The bioenergetic costs of a gene

Michael Lynch[1] and Georgi K. Marinov

Department of Biology, Indiana University, Bloomington, IN 47401

Edited by W. Ford Doolittle, Dalhousie University, Halifax, Canada, and approved October 6, 2015 (received for review July 29, 2015)

An enduring mystery of evolutionary genomics concerns the mechanisms responsible for lineage-specific expansions of genome size in eukaryotes, especially in multicellular species. One idea is that all excess DNA is mutationally hazardous, but weakly enough so that genome-size expansion passively emerges in species experiencing relatively low efficiency of selection owing to small effective population sizes. Another idea is that substantial gene additions were impossible without the energetic boost provided by the colonizing mitochondrion in the eukaryotic lineage. Contrary to this latter view, analysis of cellular energetics and genomics data from a wide variety of species indicates that, relative to the lifetime ATP requirements of a cell, the costs of a gene at the DNA, RNA, and protein levels decline with cell volume in both bacteria and eukaryotes. Moreover, these costs are usually sufficiently large to be perceived by natural selection in bacterial populations, but not in eukaryotes experiencing high levels of random genetic drift. Thus, for scaling reasons that are not yet understood, by virtue of their large size alone, eukaryotic cells are subject to a broader set of opportunities for the colonization of novel genes manifesting weakly advantageous or even transiently disadvantageous phenotypic effects. These results indicate that the origin of the mitochondrion was not a prerequisite for genome-size expansion.

gene cost | transcription | translation | cellular bioenergetics | evolutionary genomics

Although the idea that there is an intrinsic advantage to both cellular complexity and multicellularity is often taken to be self-evident, there is no direct evidence that either feature has been promoted by natural selection. Arriving at specific evidence to the contrary is also difficult, but plausible hypotheses based on mutation pressure and random genetic drift exist (1–3). Moreover, given that all extant organisms are temporally equidistant from the last universal common ancestor, the fact that multicellularity involving large numbers of cell types is only represented by two eukaryotic lineages (metazoans and land plants) raises additional questions about the global advantages of such body plans (1, 4). To help explain the absence of morphological complexity in prokaryotes, Lane and Martin (5) introduced the cost of a gene as an argument for the impossibility of high levels of cellular/developmental complexity without a power-generating mitochondrion, although an explicit evolutionary definition of such a cost was not provided.

Regardless of the intrinsic advantages/disadvantages of cellular complexity, understanding the evolutionary mechanisms that promote vs. discourage the establishment of various cellular features ultimately requires insight into the energetic costs of such structures. Here, we focus specifically on the cumulative cost of a gene, subdividing this into expenses at the genomic, transcriptional, and protein levels. Although these issues have garnered some prior attention (6, 7), to put these costs into proper context, it is also necessary to understand the lifetime energetic requirements of a cell, which we define in units of numbers of ATP → ADP energy-releasing hydrolysis events. Given the total lifetime energy requirements for a cell, the proportional contribution of each subsidiary component can then be defined.

The general logic underlying this treatment is that the selective consequences of modifying a particular genic feature (e.g., number and size of introns, expression level, amino acid use, gene-copy number, etc.) is a function of the degree to which the overall energy budget is altered. Based on its phenotypic manifestations, a gene may have a multiplicity of advantages, but the energetic cost of replication, maintenance, and expression represents a minimum burden that must be overcome to achieve a net selective advantage. If a genic variant or a novel gene is to be efficiently promoted by natural selection, the net selective advantage (beyond the energetic cost) must exceed the power of drift (defined as $1/N_e$ for a haploid organism, where N_e is the effective population size) (2, 8).

The Lifetime Energy Requirement of a Cell

Although a common route to estimating the total energy demand per cell cycle is to simply take the product of the metabolic rate per unit time and the cell-division time, this fails to distinguish the requirements for cellular reproduction from those associated with basal (non-growth-related) maintenance. By basal maintenance, we refer not to the energetic requirements of dormant cells, but to an entire menu of cellular transactions in actively growing cells that do not directly lead to growth of cell parts, including metabolic reactions, cell movement, intracellular transport, maintenance of membrane potentials, turnover of biomolecules, and so on. Whereas the production costs of the components of new daughter cells should be relatively constant per cell cycle for nondormant cells, maintenance costs cumulatively increase with cell-division time.

