Intron evolution as a population-genetic process

Michael Lynch*

Department of Biology, Indiana University, Bloomington, IN 47405

Edited by Barbara A. Schaal, Washington University, St. Louis, MO, and approved February 7, 2002 (received for review November 7, 2001)

Debate over the mechanisms responsible for the phylogenetic and genomic distribution of introns has proceeded largely without consideration of the population-genetic forces influencing the establishment and retention of novel genetic elements. However, a simple model incorporating random genetic drift and weak mutation pressure against intron-containing alleles yields predictions consistent with a diversity of observations: (*i*) the rarity of introns in unicellular organisms with large population sizes, and their expansion after the origin of multicellular organisms with reduced population sizes; (*ii*) the relationship between intron abundance and the stringency of splice-site requirements; (*iii*) the tendency for introns to be more numerous and longer in regions of low recombination; and (*iv*) the bias toward phase-0 introns. This study provides a second example of a mechanism whereby genomic complexity originates passively as a "pathological" response to small population size, and raises difficulties for the idea that ancient introns played a major role in the origin of genes by exon shuffling.

exon shuffling | genome complexity | genome evolution

The widespread occurrence of introns in eukaryotes has provoked substantial debate over the timing and mechanisms of their origin, degree of positional stability, and adaptive significance. The most extreme form of this debate is manifested in the introns-early vs. introns-late controversy. The introns-early school argues that a large pool of introns in an ancestral genome facilitated the creation of early genes by exon shuffling and that the near absence of introns in today's prokaryotes is a secondary consequence of selection for streamlined genomes (1–3). The introns-late school postulates that the vast majority of introns arose within multicellular eukaryotes and were inserted more or less randomly into preexisting genes, although a subsequent role in the adaptive evolution of proteins is not ruled out (4–6). The extreme views of these two camps have softened somewhat, but a great deal of controversy over the evolutionary biology of introns remains.

The idea that massive parallel loss of introns occurred in all ancestral lineages of modern-day prokaryotes raises obvious logical difficulties, and fails to address why selection for reduced genome size was not a priority earlier in evolution. On the other hand, the early origin of introns can no longer be denied, because they are now known to be present (although rare) in prokaryotes (7). Prokaryotic introns are of a rather different nature than those found in the nuclear genes of today's eukaryotes, but plausible arguments have been made that the two groups are evolutionarily related (4, 8). Moreover, the recent discovery that two types of spliceosomal introns coexist in some eukaryotes suggests that nuclear introns may have arisen near the base of the eukaryotic lineage (9). Thus, much of the debate about introns has refocused on issues regarding intron proliferation and retention. One view of the current phylogenetic distribution of introns is that microbial genomes have never harbored more than a few introns despite their potential for expansion. But even if this view is correct, is selection for a permanently streamlined genome the explanation for the rarity of microbial introns, or are more fundamental issues involved? And why are introns much more prominent genomic features of multicellular than of unicellular eukaryotes?

The establishment and retention of introns is a matter of survival in the face of opposing evolutionary forces. Like all novel genetic elements, introns must initiate as single mutational changes in single members of a population. To be successful in the short-term, a new intron must navigate a trajectory toward fixation under the joint influence of mutation, random genetic drift, and oftentimes opposing selection. To be successful in the long-term (postfixation), sufficiently positive selective forces must exist for the retention of the intron in the face of subsequent mutational challenges. The goal of this study is to illustrate how simple population-genetic principles may help guide our understanding of the phylogenetic and genomic distribution of introns. The primary focus will be on models that assume random genetic drift and mutation to be the only relevant evolutionary forces. Such models provide a useful benchmark against which to evaluate the necessity of invoking adaptive explanations for the abundance and distribution of introns, and, as will be shown below, they provide a potentially unifying explanation for several lingering puzzles about introns.

Intron Proliferation and Maintenance

Conditions for the Establishment of a New Intron. Although the specific mechanisms by which introns arise are not entirely understood, plausible mechanisms can lead to the birth of an intron fully endowed with the necessary features for excision during mRNA processing (5, 10). We will start with the assumption that such properties are a precondition for initial establishment, because inserts that cannot be accurately removed from a transcript are expected to have highly deleterious effects. Even if fully proficient with respect to excision, an intron may impose a weak selective disadvantage on its host allele as a consequence of reduced transcriptional efficiency. However, we will ignore this issue for the time being, focusing instead on a more fundamental disadvantage of intron-containing alleles—an elevated mutation rate to nonfunctional alleles.

Nuclear introns have specific short terminal sequences required for proper excision, almost always bearing a GT at the 5′ end and an AG at the 3′ end. In animals, most introns have a polypyrimidine tract of >10 nucleotides before the canonical AG, and an interior branch-point A is essential for the splicing reaction. *Caenorhabditis elegans* introns have a highly conserved sequence immediately adjacent to the 3′ splice site (11). In *Saccharomyces cerevisiae* and many other ascomycetes (12), several essentially invariant nucleotides surround the branch-point, and this region also exhibits significant conservation in some animals. Nuclear introns in plants exhibit a strong bias toward T proximal to the AG acceptor, and more generally exhibit an A/T bias above that in adjacent exons (13). The key point here is that mutations that alter the sequences of any nucleotide sites critical to intron processing can completely eliminate the capacity to splice (14). Such intron-debilitating mutations will result in either the expansion or contraction of the intron, with the new protein either losing or gaining amino acids, often with an accompanied change in reading frame and loss of gene function.

