. Second, I will summarize several sets of empirical data that show that the efficiency of natural selection declines dramatically between prokaryotes, unicellular eukaryotes, and multicellular eukaryotes. Third, I will demonstrate that theory and empirical observations are mutually consistent in pointing to a central role for nonadaptive processes in the origins of many of aspects of eukaryotic gene structure.

## The Role of Chance in Evolution

Evolution is an inherently stochastic process, starting from the chance events that produce single mutations and proceeding through a series of fortuitous steps that gradually lead to the spread of some mutations to every member of the descendant population. The usual conceptual point of departure here is the classic Wright-Fisher model, which assumes a population of diploid individuals, each contributing equally and synchronously to an effectively infinite gamete pool. The idealized mating system consists of randomly mating hermaphrodites and ignores complexities associated with overlapping generations, separate sexes, spatial structure, nonrandom variation in family sizes, and so on. However, most of these complications can be dealt with by equating the genetic "effective" size of a population ($N_e$) to the size of an ideal Wright-Fisher population yielding equivalent gene-frequency dynamics. The effective size of a population is a fundamental determinant of nearly all aspects of evolution as it determines the probability of (and times to) fixation or removal of mutant alleles.

Most deviations from the assumptions of the Wright-Fisher model cause $N_e$ to be less than the total number of adult individuals ($N$) (Caballero 1994; Whitlock and Barton 1997; Rousset 2003). For example, if adults differ in the number of gametes produced, either because of selection or chance ecological events, $N_e$ will be reduced simply because some individuals contribute little or nothing to the following generation. A sex ratio that deviates from 1:1 reduces $N_e$ because the rarer sex (which necessarily contributes half of the genes to the next generation) acts as a population bottleneck. Population-size fluctuations reduce long-term $N_e$ relative to the arithmetic average because the losses of variation during population bottlenecks exceed the preservational effects of equal population-size expansions. A number of procedures have been developed to estimate $N_e$ from pedigree data or by relating temporal fluctuations in allele frequencies to the expectations from sampling $2N_e$ gametes. Studies of this sort, mostly confined to vertebrates, suggest an average $N_e/N$ of $\sim 0.1$ (Frankham 1995).

To see how population size influences the long-term rate of evolution, consider a newly arisen mutation in a diploid population containing $2N$ gene copies at each locus. If the mutation is neutral, with no external forces favoring one allele over another, the probability of eventual fixation is always equal to the initial frequency, $p_0 = 1/(2N)$. Because this fraction is inversely proportional to $N$, one might expect neutral changes to accumulate more slowly in larger populations. However, the expected number of new mutations arising at a locus per generation is $2N\mu$, where $\mu$ is the rate of origin of neutral mutations per gene, and the long-term rate of evolution is equal to the product of the rate of origin of mutations and their probability of fixation. Thus, the rate of neutral evolution reduces to the genic mutation rate $\mu$, which is entirely independent of the effective and absolute population size (Kimura 1983).

Intuition suggests that the fixation probability of a beneficial allele must exceed the neutral expectation $1/(2N)$, whereas that of a detrimental allele must be $<1/(2N)$, but how much so? To see that a beneficial mutation is never guaranteed to go to fixation, no matter how favorable, consider a new mutation that improves the fitness of its initial carrier by a fraction $s$ (the selection coefficient). Such an allele has expected frequency $p_0 = (1 + s)/(2N)$ in the gamete pool leading to the next generation, and the probability that it is not successfully inherited by at least one offspring, $(1 - p_0)^{2N}$, is closely approximated by $(1 - s)e^{-1}$ for small $s$. Thus, relative to the situation for a neutral allele ($s = 0$), selection only reduces the probability of a rapid initial exit for a beneficial allele by a fraction $s$. This shows that until a favorable mutation has avoided chance elimination in the first few generations and increased its frequency in doing so, there is little assurance that it will successfully go to fixation. For mutations with additive effects on fitness (increasing homozygote fitness by $2s$), the probability of fixation is $p_f \approx (2sN_e/N)/(1 - e^{-4N_e s})$ (Kimura 1962). Thus, even in very large populations ($N_e \to \infty$), the high degree of stochasticity in the early phase of mutation establishment still restricts the probability of fixation to an upper limit of $2sN_e/N$. Given the arguments presented above, this means that the probability of fixation of a favorable mutation is almost always $<2s$.

Letting $\mu_b$ be the beneficial-mutation rate per gene and $2N\mu_b$ be the rate at the population level, the above results imply that the upper limit to the rate of incorporation of beneficial mutations is $4N_e\mu_b s$, which unlike the situation for neutral mutations increases with $N_e$. In contrast, because detrimental mutations with $-0.3 < N_e s < 0.0$ have fixation probabilities at least half as great as the neutral expectation, if the rate of origin of mutations in this range of effects is sufficiently high, a considerable load of mildly deleterious mutations can accumulate in populations of sufficiently small size (Ohta 1973, 1974). Modifications to this theory for situations in which populations are subdivided or changing in size do not change these basic scaling properties (Otto and Whitlock 1997; Whitlock 2003).

These results yield the robust prediction that the ability of a population to incorporate beneficial mutations and to purge deleterious mutations should scale positively with population size, assuming that $N_e$ scales positively with $N$. However, something beyond the demographic features of a population, the physical structure of the genome itself, will generally limit the growth of $N_e$ with $N$ in the largest of populations. Because tightly linked nucleotide sites are transmitted across generations as a unit, to a degree that depends on the rate of recombination, the fate of any new mutation depends on the selective forces operating on all linked loci. On average, this causes the fixation rates of beneficial mutations to be lower and detrimental mutations to be higher than the single-locus predictions suggested above (Hill and Robertson 1966). For example, a beneficial mutation that rapidly sweeps through a population will necessarily drag along any deleterious alleles at tightly linked loci with which it is associated at the time of origin, whereas the selective removal of deleterious alleles can impede adaptive evolution at linked loci. Even mutually advantageous alleles will interfere with each other's fixation when linked. Consider a beneficial allele A segregating at one locus, with a second beneficial mutation B arising at a tightly linked locus on an a-bearing chromosome. If the advantages of each mutation were the same, then the Ab and aB linkage groups would compete with each other in the fixation process, with one eventually excluding the other. Linkage need not be absolute for these effects to be important, but the stronger the degree of linkage the greater the degree of selective interference.

Gillespie (2000) presented an elegant argument relating the influence of linkage to the effective population size of a chromosomal region. The key issue is that the effective size of a population defines the variance of allele-frequency change from generation to generation, which for a neutral locus is $p(1 - p)/(2N_e)$, where $p$ is the current allele frequency, and $2N_e$ is the effective number of genes sampled per locus. Now imagine a neutrally evolving site completely linked to another site that is experiencing selective sweeps at rate $\delta$. On average, selective sweeps do not influence which alleles go to fixation at linked neutral sites because the probability that a beneficial mutation destined to fixation will arise in association with a particular allele is simply equal to that allele's frequency. However, selective sweeps do magnify the fluctuations of allele frequencies at linked neutral sites. Assuming as a first approximation that sweeps cleanse a population of linked variation essentially instantaneously, then conditional on a sweep occurring, the variance in allele-frequency change at the neutral locus is $p(1 - p)$. Thus, for a neutral locus in an ideal randomly mating population, the average variance of allele-frequency change is approximately $p(1 - p)\{[(1 - \delta)/(2N_e)] + \delta\}$. Equating the right-hand quantity to $1/(2N_l)$, the long-term effective population size is $N_l \approx N_e/(1 + 2N_e\delta)$, where $N_e$ is now defined to be the short-term effective size during phases free of selective sweeps. (Maruyama and Birky [1991] obtained essentially the same result by a different method.) Recombination reduces the likelihood that a selective sweep will completely purge the variation at a linked locus, but Gillespie (2000) showed that this simply modifies the preceding expression to $N_l \approx N_e/(1 + 2N_eC\delta)$, where $C$ is the average squared frequency of the neutral hitchhiking allele after the completion of a sweep (previously assumed to be equal to one).

An unresolved issue is the way in which the rate of selective sweeps in a tightly linked region, $C\delta$, scales with population size. If $C\delta$ were completely independent of population size, then $N_l$ would increase with $N_e$ at a decreasing rate, eventually reaching an upper limit of $1/(2C\delta)$. Because larger populations contain more targets for rare beneficial mutations and also experience more recombination events, the rate of sweeps ($\delta$) is expected to increase and the breadth of sweeps ($C$) to decrease with increasing $N_e$, so in principle, these opposite patterns of scaling might fortuitously balance such that the product $C\delta$ is indeed independent of $N_e$. However, in the absence of any direct observations on this matter, a more general approximation is to treat $C\delta$ as a power function of $N_e$, yielding a relationship of the form $N_l = N_e/(1 + \alpha N_e^{\beta})$. Here, $\beta = 1$ describes the case in which $C\delta$ is independent of $N_e$, whereas with $\beta = 2$ the rate of selective sweeps is proportional to $N_e$, as in the single-locus result described above. The latter condition is clearly too high to be biologically realistic as the rate of selective sweeps per locus eventually exceeds one per generation at large $N_e$. Thus, $\beta$ is likely to fall in the range of 1–2, although this remains to be formally demonstrated. Assuming that short-term $N_e$ scales linearly with absolute population size ($N$), these qualitative arguments suggest that long-term $N_l$ will also scale linearly with $N$ for small to moderate $N$ where random genetic drift is the predominant stochastic force. However, the degree of scaling is expected to be progressively reduced at larger $N$, with an asymptotic limit to $N_l$ possibly being reached at very large $N$ where stochastic fluctuations in allele frequencies are primarily a function of the chromosomal nature of the genome (Gillespie's genetic draft).

To gain an appreciation of the power of random genetic drift and draft to compromise the efficiency of natural selection, it is useful to consider the ratio of the single-locus fixation probability, $p_f$, and the neutral expectation, $1/(2N)$. This is a simple function of $4N_ls$, $\theta_f \cong 4N_ls/(1 - e^{-N_ls})$, where $N_l$ is defined by the function in the preceding paragraph (fig. 1). If the strength of selection is sufficiently large relative to the power of random genetic drift ($4N_ls > 1$), the fixation probability for an advantageous allele is inflated by a factor of $4N_ls$ relative to the neutral expectation, whereas the fixation probability of a deleterious allele asymptotically approaches zero as $4N_ls \rightarrow -\infty$. However, if $|4N_ls| < 0.2$, the probability of fixation is within 10% of the neutral expectation, and if $|4N_ls| < 0.02$, the deviation from neutrality is no more than 1%. Thus, for any long-term effective population size, there exists a range of deleterious mutations whose selective disadvantages are overwhelmed by stochastic forces. Such alleles are said to be effectively neutral.

## The Three Genomic Perils of Increased Organism Size

A central premise of this paper is that there is a general reduction in the efficiency of selection between prokaryotes, unicellular eukaryotes, and multicellular species. We now take a more empirical look at this issue, showing that all three major factors responsible for reductions in $N_l$—small population size, tight linkage, and high background mutational activity—are jointly exacerbated as organisms increase in size, producing a synergism that causes substantial reductions in the efficiency of natural selection.

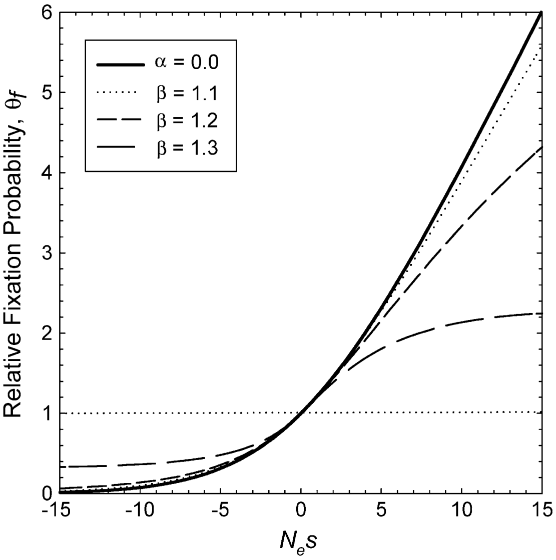

FIG. 1.—The probability of fixation of a new mutant allele relative to the neutral expectation of 1/(2*N*), given as a function of the product of the short-term (drift determining) effective population size ($N_e$) and the selection coefficient (*s*). For the dashed and dotted curves, the long-term effective population size ($N_l$) is defined by the function in the text with $\alpha = 10^{-9}$, whereas the solid line denotes $N_l = N_e$, which assumes an absence of selective sweeps ($\alpha = 0$). The horizontal dotted line denotes the neutral expectation. Negative values of $N_e s$ denote deleterious mutations.