These two contributions can be separated (and converted to ATP equivalents) by growing cells in defined environments with an energy-limiting resource in a continuous-flow chemostat. The rate of resource consumption per cell is determined from the difference in resource concentration between the inflow and outflow and the known cell density (which reaches an equilibrium in the growth chamber). The yield of ATP per unit resource consumption is obtained from knowledge of the metabolic pathways through which the substrate passes. If the rate of resource consumption is determined at several cell-division rates (equivalent to the dilution rate of the chemostat), a plot of the former vs. the latter is expected to yield a straight line, with the slope equaling the amount of resource consumed to produce a new cell and the intercept (denoting the point at which there are no excess resources for growth) providing a measure of the baseline metabolic rate (9–12).

Significance

A long-standing mystery in evolutionary genomics concerns the lineage-specific expansions of genome size in eukaryotes relative to prokaryotes. One argument is that the cellular complexity and elevated gene numbers in eukaryotes were impossible without a mitochondrion. However, the energetic burden of a gene is typically no greater, and generally becomes progressively smaller, in larger cells in both bacteria and eukaryotes, and this is true for costs measured at the DNA, RNA, and protein levels. These results eliminate the need to invoke an energetics barrier to genome complexity.

Author contributions: M.L. and G.K.M. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: milynch@indiana.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1514974112/-/DCSupplemental.

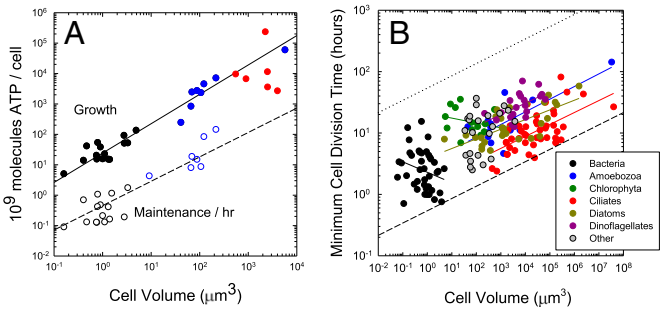

Fig. 1. (*A*) The costs associated with maintaining and producing a cell for a variety of bacteria (black) and unicellular eukaryotes (blue). The red points, which denote data for cultures of cells from multicellular species, are included for comparative purposes but were not used in the regressions. (*B*) Minimum cell-division times for unicellular species, normalized to 20 °C, with significant regression lines shown for individual phylogenetic groups. The upper dotted line denotes cell-division times that are expected to result in 50% of the cellular energy budget being allocated to maintenance; the dashed line demarcates the apparent absolute lower bound to volume-specific cell-division times across the tree of life. Data sources are provided in the tables in *SI Appendix*.

This general approach has been applied to enough organisms to reveal two generalizations. First, basal metabolic rate (here, normalized to a constant temperature of 20° C for all species) scales with cell volume (Fig. 1*A*) with a power-law relationship of

$$C_M = 0.39V^{0.88}, \quad \textbf{[1a]}$$

where C_M is in units of 10^9 molecules of ATP per cell per hour, and cell volume V is in units of cubic micrometers (*SI Appendix*). Because the SE of the allometric coefficient is 0.07, this relationship does not deviate significantly from linearity ($r^2 = 0.88$). Although the ranges of data are distinct, there is no discontinuity in the pattern across the prokaryotic–eukaryotic divide.

Second, the growth requirements per cell scale with cell volume with the power-law relationship

$$C_G = 26.92V^{0.97}, \quad \textbf{[1b]}$$

where C_G is in units of 10^9 molecules of ATP per cell (Fig. 1*A*). Because the SE of the allometric coefficient is 0.04, this highly significant regression ($r^2 = 0.96$) also implies an effectively linear relationship between the energetic requirements for growth and cell volume. Again, the relationship seems continuous over four orders of magnitude of cell size, spanning bacteria and eukaryotes, so there is no justification for invoking different metabolic scalings between these two groups. Estimates of the growth requirements for mammalian and land-plant cells also follow the pattern just noted, but the data are not applied to the regressions because of concerns with artifacts involving cells experiencing out-of-body metabolism (13).