This elevated rate of mutation to nulls exclusive to intron-

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: NMD, nonsense-mediated decay.

*E-mail: mlynch@bio.indiana.edu.

www.pnas.org/cgi/doi/10.1073/pnas.092595699

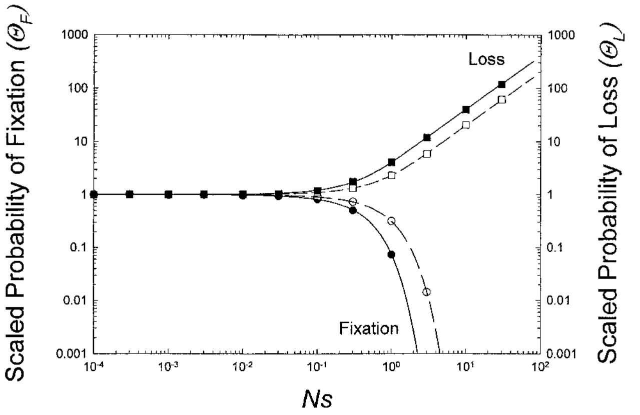

Fig. 1. Scaled probabilities of establishment (left axis, and lower curves) and loss (right axis, and upper curves) of introns in diploid (filled points and solid lines) and haploid populations (open points and dashed lines). The curved lines are the theoretical results outlined in the text, whereas the plotted values are results from computer simulations. For this particular set of simulations, $n = 10^{-5}$ and $s = 10^{-6}$.

containing alleles (s) is equivalent to a form of weak selection, because each such mutation eliminates an intron-containing allele from the total pool of functional alleles. This argument suggests that a modification of the diffusion approximation for the probability of fixation *(u_F)* of a deleterious allele with additive effects (15) should provide a reasonable description of the probability of intron establishment. For a diploid sexual population,

$$u_F \simeq \frac{2s}{e^{4Ns} - 1}, \qquad \textbf{[1a]}$$

where N is the effective population size, whereas, for a haploid population,

$$u_F \simeq \frac{2s}{e^{2Ns} - 1}. \qquad \textbf{[1b]}$$

The scaled probability of fixation, $\theta_F = 2Nu_F$ for diploids and $\theta_F = Nu_F$ for haploids, is a simple function of Ns, with $\theta_F = 1$, implying a fixation probability equal to the neutral expectation, $1/(2N)$ for diploids and $1/N$ for haploids.

The validity of these approximations was checked by extensive computer simulations for a range of sizes of randomly mating populations, letting the mutation rate to nulls be n for intron-free alleles and (n + s) for intron-containing alleles. All genotypes were assumed to have equal fitness, except those containing only null alleles, which were assumed to be lethal. The agreement between the theory and simulated results is excellent for the full range of N (Fig. 1). If $Ns < 0.1$, the excess mutation pressure to nulls for the intron-containing allele is sufficiently weak relative to the force of random genetic drift that the locus behaves in an effectively neutral manner ($\theta_F \cong 1$). On the other hand, if $Ns > 10$, there is essentially no chance of an intron becoming fixed in either a haploid or diploid population ($\theta_F \cong 0$).

How large is s likely to be? For a range of multicellular plants and animals, the mutation rate per nucleotide site is thought to fall between 3.5 and 16×10^{-9} substitutional changes per year (16). Assuming that proper intron processing requires specific nucleotides at $\approx$10 sites (roughly the case for nuclear U2-type introns) and also recognizing that insertion/deletions at other sites may alter the spatial requirements for splicing, s is expected to be on the order of 10^{-7} to 10^{-6} per generation for an organism with an annual life cycle, and the scaling with generation time may be approximately linear (17). Combined with the pattern illustrated in Fig. 1, these crude estimates suggest that organisms with $N > 10^{10}$ or so will be essentially immune to colonization by introns, whereas those with $N < 10^7$ or so will fix newly arisen introns at the neutral rate or slightly less. Global population sizes of $\approx 10^{10}$ are not particularly large for unicellular organisms, but are unrealistically high for most large multicellular organisms, so these approximate boundaries appear to be quite relevant to the phylogenetic distribution of introns.

Intron Stability. Once fixed in a population, an intron can take on important functional roles through the incorporation of regulatory elements and/or sites involved in alternative splicing, in which case positive selection could promote essentially indefinite intron retention. However, for the potentially large fraction of introns that never experience such preservational changes, initial establishment does not guarantee permanent residence. Observations among closely related species consistently point to the occurrence of intron turnover (18–23). A suspected mechanism by which an intron can be cleanly lost involves reverse transcription of a mature mRNA followed by homologous recombination with an intron-containing allele. Although there is no quantitative information on the rate of occurrence of such events, some qualitative insight into the expected longevity of an intron can be acquired by reversing the logic outlined above. Until an intron experiences a preservational event, it will be vulnerable to displacement by derived intron-free alleles, which have a weak selective advantage s because of the absence of critical intron-specific sites. The expected probability of fixation of a newly arisen intron-free allele (u_L), obtained by use of Eqs. **1a** and **1b** after substituting $-s$ for s, is again in excellent agreement with results from computer simulations (Fig. 1). For sufficiently small populations ($Ns < 0.1$), a derived intron-free allele will fix with a probability equal to the neutral expectation, and when $Ns > 10$, the probability of fixation of the intron-free allele is $\cong 2s$, independent of population size.

