Streamlining and Simplification of Microbial Genome Architecture

Michael Lynch

Department of Biology, Indiana University, Bloomington, Indiana 47405;
email: mlynch@indiana.edu

Key Words
genome evolution, genomic streamlining, mutation, prokaryotes, random genetic drift, recombination

Abstract

The genomes of unicellular species, particularly prokaryotes, are greatly reduced in size and simplified in terms of gene structure relative to those of multicellular eukaryotes. Arguments proposed to explain this disparity include selection for metabolic efficiency and elevated rates of deletion in microbes, but the evidence in support of these hypotheses is at best equivocal. An alternative explanation based on fundamental population-genetic principles is proposed here. By increasing the mutational target sizes of associated genes, most forms of nonfunctional DNA are opposed by weak selection. Free-living microbial species have elevated effective population sizes, and the consequent reduction in the power of random genetic drift appears to be sufficient to enable natural selection to inhibit the accumulation of excess DNA. This hypothesis provides a potentially unifying explanation for the continuity in genomic scaling from prokaryotes to multicellular eukaryotes, the divergent patterns of mitochondrial evolution in animals and land plants, and various aspects of genomic modification in microbial endosymbionts.
INTRODUCTION

Although it has long been known that unicellular species generally have much smaller genomes than multicellular species, the underlying reasons for these differences have only recently become clear. From the completely sequenced genomes of ∼300 prokaryotes and several dozen eukaryotes, we now know that genome-size expansion in multicellular eukaryotes is largely a consequence of noncoding-DNA proliferation (86, 87). It is commonly assumed that genomic expansion in eukaryotes is a reflection of natural selection promoting various phenotypic effects, and that genomic streamlining in microbes is a consequence of selection for metabolic efficiency. However, there is yet no direct evidence that adaptive processes are sufficient to explain genome architectural diversity, and there are compelling reasons to think that they are not.

If expansions in genome size have adaptive roots, given that unicellular species have enormous population sizes by multicellular standards and hence a higher population-level potential for adaptive mutation, why are the genomes of such species kept at such diminutive sizes? If there is a metabolic advantage to small genome size, why are large multicellular species relatively immune to such costs? Given the exceptional requirements for gene-expression coordination in developmentally complex organisms, why have nearly all multicellular eukaryotes abandoned the use of operons? And if there is a selective advantage to bulk DNA, why are most increases in genome size accomplished largely by expansions of mobile elements and introns, both of which are mutational, transcriptional, and metabolic burdens? Although the accumulation of further data from poorly studied groups will undoubtedly promote refinements in our hypotheses on these matters, the candidate mechanisms responsible for the phylogenetic diversification of genome architecture must reside in the universal population-genetic processes that drive evolution—mutation, recombination, random genetic drift, and natural selection.

This review (a) summarizes the existing information bearing on the issues outlined above, (b) highlights the limitations of adaptive explanations for genome-size variation, and (c) demonstrates why a mature field of evolutionary genomics will ultimately require an integration of population-genetic principles. Comparative genomics is a wide-ranging field, and it is impossible to cover the full menu of its target subjects here. The goal instead is to review a subset of observations
bearing on the development of a unifying theory for the structural features of prokaryotic and eukaryotic genomes. Microbial genomes occupy an extraordinarily broad range of population-genetic environments and hence take center stage in such considerations.

THE SCALING OF GENOME CONTENT AND GENOME SIZE

Across all cellular life, including DNA viruses, there is a general increase in both coding and noncoding DNA with total genome size, but the proportional contributions of these components behave differently (Figure 1). With few exceptions, viral and prokaryotic genomes consist of >85% coding DNA, and a similar situation is found in the smallest unicellular eukaryotic genomes. In contrast, the largest well-characterized genomes of unicellular eukaryotes contain up to 60% noncoding DNA, with a particularly dramatic increase in this contribution occurring with genomes >10 Mb in size. For the largest genomes of vertebrates and land plants, the amount of coding DNA plateaus at ~100 Mb, leading to a progressive decline in the fractional contribution of this source to <5% of the total genome. Both intronic and intergenic DNA contribute to eukaryotic genome growth. A central challenge for evolutionary genomics is to determine the extent to which the expansion of eukaryotic gene and genomic complexity was a necessary prerequisite or an indirect consequence of the evolution of complex morphologies (86, 87).

THE EVOLUTIONARY MAINTENANCE OF COMPACT GENOMES

Genome-size expansion and contraction is a function of the distribution of insertion and deletion sizes produced by mutation and the subsequent differential filter imposed by natural selection. If natural selection were entirely ineffective, genome size would progressively expand if the rate of insertion exceeded that of deletion and would progressively decline to a minimum level compatible with the maintenance of gene function if deletions were numerically dominant. If, however, virtually every insertion and deletion were subject to natural selection, the size of a genome (and the architecture of each gene within it) would evolve toward an optimal state for organismal fitness. Thus, one of the keys to understanding genome-size evolution resides in the spectrum of changes produced by mutation.

Directional Mutation Pressure

In many species, mobile-genetic element activity encourages genome expansion. Transposons and retrotransposons are capable of proliferating to new genomic sites at rates >10^{-5} per element per generation (48, 93, 99) and only rarely are lost by spontaneous
excision. Retrotransposon activity can also result in the insertion of pseudogenes (dead-on-arrival copies of otherwise normally functioning genes) either via accidental incorporation of downstream sequence into element transcripts (95) or via reverse transcription and genomic integration of normal mRNAs (38). With these kinds of large-scale (generally > 1 kb) insertions arising on a stochastic basis, the prevention of runaway genome expansion
requires the direct opposition of the fixation of insertions by natural selection and/or additional mutational mechanisms for subsequent deletion. Some forms of cellular activity, such as the repair of double-strand breaks by non-homologous end joining, may bias the mutational spectrum in the direction of deletions (29, 107), but numerous other pathways lead to small-scale insertions (110, 111).