The total cost of building a cell is

$$C_T \simeq C_G + tC_M, \quad \textbf{[1c]}$$

where t is the cell-division time in hours. Substituting Eqs. **1a** and **1b** and using an average exponent of 0.91 shows that provided t is smaller than $\sim 69V^{0.09}$ hours (assuming 20° C), the contribution from cell growth exceeds that associated with maintenance. Because minimum cell-division times for unicellular species are one to two orders of magnitude below this benchmark (Fig. 1*B*), setting $t = 0$ in Eq. **1c** provides a close approximation of total cellular ATP requirements at maximum growth rates.

Gene Structural Costs

The total cost of an individual gene involves three levels of investment: replication and chromosome maintenance; transcription and transcript processing; and translation and protein assembly. Each of these layers comprises several subcategories, which we have attempted to rank-order in terms of energetic requirements. For some subsidiary components, assumptions need to be made about the underlying biochemistry, but the processes are well enough understood to achieve approximations sufficient for the following analyses. Given the shortage of information on archaebacteria, attention is confined to bacterial and eukaryotic cells. Throughout, the costs inferred are in terms of numbers of phosphate bonds hydrolyzed (denoted below as P), with functions paid in units of GTP being treated as equivalent to ATP. We first describe the general features of the model and then apply it to existing data to obtain direct quantitative insight. All details can be found in *SI Appendix*.

Chromosome Level. During the lifetime of a cell, both strands of DNA must be replicated once per cell cycle, and we start with the assumption that this involves de novo synthesis of the requisite nucleotides. Nucleotide recycling can occur within cells, but permanent sequestration to a new genome ultimately requires the acquisition of new dNTPs. The cost of nucleotide synthesis includes the cost of synthesizing the intermediate metabolites (e.g., phosphoribosyl pyrophosphate and amino acids) used to make purines and pyrimidines, because the energy invested in such subunits would otherwise be available for alternative cellular functions. There is only slight variation in the biosynthetic costs of the four nucleotides, each being ~50 ATPs per nucleotide, so the biosynthetic cost of replicating a span of L_g nucleotides is $\sim 2 \cdot 50 \cdot L_g = 100L_g$ P.

Additional costs at the DNA level are small relative to nucleotide synthesis. Although chain polymerization involves the loss of a diphosphate for each base extension, this has already been incorporated into the cost of dNTPs. Unwinding of the parental double helix requires $\sim L_g$ P per gene, and the summed cost associated with the RNA primers used for replicate-strand extension and the ligation of Okazaki fragments is $\sim 0.3L_g$ and $\sim 3L_g$ P per gene in bacteria and eukaryotes, respectively. Costs associated with opening of origins of replication, clamp loading, proofreading, and DNA repair are likely to be an order of magnitude or so smaller than those just noted and can be ignored for purposes herein. (The basis for this and all other conclusions on DNA-level costs are elaborated on in *SI Appendix*.)

The highly ordered, dense nucleosome packaging of DNA presents a substantial chromosome-level cost specific to eukaryotes, although some nucleoid-associated proteins exist and must entail a low level of cost in some bacteria (14). Eukaryotic nucleosomes contain two heterotetrameric histone complexes followed by a linker histone. Throughout eukaryotes, each nucleosome wraps 147 bp, and with an average linker length between nucleosomes of 33 bp, there is on average one nucleosome per 180-bp interval. Weighting by the cost of synthesizing the amino acids that comprise histone proteins and the cost of translating such proteins, the total nucleosome-associated cost is $\sim 160L_g$ P. The range of variation for this cost among eukaryotes is of minor significance for the types of issues being evaluated here (*SI Appendix*, Supplementary Table 2).

Taking all of the above into consideration, in units of ATP hydrolyses, we estimate the summed replication-associated costs of a bacterial gene to be

$$C_{\mathrm{DNA},b} \simeq 101L_g, \quad \textbf{[2a]}$$

whereas for a haploid eukaryote

$$C_{\mathrm{DNA},h} \simeq 263L_g. \quad \textbf{[2b]}$$

Doubling the preceding cost for a diploid eukaryote yields

$$C_{\mathrm{DNA},d} \simeq 526L_g, \quad \textbf{[2c]}$$

or ~5.2 times the cost of a prokaryotic gene of equivalent length.