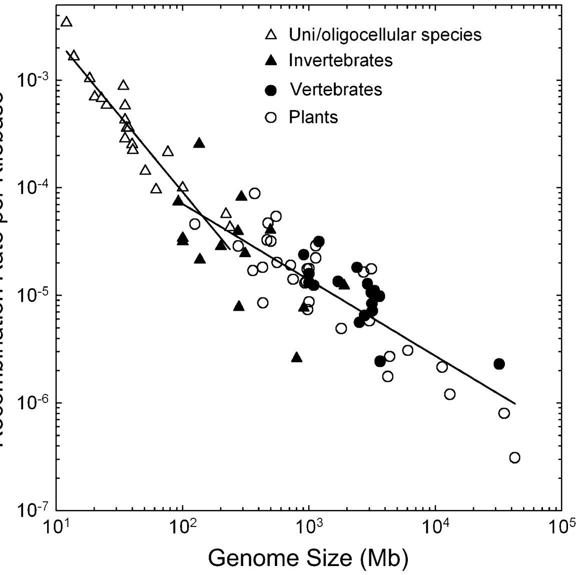

FIG. 2.—The negative relationship between recombination rate per physical distance (*c*) and total genome size (*G*) in eukaryotes, scaling as $c = 0.0019G^{-0.71}$ for animals and land plants, with the exponent having a standard error (SE) of 0.06, and as $c = 0.045G^{-1.32}$ for fungi and other unicellular/oligocellular species, with the exponent having a SE of 0.12. In the two respective cases, genome size accounts for 70% and 80% of the variance in recombination rate. (For further details, see Supplementary Table 1, Supplementary Material online).

### *Body Size and Population Size*

A typical prokaryote is five to seven orders of magnitude smaller than the average single-celled eukaryote, with a similar disparity existing between unicellular and multicellular eukaryotes (Bonner 1988). Such massive differences in size impose numerous ecological and physiological constraints and opportunities, but the implications for the population-genetic environment are equally pronounced. All other things being equal, the genetic effective size of a population should generally increase with the actual number of breeding adults (*N*), and one of the few well-established laws in ecology is that a primary determinant of *N* is the average size of members of the population. Eukaryotes generally show an inverse relationship between population density per unit area and average individual body mass within a species, with the extreme values ranging from $\sim 10^{-7}$ individuals/M$^2$ for the largest vertebrates to $\sim 10^{11}$ individuals/M$^2$ for the smallest unicellular eukaryotes (Damuth 1981; Schmid, Tokeshi, and Schmid-Araya 2000; Enquist and Niklas 2001; Carbone and Gittleman 2002; Finlay 2002).

Ecological factors unique to individual species can cause local deviations around this pattern, and an inverse scaling between population density and organism size need not reflect the pattern for total population size as it does not account for total species ranges. However, the geographic area occupied by vertebrate species is negligibly to weakly positively correlated with average body size (Gaston and Blackburn 1996; Diniz and Torres 2002; Housworth, Martins, and Lynch 2003), and the geographic ranges of unicellular species appear to be substantially greater than those for multicellular taxa (Finlay et al. 2001; Finlay, Monaghan, and Maberly 2002; Green et al. 2004; Horner-Devine et al. 2004). Thus, in a broad phylogenetic sense, there is little doubt that the total number of individuals within a species declines with increasing organism size, and the total range in *N* over all species certainly exceeds 20 orders of magnitude. Assuming $\sim 10^{30}$ prokaryotic cells inhabiting the earth (Whitman, Coleman, and Wiebe 1998) and $10^7$ being an upper-bound estimate for the number of prokaryotic species (Hammond 1995), *N* for an average prokaryote would be $\sim 10^{23}$.

### *Reduced Recombination in Large Genomes*

High-density genetic maps allow the estimation of the average amount of meiotic crossing over for numerous eukaryotes. The magnitude of recombination per physical distance scales negatively with genome size, ranging from $3 \times 10^{-10}$/bp/generation in *Pinus sylvestris* to $3 \times 10^{-6}$/bp/generation in *Saccharomyces cerevisiae* (fig. 2). Such scaling is due mostly to the simple fact that most species experience between one and two meiotic crossover events per chromosome. Because chromosome number is uncorrelated with genome size, the intensity of recombination per nucleotide position naturally increases in smaller genomes (with smaller average chromosome lengths). Less clear is why the recombination rate declines with increasing genome size twice as rapidly in unicellular as in multicellular species (fig. 2). In any event, because genome size increases with organism size, these results imply that increases in organism size are accompanied by decreases in the intensity of recombination. Not only can a selective sweep in a multicellular eukaryote drag along up to 10,000-fold more linked nucleotide sites than is likely in a unicellular species, but species with small genomes also experience increased levels of recombination on a per-gene basis. For example, the rate of recombination over the entire physical distance associated with an average gene (including intergenic DNA) is ~0.007 in *S. cerevisiae* versus ~0.001 in *Homo sapiens*, and the discrepancy is greater if one considers just coding exons and introns, 0.005 versus 0.0005. The consequences of reduced recombination rates are particularly clear in the human population, which harbors numerous haplotype blocks, tens to hundreds of kilobases in length, with little evidence of internal recombination (Daly et al. 2001; Reich et al. 2001; Dawson et al. 2002; Gabriel et al. 2002; Greenwood, Rana, and Schork 2004; McVean et al. 2004).

### *The Rate of Mutation*

Because of the rarity of mutations at individual nucleotides sites, there are enormous challenges to estimating the rate at which mutations arise at the molecular level. Most estimates are derived either from surveys of visible mutations at reporter loci or of dominant genetic disorders, followed by sequence analysis of individuals exhibiting a phenotype. These approaches are not without problems, as corrections must be made for the incidence of undetectable mutations. More indirect attempts to estimate the mutation rate are based on comparisons of distantly related species, using DNA sequences thought to be free of natural selection and making assumptions about times of interspecific divergence and species-specific generation times (e.g., Keightley and Eyre-Walker 2000; Kumar and Subramanian 2002). Only a single attempt has been made to estimate the per-nucleotide mutation directly by brute-force assays of random sequences in unselected lines (Denver et al. 2004). To minimize the assumptions involved, the following survey largely relies on data derived from the more direct approaches.

Across a phylogenetically diverse set of a species, there is a strong correlation between the mutation rate per generation and genome size (fig. 3). The range for the base-substitution mutation rate is approximately two orders of magnitude, and again exhibits a gradient with organism size, the extremes being $5.0 \times 10^{-10}$ and $5.4 \times 10^{-8}$/bp/generation for prokaryotes and vertebrates, respectively. Despite the uncertainties in each estimate contributing to this pattern, the validity of the overall relationship is supported by two observations. First, the estimate for the nematode *Caenorhabditis elegans* obtained by direct sequence analysis (highest red point in the plot) is consistent with the remaining data obtained via reporter constructs. Second, estimates of the human mutation rate obtained from observations on dominant genetic disorders (Kondrashov 2003) are very similar to those obtained from comparisons of pseudogene sequences in humans and chimpanzees (Nachman and Crowell 2000), $2.6 \times 10^{-8}$ and $2.2 \times 10^{-8}$/bp/generation, respectively. Based on more limited data, Drake (1991) concluded that the mutation rate per nucleotide per generation is inversely related with genome size in microbial species, but the results in figure 3 suggest the opposite pattern, even within the subset of unicellular species.

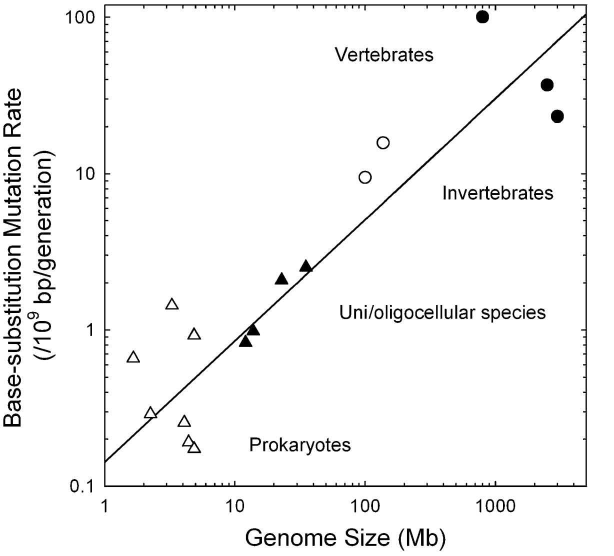

FIG. 3.—The relationship between the mutation rate for base substitutions per nucleotide site per generation ($u$) and genome size ($G$) scales as $u = 0.14G^{0.78}$ ($\times 10^{-9}$), with the exponent having a SE of 0.10 and genome size accounting for 79% of the variance in the mutation rate. (For further details, see Supplementary Table 2, Supplementary Material online).

### *The Global Effective Population Sizes of Species*

The preceding results show that three factors (low population sizes, low recombination rates, and high mutation rates) conspire to reduce the efficiency of natural selection with increasing organism size, although it is difficult to predict the magnitude of decline in $N_l$ from these three factors alone. For example, fluctuations in population size can result in a substantial depression of $N_l$ below average $N$, and it is unclear whether the magnitude of such fluctuations varies with organism size. In addition, as discussed above, hitchhiking effects should depress the $N_l/N$ ratio much more in large populations, but because the recombination rate (per meiosis) is substantially greater and the mutation rate is substantially lower in species with large $N$, this decline could be weaker than otherwise expected. Given the many additional factors that can influence $N_l$, the degree to which $N_l$ varies with organism size is best resolved by direct empirical observation.

One way to accomplish this task is to consider the amount of nucleotide-sequence variation at silent sites in protein-coding genes within natural populations. Under the assumption that mutations at such sites escape the eyes of natural selection, the amount of silent-site variation has a simple interpretation. The rate of introduction of new variation per nucleotide site in two randomly compared alleles is $2u$ (twice the base-substitution mutation rate per nucleotide), while the expected rate of loss of variation by genetic drift is $1/(2N_l)$. At equilibrium, the average number of nucleotide substitutions separating individual neutral sites in two randomly sampled alleles is the ratio of these two rates, $4N_lu$. For a haploid species, the rate of random genetic drift is $1/N_l$, and the equilibrium divergence among neutral nucleotide sites becomes $2N_lu$. Both results have the same meaning—at mutation-drift equilibrium, the amount of within-species nucleotide variation at silent sites is equal

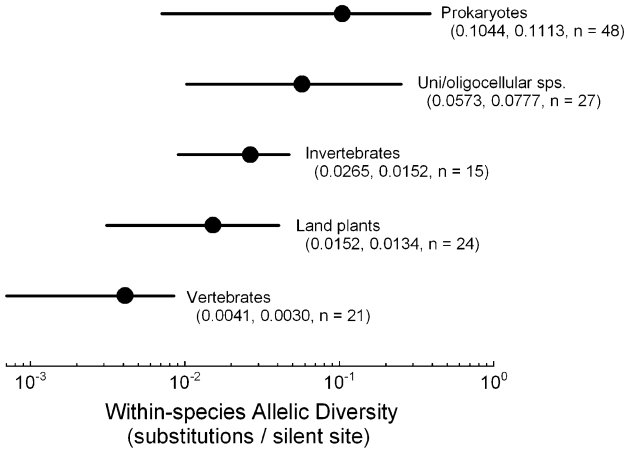

FIG. 4.—Average levels of within-species nucleotide variation (measured as the number of substitutions per site) for silent sites in protein-coding genes from a phylogenetically diverse assemblage of eukaryotic and prokaryotic species. The sampling unit is the average estimate for members of a genus, where information for multiple species was available. Means, standard deviations, and number of genera are given in parentheses; horizontal bars provide estimates of the observed ranges of values, using 5% and 95% limits to reduce the effects of outliers resulting from sampling error. Almost all data are derived from geographically wide surveys of multiple protein-coding genes. (For further details, see Supplementary Tables 3 and 4, Supplementary Material online).

to twice the effective number of gene copies at the locus times the per nucleotide mutation rate. An estimate of this composite parameter is provided by the observed level of silent-site variation within a species (hereafter $\pi_s$). Although the complex nature of the definition of $N_l$ introduces some interpretive issues with $\pi_s$ (Laporte and Charlesworth 2002), a fully general definition can be described in terms of allelic ancestry. $\pi_s$ is equal to the average age of random pairs of sequences times twice the base-substitutional mutation rate per nucleotide site. As will be seen below, the fact that $\pi_s$ is a function of the product of $N_l$ and $u$ is very useful because many aspects of genome evolution depend directly on this product.

Information on $\pi_s$ now exists for a wide enough phylogenetic range of species that some general statements can be made. Drawing from a substantially larger database than presented in the earlier survey of Lynch and Conery (2003), there is a striking inverse relationship between organism size and silent-site variation (fig. 4). For prokaryotes, $\pi_s$ lies in the broad range of 0.0071–0.3881, with an average value of 0.1044. This is nearly twice the average value for unicellular eukaryotes (0.0573), although the range of values among the latter taxa is again very high (0.0103–0.2522). For the still larger invertebrates, there is a further reduction in average $\pi_s$ to 0.0265, with a range of 0.0090–0.0473. The average value of $\pi_s$ for plants (0.0152) is still lower and that for vertebrates (0.0038) is even lower.