Comparative surveys of pseudogenes have been used to infer the relative rates and sizes of small (<500 bp) insertions and deletions. In animals, inferred deletions outnumber insertions by factors of two to six, with the average sizes of each being approximately equal (9, 66, 102, 105, 127, 129). A substantially higher rate of deletion than insertion has also been observed in pseudogenes in prokaryotes (94). Taken at face value, such observations suggest that, with no assistance from selection, small-scale deletions might be sufficient to prevent runaway genome expansion. However, this is the case only if the net erosion of DNA by small-scale deletion exceeds the rate of expansion by large-scale events, and even in this case, an equilibrium level of noncoding DNA would be defined by the relative intensities of these two activities.

One concern with these indirect studies involves the assumption that observations on pseudogenes provide an unbiased perspective of the mutation process. Any insertion-associated disadvantages and/or deletion-associated benefits will tilt the observed spectrum of effects toward deletions relative to the input by mutation (20). As discussed below, the energetic effects of small insertions or deletions on a pseudogene may be so small that they are immune to selection, but the primary burden of excess DNA need not involve metabolic effects but rather the potential for causing gene-expression problems. Even if insert spacer DNA is immune to selection against loss-of-function mutations, this need not imply an absence of selection against gain-of-function mutations. For example, noncoding regions are known to be depauperate in short motifs with potential for generating inappropriate transcription-factor binding (59), microRNA hybridization (40), and translation initiation (90, 112), and sequences contained within defective mobile elements can influence the regulation of adjacent genes (71, 84, 117, 119). That the majority of eukaryotic genomic DNA may be transcribed (18, 67), at least at low levels, raises the question of whether any segment of nonfunctional DNA is truly neutral.

Are such potential selective effects likely to be large enough to significantly bias our view of the insertion or deletion process? The tentative answer is yes, given that more direct estimates of the mutational spectrum (less biased by selection) often reveal an excess of insertions. For example, a massive sequence analysis of long-term mutation-accumulation lines of the nematode Caenorhabditis elegans taken through individual bottlenecks each generation revealed a 15:4 ratio of insertions to deletions (none of which were associated with mobile-element activity) (33). In Drosophila melanogaster, spontaneous insertions <4 kb in length are approximately four times more abundant than deletions (128), and reporter-construct experiments in the yeast Saccharomyces cerevisiae suggest a similar insertion and deletion disparity (60, 72, 101). These direct assays imply an innate mutational tendency for genome-size expansion above and beyond that caused by mobile-element activity and segmental duplications. Such observations may not be universal, however, as reporter-construct studies in Escherichia coli suggest a slight deletion bias (115, 116), as does a survey of de novo dominant mutations for human genetic disorders (70), although these studies will be subject to bias if deletions and insertions are not equally likely to produce the phenotype used to infer the presence of a mutation. Despite the need for more data, these observations shed significant doubt on the notion that mutational deletion processes are universally sufficient to prevent runaway genome size. If this view is correct, then some form of natural selection is necessary for genome-size stabilization.
The Selfish-DNA Hypothesis

Soon after the ubiquity of mobile-genetic elements became apparent, Doolittle & Sapienza (35) and Orgel & Crick (103) suggested that noncoding DNA is largely a byproduct of “selfish” elements proliferating within host genomes until opposed by natural selection. However, although the verbal logic underlying this idea appears to be formally correct with respect to transposons and retrotransposons (19), the selfish-DNA hypothesis fails to explain the large fraction of noncoding DNA that is incapable of self-replication (including spliceosomal introns and small repetitive DNAs), nor does it explain why nearly all types of excess DNA mutually expand in some genomes (e.g., multicellular eukaryotes) but not in others.

It is commonly argued that the compact genomes of microbial species is a consequence of selection for high replication rates, resulting in the elimination of all nonessential DNA (17, 35, 54, 103, 108, 113). The implicit assumption here is that the cost of maintaining and replicating a DNA segment of a few base pairs (the typical size of an intergenic insertion or deletion) is significant enough to be perceived by natural selection. As discussed below, the large effective population sizes of most unicellular species enhance the efficiency of natural selection, so this possibility cannot be ruled out entirely. However, it is unclear whether the rate of cell replication is ever limited by DNA metabolism.

First, within and among prokaryotic species, there is no correlation between cell division rate and genome size (6, 94). Second, during rapid-growth phases, prokaryotic chromosomes are often present in a nested series of replication stages (15), and some species harbor tens to hundreds of chromosomal copies at various stages of the life cycle (69, 92). Third, in E. coli and other eubacteria, DNA replication forks proceed approximately 10 to 20 times faster than mRNA elongation rates, and both processes occur simultaneously (11, 27, 45). Finally, energetic arguments suggest that cell replication is limited by factors other than DNA metabolism. Even after accounting for genomic multiplicity, DNA constitutes ~2% to 5% of the total dry weight of a typical prokaryotic cell (26, 27), and the estimated cost of genomic replication relative to a cell’s entire energy budget is even smaller (64). Similar conclusions emerge for eukaryotic cells (114).

Bulk DNA as a Phenotypic Modulator

The selfish-DNA hypothesis represented a significant break with the common (and unjustified) tradition of attempting to explain all patterns of biodiversity in terms of adaptive processes. Indeed, prior to the development of this hypothesis, a number of authors had suggested that bulk DNA in eukaryotes is selectively promoted for reasons entirely unassociated with gene content (5, 16, 23). Drawing from an impressive number of correlations between nuclear volume, cell size, and cell division rates, Cavalier-Smith (16) suggested that the evolution of cell size imposes secondary selection on nuclear genome size, which in turn modulates the area of the nuclear envelope and hence the flow of transcripts to the cytoplasm. Noting that cell division rates are related to life-history features such as developmental rate and size at maturity, it was a small step to conclude that selection on fitness-related traits encourages the expansion and contraction of noncoding DNA, just as it drives advantageous nucleotide substitutions in coding regions. Because the postulated effects of bulk DNA are irrelevant to species without a nuclear envelope, with its adherence to adaptive arguments, the bulk-DNA hypothesis again invokes metabolic efficiency arguments to explain the streamlined genomes of prokaryotes (17).