These results provide a quantitative basis for understanding the evolutionary maintenance of highly streamlined bacterial genomes, which typically have <5% intergenic DNA and generally few (if any) introns or mobile elements (15), in contrast to the bloated genomes of multicellular species, which typically contain <5% coding DNA and harbor massive numbers of large introns and mobile elements (16). Because replication is essentially a one-time investment in the life of a cell, the maximum fractional contribution of the DNA-level cost of a gene to the total lifetime energy requirements occurs at minimum cell-division times (because a longer cell-division time increases the total basal metabolic requirement). Thus, the maximum proportional cost of DNA can be determined by scaling the preceding expressions against Eq. **1c** with $t = 0$.

Using this approximation, a bacterial cell with a representative volume of 1 μm^3 will have a replication-associated cost of DNA $\simeq (4 \times 10^{-9}) L_g$, so the fractional drain on the total cellular energy budget can be as high as (4×10^{-8}) for a small 10-bp insertion and (4×10^{-6}) for a gene-sized insertion of 1,000 bp. To put this into perspective, free-living bacteria typically have effective population sizes $>10^8$, often in the range of 10^9 to 10^{10} (2, 16). Thus, when growing at maximum rates, bacteria experience efficient enough selection to remove insertions as small as 10 bp (and even smaller when $N_e > 10^8$).

In contrast, for a unicellular eukaryote with a moderate-sized 100 μm^3 cell containing a haploid genome (e.g., a yeast), the fractional cost of DNA is $\simeq 10^{-10} L_g$, yielding relative chromosome-level costs of 10^{-9} and 10^{-7} for 10- and 1,000-bp segments of DNA, respectively. Unicellular eukaryotes often have $N_e < 10^8$, sometimes ranging down to 10^6, implying that insertions of small to moderate size will frequently be unmovable by natural selection. For a larger cell size of 2,500 μm^3, more typical of a multicellular eukaryote, and a diploid genome, the relative cost of DNA declines to $\simeq 10^{-11} L_g$, so even a 10^5-bp segment of DNA has a relative cost of just 10^{-6}. The effective population sizes of invertebrates tend to be in the neighborhood of 10^6, with that of some vertebrates (including humans) ranging down to 10^4, and in such cases the power of random genetic drift is sufficient to overwhelm the ability of natural selection to eliminate quite large insertions on the basis of DNA-level costs.

Transcript Level. Unlike replication, which involves a single investment per cell division, the total cost of transcription depends on the lifespan of a cell, because transcripts are typically degraded and replaced on time scales shorter than cell-division times. Several transcription-associated costs are general across bacteria and eukaryotes. The primary investment is the synthesis of ribonucleotides, which requires one less step than that for deoxyribonucleotide synthesis, leading to an average cost of ~ 46 P per ribonucleotide monophosphate. Assuming that ribonucleotides are efficiently recycled (meaning that, because nuclease activity does not consume ATP, the cost associated with turnover consists of the two phosphates needed to recharge them; discussed below), the total number of de novo nucleotide syntheses associated with a particular gene is a function of the average steady-state number of mRNAs (N_r) and the length of the mature mRNA ($L_{r,mat}$) (including the polyadenylation tail, which when present often has a length on the order of ~100 nucleotides), yielding a total cost of biosynthesis of $46 \cdot N_r L_{r,mat}$ P.

The second major cost is associated with the turnover of transcripts within the lifespan of the cell. The number of transcription cycles is given by the product of the transcription rate R_r and the cell-division time t. Because two high-energy phosphates must be expended for each chain-elongation step, the total investment in transcript turnover is therefore $2 \cdot R_r t L_{r,pre}$ P, where $L_{r,pre}$ is the length of the precursor mRNA. The latter includes introns, the difference between the initial (~250 nucleotides) and mature (~100 nucleotides) length of the poly(A) tail, and any transcriptional read-through past the termination point, which can also be a few hundred nucleotides in length in some eukaryotes. Note that it is assumed here that introns and other intermediate transcription products are rapidly and efficiently recycled. At steady state, $R_r = \delta_r N_r$, where δ_r is the degradation rate, so the cost of transcript replacement becomes $2 \cdot N_r \delta_r t L_{r,pre}$ P. There are additional costs associated with the activation and initiation of transcription, but these are small enough to be ignored for purposes herein (*SI Appendix*).