The long-term rate of allelic turnover in a population will depend on the rates of origin and loss of intron-containing alleles as well as their fixation probabilities. Focusing on a particular site of potential occupancy, the simplest situation arises when the population size is sufficiently small to satisfy the conditions of effective neutrality ($Ns < 0.1$). The rates at which intron-free alleles gain introns by insertion (b) and at which intron-containing alleles lose introns by deletion (d) will then be the sole determinants of the allele-frequency distribution. Under these conditions, a diploid population fixed for intron-free alleles is expected to retain that status for an average of $[(1/b) + 4N]$ generations, whereas a population fixed for an intron is expected to retain that state for $[(1/d) + 4N]$ generations, where $4N$ is the mean time to fixation for a neutral mutation (15). Provided $(1/b)$ and $(1/d)$ are $<4N$, periods of presence/absence polymorphism are expected to be very rare, and the relative probabilities of the population being in the intron-free or the intron-containing state are $d/(b + d)$ and $b/(b + d)$, respectively. No direct estimates are available for b and d, but we expect that both quantities must be very small, perhaps $\ll 10^{-6}$, and the rate of intron deletion probably greatly exceeds the rate of origin of an intron at a specific site ($b \ll d$). Thus, provided $Ns < 0.1$, the expected long-term average site occupancy (number of introns/number of nucleotides in the coding region) is $\cong b/d$. Multiplying the probability of each population state by the number of mutations arising per generation and the probability of fixation, we find that the expected rate of transition of the population between the two pure states is simply equal to the harmonic mean of the rates of birth and death,

$$\delta = \left[\frac{b}{b+d} \cdot 2Nd \cdot \frac{1}{2N}\right] + \left[\frac{d}{b+d} \cdot 2Nb \cdot \frac{1}{2N}\right] = \frac{2}{(1/b) + (1/d)}. \qquad \textbf{[2]}$$

These results can be generalized to larger population sizes in the following way. Considering just the functional alleles at the

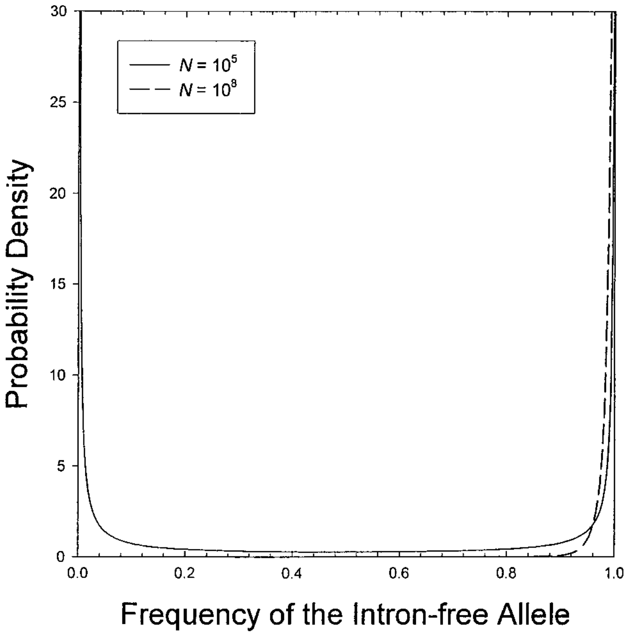

Fig. 2. Steady-state distributions for the frequency of the intron-free state obtained by use of Eq. **3**. Results are given for two population sizes, with $b = d = 10^{-10}$ and $s = 10^{-7}$. The distribution can be viewed as the probability that the population has a particular gene frequency at a particular time or as the fraction of time that the population spends in a particular state.

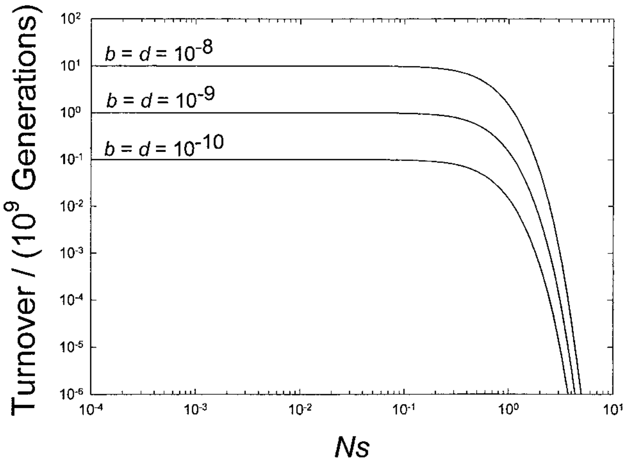

Fig. 3. The equilibrium rate of turnover for introns at a specific site, with b and d representing the per generation rates of origin (by insertion) and elimination (by deletion) of introns.

locus, the long-term steady-state distribution for the frequency of the intron-free allele is

$$\phi(p) = Ce^{4Nsp}p^{4Nd-1}(1-p)^{4Nb-1}, \qquad [3]$$

where C is a normalization constant (from equation 13.51 in ref. 24), showing that a population is transiently monomorphic for the intron-free allele with probability

$$Pr(p = 1) = \int_{1-(1/2N)}^{1} \phi(p)dp \simeq Ce^{4Ns}(4Nb)^{-1}(2N)^{-4Nb}, \qquad [4a]$$

and for the intron-containing allele with probability

$$Pr(p = 0) = \int_{0}^{1/2N} \phi(p)dp \simeq C(4Nd)^{-1}(2N)^{-4Nd}. \qquad [4b]$$