A major difficulty with the bulk-DNA hypothesis became clear with the annotation of fully sequenced genomes. Invariably, a substantial fraction of the growth of large
genomes is derived from the proliferation of mobile-genetic elements (87). Given the considerable mutational burden imposed by such activity (48, 93, 99), this form of genomic growth is more likely to be opposed than encouraged by positive selection at the level of the host genome (19).

A second concern with the bulk-DNA hypothesis involves the matter of causality. The known mechanisms for controlling cell growth include modifications of gene-copy number, ribosome number, transcription rates, nuclear membrane porosity and export rates, and transcript longevity. The bulk-DNA hypothesis provides no explanation for why natural selection on life-history features would be forced to accomplish cell metabolic changes via risky expansions or contractions of noncoding DNA rather than through conventional allelic changes in gene expression and/or function. An alternative possibility is that there is no direct causal connection between genome size and cell features, but rather that genome-size expansion is largely a passive response to life-history features that promote population-genetic properties that reduce the ability of natural selection to eradicate excess DNA (86, 87).

A key point to be detailed below is that the long-term effective size of a population is a fundamental determinant of the possible pathways of genomic evolution in a lineage. Suppose that natural selection favors an increase in cell or body size in a particular lineage. In general, a doubling in organismal size results in a ~50% reduction in the number of individuals within a species (below), and by increasing the power of random genetic drift, this reduces the efficiency of natural selection, thereby facilitating the accumulation of mildly deleterious insertions. Thus, genome size is expected to expand in larger organisms, not necessarily because of an intrinsic insensitivity to excess DNA (as postulated by the selfish-DNA hypothesis) or because of a direct advantage of such DNA (as postulated by the bulk-DNA hypothesis), but because of a reduced ability to eradicate it.

The Mutational-Burden Hypothesis

An alternative explanation for the phylogenetic diversity of gene and genome architecture, with broad explanatory power, derives from fundamental principles of population genetics and the known constraints on gene function. As noted above, nearly all forms of excess DNA are a mutational burden, even when they have no direct functional significance. For example, genes carrying introns must retain specific sequences for proper spliceosomal recognition of splice sites (85). Upstream regions of genes that are transcribed but not translated (5' UTRs) provide substrate for the mutational origin of premature translation-initiation codons, which can result in defective frameshifted products (90). Complex, modular regulatory regions increase a gene's mutational target size relative to simpler, overlapping regulatory elements (44), and nonfunctional DNA can also increase the rate of origin of harmful regulatory mutations (40, 59).

Under the hypothesis that the primary burden of excess DNA is its mutational liability, minimization of genome size (within the constraints dictated by the maintenance of gene function) is expected to be uniformly favorable (86). However, the prevailing population-genetic environment will determine which lineages are most capable of purging excess DNA by natural selection. Before considering a semiformal explanation for this hypothesis, two additional points are worth making. First, although mutation and selection are often viewed as independent processes, variation in the degenerative mutation rate among alleles is equivalent to a selective process, as the descendants of some alleles are removed from the population more rapidly than others. Second, although random genetic drift is often viewed as evolutionary noise that simply renders the outcome of natural selection less predictable, this is a gross oversimplification of the situation at the genomic level, where the degree of drift predictably defines the types of pathways that are open to evolution.
Imagine a newly arisen mutation that reduces the fitness of its carriers by the fraction $s$ (the selection coefficient). In an infinite population, such a mutation would virtually always be eliminated by natural selection, but finite populations experience random temporal fluctuations in allele frequency owing to the chance variation in gamete or zygote sampling. The magnitude of such fluctuations is determined by the effective number of genes residing at a locus at the time of reproduction ($N_g$), which is equivalent to the effective number of individuals in a haploid population and roughly twice that in a randomly mating diploid population. The stochastic component of allele-frequency change is defined by $1/N_g$, and natural selection will be effective at eliminating a deleterious allele only if the directional change dictated by $s$ exceeds this value. If $1/N_g$ is substantially larger than $s$, a mutant allele will behave as though it were entirely neutral, as its dynamics are almost entirely dictated by chance events.

In the context of long-term genome evolution, the most relevant consideration is the probability that a new mutation will eventually displace the descendants of all other gene copies in the population. For a completely neutral mutation, this fixation probability is simply equal to the initial frequency, as all gene copies have equal chances of stochastic elimination. A key to understanding evolutionary processes is the degree to which $2N_g s$ varies among lineages.

Taken at face value, this simple theoretical principle would appear to introduce insurmountable practical difficulties, as the technical barriers to obtaining direct estimates of both $N_g$ and $s$ are enormous. However, progress can be made by recalling the mutational basis of the selective disadvantage of excess DNA. Suppose that the addition of a segment of DNA within or around an essential gene imposes the requirement that $n$ nucleotides must remain unaltered to maintain gene function, e.g., the requirements for spliceosomal-intron recognition imply $n \approx 30$ (85). If each of the $n$ hazardous sites mutates at rate $u$ (per generation), a piece of excess DNA will elevate the rate of production of defective alleles by an amount $nu$. Thus, substituting $nu$ for $s$, the criterion for effective neutrality of a hazardous DNA segment becomes $2N_g nu \ll 1$ (86). Rearranging, a new DNA insertion that imposes $n$ hazardous sites at a locus will be capable of drifting to fixation in an effectively neutral fashion provided $2N_g u \ll 1/n$, whereas lineages with $2N_g u \gg 1/n$ are essentially immune to such colonization.

This simple relationship has considerable utility because the composite quantity $2N_g u$ is readily estimated by measurements of nucleotide diversity at silent sites in natural populations (below). Moreover, the quantity $2N_g u$ has a simple mechanistic interpretation, as it is equivalent to twice the ratio of the forces of mutation ($u$) and genetic drift ($1/N_g$) operating at a nucleotide position. The remainder of this paper explores whether $2N_g u$ in microbes is sufficiently large to insure a path toward genomic streamlining as postulated by the mutational-burden hypothesis.