Summing the expenditures on ribonucleotide synthesis and chain elongation, the cost of transcribing a bacterial gene is

$$C_{\mathrm{RNA},b} \simeq 2 N_r L_g (23 + \delta_r t), \qquad \textbf{[3a]}$$

because in such species $L_g \simeq L_{r,pre} \simeq L_{r,mat}$. Comparison with Eqs. **2a–2c** shows that in bacteria, the cost of transcription exceeds that for replication provided the steady-state number of transcripts $N_r > 2$.

The cost of transcription is more involved in eukaryotes, because there is a significant difference between the lengths of premature and mature mRNAs owing to the presence of introns and poly(A) tails and other aspects of processing, and additional energy is expended on posttranscriptional modifications of nucleosomes and the RNA polymerase itself during transcriptional cycles. The cost of the latter likely sums to $<0.17 N_r L_{r,pre} \delta_r t$ (*SI Appendix*), and there are also relatively minor costs associated with splicing, 5′ capping and mRNA export. Summing the predominant components, the total cost associated with transcription of a eukaryotic gene is

$$C_{\mathrm{RNA},e} \simeq N_r \left(46 \cdot L_{r,mat} + 2.17 \cdot \delta_r t L_{r,pre}\right). \qquad \textbf{[3b]}$$

Note that in Eqs. **3a** and **3b** the total cost associated with transcription is subdivided into two components, the first defining the baseline requirement for building a cell, and the second being a function of the cell-division time.

Single-cell assays provide quantitative insight into the parameters in these formulations. First, although data are available for only four species, power-law relationships describe the relationships between total numbers of mRNAs per cell and mean and median numbers of transcripts/active genes and cell volume (Fig. 2). Second, results from several bacterial species imply average mRNA decay rates on the order of 10 per hour, whereas those in yeast and mammalian cells are on the order of 4 and 0.1 per hour, respectively (details are given in *SI Appendix*).

With these formulations and estimates of background variables in mind, the cost of a gene at the transcriptional level can be determined. Consider, for example, a typical bacterial gene with a 1.0-kb transcript (17). At 20° C, the minimum division time for bacterial cells is ~0.7 h (Fig. 1*B*), so the minimum cost of transcription is $\sim 2{,}000 N_r (23 + 0.7\delta_r)$, which with a decay rate of 10 per hour, becomes $\sim (6 \times 10^4)$ P for a lowly expressed gene with $N_r = 1$ and 100 times higher for a gene with $N_r = 100$. With the lifetime growth requirements of bacterial cells being generally in the range of 10^{10} to 10^{11} P, this shows that the fractional cost of even a weakly expressed bacterial gene, $\sim (6 \times 10^{-6})$ under high-growth-rate conditions, is sufficiently large to be opposed by selection in bacteria with $N_e > 10^8$ if not offset by substantial fitness advantages.

Now consider a typical intron-free eukaryotic gene. The average lengths of eukaryotic proteins are ~50% longer than those in bacteria (17), and eukaryotic transcripts harbor 5′ and 3′ UTRs with average respective lengths of ~150 and ~350 bp (2), so a 2-kb transcript length is a reasonable baseline value. After also accounting for polyadenylation, letting $\delta_r = 3$ per hour, and assuming a minimum cell division time of $t = 4$ h, we arrive at an approximate cost of $(1.5 \times 10^5) N_r$ P, which is ~2.5 times the minimum cost of transcription in bacteria.

Most eukaryotic genes contain introns, commonly with more than five per gene, and in multicellular species often averaging more than 1 kb in length. With a total intron length of 5 kb/gene, $L_{r,pre}$ increases to 7,000, and the transcription-associated cost

rises to $\sim(2.8\times10^5)N_r$ P, showing that intron addition can substantially increase the transcriptional cost of a gene.