A significant caveat with respect to these data is that the bulk of existing surveys on nuclear variation in unicellular species have focused on pathogens, which because of the demographic dependence on their host species, probably have lower $N_e$ than free-living species (Hartl et al. 2002). Two of the three prokaryotes with $\pi_s < 0.01$ are *Serratia* (a human pathogen) and *Buchnera* (an obligate endosymbiont of aphids). The six lowest estimates of $\pi_s$ for unicellular eukaryotes are all derived from pathogens (*Candida*, *Coccidioides*, *Encephalitozoon*, *Fusarium*, *Phytophthora*, and *Plasmodium*), all other taxa (including some pathogens) having $\pi_s > 0.02$.

From these estimates of $\pi_s$, $N_l$ can be disentangled from $u$ by applying the mutation-rate estimates described above. For example, using the average observed value of $u \approx 5.0 \times 10^{-10}$ for prokaryotes to factor $u$ out of $2N_lu$, the estimated average $N_l$ for prokaryotes is $\sim 10^8$. After removal of the six lowest values of $\pi_s$ for eukaryotic parasites, application of the average mutation rate for unicellular eukaryotes ($1.6 \times 10^{-9}$) yields an average $N_l$ of $\sim 10^7$ for this group. Similar analyses for invertebrates and vertebrates yield average $N_l$ estimates of $10^6$ and $10^4$, respectively. Finally, a phylogenetically based mutation-rate estimate for plants of $7.3 \times 10^{-9}$/bp/year for plants (Lynch 1997) yields an average $N_l$ estimate of $\sim 10^6$ for annual species and assuming a generation time of 20 years, $\sim 10^4$ for trees.

Given the rough nature of the preceding calculations, they are intentionally reported to just an order of magnitude. Nevertheless, it is likely that the range in $N_l$ from prokaryotes to multicellular eukaryotes exceeds the four orders of magnitude just noted. Any selection on silent sites associated with codon-usage bias and/or mRNA processing features will bias $\pi_s$ below the neutral expectation, and the magnitude of bias will be greatest in large populations where selection is most efficient. Several observations suggest that this issue is of significance (Bustamante, Nielsen, and Hartl 2002; Hellmann et al. 2003; Chamary and Hurst 2004; Desai et al. 2004; Halligan et al. 2004; Sharp et al. 2005), and because the divergence rate of silent sites in prokaryotes may be at least ten times lower than the mutation rate (Ochman 2003), the prokaryotic $N_l$ estimates given above could be underestimated by at least tenfold. Despite these uncertainties, it is clear that the disparity in $N_l$ across all domains of life is nearly 20 orders of magnitude less than the disparity in absolute numbers, a pattern that is consistent with a significant stochastic role of genetic draft in large populations.

Because of their potential for considerable clonal structure, prokaryotic species may be particularly vulnerable to selective sweeps, but the breadth of such sweeps remains unclear. Moreover, because prokaryotes have a number of mechanisms for uptake and exchange of exogenous DNA, the absence of meiosis need not imply exceptionally low levels of recombination. Some insight into this matter can be acquired by considering the statistical associations between locus-specific allelic variants that develop stochastically as a consequence of random genetic drift. At drift-recombination balance, the amount of linkage disequilibrium in a population is a function of the product $N_lc$, where $c$ is the rate of recombination per nucleotide site (Ohta and Kimura 1971; Hill 1975). This quantity can be estimated by evaluating the rate at which the level of disequilibrium declines with the physical distance between nucleotide sites in samples of gene sequences. When joint estimates of $N_lu$ and $N_lc$ are available, their ratio eliminates $N_l$, providing an estimate of the relative rates of recombination and mutation ($c/u$).

**Table 1**
**Estimates of *c*/*u*, the Ratio of the Recombination Rate to the Mutation Rate Per Base Pair, for Various Species**

| Species | *c*/*u* | Reference |
|---|---|---|
| Animals | | |
| *Homo sapiens* | 0.6 | Ptak, Voepel, and Przeworski (2004) |
| *Chorthippus parallelus* | 2.5 | Ibrahim, Cooper, and Hewitt (2002) |
| *Drososphila* spp. | 3.8 | Hey and Wakeley (1997);<br>Machado et al. (2002) |
| Land plants | | |
| *Arabidopsis* spp. | 0.7 | Wright, Lauga, and Charlesworth (2003) |
| *Brassica nigra* | 0.3 | Lagercrantz, Kruskopf Osterberg, and Lascoux (2002) |
| *Cryptomeria japonica* | 3.0 | Kado et al. (2003) |
| *Pinus taeda* | 0.3 | Brown et al. (2004) |
| *Zea mays* | 1.6 | Tenaillon et al. (2004) |
| Prokaryotes | | |
| *Neisseria gonorrheae* | 1.0 | Posada et al. (2000) |
| *Neisseria meningitidis* | 4.8 | Feil et al. (2001) |
| *Pseudomonas syringae* | 0.3 | Sarkar and Guttman (2004) |
| *Staphylococcus aureus* | 6.5 | Feil et al. (2001) |
| *Streptococcus pneumoniae* | 8.9 | Feil et al. (2001) |

For eukaryotes, the results from figures 2 and 3 can be used to reveal *c*/*u* more directly, yielding average relationships of $c/u \approx 0.321G^{-2.1}$ for unicellular species and $0.014G^{-1.5}$ for multicellular species, which implies expected values of 17 and 0.4 for genomes 100 and 1,000 Mb in size, respectively. The few available estimates for *c*/*u* for eukaryotes from polymorphism studies are in rough accord with these predictions, the average being 2.3 (0.9) for animals and 1.2 (0.5) for land plants (table 1), whose genomes are generally in the vicinity of a few hundred to several thousand megabases. The few available estimates of *c*/*u* for prokaryotes are of the same order of magnitude as those for multicellular eukaryotes, although not as high as expected for unicellular eukaryotes, averaging 4.3 (1.6). Thus, relative to the background rate of mutation, recombination at the nucleotide level is not exceptionally low in prokaryotes. Horizontal transfer across species boundaries can also expand the genomic resources available to prokaryotes (Ochman, Lawrence, and Groisman 2000), and hence the efficiency of selection, although this source of diversity has been avoided in the preceding analyses.

In summary, all lines of evidence point to the fact that the efficiency of selection is greatly reduced in eukaryotes to a degree that depends on organism size. However, as suggested by Gillespie (2000), the physical nature of chromosomes puts an ultimate limit on $N_l$ even in species with enormous numerical abundances. Indeed, the highest estimate of $N_l$ derivable from the results in figures 3 and 4 is $\sim 2 \times 10^9$ for *Helicobacter pyogenes*, a highly recombining member of the eubacteria. After accounting for the probable downward bias of this estimate, the upper limit to $N_l$ for all species dictated by the unavoidable constraints of linkage and selective sweeps may be on the order of $10^{10}$–$10^{11}$. These numbers are relevant because, as will be shown below, they are just a few orders of magnitude above the point at which the population-genetic environment for gene-structure evolution becomes significantly altered.

One caveat with respect to estimates of $N_l$ derived from polymorphism data is that they apply only over the time span necessary for the fixation of an average neutral mutation, $4N_l$ and $2N_l$ generations for diploids and haploids, respectively, which necessarily increases in species with larger $N_l$. Because many of the gross features of genomes may require tens to hundreds of millions of years to emerge, the short-term estimates of $N_l$ for any particular species are likely to frequently misrepresent longer term conditions relevant to genome evolution. For example, newly emergent pathogenic bacteria, which often harbor almost no genetic variation (Daubin and Moran 2004), are not expected to exhibit a signature of random genetic drift at the level of genomic architecture. On the other hand, averages of $N_l$ over the members of broad taxonomic/functional groups (fig. 4) eliminate outliers resulting from sampling error and stochastic temporal fluctuations in population size, thereby providing more meaningful estimates of long-term conditions.

## Drift, Mutation Pressure, and the Emergence of Eukaryotic Gene Complexity

Associated with reductions in $N_l$ in eukaryotes are dramatic expansions in genome size, most of which reflect changes in noncoding regions: introns, mobile elements and their remnants, and other forms of intergenic DNA (fig. 5). A notable feature of these scalings is their continuity over all forms of life, even across the prokaryote-eukaryote boundary. This strongly suggests that neither cellular nor physiological changes associated with phylogenetic transitions are major determinants of genome size. We have previously suggested that the types of genomic evolution that are possible in various lineages are instead largely defined by the population-genetic environment, in particular by the effective number of individuals within a species (Lynch and Conery 2003). In the remainder of this paper, these ideas

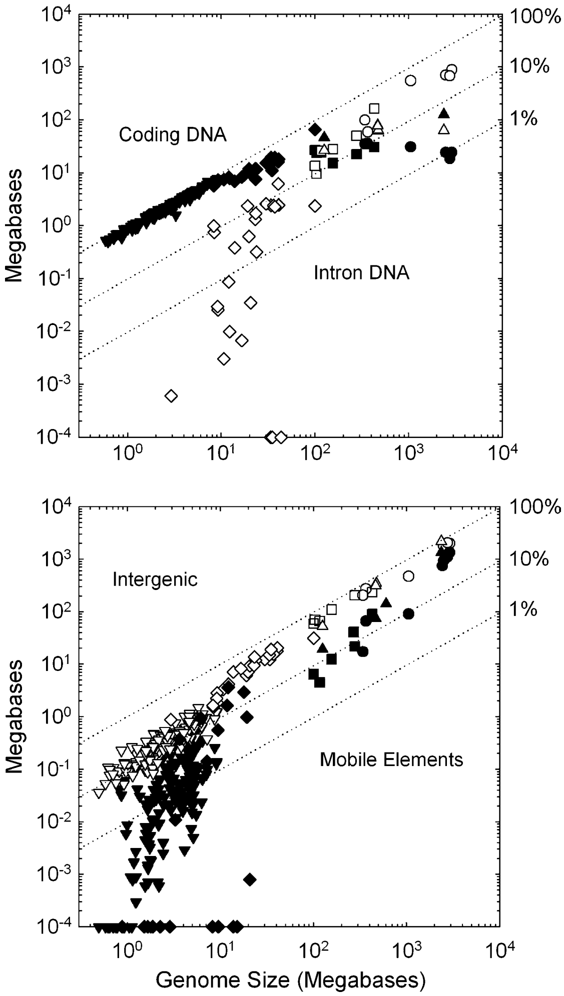

FIG. 5.—Scaling of genome content with genome size in prokaryotes (inverted triangles) and various eukaryotic groupings: unicellular and oligocellular species (diamonds), invertebrates (squares), vertebrates (circles), and land plants (triangles). Diagonal dashed lines denote points of equal proportional contributions to total genome content. Note that mobile-element associated DNA (which includes retrotransposons and DNA-based transposons) may be found in introns as well as intergenic regions and that the intronic DNA depicted here does not include introns in noncoding exons (UTRs). The data were obtained from the various whole-genome sequencing projects.

will be extended to show how many aspects of eukaryotic gene structure may have arisen by nonadaptive processes.

Prokaryotic genes generally have remarkably simple structures—a single continuous coding region with one or two transcription-factor binding (TFB) sites residing just a few nucleotides upstream. Often, a single transcription-initiation site services several downstream prokaryotic genes, which are jointly transformed into a single polycistronic mRNA (an operon). In contrast, the coding regions of eukaryotic genes are often dissected by introns, which are transcribed into precursor mRNAs and then subsequently eliminated by splicing (fig. 6). In multicellular species, dozens of introns may occupy a single gene, and each intron can be many times longer than its surrounding exons. Eukaryotic genes also often have complex sets of regulatory elements distributed over large distances upstream (and sometimes internally or downstream) of the coding region, and with few exceptions, eukaryotic genes are transcribed as single monocistronic units. Finally, eukaryotic gene transcripts are generally flanked by extensive UTRs, which may harbor additional introns. Understanding how such modifications of gene structure emerged is a major challenge for evolutionary genomics because each additional layer of gene complexity entails a cost in terms of mutational vulnerability.

Population geneticists have historically treated selection and mutation as separate forces in the dynamics of evolutionary change, with mutation producing the variation upon which natural selection acts but having no further influence on the fates of alleles. However, in the context of gene architectural features, there are numerous ways in which mutation can act indirectly as a selective agent. Consider a pair of alleles with different forms of gene architecture but otherwise identical functions. Aside from any energetic burden associated with the maintenance of larger numbers of nucleotides, as a larger mutational target, the more complex allele will experience a greater rate of transformation to defective copies. For example, intergenic DNA has the potential to incur mutations that produce spurious TFB sites that cause inappropriate patterns of gene expression; introns necessitate the maintenance of localization signals at the nucleotide level to insure proper mRNA splicing; and 5′ UTRs can acquire premature translation-initiation codons that cause downstream frameshifts. In this sense, most aspects of gene-architectural complexity impose an intrinsic mutational burden. The selective disadvantage associated with any single aspect of gene complexity need not be very large, as it is roughly equivalent to the product of the mutation rate per nucleotide per generation ($u$) and the excess number of nucleotide sites in the more complex allele critical to gene function ($n$) (Lynch 2002), and arguments presented below suggest that $n$ often falls in the approximate range of 1–50. This implies that expansions of gene-architectural complexity are unusually vulnerable to fixation by random genetic drift in populations with small genetic effective sizes.