THE POPULATION-GENETIC ENVIRONMENT OF MICROBIAL SPECIES

A major gap in our understanding of molecular genetics concerns the rate at which mutations arise. Most estimates for microbes are derived from laboratory reporter constructs, whereas most estimates for multicellular
species are derived from phylogenetic comparisons of silent sites in protein-coding genes. Neither approach is without problems: the first relying on various assumptions regarding mutational detectability, and the second on assumptions regarding neutrality. Nevertheless, with these caveats in mind, some qualitative generalities have begun to emerge for \( n \) (Table 1). On a per-cell division basis, prokaryotes have very low mutation rates, approximately three times lower than those for unicellular eukaryotes. The per-generation mutation rates of multicellular eukaryotes are approximately 10 to 50 times greater than those for unicellular eukaryotes, a degree that correlates with the number of cell divisions per generation in the former (86). The rates for DNA viruses are enormous, approximately 10 to 100 times greater than those for prokaryotes.

Put in the context of the previous theory, these estimates provide some indication of the population-size constraints on genomic evolution. For example, noting that the threshold \( N_g \) below which hazardous excess DNA is likely to accumulate is \( \sim 1/(2n) \), for a moderately sized insert with \( n = 10 \), the threshold \( N_g \) is \( \sim 10^6 \) for viruses and multicellular eukaryotes, \( \sim 10^7 \) for unicellular eukaryotes, and \( \sim 10^8 \) for prokaryotes. Because \( 10^8 \) is a minuscule population size for a unicellular species (41–43), if taken at face value, these benchmarks might suggest that few microbes should ever be on a road to genomic obesity. However, \( N_g \) is not simply the absolute number of individuals within a species, but rather reflects the genetic effective population size. The latter is influenced by numerous demographic and ecological factors as well as by the magnitude of linkage within a genome, which together lead to a pronounced reduction of \( N_g \) below the numerical abundance of a species (12, 21, 52).

Although there are well-established methods for estimating \( N_g \) from observations on pedigree structure, life-history features, and/or temporal fluctuations in allele frequencies, none of these approaches incorporate the long-term effects of linkage. The matter is important because nucleotide positions that are tightly linked to other selected sites are subject to reduction in \( N_g \) via Hill-Robertson (62) effects. This is because selection reduces the effective number of segments of a particular chromosomal region transmitted to the next generation, decreasing the efficiency of selection throughout the entire region (52). In principle, a large microbial population could have a small \( N_g \) if selective sweeps during prolonged periods of asexual reproduction periodically purged most lineages from the population.

The only known way to estimate a species \( N_g \) relevant to long-term genomic evolution is to evaluate the amount of standing variation at nucleotide sites assumed to be neutral. With new divergence between sites in two random alleles arising at rate \( 2n \) per generation, and

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Mutation rate estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (SE)</td>
</tr>
<tr>
<td>Mammalian virus</td>
<td>12.41</td>
</tr>
<tr>
<td>Bacteriophage</td>
<td>37.35 (18.63)</td>
</tr>
<tr>
<td>Prokaryotes</td>
<td>1.60 (0.41)</td>
</tr>
<tr>
<td>Unicellular eukaryotes, nu</td>
<td>6.62 (3.17)</td>
</tr>
<tr>
<td>Invertebrates, nu</td>
<td>51.64 (23.88)</td>
</tr>
<tr>
<td>Mammals, nu</td>
<td>7.30</td>
</tr>
<tr>
<td>Land plant, nu</td>
<td>193.04 (57.67)</td>
</tr>
<tr>
<td>Vertebrate, mt</td>
<td>0.91 (0.33)</td>
</tr>
<tr>
<td>Land plant, mt</td>
<td>0.91 (0.33)</td>
</tr>
</tbody>
</table>

Rates are given for base substitution mutations, in units of \( 10^{-9} \) per nucleotide site per generation, where a generation is equivalent to a single-genome replication in the case of microbes and to a complete generation for multicellular species (including organelles). The sample size \( (n) \) denotes the number of species for which data are available. All estimates for nonmulticellular species are derived from laboratory experiments with reporter constructs. All but one estimate (a direct sequence-based estimate from a mutation-accumulation experiment) for animal nuclear genomes are derived from phylogenetic comparisons of silent sites in protein-coding genes. The land-plant estimate for the nuclear genome is based on a broad phylogenetic comparison, assumes an annual life cycle, and should be multiplied by \( x \) for species with generation times of \( x \) years. Estimates for organelles are obtained from phylogenetic comparisons. Data for double-stranded DNA viruses are taken from Drake (36) and Drake & Hwang (37), while those for the remaining taxa are derived from compilations in Lynch (86) and Lynch et al. (89). nu, nuclear; mt, mitochondrion.
old heterozygosity being lost at rate 1/Ng, the average divergence at neutral sites is 2Ngμ at mutation-drift equilibrium. The traditional estimator of this quantity is the average number of substitutions separating the silent sites of randomly sequenced protein-coding genes (πs), with the most notable studies of this sort in microbiology being MLST (multi-locus sequence typing) surveys of global isolates of prokaryotes and fungi (24, 51). On average, πs is ~0.104 in prokaryotic species, and ~0.057 and ~0.014 in the nuclear genes of unicellular and multicellular eukaryotes, respectively (Table 2). Substantial variation among estimates exists within each of these groups as a consequence of sampling error, but the ranking of average values is highly significant, and because most species can be expected to fluctuate in density over long timescales, these values are much more meaningful than individual estimates. There are, nevertheless, a number of caveats with respect to these data.

First, because the mean time to fixation of a neutral mutation is 2Ng generations, for constant μ, larger 2Ngμ estimates integrate the series of events within a lineage over longer time spans. Second, a substantial fraction of 2Ngμ estimates for microbial species (prokaryotic and eukaryotic) are derived from pathogens, the population structure of which must be largely dictated by their substantially less abundant multicellular hosts. As a consequence, pathogens have reduced levels of variation relative to free-living species (86). Third, numerous difficulties exist in identifying species boundaries of microbial species. Although microbial sequence surveys often subdivide species into clades, and these were generally adhered to with the estimates leading to the results in Table 1, such distinctions are often no more defensible than drawing phylogenetic boundaries between populations of multicellular species differing in local adaptations. Fourth, the interpretation of silent-site diversity as an estimate of 2Ngμ critically depends on the assumption of neutrality of synonymous sites of protein-coding genes. This is a significant concern because the composition of silent sites of mRNAs may be under selection for features influencing transcript processing and/or translation efficiency. Any downward bias in estimates of 2Ngμ imposed by such selection may be greatest in populations with large Ng, and the fact that prokaryotic divergence rates of silent sites can be 5 to 10 times lower than actual mutation rates may reflect such an effect (81, 100).