Most cells only rarely reside in environments that permit maximum growth rates, often experiencing strongly nutrient-limiting conditions. It is therefore relevant to consider the limits of RNA-level costs as cell-division time $t\rightarrow\infty$, which approach $\sim 2N_rL_{r,pre}\delta_r t$ in both bacteria and eukaryotes. To understand what this means for cells of different size, the key issue is how the composite parameter $N_r\delta_r$, the rate of mRNA turnover, scales relative to $V^{0.88}$, the scaling of metabolic rate with cell volume. Given the estimates of N_r and δ_r noted above, the product $N_r\delta_r$ seems to decline with cell volume, implying that transcription-related energetic costs of a gene relative to the total cellular energy budget decline with increasing cell volume under nutrient-limiting conditions, and perhaps strongly so.

Protein Level. As in the case of replication and transcription, several subcategories of costs are associated with protein production and management, although the overwhelming contributions per cellular lifetime are associated with three functions. First, the cost associated with replacement of the standing level of protein is $N_pL_p\bar{c}_{AA}$ P, where N_p is the steady-state number of protein molecules derived from the gene per cell, L_p is the number of amino acids per protein, and $\bar{c}_{AA}$ is the average cost of synthesis per amino acid residue in the protein (a function of a protein's amino acid content). Second, the cost associated with chain elongation is $4N_pL_p\delta_p t$, where δ_p is the rate of protein decay. Third, degradation of proteins imposes an approximate cost of $N_pL_p(\delta_p t-1)$ P (*SI Appendix*). Additional costs small enough to be ignored are associated with translation initiation and termination, posttranslational modifications, ribosomal proofreading, and protein folding. Summing up, the total protein-level cost of a gene in both bacteria and eukaryotes is

$$C_{PRO}\simeq N_pL_p\left[(\bar{c}_{AA}-1)+5\delta_p t\right], \qquad [4]$$

where again the first term represents a one-time cost incurred regardless of the length of the cell cycle, and the second term represents the cumulative cost resulting from protein turnover and replacement.

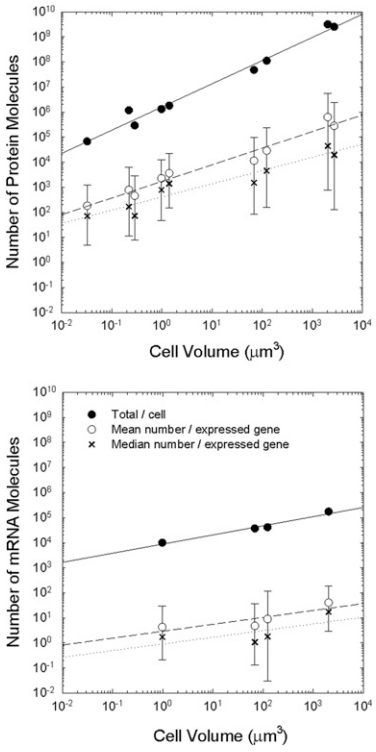

Fig. 2. Numbers of protein and messenger RNA molecules per cell, with the five left-most points being for bacterial species, the intermediate two for yeasts, and the two right-most points for mammalian cells. Total numbers of molecules per cell (summed over all genes) are given by the closed points, with the solid-line regression. The brackets for numbers of molecules per gene denote the lower 2.5% and upper 97.5% cutoffs in the overall distributions; the dashed and dotted lines denote the regressions involving the means and medians. For transcripts, the total number per cell and the average number per active gene scale with cell volume (V) as $8{,}831V^{0.36}$ ($r^2=0.97$) and $2.93V^{0.28}$ ($r^2=0.71$), respectively. For proteins, the total number per cell and the average number per active gene scale with cell volume (V) as $1{,}588{,}547V^{0.93}$ ($r^2=0.98$) and $1{,}698V^{0.66}$ ($r^2=0.96$), respectively.

Observations from high-throughput proteomics provide insight into the key parameters appearing in the preceding formulation. First, results from nine species indicate that cellular abundances of proteins (N_p) are much higher than those for their cognate mRNAs and are adequately represented by power-law relationships with cell volume at both the level of total proteins per cell and average number of molecules per active gene (Fig. 2). The exponents of the power functions are nearly three times greater than those for mRNAs, implying an elevated investment in protein number per unit volume. Second, decay rates of proteins are typically much lower than those of cognate mRNAs. Protein decay rates are generally in the range of 0.1–0.9 per hour in bacteria, on the order of 1.0 per hour in yeast, and in the range of 0.1–1.0 per hour in eukaryotes (*SI Appendix*).