As $N \to 0$, $\Pr(P = 1) \to C/(4Nb)$ and $\Pr(P = 0) \to C/(4Nd)$, verifying the contention in the previous paragraph that sufficiently small populations spend almost all of their time fixed for intron-free or intron-containing states (Fig. 2), in relative proportions d:b. On the other hand, as $N \to \infty$, $\Pr(P = 1) \to 0$ and $\Pr(P = 0) \to 0$, and a stable state of approximate selection-mutation balance is reached, with the frequency of the intron-containing allele being kept at a low level defined by the relative values of s, b, and d (Fig. 2). For intermediate population sizes, the rate of flux between states dominated by the intron-free or the intron-containing allele can be obtained in the same manner as Eq. **2**, after accounting for the different probabilities of fixation,

$$\delta = \frac{4N(bu_{\mathrm{F}})(du_{\mathrm{L}})}{(bu_{\mathrm{F}}) + (du_{\mathrm{L}})}. \qquad [5]$$

Provided $Ns < 1$, the rate of turnover is essentially independent of N, but beyond $Ns = 4$, δ rapidly approaches zero (Fig. 3). Thus, for nearly the full range of conditions under which introns are likely to be at detectable levels, the rate of intron turnover is adequately described by Eq. **2**. Moreover, the turnover rate is quite small, on the order of $2b$ when $d \gg b$, $2d$ when $b \gg d$, and b when $b \cong d$. These results apply only to introns that have not taken on important functional roles preserved by positive selection.

Intron Sliding and Phase Bias. Although considerable empirical attention has been given to the physical distribution of introns, any evolutionary interpretation of such observations depends on the degree to which introns remain stably in their ancestral sites. Intron locations sometimes differ among species by only a few nucleotides, but because the rate of sequence divergence tends to be quite high for introns, the rapid loss of homology makes it difficult to infer the source of such spatial differences between distantly related species. The introns-early school tends to interpret small positional shifts as consequences of intron sliding (3, 25). Others have argued that most shifts in intron positions are due to loss/gain of introns in independent lineages (18, 26), but convincing cases of intron sliding have been recorded (27–29). Quantitative resolution of the relative contribution of intron sliding to patterns of intron position will ultimately be a matter of empirical observation, but some theoretical insight into the relevant issues can be obtained by the following reasoning.

As pointed out by Stoltzfus *et al.* (26), intron sliding will generate a frameshift in the downstream exon unless reciprocal changes occur in both flanking exons. Because the simultaneous occurrence of two rare and specific mutational events is miniscule, a more likely path to intron sliding is a series of two or more expansion/contraction events involving intron–exon boundaries. Consistent with this view is the observation that insertions and deletions in coding sequence are often associated with intron–exon boundaries (25). Movement of an intron–exon boundary is likely to occur whenever a nucleotide in one of the splice sites is altered, and, depending on the positions of alternative splice sites, the mutant allele will experience an expansion, contraction, or elimination of the previous intron. The resultant loss or increase in flanking exon sequence will be accompanied by a frameshift in all downstream codons unless the number of nucleotides involved in the shift to the new intron–exon boundary is a multiple of three (Fig. 4). For this simple reason, we expect intron expansion/contraction events involving $3n$ nucleotide shifts, where n is an integer, to be more easily fixed in a population than $3n + 1$ or $3n + 2$ shifts. Moreover, the selective consequences of $3n$ shifts will depend on the phase of the original intron. If the original intron is in phase 1 or 2 (i.e., splitting an ancestral codon), expansion/contraction by $3n$ nucle-

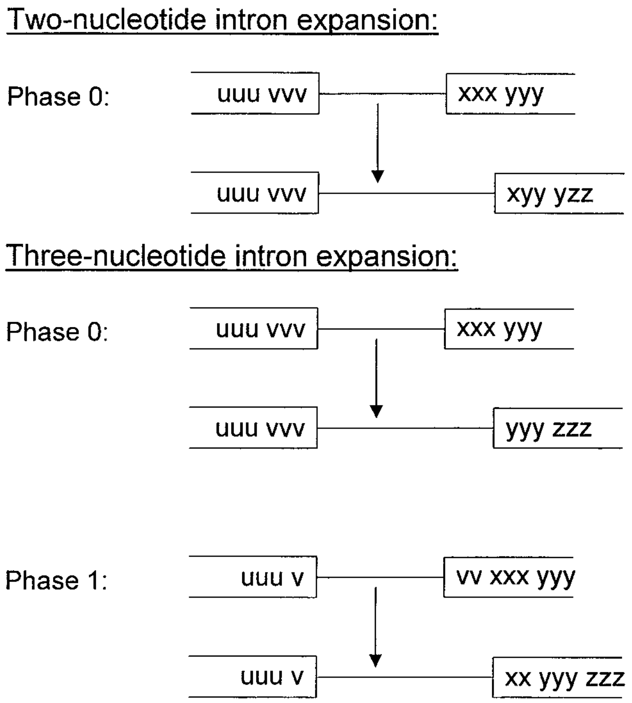

Fig. 4. The consequences of an intron-expansion event for the coding sequence of two flanking exons. (*Upper*) Regardless of the intron phase, a 2-nucleotide (or 1-nucleotide) expansion will result in a frameshift in all downstream codons. (*Lower*) A 3-nucleotide expansion results in the loss of a codon if the intron is in phase 0, but a loss of a codon plus one altered codon sequence if the intron is in phase 1 or 2.

otides, while maintaining the downstream frame, will induce a codon change by splicing portions of the codons flanking the expansion (Fig. 4). Because 3/64 of such codon-splicing events will produce a stop codon and other nonsynonymous changes are likely to be mildly deleterious, intron-sliding events are expected to favor phase-0 introns over phase-1 and -2 introns.