As noted above, if a costly modification of gene architecture is to evolve in an effectively neutral manner, $4N_gs$ must be smaller than $\sim$1.0. Because $s = nu$ and $\pi_s$ is a function of $N_gu$, this criterion is equivalent to $\pi_sn < 1.0$. Thus, recalling the average estimate of $\pi_s$ for prokaryotes (fig. 4) and its likely downward bias, the population-genetic environment of prokaryotic species may only rarely be conducive to expansions in gene-architectural complexity. In contrast, the extremely low levels of $\pi_s$ for multicellular eukaryotes create situations that are highly permissive to the accumulation of gene architectural changes with weak mutational disadvantages, which are easily overwhelmed by the power of random genetic drift. With their wide ranges of $N_g$, the various lineages of unicellular eukaryotes are expected to fall between these two extremes. It should be noted, however, that although organism size appears to be the primary determinant of $N_g$, it is the latter that ultimately governs the genetic properties of populations. It is conceivable that some unicellular species may reside at a sufficiently low $N_g$ for long enough periods to promote genomic expansion but exceedingly unlikely that any

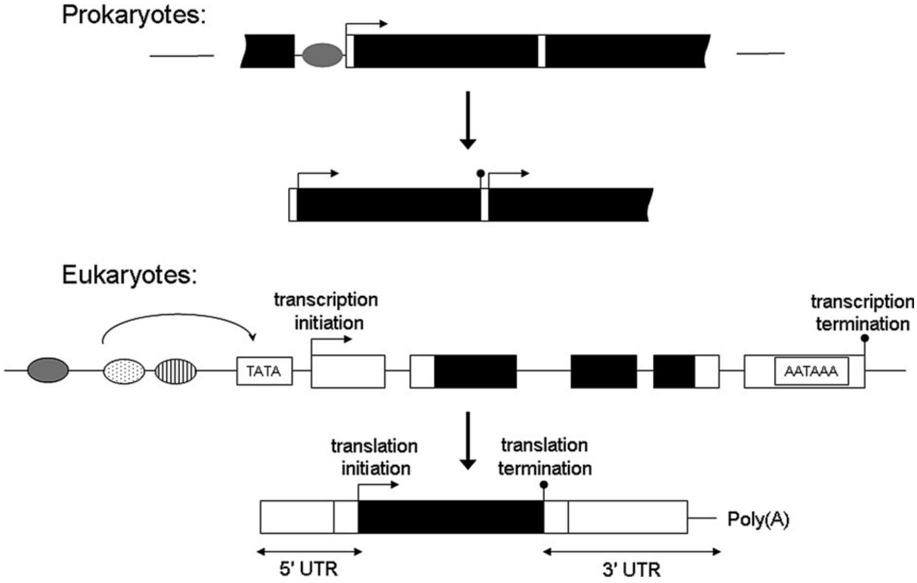

FIG. 6.—Generalized structures of prokaryotic and eukaryotic genes and transcription units. Note that the lengths of various gene parts are not necessarily to scale, for example, eukaryotic genes can harbor as many as several dozens of introns, each of which greatly exceeds the length of its surrounding exons, and TFB sites (small ovals) can be much more numerous and widely distributed than illustrated. TATA denotes one of the possible eukaryotic core promoter elements, to which transcription factors convey information, and poly(A) denotes the posttranscriptional addition of a poly(A) tail. Black bars denote coding DNA, open bars denote transcribed but untranslated DNA, and thin lines within transcribed regions denote introns.

multicellular species ever achieves prokaryote-like levels of $N_I$.

It is clear that the emergence of the complexities of eukaryotic gene structure offered novel opportunities for the evolution of organismal diversity, many of which have been exploited by multicellular eukaryotes, for example, increased regulatory-region complexity and alternative splicing associated with introns. Less certain is whether multiple cell types and mechanisms of cell signaling are advantageous in a formal fitness sense. In any event, because any such adaptive modifications are highly unlikely to have arisen de novo, alternative explanations are needed for the first steps in the retailoring of the eukaryotic genome. Specific examples are now given on how three aspects of eukaryotic gene complexity may have emerged despite their initial intrinsic disadvantages. Each scenario discussed is quantitatively consistent with the theory presented above and shows how a reduction in $N_I$ can passively promote the evolution of gene architectural changes that ultimately facilitate the evolution of organismal complexity by descent with modification.

### *Introns*

As noted above, introns impose a burden on their host genes, in that specific nucleotide signatures must be reserved to insure precise recognition of each exon-intron junction by the spliceosome. The most conserved nucleotide sites are located at the ends of introns and at internal intronic branch points (Burge, Tuschl, and Sharp 1999; Lorković et al. 2000; Bon et al. 2003). However, this information is often insufficient for proper spliceosomal recognition, particularly in the case of large introns containing numerous spurious recognition sites (Mount et al. 1992; Burge, Tuschl, and Sharp 1999; Long and Deutsch 1999). Supplemental information often resides within the surrounding exons in the form of exon splicing enhancers and exon splicing silencers (ESSs), each typically four to ten nucleotides in length (Liu, Zhang, Krainer 1998; Schaal and Maniatis 1999; Blencowe 2000). In mammals, ~2%–4% of exonic sequences match the signatures of known ESSs, with ~5 such clumps per exon (Fairbrother et al. 2002), and the maintenance of such motifs by transcription-related processes is supported by the significantly different frequencies of various oligomers in intron-containing versus intron-free genes (Federov et al. 2001). The most direct evidence for the increased mutational vulnerability associated with introns derives from the observation that about a third of human genetic disorders is attributable to mutations causing defective splice-site recognition (Culbertson 1999; Frischmeyer and Dietz 1999; Philips and Cooper 2000), many of which are located in exons (including substitutions at synonymous sites) (Cooper and Mattox 1997; Nissim-Rafinia and Kerem 2002).

Based on the known molecular requirements for spliceosomal recognition, one may surmise that the equivalent of $n = 20$ to 40 nucleotide sites are required for the precise removal of each intron (Lynch 2002), and indirect estimates based on the incidence of splicing-defective alleles among new mutations are consistent with this prediction (Lynch, Hong, and Scofield 2005*a*). Recalling the theory presented above, a permissive environment for intron colonization requires that $N_Is = N_Inu$ be smaller than ~0.25, or equivalently $\pi_s < 1/n$. Thus, as a first-order approximation, populations with silent-site nucleotide diversities greater than ~0.05 are expected to be nearly immune to intron colonization. Essentially, the full range of variation in observed $\pi_s$ for animals and land plants is well below this threshold value (fig. 4), and all members of these groups have an average of four to seven introns per protein-coding gene (Lynch and Conery 2003). In contrast, the average value of $\pi_s$ for unicellular species (0.06) slightly exceeds the expected threshold, yielding the prediction that the demographic features of such species often place them in close proximity to the barrier to intron colonization (and maintenance). Consistent with this prediction is the broad range of variation in intron numbers in unicellular eukaryotes, ranging from a few dozen or less in the entire genomes of some species (e.g., trypanosomes, the diplomonad *Giardia*, the red alga *Cyanidioschyzon*, and some fungi) to numbers approaching those in animals and land plants in other fungi (fig. 5). For prokaryotes, which are devoid of spliceosomal introns, average $\pi_s$ is more than twice the threshold for intron colonization.

The phylogenetic distribution of introns and the components of the spliceosome make it quite clear that the stem eukaryote harbored introns (Lynch and Richardson 2002; Collins and Penny 2005), and perhaps a substantial number of them (an average of up to three per protein-coding gene being plausible) (Rogozin et al. 2003; Roy and Gilbert 2005*a*; see Qiu, Schisler, and Stoltzfus 2004 for an alternative view). Thus, because there is no evidence of the prior existence of a spliceosome in any prokaryote, the stem eukaryote must have provided a highly permissive environment for intron colonization, with some subsequent lineages then experiencing conditions that favored intron loss and others favoring further intron gain. Could the stem eukaryote have had a sufficiently small population size to allow the accumulation of a substantial intron population via effectively neutral processes alone? Two different approaches have led to the conclusion that the birth rate of introns within the past ~100 Myr is ~0.001/nucleotide site/Byr in invertebrates (Lynch and Richardson 2002; Roy and Gilbert 2005*b*). At this rate, ~1.0 Byr would be required since the origin of the spliceosome for the protein-coding genes of the stem eukaryote to acquire an average of ~1.0 introns, assuming an average coding length of ~1.0 kb as in common in today's eukaryotes. Thus, because the time span between the origin of life and the origin of eukaryotes is ~1.0 Byr (Knoll 1992; Furnes et al. 2004), the passive acquisition of more than one intron per protein-coding gene in the stem eukaryote is just barely plausible, unless the physical rate of intron birth was substantially higher than in today's species.

A more rapid early proliferation of introns could have occurred if some form of positive selection offset the intrinsic disadvantages associated with elevated mutational vulnerability as this would increase the rate of fixation beyond the neutral expectation. One possibility involves the nonsense-mediated decay (NMD) pathway, an mRNA surveillance mechanism for detecting and eradicating transcripts harboring premature termination codons (PTCs). The details of this process have been worked out in only a few organisms, but at least in mammals NMD often uses a protein complex laid down at splicing junctions (the exon junction complex [EJC]) to discriminate PTCs from proper termination codons. If a termination codon is detected upstream of an EJC, the transcript is generally targeted for destruction, a process that works so long as the true termination codon generally resides in the final exon, as is usually the case in mammals (Maquat 2004). Aside from the transcription of mutant alleles, there are many different routes to the stochastic production of PTC-containing transcripts, including base misincorporation, sloppy points of transcription initiation, and erroneous splicing (reviewed in Lynch, Hong, and Scofield 2005*a*). Thus, most cells are regularly confronted with the need to eliminate transcripts that could lead to harmful truncated proteins, and the benefits of doing so via NMD are well documented (Hodgkin et al. 1989; Leeds et al. 1992; Dahlseid et al. 1998; Mendell et al. 2000; Medghalchi et al. 2001).

Phylogenetic analysis suggests that both NMD and the EJC were present in the stem eukaryote (Lynch, Hong, and Scofield 2005*a*). Thus, an early functional association of NMD with introns could have elevated the rate of intron proliferation beyond the neutral expectation (Lynch and Kewalramani 2003). Under this hypothesis, the first intron to colonize a gene would provide a basis for eliminating transcripts with the subset of upstream PTCs. However, because the spatial locations of initially colonizing introns must be largely random, because introns themselves encourage the production of erroneous transcripts via splicing errors, and because some PTCs may be unable to elicit NMD if the nearest EJC is too far downstream, once this coevolutionary process initiated, further colonization of introns would be encouraged. In this manner, some of the earliest colonizing introns (those in locations that allowed sufficient PTC detection) may have had a selective advantage that offset the cost of increased mutational susceptibility. The overdispersed distributions of introns in the genes of multicellular species support the hypothesis that selection favors a uniform coverage of coding regions with introns (Lynch and Kewalramani 2003). Other factors may encourage the colonization of introns (Lynch and Richardson 2002), but the central point here is that once introns became a reliable aspect of a substantial fraction of eukaryotic genes, they served as a natural substrate for secondary adaptive evolution.

Despite the obvious benefits of an mRNA surveillance system, a few species appear to have lost the NMD pathway (Lynch, Hong, and Scofield 2005*a*). These include the kinetoplastids *Trypanosoma* and *Leishmania*, the unicellular red alga *Cyanidioschyzon*, the microsporidian *Encephalitozoon*, and the diplomonad *Giardia*. Remarkably, each of these lineages is almost entirely devoid of introns, and with the exception of *Leishmania*, they all appear to have lost the EJC apparatus. It is tempting to conclude that the loss of introns and NMD must go hand in hand simply because of the latter's functional requirement for an EJC. However, because NMD operates on some genes in an intron-independent manner in a phylogenetically broad group of species (Ruiz-Echevarria, González, and Peltz 1998; Hilleren and Parker 1999; Gatfield et al. 2003; Amrani et al. 2004), it is not clear that introns are an absolute requirement for the maintenance of NMD. An alternative explanation for NMD losses is that the species involved have had very large historical effective sizes, which facilitated the elimination of all forms of extraneous DNA, including mobile elements and most intergenic DNA. As the degree of genomic streamlining increases, the production of erroneous transcripts may eventually decline to the point at which the selective advantage of an intron-based NMD system is no longer sufficient to insure its evolutionary stability (Lynch, Hong, and Scofield 2005*a*).

A central unresolved issue with respect to introns is whether intron numbers have reached a steady-state equilibrium, and if so, what prevents runaway intron colonization. The equilibrium occupancy of introns is a function of the ratio of birth (*b*) to death (*d*) rates per coding nucleotide site (Lynch 2002), but no equilibrium is possible if $b > d$ for all levels of occupancy. The NMD hypothesis provides a potential density-dependent mechanism that could stabilize intron numbers. As a sufficiently well-distributed population of introns is established, the NMD-associated advantages of additional introns will progressively decline until a point is eventually reached at which further intron colonization imposes a net disadvantage. Such a scenario would provide a natural barrier to runaway intron colonization only if the net selective disadvantage of each additional intron exceeded the power of random genetic drift or if the physical rate of intron removal somehow increased with intron number. The average number of introns per protein-coding gene in vertebrates ranges from 5.2 (*Fugu*) to 7.9 (*Gallus*), whereas the range for invertebrates is nearly nonoverlapping, 3.1 (*Drosophila*) to 5.5 (*Bombyx*). Thus, it is clear that the animal lineage experienced a basal increase in intron number, although it is an open question as to whether the numbers in vertebrates continued to expand (perhaps even today) as a consequence of the reduced efficiency of selection associated with low $N_l$. One analysis is consistent with the latter interpretation (Rogozin et al. 2003).