Finally, although prokaryotes do not engage in conventional meiotic recombination, which could greatly magnify their vulnerability to selective sweeps, such species do engage in a number of parasexual processes, including conjugation, transformation, and transduction. The net effects of such activities will determine whether prokaryotes are more vulnerable to hitchhiking effects than multicellular eukaryotes. Insight into this matter can be gained from the levels of linkage disequilibrium among silent sites in natural populations, which provide an estimate of 2Ngμc, where c is the rate of recombination per site per generation. Dividing such a measure by 2Ngμ provides an estimate of c/μ, the recombination rate scaled to the mutation rate, a parameter that dictates the vulnerability of a species to

### Table 2 Estimates of 2Ngμ

<table>
<thead>
<tr>
<th>Category</th>
<th>Mean (SE)</th>
<th>Standard deviation</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prokaryotes</td>
<td>0.1044 (0.0161)</td>
<td>0.1113</td>
<td>48</td>
</tr>
<tr>
<td>Unicellular eukaryotes, nu</td>
<td>0.0573 (0.0150)</td>
<td>0.0777</td>
<td>27</td>
</tr>
<tr>
<td>Animals, nu</td>
<td>0.0134 (0.0025)</td>
<td>0.0149</td>
<td>36</td>
</tr>
<tr>
<td>Land plants, nu</td>
<td>0.0152 (0.0027)</td>
<td>0.0134</td>
<td>24</td>
</tr>
<tr>
<td>Unicellular eukaryotes, mt</td>
<td>0.0175 (0.0072)</td>
<td>0.0177</td>
<td>6</td>
</tr>
<tr>
<td>Animals, mt</td>
<td>0.0345 (0.0036)</td>
<td>0.0288</td>
<td>66</td>
</tr>
<tr>
<td>Land plants, mt</td>
<td>&lt;0.0004 (0.0004)</td>
<td>0.0008</td>
<td>4</td>
</tr>
</tbody>
</table>

All estimates are of intraspecific diversities for silent sites of protein-coding genes, taken from compilations in Lynch (86) and Lynch et al. (89). The unicellular eukaryote category includes some taxa with a small number of cell types (e.g., filamentous fungi and slime molds). The sample size (n) denotes the number of genera in the analysis, with the estimate for each genus being an average of several species in some cases. nu, nuclear; mt, mitochondrion.
selective sweeps. The average estimates of this quantity in animals and land plants, 2.3 (0.9) and 1.2 (0.5), respectively, are comparable to the average for the five prokaryotic species with available data, 4.3 (1.6) (86). Thus, although there is considerable room for more work in this area, there appears to be no formal justification for the idea that prokaryotic species are unusually vulnerable to hitchhiking effects.

Taking all of these issues into consideration, it is difficult to escape the conclusion that the average $2N_{eu}$ for prokaryotes is at least 10 times that in multicellular eukaryotes, despite the relatively low prokaryotic mutation rate. It is notable, however, that even if $2N_{eu}$ in prokaryotes is 10 times greater than the estimates in Table 2 suggest, the average prokaryotic $N_e$ would still be only $\sim 10^{10}$. As the total numbers of individuals within species increase nearly 20 orders of magnitude from large vertebrates to miniscule prokaryotes, this finding demonstrates that linkage effects are much more important than demographic effects in defining long-term genetic effective population sizes.

CONSEQUENCES OF EFFICIENT REMOVAL OF MUTATIONALLY HAZARDOUS DNA

In an earlier review on the hazardous nature of insertional DNA, Lawrence et al. (78) suggested that prokaryotes have evolved mutational processes biased toward deletions to prevent such accumulations. However, as noted above, there is yet no convincing evidence for deletion bias at the level of mutation in any organism, and were such a bias to exist, it would impose a substantial challenge to the maintenance of essential genes in a compact genome. The argument promoted here, with two examples, is that mutational deletion bias need not be invoked at all to explain the sanitized genomes of microbial species, as $N_e$ in most such species is large enough for natural selection to inhibit the accumulation of excess DNA. Moreover, purifying selection in prokaryotes appears to be so efficient that evolutionary pathways that are unavoidable in multicellular eukaryotes are essentially unavailable for exploration. Contrary to the argument that the relative constancy in genome architecture among free-living prokaryotes is a deep mystery (58), this pattern is entirely compatible with population-genetic theory.

The Immunity of Microbial Species to Intron and Mobile-Element Proliferation

Soon after the discovery that eukaryotic genes are often subdivided into pieces, it was proposed that introns contributed to the formation of new proteins by promoting the recombinational shuffling of modular domains into combinations with novel functions (30, 34, 50). However, although it is now clear that every well-studied eukaryotic genome contains introns as well as the spliceosomal machinery for removing them from precursor mRNAs (88), it is equally clear that all prokaryotic genomes are devoid of them. There is no direct evidence that prokaryotes ever harbored spliceosomal introns, and the theory presented above suggests why.

Because the known molecular requirements for spliceosomal processing imply the reservation of $n \approx 20$ to 40 nucleotide sites for the precise recognition of each intron (85, 86), the approximate condition for the maintenance of an intron is $2N_{eu} \ll 1/20$. Thus, because the average estimate of $2N_{eu}$ for prokaryotes, $\pi_s = 0.10$, is probably downwardly biased, natural selection appears to be efficient enough to effectively immunize prokaryotic genomes from intron colonization. In contrast, average $\pi_s$ for unicellular eukaryotes (0.06) is close to the threshold for spliceosomal intron colonization, and such species exhibit a wide range of variation in intron numbers, with some approaching zero (Figure 1). Finally, with average $\pi_s \approx 0.01$, multicellular species provide a highly permissive environment for intron colonization, and
all such species average four to seven introns per protein-coding gene (86, 87).