Although a quantitative statement cannot be made as to the magnitude of the selective advantage of phase-0 introns, two qualitative predictions can be made—phase-0 introns are expected to be more abundant than either phase-1 or -2 introns, and phases 1 and 2 should be approximately equal in frequency. These predictions are fairly consistent with observational data. Averaging over several thousand introns in several species, the relative frequencies of phases 0, 1, and 2 are approximately 50%, 27%, and 23% (30, 31). The actual frequencies vary by a few percent from study to study, with phase-1 introns consistently outnumbering those in phase 2, although, in *Arabidopsis thaliana, Schizosaccharomyces pombe*, and *Drosophila melanogaster,* this excess is only 1 to 2% (31).

These predictions regarding bias in the distribution of intron phases are quite germane to the introns early-late debate. The introns-early proponents, who invoke intron-sliding as the explanation for observed modifications of intron position, also view the bias toward phase-0 introns as compelling support for an early evolutionary period of exon shuffling (30–32). However, the validity of this conclusion rests on the incorrect assumption that the introns-late model predicts equal frequencies of all three phases. Even if one accepts that new insertions of functional introns should be unbiased with respect to phase, those that initiate in phase 0 are expected to enjoy the greatest longevity provided events involving intron sliding occur. The view that the tendency for adjacent introns to be in the same phase supports the introns-early hypothesis (30, 33) has similar problems, because it ignores the weak selection against alleles whose intron phases are uncorrelated. If, for example, an exon is spliced out by accident during pre-mRNA processing, the downstream coding sequence will experience a frameshift unless the introns at the ends of the missing exon are in the same phase. Thus, unless postinsertional events are negligible, observations on intron-phase distributions appear to be of little relevance to issues concerning the age of introns.

Intron Sliding Facilitated by Nonsense-Mediated Decay. Despite their apparent disadvantages, it is clear that some intron-sliding events involve $3n + 1$ or $3n + 2$ nucleotide shifts (28). This finding suggests that successful intron-sliding events are sometimes preceded by an intermediate stage involving a frame-shifted allele. Such transitions may be facilitated by an mRNA surveillance mechanism known to be specifically associated with intron-containing loci. In animals, nonsense-mediated decay (NMD) depends on proteins that demarcate exon–exon junctions to identify premature stop codons in processed mRNAs, with selective degradation of the mRNA occurring when a stop codon appears more than about 50 nucleotides upstream of the final exon–exon junction (34). For a haplosufficient locus, NMD will render a frameshift-containing allele completely recessive (provided it contains a premature stop codon), thereby enhancing its retention time in a population and increasing the likelihood of a compensatory mutation to a restored reading frame.

To evaluate the probability of a successful intron slide via this two-step process, computer simulations were initiated with a population in which all but one of the alleles contained a functional intron. The mutant frame-shifted allele was assumed to be kept silent and completely recessive by NMD. Both allelic types mutated to permanent null alleles with probability u, but copies of the allele containing the premature stop codon were also restored to wild type at the compensatory mutation rate r. Under this scenario, the mutant allele is ultimately either lost or drifts to fixation after the acquisition of a compensatory mutation in a descendent member of its lineage.

An analytical approximation for the probability of a successful two-step intron-sliding event (involving an intermediate null state) can be obtained in the following way. Silent mutant alleles in the first step of the sliding process are selectively removed from the population through encounters with other null alleles, the expected frequency of which is $p_0 = \Gamma(2Nu + 0.5)/\sqrt{(2N)}\Gamma(2Nu)$ under drift-mutation-selection equilibrium (15), where Γ denotes a gamma function. The average number of generations that a specific null allele is retained in a population is approximately p_0/u generations, and, if we assume that a mutant allele in the first step of a sliding process is typically present in the population as only a single copy before its elimination, then the mean number of descendant copies restored to the proper reading frame is $\cong rp_0/u$. Each of these second-step mutations is neutral with respect to the original (functional) allele and drift to fixation with probability $1/(2N)$. Thus, for sufficiently small populations, the probability that a frame-shifted allele kept silent by NMD is restored to frame by an intron slide that becomes fixed is $rp_0/(2Nu)$.

In large populations, the additional possibility exists that all of the descendants of the two-step mutant are silenced by null mutations in the coding region while the lineage drifts toward fixation. A crude correction for this secondary effect can be obtained by noting that, of the average $4N$ generations required for the coalescence of a neutral locus, $2N$ are expected at the base of the gene genealogy (i.e., the final coalescence). If a neutral allele destined to fixation is going to be completely silenced before fixation, the silencing mutations will generally need to be acquired in this basal coalescence, because the likelihood of multiple independent mutations in shorter descendant branches in the genealogy becomes diminishingly small. The probability that both basal branches do not acquire null mutations is $K = 1 - (1 - e^{-2Nu})^2$. Thus, the corrected