Finally, it is worth considering the origin of the complex molecular machine that makes introns possible, the spliceosome. The most credible hypothesis involves descent from a group II intron (Sharp 1985; Cech 1986; Lambowitz and Zimmerly 2004). Although these "self-splicing" introns have never been found in nuclear genes, their presence in eubacteria, archaea, and the organelles of plants, fungi, and numerous protists (Bonen and Vogel 2001; Dai and Zimmerly 2002, 2003; Rest and Mindell 2003) makes plausible the idea that they were present in the stem eukaryote. Deriving further support from the numerous structural and functional similarities between the excision mechanisms for group II introns and spliceosome-dependent introns (Michel and Ferat 1995; Hetzer et al. 1997; Burge, Tuschl, and Sharp 1999; Sontheimer, Gordon, and Piccirilli 1999; Shukla and Padgett 2002; Valadkhan and Manley 2002), the group II seed hypothesis postulates that the five small RNAs at the heart of the spliceosome are direct descendants of the major subunits of the catalytic core of a group II intron. However, the transition from a self-splicing group II intron to a large population of eukaryotic spliceosome-dependent introns would have involved a number of evolutionary challenges (Stoltzfus 1999; Lynch and Richardson 2002), not the least of which is the reassignment of functional fragments from group II introns associated with specific genes to a more generalized splicing mechanism servicing hundreds to perhaps thousands of genes. The proposed evolutionary pathway to group II intron fragmentation involves a series of effectively neutral steps (Cavalier-Smith 1991; Stoltzfus 1999) that are intrinsic to the subfunctionalization process (Force et al. 1999). However, subfunctionalization is exceedingly unlikely in large populations, which impose difficulties for the establishment of functional fragments without mutational deterioration during the long time required for fixation (Lynch and Richardson 2002). Thus, if the group II seed hypothesis is correct, it reinforces the view that the emergence of eukaryotes was accompanied by a reduction in $N_l$.

### *5′ UTRs*

Like introns, the 5′-untranslated leader sequences of mRNAs are liabilities for genes because they increase the size of the mutational target. Most notably, the 5′ UTR serves as substrate for the mutational appearance of premature translation start codons (PSCs), which because of the scanning mechanism for translation initiation in eukaryotes (Kozak 1994) can lead to N-terminal expansion of the protein product (in ~1/3 of cases) and a shift in the reading frame and protein truncation (in ~2/3 of cases). A deficit of ATG triplets in the exons of 5′ UTRs but much less so in their introns provides compelling evidence of the negative translation-associated consequences of such mutations (Rogozin et al. 2001; Lynch, Scofield, and Hong 2005*b*), as does the incidence of human genetic disorders associated with the appearance of PSCs (Kozak 2002). This raises questions not only as to why 5′ UTRs are present, but why they are so long. Contrary to the situation for introns, which vary in average size by over two orders of magnitude in different phylogenetic groups, the average 5′-UTR lengths of most eukaryotic lineages are remarkably constant, falling in the narrow range of 100–200 bp (Lynch, Scofield, and Hong 2005*b*).

Messenger RNAs are unlikely to require leader sequences as physical landing pads for the ribosome. For example, although archaebacteria employ transcriptional mechanisms similar to those of eukaryotes (Bell and Jackson 2001), their 5′ UTRs are generally no more than a dozen nucleotides in length and in some cases are completely absent (Slupska et al. 2001). Similar situations are observed in eubacteria (Weiner, Herrmann, and Browning 2000; Moll et al. 2002) and in mitochondria (Gillham 1994; Taanman 1999). Moreover, 5′ UTRs in the diplomonad *Giardia* often consist of just a single nucleotide (Iwabe and Miyata 2001), and a substantial fraction of those in a variety of other unicellular eukaryotes are <25 bp, including those in the ciliate *Euplotes crassus* (Ghosh et al. 1994), the amoeba *Entamoeba histolytica* (Singh et al. 1997), and the trichomonad *Trichomonas vaginalis* (Liston and Johnson 1999). Experimental evidence suggests that such diminutive leader sequences are sufficient to support translation in mammals and yeast, although the efficiency of translation can be reduced with UTRs shorter than ~30 bp (van den Heuvel et al. 1989; Maicas, Shago, and Friesen 1990; Hughes and Andrews 1997).

To evaluate whether the expansion of eukaryotic 5′-UTR lengths might be a simple consequence of the reduced efficiency of selection associated with small $N_l$, a simple null model for the stochastic growth and contraction of

UTRs based on mutational gains and losses of PSCs and transcription-initiation signals (TISs, e.g., the TATA box) has been developed (Lynch, Scofield, and Hong 2005*b*). Under this model, all alleles are assumed to be effectively neutral with respect to each other, with the exception of two defective classes: alleles containing a harmful PSC between the TIS and the true translation start codon, and alleles for which the TIS has moved so close to the coding region that transcription initiates beyond the translation-start point. As ATG triplets are free to accumulate upstream of currently utilized transcription-initiation sites, this process results in a natural barrier to excessive growth of 5′ UTRs, which can only expand in a 5′ direction if the extension is devoid of harmful PSCs. Over time, the stochastic winking on and off of PSCs and TISs upstream of the true translation-start site results in an equilibrium L-shaped distribution of 5′-UTR lengths with a mean and variance that are quite similar to those observed within a wide variety of species, largely independent of the assumed length of the TIS (Lynch, Scofield, and Hong 2005*b*). This result suggests that the reduced mutational burden of short 5′ UTRs may be inconsequential in comparison to the power of random genetic drift in most eukaryotes.

To understand the mechanistic underpinnings of such a condition, it is necessary to consider the long-term excess mutation rate to defective alleles for an allelic lineage in which UTR lengths simply stochastically grow and contract (as in the above model) without selective discrimination among functional alleles, in contrast to the situation in which minimal length UTRs are maintained by positive selection. The mutational disadvantage of stochastically generated 5′ UTRs increases with the length of the TIS ($n$) not only because of the larger target size of the TIS itself, but also because a larger $n$ implies that in the event of a TIS loss, the nearest upstream alternative TIS will be increasingly likely to reside beyond a harmful PSC (because of the rarity of complex sequences). This is a potential concern because of the limited information available on the length of core promoter sequences. However, the key result is fairly robust to the length of the TIS—for $n$ in the range of 4 (e.g., TATA) to 8 (which is beyond the complexity of known TISs), the long-term mutational advantage of a minimal-length 5′ UTR is in the narrow range of $u$ to $4u$ (Lynch, Scofield, and Hong 2005*b*), substantially smaller than the mutational penalty associated with an intron.

Again following the logic outlined above, this result implies that a permissive environment for the stochastic expansion of 5′ UTRs is $\pi_s < 0.25$. Thus, because all estimates of $\pi_s$ for multicellular eukaryotes are well below 0.10 (fig. 4), the power of genetic drift in all such species is far too great for the maintenance of minimal-length 5′ UTRs by positive selection. The estimates of $\pi_s$ for essentially all unicellular eukaryotes are also below the threshold value of 0.25, although as discussed above, some of these estimates are likely to be downwardly biased, raising the possibility that the long-term $N_l$, of some microbial eukaryotes and certainly many prokaryotes is adequate to allow the selective promotion of alleles with 5′ UTRs of minimal length. Although the data are limited, a few protist lineages do appear to have average 5′-UTR lengths too small to be accommodated by the null model (Ghosh et al. 1994; Singh et al. 1997; Liston and Johnson 1999; Yee et al. 2000; Adam 2001).

If the preceding hypothesis is correct, then selection for gene-specific regulatory features need not be invoked to explain either the expansion of eukaryotic 5′ UTRs relative to the situation in prokaryotes or the thousandfold range of 5′-UTR lengths among genes within species. Nevertheless, once permanently established, expanded 5′ UTRs may have provided novel substrate for the evolution of posttranscriptional mechanisms for regulating gene expression. For example, numerous features of 5′ UTRs (including their lengths) in today's eukaryotes can influence the rate of protein synthesis by modifying the efficiency of translation (de Moor and Richter 2001; Pickering and Willis 2005), and some have argued that upstream open reading frames (uORFs) serve an adaptive function by causing the ribosome to terminate and/or reinitiate, thereby slowing the rate of translation (Morris and Geballe 2000; Meijer and Thomas 2002; Vilela and McCarthy 2003). However, the extent to which any features of 5′ UTRs are adaptive remains an open question. For example, uORFs are generally ~20 codons in length, approximately what is expected by chance, and these may simply exist because their stop codons have neutralized the effects of a PSC, enabling the downstream coding domain to perform at normal levels.

### *Modularization of Regulatory Regions*

The ability to regulate the expression of genes in tissue-, developmental stage-, and/or environment-specific manners is the hallmark of organisms with multiple cell types. To accomplish such tasks, genes generally harbor numerous *cis*-acting regulatory elements that cooperatively interact with multiple *trans*-acting factors to finely tune levels of transcription. These regulatory elements are often organized into modular units such that mutations to individual gene subfunctions have restricted phenotypic effects (Force et al. 1999; Prince and Pickett 2002). It has often been argued that this particulate form of gene regulation was a fundamental prerequisite for the origin of developmental modules capable of independent evolutionary trajectories, and some have suggested that such organization enhances the evolvability of lineages (Raff 1996; Wagner 1996; Wagner and Altenberg 1996; Gerhart and Kirschner 1997; Hartwell et al. 1999; Niehrs and Pollet 1999; Carroll, Grenier, and Weatherbee 2001; Davidson 2001). However, it is an open question as to whether evolvability, a population-level phenomenon, is ever promoted by natural selection.

Each essential regulatory element elevates a gene's susceptibility to inactivating mutations, and it is unclear whether adaptive processes are sufficient to explain the *origin* of modular gene structure and regulatory logic. Therefore, an essential first step in understanding the initial establishment of modular gene-regulatory structure is a consideration of the situation in which selection is a negligible force (Zuckerkandl 2001; Force et al. 2005). The following discussion outlines how the modular organization of regulatory regions can arise passively in response to drift and mutation processes alone. To simplify the presentation, new subfunctions will be assumed to be defined by TFB sites (or integrated regions of such sites) that are potentially separable from other such sites, both mutationally and functionally (Yuh, Bolouri, and Davidson 1998; Force et al. 1999; Arnosti 2003). However, the principles outlined below are broadly generalizable, as a gene subfunction may also correspond to alternative splice sites, transcription-initiation sites, polyadenylation signals, etc.

Before exploring the population-genetic issues, a brief overview of eukaryotic transcriptional control will be useful. Transcription-factor genes are generally expressed at specific stages of development at a hierarchy of levels of organization. Some transcription factors are organ specific, some are specific to individual tissues within an organ, and still others are expressed in specific cell types within a tissue. By virtue of their ability to bind DNA and/or to act as cofactors in binding each other together in functionally significant ways, the full suite of transcription factors and their combinatorial expression provides a heterogeneous informational network within the organism. A gene's pattern of expression is then defined by the match between its TFB sites and the local transcription-factor environment. Although the entire regulatory domain of a gene can sometimes rival the length of the coding region and individual genes can harbor dozens of regulatory elements, individual TFB sites are usually no more than a dozen base pairs in length, and such sites are often partially overlapping or entirely embedded. Because small regulatory elements can arise by a fortuitous series of de novo base substitutions and/or by segmental insertions, there are many potential ways in which the numbers, locations, and types of TFB sites for a gene can become modified (Brosius 1999; von Dassow and Monro 1999; Edelman et al. 2000, Stone and Wray 2001; Rockman and Wray 2002; Jordan et al. 2003; Wray et al. 2003; MacArthur and Brookfield 2004). However, the evolution of a new form of gene regulatory architecture will generally require a period of redundancy, as this will allow the modification of one set of TFB sites before the other is abandoned, thereby insuring overall constancy of gene expression during the transition phase. Because the origin of tandem duplicated stretches of DNA is quite high, especially for small segments (Lynch and Conery 2000; Katju and Lynch 2003), such situations appear to be very common in eukaryotes, and various permutations of TFB sites often yield functionally equivalent patterns of gene expression (Bonneton et al. 1997; Hancock et al. 1999; Ludwig et al. 2000; Dermitzakis and Clark 2002; Shaw et al. 2002; Arnosti 2003; Dermitzakis, Bergman, and Clark 2003).