Some prokaryotic genomes do contain Group-II “self-splicing” introns, which, instead of relying on an extrinsic spliceosome, moderate their own excision by use of secondary RNA structures in combination with self-encoded multifunctional proteins (74). Because of the larger number of nucleotide sites involved in Group-II intron function, such elements are expected to impose a substantially greater mutational burden on host genes than do spliceosomal introns and should be easily purged from prokaryotic genomes. In fact, the few Group-II introns that do reside in prokaryotes avoid this fate by inhabiting intergenic regions (in which case, they are not technically introns at all) or mobile elements (which may facilitate horizontal transfer) (28, 109, 124, 130).

The phylogenetic distribution of mobile elements closely resembles that of introns. Although prokaryotic genomes can harbor various forms of transposons (often referred to as insertion sequences), most of them contain 10 or fewer elements, which together usually comprise <0.1% of the host genome (91). Retrotransposons are entirely unknown in prokaryotes, although a poorly understood type of element is found at very low levels in some species (75). In eukaryotes, there appears to be a threshold genome size of ∼10 Mb below which mobile elements are nearly absent, an intermediate range in which only a fraction of species harbor them, and an upper threshold (∼100 Mb) above which all species (entirely multicellular) are heavily infected with multiple element families (87).

This general pattern is readily understood in terms of the demographic requirements for the maintenance of a mobile-element family. Once inserted into a host genome, all mobile elements accumulate degenerative mutations in an effectively neutral manner, such that the long-term viability of an element family within a host species requires that an average element produce at least one successful offspring insertion prior to complete inactivation. Because mobile-element insertions have a wide range of deleterious effects on the host (8, 25, 63), many of them are eliminated by natural selection, the exact number depending on the host’s $N_g$. Insertions with effects on host fitness >1/$N_g$ are rapidly purged, whereas those with effects <<1/$N_g$ are free to go to fixation. Thus, for any mobile-element family, there exists a critical host $N_g$ above which the maintenance of insert number is unsustainable.

The Demise of Operons and the Emergence of Modular Gene Structure in Eukaryotes

Coregulation of the expression of genes mutually engaged in metabolic and developmental processes is critical to all organisms. Prokaryotes often achieve this end by using operons to cotranscribe adjacent genes, a process only rarely seen in eukaryotes. Nearly all genes in trypanosomes are transcribed into poly-cistronic units prior to trans splicing (13), but the genomes of such species are remarkably prokaryote-like in that they are nearly devoid of introns and mobile elements. Some genes of nematodes (10), flatworms (31), and nucleomorphs of chlorarachniophyte algae (53) are also arranged into operons, but the vast majority of well-studied eukaryotes (including fungi, plants, and other metazoans) appear to lack operons entirely. Given that the eukaryotic lifestyle does not impose an insurmountable barrier to operon function and the potential advantages of operons for coordinated gene expression, why have operons become disassembled in most eukaryotic genomes?

Following the logic developed above, a case can be made that the loss of eukaryotic operons was not a consequence of positive selection for single-gene control, but a passive response to a reduction in the efficiency of selection for operon maintenance. Operons minimize the mutational vulnerability of genetic pathways by eliminating the necessity of maintaining gene-specific regulatory elements. However, because small local
duplications are common and most transcription factor binding sites are simple enough (<10 bp in length) to arise at an appreciable rate by single-base substitutions, all genes are vulnerable to the chance invasion of regulatory elements capable of driving independent expression. By increasing the size of the mutational target, the accumulation of gene-specific regulatory elements imposes a selective disadvantage equivalent to \( \sim 10u \) for each regulatory element (assuming 10 nucleotides per element). Thus, unless \( 2N_eu > 0.1 \), which is rarely the case in eukaryotes, selection is incapable of preventing the fixation of a mutant self-regulated allele. In contrast, prokaryotic populations, where \( 2N_eu > 0.1 \) is common, are resistant to the break up of operons by drift-mutation processes.

Thus, direct selection for genomic streamlining need not be responsible for the assembly and maintenance of operons, which instead may be simple consequences of population-genetic environments that discourage the emergence of single-gene control by mutational processes. A similar idea has been suggested by Price et al. (106), although these authors put more emphasis on positive selection for the coordinated expression of genes than on the mutational cost of maintaining redundant regulatory machinery for individual genes.

This hypothesis is a radical departure from the selfish-operon model of Lawrence & Roth (79), which postulates that operons evolve to facilitate the joint horizontal transfer of functionally integrated genes to alternative host species. The selfish-operon model was motivated by the belief that (a) the metabolic advantages of operons are too weak to maintain them in the face of stochastic microchromosomal rearrangements, and (b) the chance origin of gene-specific promoters will eventually overwhelm the benefits of coregulation. However, the fact that functionally related genes can remain aggregated to some extent even outside operons, in both prokaryotes (76, 104) and eukaryotes (22, 80, 82, 83, 120), is inconsistent with the first point, and the ability of selection to oppose the mutational cost of independent gene regulation in microbes is inconsistent with the second point.

**INSIGHTS FROM ENDOSYMBIONTS**

Prokaryotes have entered stable coevolutionary relationships with larger eukaryotic hosts on numerous occasions. Such interactions can dramatically alter the population-genetic environment by changing the colonist’s transmission dynamics and/or mutational features. Thus, the genomes of endosymbionts provide unique resources for evaluating the general principles outlined above. The degree to which an endosymbiont’s mutation rate is altered depends on unpredictable gains and/or losses of DNA-repair pathways experienced subsequent to colonization as well as on modifications in environmental mutagenicity imposed by host cells, but in general endosymbionts are expected to experience substantial effective population size reductions. Once such species have become so entrenched in their host’s biology that transmission is entirely vertical (strictly between host parent and offspring), \( N_e \) is expected to approximate that of the host species (7).