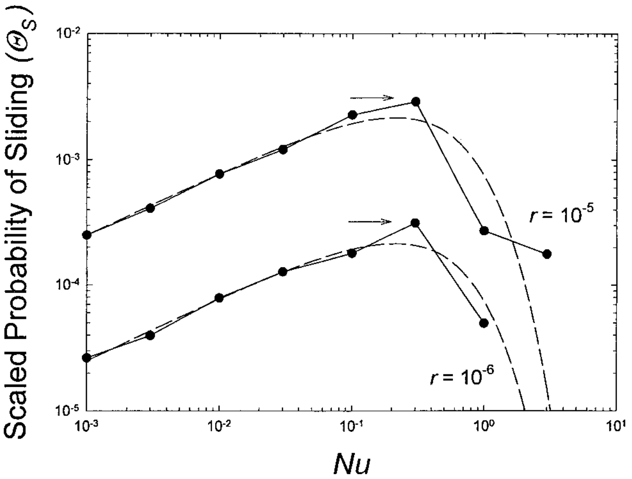

Fig. 5. The scaled probability of intron-sliding for a two-step process, as described in the text. The mutation rate to null alleles is u (in all plotted cases equal to 10^{-5}), whereas the rate of compensatory mutation is r. The filled points are the results from computer simulations, whereas the dashed lines are the analytical approximations, and the arrows denote the uncorrected large-population size result, $r/\sqrt{u}$.

probability of successful sliding becomes $Krp_0/(2Nu)$, and $2N$ times this quantity (the scaled probability of a successful two-step intron slide) is $\theta_S = Krp_0/u$.

This simple expression yields predictions that are quite close to those obtained by simulations, and shows that, although intron-sliding events involving an intermediate null mutation are nonnegligible, they are quite rare (Fig. 5). For biologically reasonable values for u and r, the scaled probability of a successful two-step intron slide involving a null intermediate is at least two orders of magnitude less than the neutral expectation of $\theta_S = 1$. When $Nu < 0.1$, $p_0 \cong u\sqrt{(2\pi N)}$, $k \cong 1$, and $\theta_S \cong r\sqrt{(2\pi N)}$, showing that the probability of a successful intron slide initially scales with the square root of the effective population size. Under these conditions, the number of opportunities for acquiring the second-step compensatory mutation required for frame restoration is limiting. On the other hand, for larger Nu, the time required for fixation of the restored allele becomes so large that secondary mutations to nulls start to become a barrier. Provided $0.1 < Nu < 1.0$, this phenomenon is not a major problem, p_0 approaches $\sqrt{u}$, and θ_S approaches a maximum value $\cong r/\sqrt{u}$, but for $Nu > 1.0$, θ_S rapidly declines to zero. In the absence of NMD, the likelihood of intron sliding is necessarily smaller than the preceding expectations, unless the intermediate null allele is intrinsically completely recessive.

Discussion

The theory presented above treats the birth of an intron as an autonomous event, independent of the presence of other introns within the gene of interest or elsewhere in the genome. Little is known about the source of new nuclear introns, and it is conceivable that successful origins simply represent a small subset of the random insertions that arise within a genome by the capture of nascent DNA fragments by double-strand breaks (35). To serve as a successful nuclear intron, such insertions would need to be fortuitously endowed with all of the key sequences necessary for proper splicing. Although such events may be very infrequent, it is clear that *de novo* introns can arise. In plants, for example, insertions with an adequate AT-richness serve as the nuclei of new introns that extend to adjacent GT/AG boundaries preexisting within coding regions (36). Reverse transcripts of spliced introns presumably have some opportunity to reinsert themselves into the nuclear genome, and these transcripts would be naturally endowed with recognition sequences. So far, however, only a single study has provided evidence for the origin of a nuclear intron by intragenomic duplication of a preexisting intron (21). Transposable-element insertions may sometimes give rise to functional introns (37), but there is as yet no compelling evidence that this is a common mode of origin of successful introns either. Regardless of the mechanism by which functional introns arise, the population-genetic treatment of intron evolution as a birth-death process yields qualitative predictions that are consistent with an array of observations on the phylogenetic and genomic distribution of introns.

Intron Abundance and Effective Population Size. Advocates of the introns-early hypothesis generally invoke selection for a streamlined genome to explain the rarity of introns in modern-day prokaryotes. However, the simple population-genetic arguments presented above challenge this view. Broadly speaking, unicellular organisms have larger population sizes than multicellular species, and the majority of unicellular species may have global effective population sizes $\gg 10^{10}$, which is large enough for the very weak mutation pressure against newborn introns to become evolutionarily significant. If intron processing has an additional cost, either in terms of energetic requirements or in terms of errors induced by the splicing process, the upper population size limits permissive to intron establishment will be even lower. Moreover, for early life forms without refined DNA-repair pathways, s was probably substantially larger than it is today, reducing the critical effective size still further, and rendering the idea that the creative assembly of early genes was facilitated by introns even less plausible.

Thus, a parsimonious explanation for the rarity of introns in prokaryotes is that weak mutation pressure prevents the establishment of functional introns in taxa for which the average population size is of sufficiently large size. If this hypothesis is correct, then we would also expect introns to be rare in unicellular eukaryotes, again not because of a failure to invent introns but because of the mutational constraints on their proliferation. Generally speaking, the nuclear genomes of unicellular protists and fungi do contain relatively low numbers of introns compared with those of multicellular plants and animals (38–40). The average sizes of introns in unicellular eukaryotes are also much smaller than those in multicellular species (41), consistent with expectations if insertions into introns are more deleterious than deletions (see next section). The presence of at least a few nuclear introns in most of the deeply diverging basal groups of eukaryotes (42) implies that the spliceosomal apparatus may have been present in the common ancestor to all eukaryotes, and, if this proves to be the case, it would further bolster the idea that the necessary population-size/mutation-rate requirements for intron proliferation were met only after the origin of multicellularity and the associated reduction in effective population size. The elevated abundance of introns in many lineages of organelle genomes (chloroplasts and mitochondria) of plants and fungi (relative to that in their ancestral prokaryotic states; ref. 5) is also consistent with this idea, in that the effective population size of an organelle is 25–100% of that of its eukaryotic host, depending on the mode of organelle inheritance.