Before proceeding, it is worth emphasizing that the goal here is not to explain the expression of a gene in a new temporal or spatial context, but to understand how a gene that is initially under control of a ubiquitously expressed transcription factor comes to be regulated by spatially and/or temporally restricted transcription factors while initially retaining the same overall expression pattern. The process envisioned, subfunction fission, invokes gradual structural modifications of preexisting enhancers within a gene (descent with modification) rather than the saltatory appearance of entirely new regulatory modules (Force et al. 2005). Subfunction fission involves consecutive phases of regulatory-region expansion and contraction (fig. 7).

The first step toward the modularization of a gene's regulatory apparatus, accretion and degeneration, involves the acquisition of new semi-redundant TFB sites combined with the degeneration of one or more ancestral sites. This converts a universally used set of TFB sites to a semi-independent enhancer with a pair of partially overlapping regulatory elements (series a–d, upper left of fig. 7). If the alternative TFB sites are subject to both mutational gain and loss, there is not necessarily a permanent allelic state under this model, as the alternative alleles are free to mutate back and forth (hence, the two-way arrows in fig. 7). Nevertheless, it is instructive to know the population-genetic conditions under which the semi-independently regulated allele (d) is most likely to accumulate as this is an essential first step in the transition to an allele with two entirely independent subfunctions. Letting $\mu_b$ and $\mu_d$, respectively, denote the rates of birth and loss of individual TFB sites, it is possible to estimate the average time that transpires between a transient state fixed for an allele with shared regulatory sites (a) and an allele with a semi-independent enhancer (d) (Force et al. 2005). Provided $4N_l\mu_d \ll 1$, the transition time to the semi-independently regulated state is independent of population size because the waiting times for mutational changes, $1/\mu_b$ and $1/\mu_d$, are much greater than the time to drift to fixation, $4N_l$. In contrast, as $N_l$ approaches the point at which $4N_l\mu_d = 1$, the transition time to the semi-independently regulated allele begins to scale linearly with $N_l$. Thus, as in the case of introns and 5′ UTRs, the key condition for maximizing the likelihood of a transition from a simple to more complex (semi-independently regulated) regulatory region, $4N_l\mu_d \ll 1$, is a function of the ratio of the power of a mutational force ($2\mu_d$ being the rate of loss of TFB sites per individual) and the power of drift ($1/2N_l$).

Although the net addition of a regulatory element to a gene imposes a weak selective disadvantage by increasing the mutation rate to defective alleles, selection is ineffective except in very large populations. Given the small size of typical TFB sites, $4N_l\mu_d$ is expected to be on the order of $10\pi_s$ (or perhaps somewhat higher, after accounting for bias and also allowing for inactivation by insertion/deletion mutations). In any event, again recalling the results in figure 4, the condition for accretion and degeneration is probably fulfilled in all multicellular eukaryotes as well as in numerous unicellular eukaryotes and probably violated in most prokaryotes. This implies that that as soon as transcription factors with spatially and/or temporally restricted patterns of expression evolved, large numbers of broadly expressed genes in eukaryotes would have become subject to the passive accumulation of alleles with semi-independent regulation.

In the second step toward complete modularization (duplication, degeneration, and complementation), the local duplication of a semi-independent enhancer provides an opportunity for the formation of two entirely independent regulatory elements (lower left of fig. 7). The events during this phase are conceptually identical to those underlying the subfunctionalization model for the preservation of duplicate genes (Force et al. 1999; Lynch et al. 2001), although in this case the process involves smaller regulatory elements. Progression to the modularized state during this phase

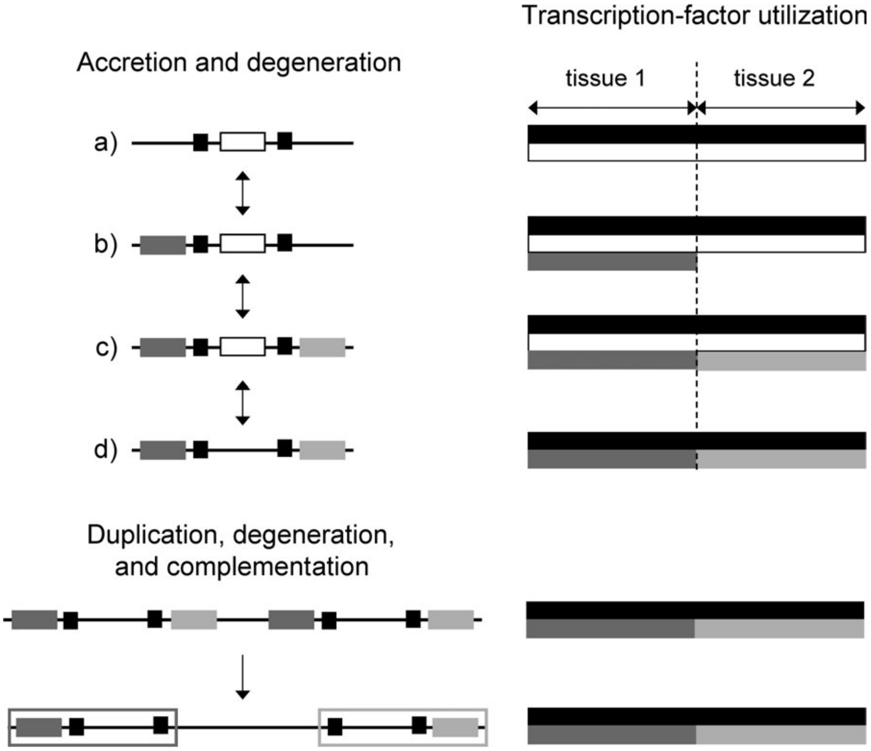
Transcription-factor utilization
tissue 1
tissue 2
Accretion and degeneration
a)
b)
c)
d)
Duplication, degeneration, and complementation

FIG. 7.—A hypothetical scenario in which an allele with two independently mutable subfunctions can arise from an allele with a single generalized expression mechanism. Regulatory regions are depicted on the left, with each regulatory element color coded according to the transcription factor that binds to it. On the right, the allele-specific utilizations of transcription factors are depicted. Transcription factors denoted by black and white are ubiquitously expressed, whereas those denoted by two shades of grey are each expressed in single, nonoverlapping tissues. In the first phase of allelic evolution (accretion and degeneration), the regulatory region undergoes sequential accretion of the dark and light grey elements, which together are redundant with respect to the white (but not the black) element. The redundant white element is then lost, yielding a descendant allele will a semi-independent mode of expression, as the black element is still essential to expression in both tissues. In the second phase (duplication, degeneration, and complementation), the entire enhancer region is tandemly duplicated, with each component then losing a complementary (light/dark grey) element. The resultant allele has two independently mutable subfunctions denoted by the dark and light grey open boxes, as a mutation in either region has effects that are confined to a single tissue. Note that throughout these transitions, there has been no change in the pattern of deployment of the gene, which is always expressed in both tissues; only the mechanism of achieving expression is altered.

requires that each semi-independent enhancer incurs a complementary loss of tissue-specific function prior to one of the enhancers being entirely silenced. The probabilities of these alternative outcomes depend on the relative rates of loss of tissue-specific (two shades of grey) and shared (black) TFB sites, but regardless of these mutational properties, large $N_l$ can impose a complete barrier to enhancer subfunctionalization. This population-size dependence is again a consequence of a modularized allele serving as a larger mutational target (four black binding sites at the bottom left of fig. 7, as opposed to two for allele d) (Lynch et al. 2001; Force et al. 2005).

These results have several implications. First, a modular, tissue-specific gene regulatory structure can emerge spontaneously from an initial state in which the full expression breadth of a gene is under unified control. Not only does this process proceed via entirely nonadaptive processes (random genetic drift and mutation), but because alleles with more complex regulatory architectures have an excess mutation rate to nulls, it is less likely to occur in situations where the power of natural selection exceeds that of drift. Thus, contrary to popular belief, not only may natural selection be an insufficient mechanism for the origin of genotypic modularity, but conditions where selection is most likely to be efficient (large populations) promote the opposite situation—alleles under unified transcriptional control. In this sense, the reduction in $N_l$ that accompanied the origin of eukaryotes may have paved the way for the emergence of modular gene architecture without any direct involvement of positive selection. The process will be even more powerful in multicellular species because a wider range of regulatory-region structures with mildly deleterious fitness effects (beyond the mutational disadvantage) will be effectively neutral, providing additional routes for the colonization of new TFB sites.

Second, the transition to modular gene structure may not always require the accretion of new regulatory elements. Due to the recurrent mutational production of alternative TFB sites, large populations can be expected to harbor high frequencies of semi-modular alleles (of the form c in fig. 7), which owing to the presence of redundant TFB sites (white and grey elements) have a weak mutationally induced selective advantage relative to alleles of type a (Wagner 1999, 2001). This implies that a population that has experienced a prolonged phase of large effective size, which inhibits the evolution of modularity, can nevertheless be poised to make a transition to allelic state d following a population-size reduction and the subsequent neutral loss of the shared redundant (white) TFB site from its precursor allele (c).

Finally, although the pattern of tissue-specific gene expression remains unaltered throughout the process of subfunction fission, the underlying molecular mechanisms for achieving the pattern are modified in ways that can eliminate pleiotropic constraints associated with shared regulatory elements, thereby opening up previously inaccessible evolutionary pathways. Moreover, the spontaneous emergence of modular gene organization by random drift and mutation processes has a powerful secondary effect. When a modularized gene is duplicated in a small population, which occurs frequently (Lynch and Conery 2000), the two members of the duplicate pair will have a high probability of experiencing complementary degenerative mutations that result in gene specialization (Force et al. 1999; Lynch and Force 2000; Lynch et al. 2001), eliminating still another layer of pleiotropic constraints. This joint operation of subfunction formation at the gene level and subfunctionalization at the genome level provides a passive mechanism for the gradual remodeling of entire developmental genetic pathways without any direct initial involvement of natural selection. In this sense, multicellular species may be crucibles for the origin of complex patterns of cellular and developmental patterning, not because such features are directly advantageous but because they are unavoidable consequences of small population size environments.

## Concluding Remarks

Despite the enormous progress in molecular genetics over the past 50 years, no general theory for the evolution of the basic architectural features of genes has been formulated. Many attempts have been made to explain the features of genes, genomes, and genetic networks in the context of putatively adaptive cellular and/or developmental features, but few of these efforts have been accompanied by a formal evolutionary analysis. Because evolution is a population-level process, any theory for the origins of the genetic machinery must ultimately be consistent with basic population-genetic mechanisms. However, because natural selection is just one of several forces contributing to the evolutionary process, an uncritical reliance on adaptive Darwinian mechanisms to explain all aspects of organismal diversity is not greatly different than invoking an intelligent designer.

This paper represents a first step toward the formal development of a general theory for the evolution of the gene that incorporates the universal properties of random genetic drift and mutation pressure. Although the ideas presented are unlikely to be correct in every detail, at a minimum they serve as a null model. For if verbal adaptive arguments are to provide confident explanations for any aspect of gene or genomic structure, something must be known about patterns expected in the absence of selection. This is a significant challenge because at this point it is difficult to reject the hypothesis that the basic embellishments of the eukaryotic gene originated largely as a consequence of nonadaptive processes operating contrary to the expected direction of natural selection. A significant area of future research will be to take these observations on gene and genome complexity to the next level, to evaluate whether natural selection is a necessary and/or sufficient force to explain the evolution of the cellular and developmental complexities of eukaryotes.

Finally, an issue not considered above is the potential energetic consequences of the expansion of eukaryotic genes. It is commonly argued that a significant metabolic burden associated with excess DNA is responsible for the streamlining of prokaryotic genomes by natural selection (Cavalier-Smith 2005; Giovannoni et al. 2005). However, no direct estimate of the contribution of DNA replication to a cell's total energy budget has ever been made, and the lack of correlation between prokaryotic cell division rate and genome size does not inspire confidence in the metabolic-cost hypothesis (Mira, Ochman, and Moran 2001). Castillo-Davis et al. (2002) have suggested that the negative correlation between intron size and the level of expression of genes in nematodes and humans reflects an adaptive response to the costs of transcription, although introns in highly transcribed genes might simply experience higher levels of deletion. In the only attempt to calculate the costs of transcription, Wagner (2005) concluded that a doubling of the transcription of a yeast gene might cause a fractional reduction of fitness of $\sim 10^{-5}$. Numerous assumptions had to be made to arrive at this conclusion, and incremental modifications to the complexity of transcribed portions of genes may generally only be ~1% of the total transcript length, which would imply an incremental transcriptional cost of $\sim 10^{-7}$ or perhaps lower, depending on the relationship between energetic cost and fitness. A welcome addition to this area would be a formal evaluation of the costs of DNA and mRNA production relative to the total energetic requirements of cell maintenance and replication. If such costs can be translated into a selection coefficient, the ability of natural selection to resist them can then be evaluated using the same procedures discussed above.

## Supplementary Materials

Supplementary Tables 1–4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

I am very grateful to E. Koonin and F. Kondrashov for important comments that led to significant improvement of this presentation. J. Conery helped in procuring the prokaryotic genome content data. The work has been supported by National Institutes of Health grant GM36827 and National Science Foundation grant MCB-0342431.

## Literature Cited

Adam, R. D. 2001. Biology of *Giardia lamblia*. Clin. Microbiol. Rev. **14**:447–475.