**Mitochondrial Genomes**

One of the defining features of eukaryotes is the presence of a mitochondrion, apparently acquired in the stem eukaryote via the colonization of an \( \alpha \)-Proteobacterium that subsequently propagated throughout the entire eukaryotic domain (55, 56). Relative to their eubacterial ancestors, all mitochondrial genomes are extraordinarily degenerate with respect to gene content. None contain more than 70 genes and most considerably fewer. Although the scaling of mitochondrial genome content with genome size is similar to that for nuclear and prokaryotic genomes, the placement of phylogenetic groups is dramatically different (89). Most notably, in contrast to the similar architectures of the nuclear
estimated to be nuclear mutation rates in most animals is evidence. Based on a broad survey of interspecific divergence, the ratio of mitochondrial to nuclear mutation rates in most animals is estimated to be ~10 to 25, whereas that in land plants is ~0.05, and that in unicellular species is ~1.0 (89). When measures of silent-site diversity (\(\pi_u\)) for both mitochondrial (\(m\)) and nuclear (\(n\)) genes within a species are available along with measures of silent-site divergence between species, the ratio of diversity to divergence, \(\left(2N_{gm}u_m/2N_{gn}u_n\right)/(u_m/u_n)\), provides an estimate of the ratio of \(N_g\) in the two genomic compartments, which averages ~1.3 in animals and ~0.5 in unicellular eukaryotes (89). Unfortunately, data are not available for a comparable computation in land plants, but given that \(N_g\) for nuclear genes is comparable in plants and animals, there is no compelling reason to think that land-plant mitochondria would have unusually low \(N_g\).

Thus, even though organelle genomes are present in multiple copies per host cell, the rapid loss of heteroplasmy within individuals causes the overall genealogical depth of the mitochondrion to be similar to that for nuclear genes of the host species (7). Because \(u\) for plant mitochondria is comparable to that for prokaryotes (on a per-generation basis) (Table 1), were it not for differences in \(N_g\), plant mitochondrial genomes might be expected to evolve a prokaryote-like genome structure, but the tremendous decrease (\(\geq 250\) times) in \(N_g\) in land plants causes a substantial decline in the ability of selection to remove mutational-hazardous excess DNA relative to the situation in microbes.

With the differences in \(2N_{gm}u\) between animal and land-plant mitochondria in mind, a broad array of superficially disconnected differences in the genomic features of these two lineages can be accommodated by the mutational-burden hypothesis (89). For example, although mitochondria do not harbor spliceosomal introns, land-plant mitochondria typically contain 20 to 30 Group-II introns, substantially more than the 0 to 8 found in green algal species. As noted above, such introns are expected to have stringent sequence requirements for proper splicing, and the high level of average \(2N_{gm}u\) for animal mitochondria appears to rule out the possibility of their residence. Consistent with this hypothesis, the
only animal mitochondria that harbor introns are contained within cnidarians, which apparently have low (plant-like) mitochondrial mutation rates.

**Arthropod Endosymbionts**

Although the mitochondrion represents the most spectacularly successful case of endosymbiosis, many more phylogenetically restricted endosymbionts are known, the eu-bacterial inhabitants of various insect lineages having received the most attention. Because most such species are strictly vertically inherited, like mitochondria, their $N_g$ is expected to be comparable to that of their hosts. This idea can be checked with data for the γ-Proteobacterium *Buchnera aphidicola*, an obligate endosymbiont of aphids. The estimate of $\pi$ for *B. aphidicola* derived from a wide geographic sample of its host *Uroleucon ambrosiae* is low and not greatly different from that for the host’s mitochondrion, 0.0042 versus 0.0036, whereas the ratio of silent-site divergence for *B. aphidicola* versus mitochondria in different host species is 0.60 (47, 61). Employing the approach described above, *B. aphidicola* $N_g$ is ~1.9 times that of the host mitochondrion. Thus, for this species at least, $N_g$ is indeed substantially reduced with respect to the average free-living prokaryote. However, given the exceptionally low $\pi$, for the host species, it is possible that $2N_u$ for the *B. aphidicola–U. ambrosiae* assemblage is not an accurate reflection of the historical conditions under which this genome evolved, so the survey of other such species would be useful.

All arthropod endosymbiont genomes exhibit a suite of modifications that likely reflect the consequences of small $N_g$ (97). For example, three sequenced *B. aphidicola* genomes, each from a different aphid lineage, have an extraordinarily high A:T content (~74%), and comparative analyses suggest elevated long-term rates of molecular evolution and the accumulation of mildly deleterious mutations in protein-coding and ribosomal-RNA genes (4, 61, 73, 96, 118, 121, 125). Containing just 550 to 630 protein-coding genes, *B. aphidicola* genomes are among the smallest known in prokaryotes (77), and many hundreds of genes (including those involved in DNA repair) were lost early in the establishment of the lineage, presumably because of the reduced efficiency of selection for gene maintenance in a stable host cell environment. A similar syndrome of events is recorded in other related lineages, *Blochmannia* (an endosymbiont of ants) and *Wigglerworthia* (an endosymbiont of tsetse flies) (1, 14, 32, 46, 49).

Noting that $2N_u$ for *B. aphidicola* is approximately 25 times smaller than the average for other prokaryotes and approximately 8 times smaller than the average for animal mitochondria (*Table 2*), the genomes of γ-proteobacterial endosymbionts might be expected to evolve elevated levels of noncoding DNA (58), but they are often nearly devoid of mobile elements and contain approximately the same proportion of noncoding DNA as other prokaryotes (97). Although the genomes in this group are less streamlined than animal mitochondrial genomes, neither comparison should be taken at face value as supporting or conflicting with the mutational-burden hypothesis. Because relaxed selection need not always promote the evolution of genomic obesity, particularly in this unique group of endosymbionts, the preceding observations are by no means incompatible with the idea that patterns of microbial genome evolution are inconsistent with basic population-genetic principles, contrary to the claims of some authors (58).