Intron Size and Recombination Frequency. Genome-wide analyses of *D. melanogaster* and vertebrates indicate that intron size and number increase in regions of low recombination (43–45). Carvalho and Clark (43) suggest that this relationship is an evolutionary consequence of the reduced efficiency of selection against insertion mutations in regions of low recombination (the Hill-Robertson effect). In contrast, noting that deletion mutations outnumber insertions, Comeron and Kreitman (44) argue that a selective advantage for large introns in regions of low recombination must offset the directional mutation pressure to small intron size. They surmise that, by acting as modifiers of the recombination rate, large introns can reduce the load caused by deleterious mutations otherwise expected in regions of low recombination. However, it is unclear whether the special population-genetic circumstances necessary to selectively promote recombination-frequency modification through changes in intron size exist in *D. melanogaster* or any other species (46).

The weak mutation-pressure hypothesis provides a formal explanation for the associations of intron abundance and size with recombination frequency that is consistent with Carvalho and Clark (43). First, the reduction in effective population size in regions of low recombination is conducive to the establishment and retention of introns (Fig. 1). Second, if there is a weak selective advantage for reduced intron size (due, for example, to increased rates of transcription and/or increased accuracy of splicing), the efficiency of selection for intron-size reducing mutations (and against intron-size enhancing mutations) is expected to be weaker in regions of low recombination. Regardless of the relative rates of mutations to insertions and deletions, the Hill-Robertson effect will tip the balance toward insertions in regions of low recombination. Thus, there appears to be no reason to invoke a secondary advantage of recombination-frequency modification.

Intron Abundance and Intron-Recognition Properties. Along with population-level properties (e.g., effective size) and chromosome-level properties (e.g., recombination frequency), aspects of intron-recognition sequences are likely to influence the relative incidences of introns in different lineages. More stringent intron-recognition requirements imply a higher vulnerability to intron-debilitating mutations (higher s), a lower likelihood of random intron origin (lower b), and hence a lower equilibrium incidence of introns.

A potential example of genomic differences in intron abundance resulting from differences in intron-recognition sequences is the contrast between budding yeast *(S. cerevisiae)*, which as noted above has a highly stringent seven-nucleotide branch-point sequence requirement, and fission yeast *(S. pombe)*, for which the intronic sequences are much less conserved. The incidence of introns is inflated approximately 9-fold in *S. pombe* (47), and similar inflation is seen in *Neurospora crassa,* which also has relaxed splicing requirements (48). A second potential example concerns the observation that introns in vascular-plant genes are much rarer in genomic regions with high GC content (49). Because plant introns have a high AT-content requirement for efficient splicing (13), assuming that regional differences in GC content result in part from differential mutation pressure, introns in GC-rich regions are expected to experience a higher rate of degenerative mutation and hence to have a lower incidence.

Similar logic helps explain the disparity in the relative abundances of different types of introns within the same genome. For example, although >99% of all eukaryotic introns appear to be of the classical GT-AG "U2-type," a very rare AT-AC type is known to exist in several groups of eukaryotes (including plants, cnidarians, arthropods, and vertebrates, but not fungi or nematodes; ref. 9). There is no evidence that the two types of introns are intrinsically different with respect to direct effects on individual fitness or with respect to ability to proliferate. However, U12-type introns are distinct in having a highly conserved 5′ splice site (ATATCCTT) and a highly conserved branch sequence (TCCTTAAC), quite unlike the more relaxed sequence requirements for U2-type introns. This greater degree of conservation implies that U12-type introns are incapacitated by mutation much more easily than U2-type introns, although some types of mutations simply convert the U12 type to the U2 type (9). Thus, the extremely low incidence of U12-type introns in all species (and their complete absence from some) is consistent with the expectations of the birth-death model.

Genome Complexity as a Pathological Response to Small Population Size. The results of this study suggest that the simple structure of genes in microbes relative to higher eukaryotes may have little to do with selective constraints (genomic streamlining in microbes) or adaptive requirements (the expansion of cellular complexity in multicellular eukaryotes). Rather, the genomes of species with habitually large population sizes may simply be immunized from the spread of introns by the power of secondary mutation. This is the second instance in which we have shown that genomic complexity can passively arise in response to small population size, without any direct selection for organismal complexity and without any adaptive significance. We have previously demonstrated that, whereas duplicate genes can be preserved by subfunctionalization (partitioning of gene functions) in small populations, such preservation is unlikely once a population reaches a sufficiently large size, again because the fixation of subfunctionalized alleles is opposed by the weak selection induced by secondary mutations in large populations (50). The evolution of organismal complexity may be more of a secondary consequence than a primary cause of genome complexity.

I am very grateful to C. Baer, J. Palmer, D. Rice, and A. Richardson for helpful comments. This work was facilitated by National Institutes of Health Grant GM20887 and National Science Foundation Grant DEB-0003920.