Amrani, N., R. Ganesan, S. Kervestin, D. A. Mangus, S. Ghosh, and A. Jacobson. 2004. A faux 3′-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. Nature **432**:112–118.

Arnosti, D. N. 2003. Analysis and function of transcriptional regulatory elements: insights from *Drosophila*. Ann. Rev. Entomol. **48**:579–602.

Bell, S. D., and S. P. Jackson. 2001. Mechanism and regulation of transcription in Archaea. Curr. Opin. Microbiol. **4**:208–213.

Blencowe, B. J. 2000. Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem. Sci. **25**:106–110.

Bon, E., S. Casaregola, G. Blandin et al. (11 co-authors). 2003. Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. Nucleic Acids Res. **31**: 1121–1135.

Bonen, L., and J. Vogel. 2001. The ins and outs of group II introns. Trends Genet. **17**:322–331.

Bonner, J. T. 1988. The evolution of complexity. Princeton University Press, Princeton, N.J.

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.

Brown, G. R., G. P. Gill, R. J. Kuntz, C. H. Langley, and D. B. Neale. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. Proc. Natl. Acad. Sci. USA **101**:15255–15260.

Burge, C. B., T. Tuschl, and P. A. Sharp. 1999. Splicing of precursors to mRNAs by the spliceosomes, Pp. 525–560 *in* R. F. Gesteland, T. R. Cech, and J. F. Atkins, eds. The RNA world, 2nd edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Bustamante, C. D., R. Nielsen, and D. L. Hartl. 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol. Biol. Evol. **19**:110–117.

Caballero, A. 1994. Developments in the prediction of effective population size. Heredity **73**:657–679.

Carbone, C., and J. L. Gittleman. 2002. A common rule for the scaling of carnivore density. Science **295**:2273–2276.

Carroll, S. B., J. K. Grenier, and S. D. Weatherbee. 2001. From DNA to diversity. Blackwell Science, Malden, Mass.

Castillo-Davis, C. I., S. L. Mekhedov, D. L. Hartl, E. V. Koonin, and F. A. Kondrashov. 2002. Selection for short introns in highly expressed genes. Nat. Genet. **31**:415–418.

Cavalier-Smith, T. 1991. Intron phylogeny: a new hypothesis. Trends Genet. **7**:145–148.

———. 2005. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann. Bot. **95**:147–175.

Cech, T. R. 1986. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. Cell **44**:207–210.

Chamary, J. V., and L. D. Hurst. 2004. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively driven codon usage. Mol. Biol. Evol. **21**:1014–1023.

Collins, L., and D. Penny. 2005. Complex spliceosomal organization ancestral to extant eukaryotes. Mol. Biol. Evol. **22**: 1053–1066.

Cooper, T. A., and W. Mattox. 1997. The regulation of splice-site selection, and its role in human disease. Am. J. Hum. Genet. **61**:259–266.

Culbertson, M. R. 1999. RNA surveillance: unforeseen consequences for gene expression, inherited genetic disorders and cancer. Trends Genet. **15**:74–80.

Dahlseid, J. N., J. Puziss, R. L. Shirley, A. L. Atkin, P. Hieter, and M. R. Culbertson. 1998. Accumulation of mRNA coding for the ctf13p kinetochore subunit of *Saccharomyces cerevisiae* depends on the same factors that promote rapid decay of nonsense mRNAs. Genetics **150**:1019–1035.

Dai, L., and S. Zimmerly. 2002. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. Nucleic Acids Res. **30**:1091–1102.

———. 2003. ORF-less and reverse-transcriptase-encoding group II introns in Archaebacteria, with a pattern of homing into related group II intron ORFs. RNA **9**:14–19.

Daly, M. J., J. D. Rioux, S. F. Schaffner, T. J. Hudson, and E. S. Lander. 2001. High-resolution haplotype structure in the human genome. Nature Genet. **29**:229–232.

Damuth, J. 1981. Population density and body size in mammals. Nature **290**:699–700.

Daubin, V., and N. A. Moran. 2004. Comment on “the origins of genome complexity”. Science **306**:978.

Davidson, E. H. 2001. Genomic regulatory systems: development and evolution. Academic Press, San Diego, Calif.

Dawson, E., G. R. Abecasis, S. Bumpstead et al. (29 co-authors). 2002. A first-generation linkage disequilibrium map of human chromosome 22. Nature **418**:544–548.

de Moor, C. H., and J. D. Richter. 2001. Translational control in vertebrate development. Int. Rev. Cytol. **203**:567–608.

Denver, D. R., K. Morris, M. Lynch, and W. K. Thomas. 2004. High mutation rate and dominance of length change mutations in the nuclear genome of *Caenorhabditis elegans*. Nature **430**:679–682.

Dermitzakis, E. T., C. M. Bergman, and A. G. Clark. 2003. Tracing the evolutionary history of *Drosophila* regulatory regions with models that identify transcription factor binding sites. Mol. Biol. Evol. **20**:703–714.

Dermitzakis, E. T., and A. G. Clark. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. **19**: 1114–1121.

Desai, D., K. Zhang, S. Barik, A. Srivastava, M. E. Bolander, and G. Sarkar. 2004. Intragenic codon bias in a set of mouse and human genes. J. Theor. Biol. **230**:215–225.

Diniz, J. A. F., and N. M. Torres. 2002. Phylogenetic comparative methods and the geographic range size—body size relationship in new world terrestrial Carnivora. Evol. Ecol. **16**:351–367.

Drake, J. W. 1991. A constant rate of spontaneous mutation in DNA-based microbes. Proc. Natl. Acad. Sci. USA **88**: 7160–7164.

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.

Enquist, B. J., and K. J. Niklas. 2001. Invariant scaling relations across tree-dominated communities. Nature **410**:655–660.

Fairbrother, W. G., R. F. Yeh, P. A. Sharp, and C. B. Burge. 2002. Predictive identification of exonic splicing enhancers in human genes. Science **297**:1007–1013.

Fedorov, A., X. Cao, S. Saxonov, S. J. de Souza, S. W. Roy, and W. Gilbert. 2001. Intron distribution difference for 276 ancient and 131 modern genes suggests the existence of ancient introns. Proc. Natl. Acad. Sci. USA **98**:13177–13182.

Feil, E. J., E. C. Holmes, D. E. Bessen et al. (12 co-authors). 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA **98**:182–187.

Finlay, B. J. 2002. Global dispersal of free-living microbial eukaryote species. Science **296**:1061–1063.

Finlay, B. J., G. F. Esteban, K. J. Clarke, and J. L. Olmo. 2001. Biodiversity of terrestrial protozoa appears homogeneous across local and global spatial scales. Protist **152**: 355–366.

Finlay, B. J., E. B. Monaghan, and S. C. Maberly. 2002. Hypothesis: the rate and scale of dispersal of freshwater diatom species is a function of their global abundance. Protist **153**: 261–273.

Force, A., W. A. Cresko, F. B. Pickett, S. R. Proulx, C. Amemiya, and M. Lynch. 2005. The origin of subfunctions and modular gene regulation. Genetics **170**:433–446.

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

Frankham R. 1995. Effective population size/adult population size ratios in wildlife: a review. Genet. Res. **66**:95–107.

Frischmeyer, P. A., and H. C. Dietz. 1999. Nonsense-mediated mRNA decay in health and disease. Hum. Mol. Genet. **8**:1893–1900.

Furnes, H., N. R. Banerjee, K. Muehlenbachs, H. Staudigel, and M. de Wit. 2004. Early life recorded in archean pillow lavas. Science **304**:578–581.

Gabriel, S. B., S. F. Shaffner, H. Nguyen et al. (18 co-authors). 2002. The structure of haplotype blocks in the human genome. Science **296**:2225–2229.

Gaston, K. J., and T. M. Blackburn. 1996. Range size-body size relationships: evidence of scale dependence. Oikos **75**: 479–485.

Gatfield, D., L. Unterholzner, F. D. Ciccarelli, P. Bork, and E. Izaurralde. 2003. Nonsense-mediated mRNA decay in *Drosophila*: at the intersection of the yeast and mammalian pathways. EMBO J. **22**:3960–3970.

Gerhart, J., and M. Kirschner. 1997. Cells, embryos, and evolution. Blackwell Science, Malden, Mass.

Ghosh, S., J. W. Jaraczewski, L. A. Klobutcher, and C. L. Jahn. 1994. Characterization of transcription initiation, translation initiation, and poly(A) addition sites in the gene-sized macronuclear DNA molecules of *Euplotes*. Nucleic Acids Res. **22**:214–221.

Gillham, N. W. 1994. Organelle genes and genomes. Oxford University Press, New York.

Gillespie, J. H. 2000. Genetic drift in an infinite population: the pseudohitchhiking model. Genetics **155**:909–919.

Giovannoni, S. J., et al. 2005. Genome streamlining in a cosmopolitan oceanic bacterium. Science **309**:1242–1245.

Green, J. L., A. J. Holmes, M. Westoby, I. Oliver, D. Briscoe, M. Dangerfield, M. Gillings, and A. J. Beattie. 2004. Spatial scaling of microbial eukaryote diversity. Nature **432**: 747–750.

Greenwood, T. A., B. K. Rana, and N. J. Schork. 2004. Human haplotype block sizes are negatively correlated with recombination rates. Genome Res. **14**:1358–1361.

Halligan, D. L., A. Eyre-Walker, P. Andolfatto, and P. D. Keightley. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. Genome Res. **14**:273–279.

Hammond, P. M. 1995. Described and estimated species numbers: an objective assessment of current knowledge, Pp. 29–71 *in* D. Allsopp, R. R. Colwell, and D. L. Hawksworth, eds. Microbial diversity and ecosystem function. CAB International, Wallingford, United Kingdom.

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.

Hartl, D. L., S. K. Volkman, K. M. Nielsen, A. E. Barry, K. P. Day, D. F. Wirth, and E. A. Winzeler. 2002. The paradoxical population genetics of *Plasmodium falciparum*. Trends Parasitol. **18**:266–272.

Hartwell, L. H., J. J. Hopfield, S. Leibler, and A. W. Murray. 1999. From molecular to modular cell biology. Nature **402**:C47–C52.

Hellmann, I., I. Ebersberger, S. E. Ptak, S. Paabo, and M. Przeworski. 2003. A neutral explanation for the correlation of diversity with recombination rates in humans. Am. J. Hum. Genet. **72**:1527–1535.

Hetzer, M., G. Wurzer, R. J. Schweyen, and M. W. Mueller. 1997. Trans-activation of group II intron splicing by nuclear U5 snRNA. Nature **386**:417–420.

Hey, J., and J. Wakeley. 1997. A coalescent estimator of the population recombination rate. Genetics **145**:833–846.

Hill, W. G. 1975. Linkage disequilibrium among multiple neutral alleles produced by mutation in finite population. Theor. Popul. Biol. **8**:117–126.

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

Hilleren, P., and R. Parker. 1999. mRNA surveillance in eukaryotes: kinetic proofreading of proper translation termination as assessed by mRNP domain organization? RNA **5**:71171–71179.

Hodgkin, J., A. Papp, R. Pulak, V. Ambros, and P. Anderson. 1989. A new kind of informational suppression in the nematode *Caenorhabditis elegans*. Genetics **123**:301–313.

Horner-Devine, M. C., M. Lage, J. B. Hughes, and B. J. Bohannan. 2004. A taxa-area relationship for bacteria. Nature **432**:750–753.

Housworth, E. A., E. P. Martins, and M. Lynch. 2003. The phylogenetic mixed model. Am. Nat. **163**:84–96.

Hughes, M. J., and D. W. Andrews. 1997. A single nucleotide is a sufficient 5′ untranslated region for translation in an eukaryotic in vitro system. FEBS Lett. **414**:19–22.

Huxley, J. S. 1942. Evolution: the modern synthesis. Allen and Unwin, London, United Kingdom.

Ibrahim, K. M., S. J. Cooper, and G. M. Hewitt. 2002. Testing for recombination in a short nuclear DNA sequence of the European meadow grasshopper, *Chorthippus parallelus*. Mol. Ecol. **11**:583–590.

Iwabe, N., and T. Miyata. 2001. Overlapping genes in parasitic protist *Giardia lamblia*. Gene **280**:163–167.

Jordan, I. K., I. B. Rogozin, G. V. Glazko, and E. V. Koonin. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet. **19**: 68–72.

Kado, T., H. Yoshimaru, Y. Tsumura, and H. Tachida. 2003. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae *sensu lato*). Genetics **164**:1547–1559.

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.

Keightley, P. D., and A. Eyre-Walker. 2000. Deleterious mutations and the evolution of sex. Science **290**:331–333.

Kimura, M. 1962. On the probability of fixation of mutant genes in populations. Genetics **47**:713–719.

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, United Kingdom.

Knoll, A. H. 1992. The early evolution of eukaryotes: a geological perspective. Science **256**:622–627.

Kondrashov, A. S. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. Hum. Mutat. **21**:12–27.

Kozak, M. 1994. Determinants of translational fidelity and efficiency in vertebrate mRNAs. Biochimie **76**:815–821.

———. 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene **299**:1–34.