For example, species with a strictly clonal mode of inheritance impose an exceptionally unstable habitat for mobile elements. An overly aggressive element may cause the extinction of its host, whereas an overly quiescent element eventually becomes completely inactivated by mutations (2). This filtering process will likely bias the surviving pool of asexual host lineages toward the mobile-element free state, a hypothesis consistent with the observation that prokaryotes
in early (facultative) stages of host associations generally have exceptionally high levels of mobile-element insertions (97). In addition, *Wolbachia*, a broadly distributed endosymbiont with a proclivity for horizontal transfer and recombination, contains large numbers of insertion elements (3, 65). Furthermore, genomic expansion in the face of relaxed selection requires that the mutational spectrum be biased toward insertions. Although the idea that deletion mutations outnumber insertions was questioned above, the matter remains to be resolved for *B. aphidicola* and allies, which have the unique attribute of having lost conventional mismatch repair pathways. By promoting recombination between nonorthologous sequences, defective mismatch repair enhances the deletion process (98).

Two final examples demonstrate that genome size reduction is by no means universal in arthropod endosymbionts. First, *So- dalis glossinidius*, like *Wigglesworthia* species, is an endosymbiont of tsetse flies, but, unlike the latter, has only 51% coding DNA (123), by far the lowest value for any prokaryote except for the intracellular parasite *Mycoplasma leprae* (48%). The A/T content of this species (45%) is much higher than that in other γ-proteobacterial endosymbionts, and the genome harbors a much higher density of pseudogenes than any other endosymbiont. Perhaps it is no coincidence that unlike the species described above, *S. glossinidius* has retained its DNA-repair genes.

Second, thousands of parasitoid wasp species use endosymbiotic viruses (from the *Polydnaviridae* family) to facilitate progeny development within parasitized lepidopteran hosts. Most viruses are pathogens with substantial capacity for horizontal transmission and presumably high $N_G$, as reflected in their streamlined genomes (Figure 1). However, because polydnaviral chromosomes are integrated directly into the wasp genome, they should be subject to the same drift and mutation processes as the host. Consistent with the hypothesis that relatively low $2N_Gu$ encourages the accumulation of excess DNA in animals, the genomes of these viruses are extraordinarily large (~200 to 600 kb), with only 17% to 29% coding DNA (39, 126). The non-coding DNA of these species contains numerous mobile elements, and 10% to 20% of the protein-coding genes contain spliceosomal introns (both highly unusual features for DNA viruses). These observations illustrate that when placed in a small $N_G$ environment, even viral genomes eventually succumb to a predictable suite of “pathological” responses to random genetic drift.

### SUMMARY POINTS

1. The scaling of most aspects of genomic architecture and gene structure with genome size fall on a continuum from viruses to prokaryotes to single-celled eukaryotes to multicellular eukaryotes. This suggests that general population-genetic mechanisms, transcending cellular and metabolic features, are the predominant drivers of interspecific divergence in genome architecture.

2. It is commonly thought that genomic streamlining in microbes is a direct consequence of biased mutation pressure toward deletions or an indirect consequence of selection for high metabolic efficiency, but neither hypothesis enjoys convincing empirical support.

3. An alternative hypothesis is that nonfunctional DNA is opposed by weak natural selection because it increases the degenerative mutation rate of associated genes. The extent to which such selection can operate efficiently depends on the ratio of the power of mutation (u) to that of genetic drift ($1/2N_G$), $2N_Gu$. 

---

*Annu. Rev. Microbiol. 2006.60:327-349. Downloaded from arjournals.annualreviews.org by INDIANA UNIVERSITY - Bloomington on 08/31/07. For personal use only.*
4. In microbial species, particularly prokaryotes, $2N_e u$ is generally large enough to prevent significant colonization of introns, mobile elements, and excess intergenic DNA, whereas $2N_e u$ in multicellular eukaryotes is sufficiently low to provide a highly permissive environment for the expansion of nonfunctional DNA. The quantitative nature of these results strongly implies the population-genetic environment of a species defines the genomic pathways that are open to evolutionary exploration.

5. This general theory also provides a potentially unifying explanation for genome structure modifications observed in other contexts, including the divergent evolutionary trajectories taken by animal versus land-plant mitochondria and the unique genomic attributes of many microbial endosymbionts.

ACKNOWLEDGMENTS

This work has been supported by grants from the National Institutes of Health, National Science Foundation, and the Indiana MetaCyte Initiative funded by the Lilly Foundation.

LITERATURE CITED


44. A classical paper demonstrating the role that linkage plays in defining the long-term effective population size.

52. An outline why the evolution of complex, modular structures for gene regulation requires a small effective population size.
80. Lee JM, Sonnhammer EL. 2003. Genomic gene clustering analysis of pathways in eu-
karyotes. *Genome Res.* 13:875–82
81. Lenski RE, Winkworth CL, Riley MA. 2003. Rates of DNA sequence evolution in exper-
imental populations of *Escherichia coli* during 20,000 generations. *J. Mol. Evol.* 56:498–508
43
83. Lercher MJ, Urrutia AO, Hurst LD. 2002. Clustering of housekeeping genes provides a
exon: 3’ splice-site selection in *Alu* exons. *Science* 300:1288–91
USA* 99:6118–23
88. Lynch M, Hong X, Scofield DG. 2006. Nonsense-mediated decay and the evolution of
89. Lynch M, Koskella B, Schaack S. 2006. Mutation pressure and the evolution of organelle
and distribution of transposable elements in *Drosophila melanogaster*: in situ hybridization
95. Moran JV, DeBerardinis RJ, Kazazian HHJ. 1999. Exon shuffling by L1 retrotransposi-
96. Moran NA. 1996. *Accelerated evolution and Muller’s ratchet in endosymbiotic bac-
teria.* *Proc. Natl. Acad. Sci. USA* 93:2873–78
*Curr. Opin. Genet. Dev.* 14:627–33
16

96. The first attempt to demonstrate the susceptibility of bacterial endosymbionts to the accumulation of mildly deleterious mutations.
106. Price MN, Huang KH, Arkin AP, Alm EJ. 2005. Operon formation is driven by coregulation and not by horizontal gene transfer. Genome Res. 15:809–19