1. Darnell, J. E. (1978) *Science* **202,** 1257–1260.
2. Blake, C. C. F. (1978) *Nature (London)* **273,** 267.
3. Gilbert, W. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **52,** 901–905.
4. Cavalier-Smith, T. (1991) *Trends Genet.* **7,** 145–148.
5. Palmer, J. D. & Logsdon, J. M., Jr. *(1991) Curr. Opin. Genet. Dev.* **1,** 470–477.
6. Patthy, L. (1999) *Protein Evolution* (Blackwell Scientific, Oxford).
7. Belfort, M., Reaban, M. E., Coetzee, T. & Dalgaard, J. Z. (1995) *J. Bacteriol* **177,** 3897–3903.
8. Sharp, P. A. (1991) *Science* **254,** 663.
9. Burge, C. B., Padgett, R. A. & Sharp, P. A. (1998) *Mol. Cell* **2,** 773–785.
10. Rogers, J. H. (1989) *Trends Genet.* **5,** 213–216.
11. Blumenthal, T. & Steward, K. (1997) in *C. elegans II,* eds. Riddle, D. L., Blumenthal, T., Meyer, B. J. & Preiss, J. R. (Cold Spring Harbor Lab. Press, Plainview, NY), pp. 117–145.
12. Bhattacharya, D., Lutzoni, F., Reeb, V., Simon, D., Nason, J. & Fernandez, F. (2000) *Mol. Biol. Evol.* **17,** 1971–1984.
13. Lorkoviæ, Z. J., Wieczorek Kirk, D. H., Lambermon, M. H. L. & Filipowicz, W. (2000) *Trends Plant Sci.* **5,** 160–167.
14. Moore, M. J. (2000) *Nat. Struct. Biol.* **7,** 14–16.
15. Crow, J. F. & Kimura, M. (1970) *An Introduction to Population Genetics Theory* (Harper and Row, New York).
16. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
17. Keightley, P. D. & Eyre-Walker, A. (2000) *Science* **290,** 331–333.
18. Rzhetsky, A., Ayala, F. J., Hsu, L. C., Cheng, C. & Yoshida, A. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 6820–6825.
19. Frugoli, J. A., McPeek, M. A., Thomas, T. L. & McClung, C. R. (1998) *Genetics* **149,** 355–365.
20. Robertson, H. M. (1998) *Genome Res.* **8,** 449–463.
21. Tarrío, R., Rodríguez-Trelles, F. & Ayala, F. J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 1658–1662.
22. Kent, W. J. & A. M. Zahler. (2000) *Genome Res.* **10,** 1115–1125.
23. Nikoh, N. & Fukatsu, T. (2001) *Mol. Biol. Evol.* **18,** 1631–1642.
24. Wright, S. (1969) *Evolution and Genetics of Populations.* The Theory of Gene Frequencies (Univ. of Chicago Press, Chicago), Vol. 2.
25. Craik, S. S., Rutter, W. J. & Fletterick, R. (1983) *Science* **220,** 1125–1129.
26. Stoltzfus, A., Logsdon, J. M., Jr., Palmer, J. D. & Doolittle, W. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 10739–10744.
27. Sato, Y., Niimura, Y., Yura, K. & Go, M. (1999) *Gene* **238,** 93–101.
28. Rogozin, I. B., Lyons-Weiler, J. & Koonin, E. V. (2000) *Trends Genet.* **16,** 430–432.
29. Sakharkar, M. K., Tan, T. W. & de Souza, S. J. (2001) *Bioinformatics* **17,** 671–675.
30. Long, M., Rosenberg, C. & Gilbert, W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 12495–12499.
31. Long, M. & Deutsch, M. (1999) *Mol. Biol. Evol.* **16,** 1528–1534.
32. Gilbert, W., de Souza, S. J. & Long, M. (1997) *Proc. Natl. Acad. Sci. USA* **52,** 7698–7703.
33. de Souza, S. J., Long, M., Klein, R. J., Roy, S., Lin, S. & Gilbert, W. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 5094–5099.
34. Lykke-Andersen, J. (2001) *Curr. Biol.* **11,** R88–R91.
35. Lin, Y. & Waldman, A. S. (2001) *Genetics* **158,** 1665–1674.
36. Luehrsen, K. R. & Walbot, V. (1994) *Genes Dev.* **8,** 1117–1130.
37. Purugganan, M. & Wessler, S. (1992) *Genetica* **86,** 295–303.
38. Russell, C. B., Fraga, D. & Hinrichsen, R. D. (1994) *Nucleic Acids Res.* **22,** 1221–1225.
39. Deutsch, M. & Long, M. (1999) *Nucleic Acids Res.* **27,** 3219–3228.
40. Katinka, M. D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., *et al.* (2001) *Nature (London)* **414,** 450–453.
41. Vinogradov, A. E. (1999) *J. Mol. Evol.* **49,** 376–384.
42. Logsdon, J. M., Jr. (1998) *Curr. Opin. Genet. Dev.* **8,** 637–648.
43. Carvalho, A. B. & Clark, A. G. (1999) *Nature (London)* **401,** 344.
44. Comeron, J. M. & Kreitman, M. (2000) *Genetics* **156,** 1175–1190.
45. Duret, L., Mouchiroud, D. & Gautier, C. (1995) *J. Mol. Evol.* **40,** 308–317.
46. Duret, L. (2001) *Trends Genet.* **17,** 172–175.
47. Käufer, N. F. & Potashkin, J. (2000) *Nucleic Acids Res.* **28,** 3003–3010.
48. Edelmann, S. & Staben, C. (1994) *Exp. Mycol.* **18,** 70–81.
49. Carels, N. & Bernardi, G. (2000) *Genetics* **154,** 1819–1825.
50. Lynch, M., O'Hely, M., Walsh, B. & Force, A. (2001) *Genetics* **159,** 1789–1804.