Kumar, S., and S. Subramanian. 2002. Mutation rates in mammalian genomes. Proc. Natl. Acad. Sci. USA **99**: 803–808.

Lagercrantz, U., M. Kruskopf Osterberg, and M. Lascoux. 2002. Sequence variation and haplotype structure at the putative flowering-time locus COL1 of *Brassica nigra*. Mol. Biol. Evol. **19**:1474–1482.

Lambowitz, A. M., and S. Zimmerly. 2004. Mobile group II introns. Ann. Rev. Genet. **38**:1–35.

Laporte, V., and B. Charlesworth. 2002. Effective population size and population subdivision in demographically structured populations. Genetics **162**:501–519.

Leeds, P., J. M. Wood, B. S. Lee, and M. R. Culbertson. 1992. Gene products that promote mRNA turnover in *Saccharomyces cerevisiae*. Mol. Cell. Biol. **12**:2165–2177.

Liston, D. R., and P. J. Johnson. 1999. Analysis of a ubiquitous promoter element in a primitive eukaryote: early evolution of the initiator element. Mol. Cell. Biol. **19**:2380–2388.

Liu, H. X., M. Zhang, and A. R. Krainer. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. **12**:1998–2012.

Long, M., and M. Deutsch. 1999. Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Mol. Biol. Evol. **16**:1528–1534.

Lorković, Z. J., D. A. Wieczorek Kirk, M. H. Lambermon, and W. Filipowicz. 2000. Pre-mRNA splicing in higher plants. Trends Plant Sci. **5**:160–167.

Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**:564–567.

Lynch, M. 1997. Mutation accumulation in nuclear, organelle, and prokaryotic genomes: transfer RNA genes. Mol. Biol. Evol. **14**:914–925.

———. 2002. Intron evolution as a population-genetic process. Proc. Natl. Acad. Sci. USA **99**:6118–6123.

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

———. 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., X. Hong, and D. G. Scofield. 2005*a*. Nonsense-mediated decay and the evolution of eukaryotic gene structure. *In* L. E. Maquat, ed. Nonsense-mediated decay. Landes Bioscience, Georgetown, Tex (in press).

Lynch, M., and A. Kewalramani. 2003. Messenger RNA surveillance and the evolutionary proliferation of introns. Mol. Biol. Evol. **20**:563–571.

Lynch, M., and A. O. Richardson. 2002. The evolution of spliceosomal introns. Curr. Opin. Genet. Dev. **12**:701–710.

Lynch, M., D. G. Scofield, and X. Hong. 2005*b*. The evolution of transcription-initiation sites. Mol. Biol. Evol. **22**:1137–1146.

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

MacArthur, S., and J. F. Brookfield. 2004. Expected rates and modes of evolution of enhancer sequences. Mol. Biol. Evol. **21**:1064–1073.

Machado, C. A., R. M. Kliman, J. A. Markert, and J. Hey. 2002. Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. Mol. Biol. Evol. **19**:472–488.

Maicas, E., M. Shago, and J. D. Friesen. 1990. Translation of the *Saccharomyces cerevisiae tcm1* gene in the absence of a 5′-untranslated leader. Nucleic Acids Res. **18**:5823–5828.

Maquat, L. E. 2004. Nonsense-mediated mRNA decay: a comparative analysis of different species. Curr. Genomics **5**:175–190.

Maruyama, T., and C. W. Birky Jr. 1991. Effects of periodic selection on gene diversity in organelle genomes and other systems without recombination. Genetics **127**:449–451.

McVean, G. A., S. R. Myers, S. Hunt, P. Deloukas, D. R. Bentley, and P. Donnelly. 2004. The fine-scale structure of recombination rate variation in the human genome. Science **304**: 581–584.

Medghalchi, S. M., P. A. Frischmeyer, J. T. Mendell, A. G. Kelly, A. M. Lawler, and H. C. Dietz. 2001. *Rent1*, a trans-effector of nonsense-mediated mRNA decay, is essential for mammalian embryonic viability. Hum. Mol. Genet. **10**:99–105.

Meijer, H. A., and A. A. Thomas. 2002. Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. Biochem J. **367**:1–11.

Mendell, J. T., S. M. Medghalchi, R. G. Lake, E. N. Noensie, and H. C. Dietz. 2000. Novel Upf2p orthologues suggest a functional link between translation initiation and nonsense surveillance complexes. Mol. Cell. Biol. **20**:8944–8957.

Michel, F., and J.-L. Ferat. 1995. Structure and activities of group II introns. Ann. Rev. Biochem. **64**:435–461.

Mira, A., H. Ochman, and N. A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. Trends Genet. **17**:589–596.

Moll, I., S. Grill, C. O. Gualerzi, and U. Bläsi. 2002. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. Mol. Microbiol. **43**:239–246.

Morris, D. R., and A. P. Geballe. 2000. Upstream open reading frames as regulators of mRNA translation. Mol. Cell. Biol. **20**:8635–8642.

Mount, S. M., C. Burks, G. Hertz, G. D. Stormo, O. White, and C. Fields. 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Nucleic Acids Res. **20**:4255–4262.

Nachman, M. W., and S. L. Crowell. 2000. Estimate of the mutation rate per nucleotide in humans. Genetics **156**:297–304.

Niehrs, C., and N. Pollet. 1999. Synexpression groups in eukaryotes. Nature **402**:483–487.

Nissim-Rafinia, M., and B. Kerem. 2002. Splicing regulation as a potential genetic modifier. Trends Genet. **18**:123–127.

Ochman, H. 2003. Neutral mutations and neutral substitutions in bacterial genomes. Mol. Biol. Evol. **20**:2091–2096.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**:299–304.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature **246**:96–98.

———. 1974. Mutational pressure as the main cause of molecular evolution and polymorphism. Nature **252**:351–354.

———. 1997. Selected papers on theoretical population genetics and molecular evolution. Department of Population Genetics, National Institute of Genetics, Mishima, Japan.

Ohta, T., and M. Kimura. 1971. Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. Genetics **68**:571–580.

Otto, S. P., and M. C. Whitlock. 1997. The probability of fixation in populations of changing size. Genetics **146**:723–733.

Philips, A. V., and T. A. Cooper. 2000. RNA processing and human disease. Cell. Mol. Life Sci. **57**:235–249.

Pickering, B. M., and A. E. Willis. 2005. The implications of structured 5′ untranslated regions on translation and disease. Semin. Cell Dev. Biol. **16**:39–47.

Posada, D., K. A. Crandall, M. Nguyen, J. C. Demma, and R. P. Viscidi. 2000. Population genetics of the porB gene of *Neisseria gonorrhoeae*: different dynamics in different homology groups. Mol. Biol. Evol. **17**:423–436.

Prince, V. E., and F. B. Pickett. 2002. Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. **3**:827–837.

Ptak, S. E., K. Voelpel, and M. Przeworski. 2004. Insights into recombination from patterns of linkage disequilibrium in humans. Genetics **167**:387–397.

Qiu, W. G., N. Schisler, and A. Stoltzfus. 2004. The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol. Biol. Evol. **21**:1252–1263.

Raff, R. A. 1996. The shape of life. University Chicago Press, Chicago, Ill.

Reich, D. E., et al. 2001. Linkage disequilibrium in the human genome. Nature **411**:199–204.

Rest, J. S., and D. P. Mindell. 2003. Retroids in Archaea: phylogeny and lateral origins. Mol. Biol. Evol. **20**:1134–1142.

Rockman, M. V., and G. A. Wray. 2002. Abundant raw material for *cis*-regulatory evolution in humans. Mol. Biol. Evol. **19**:1991–2004.

Rogozin, I. B., A. V. Kochetov, F. A. Kondrashov, E. V. Koonin, and L. Milanesi. 2001. Presence of ATG triplets in 5′ untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. Bioinformatics **17**:890–900.

Rogozin, I. B., Y. I. Wolf, A. V. Sorokin, B. G. Mirkin, and E. V. Koonin. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr. Biol. **13**:1512–1517.

Rousset, F. 2003. Effective size in simple metapopulation models. Heredity **91**:107–111.

Roy, S. W., and W. Gilbert. 2005*a*. Complex early genes. Proc. Natl. Acad. Sci. USA **102**:1986–1991.

———. 2005*b*. Rates of intron loss and gain: implications for early eukaryotic evolution. Proc. Natl. Acad. Sci. USA **102**:5773–5778.

Ruiz-Echevarria, M. J., C. I. González, and S. W. Peltz. 1998. Identifying the right stop: determining how the surveillance complex recognizes and degrades an aberrant mRNA. EMBO J. **17**:575–589.

Sarkar, S. F., and D. S. Guttman. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. Appl. Environ. Microbiol. **70**:1999–2012.

Schaal, T. D., and T. Maniatis. 1999. Selection and characterization of pre-mRNA splicing enhancers: identification of novel SR protein-specific enhancer sequences. Mol. Cell. Biol. **19**:1705–1719.

Schmid, P. E., M. Tokeshi, and J. M. Schmid-Araya. 2000. Relation between population density and body size in stream communities. Science **289**:1557–1560.

Sharp, P. A. 1985. On the origin of RNA splicing and introns. Cell **42**:397–400.

Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden, and R. E. Sockett. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. **33**:1141–1153.

Shaw, P. J., N. S. Wratten, A. P. McGregor, and G. A. Dover. 2002. Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. Evol. Dev. **4**:265–277.

Shukla, G. C., and R. A. Padgett. 2002. A catalytically active group II intron domain 5 can function in the U12-dependent spliceosome. Mol. Cell **9**:1145–1150.

Singh, U., J. B. Rogers, B. J. Mann, and W. A. Petri Jr. 1997. Transcription initiation is controlled by three core promoter elements in the hgl5 gene of the protozoan parasite *Entamoeba histolytica*. Proc. Natl. Acad. Sci. USA **94**: 8812–8817.

Slupska, M. M., A. G. King, S. Fitz-Gibbon, J. Besemer, M. Borodovsky, and J. H. Miller. 2001. Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. J. Mol. Biol. **309**:347–360.

Sontheimer, E. J., P. M. Gordon, and J. A. Piccirilli. 1999. Metal ion catalysis during group II intron self-splicing: parallels with the spliceosome. Genes Dev. **13**:1729–1741.

Stoltzfus, A. 1999. On the possibility of constructive neutral evolution. J. Mol. Evol. **49**:169–181.

Stone, J. R., and G. A. Wray. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. Mol. Biol. Evol. **18**:1764–1770.

Taanman, J. W. 1999. The mitochondrial genome: structure, transcription, translation and replication. Biochim. Biophys. Acta **1410**:103–123.

Tenaillon, M. I., J. U'Ren, O. Tenaillon, and B. S. Gaut. 2004. Selection versus demography: a multilocus investigation of the domestication process in maize. Mol. Biol. Evol. **21**:1214–1225.

Valadkhan, B., and J. L. Manley. 2002. Intrinsic metal binding by a spliceosomal RNA. Nat. Struct. Biol. **9**:498–499.

van den Heuvel, J. J., R. J. Bergkamp, R. J. Planta, and H. A. Raue. 1989. Effect of deletions in the 5′-noncoding region on the translational efficiency of phosphoglycerate kinase mRNA in yeast. Gene **79**:83–95.

Vilela, C., and J. E. McCarthy. 2003. Regulation of fungal gene expression via short open reading frames in the mRNA 5′ untranslated region. Mol. Microbiol. **49**:859–867.

von Dassow, G., and E. Monro. 1999. Modularity in animal development and evolution: elements of a conceptual framework for evo-devo. J. Exp. Zool. **285**:307–325.

Wagner, A. 1999. Redundant gene functions and natural selection. J. Evol. Biol. **12**:1–16.

———. 2001. Birth and death of duplicated genes in completely sequenced eukaryotes. Trends Genet. **17**:237–239.

———. 2005. Energy constraints on the evolution of gene expression. Mol. Biol. Evol. **22**:1365–1374.

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.

Weiner, J. III., R. Herrmann, and G. F. Browning. 2000. Transcription in *Mycoplasma pneumoniae*. Nucleic Acids Res. **28**:4488–4496.

Whitlock, M. C. 2003. Fixation probability and time in subdivided populations. Genetics **164**:767–779.

Whitlock, M. C., and N. H. Barton. 1997. The effective size of a subdivided population. Genetics **146**:427–441.

Whitman, W. B., D. C. Coleman, and W. J. Wiebe. 1998. Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. USA **95**:6578–6583.

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**: 1377–1419.

Wright, S. I., B. Lauga, and D. Charlesworth. 2003. Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Mol. Ecol. **12**:1247–1263.

Yee, J., M. R. Mowatt, P. P. Dennis, and T. E. Nash. 2000. Transcriptional analysis of the glutamate dehydrogenase gene in the primitive eukaryote, *Giardia lamblia*. Identification of a primordial gene promoter. J. Biol. Chem. **275**:11432–11439.

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.

Zuckerkandl, E. 2001. Intrinsically driven changes in gene interaction complexity. I. Growth of regulatory complexes and increase in number of genes. J. Mol. Evol. **53**:539–554.

William Martin, Associate Editor

Accepted October 19, 2005