Comparison of Eq. **4** with Eqs. **3a** and **3b** indicates that the protein-level cost of a gene is much greater than that at the level of RNA (see also *SI Appendix*, Supplementary Figs. 1–4). Considering, for example, the high-growth rate limit ($t\rightarrow0$), as a first-order approximation with $\bar{c}_{AA}\simeq25$ P, the ratio of protein- to RNA-level costs is $\sim0.5N_pL_p/(N_rL_r)$. Because N_p/N_r is on the order of 1,000, and L_p/L_r is on the order of 0.25, the cost at the protein level is ~125 times that at the level of RNA at the high-growth-rate limit. At the low-growth-rate limit, $t\rightarrow\infty$, this ratio is even higher, on the order of 600.

Considering a protein of moderate length, $L_p=300$, and a moderate cellular abundance of $N_p=25{,}000$ for a eukaryotic cell, at the high-growth-rate limit, the protein-level cost relative to the total cellular budget is $\sim(8\times10^{-5})$ for a cell with volume 100 μm^3 and $\sim(3\times10^{-6})$ for a cell with volume 2,500 μm^3. Given the effective population sizes noted above, these results suggest that genes that are moderately to highly translated in eukaryotic cells can sometimes impose a high enough energetic burden to be opposed by selection if they do not confer sufficient added benefits.

Again, the question as to whether the relative protein-level cost of a gene increases with cell size can be evaluated by comparing Eq. **4** with Eqs. **1a** and **1b**. Depending on the cell-division time, the total cellular energy budget scales with cell volume with exponent 0.88 (slow growth) to 0.97 (maximum growth rate). Because the average length of a protein does not vary greatly among phylogenetic lineages (17), and the results cited above suggest that both N_p and δ_p scale with cell volume with an exponent $\ll1$, it seems that the relative protein-level cost of a gene also declines with cell size.

Empirical Estimates. For four species, sufficient data exist to estimate the three classes of contributions to the total energetic costs of the full sets of protein-coding genes. In all four taxa, the DNA associated with each gene was characterized from the annotated genomes, which also provided information on intron number and size, and lengths of mature transcripts and proteins. Total numbers of mRNAs and proteins per cell were estimated from the empirically determined power-function relationships in Fig. 2 and/or primary data. It was then possible to use publicly available transcriptomic and ribosomal-profiling data (under the assumption that the translatome is an accurate representation of the proteome) to partition the total numbers of molecules into steady-state gene-specific numbers. Based on the available data (*SI Appendix*), mRNA decay rates were assumed to be 10 per hour in bacteria and 4 per hour in eukaryotes, and a protein decay rate

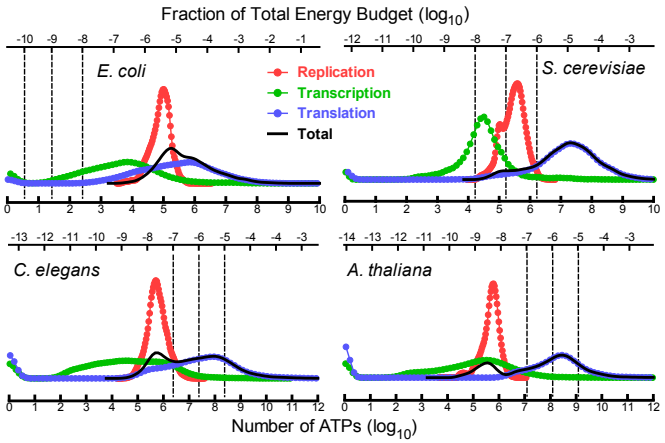

Fig. 3. Distribution of energy costs for the full sets of annotated genes in one bacterium (*E. coli*) and four eukaryotic species (*Saccharomyces cerevisiae, C. elegans, and A. thaliana*). The bottom axis shows the absolute costs in ATP units, and the upper axis shows the corresponding costs as the fraction of the cell's lifetime energy budget. The dashed vertical lines denote key positions below which the energy cost is expected to be too low to be opposed by selection (in the absence of any additional advantages for the gene); for genes to the left of a particular vertical bar (with logarithmic value x on the upper axis), the energetic cost would be effectively neutral if the effective population size (N_e) were $>10^{-x}$. The three vertical lines in each plot provide the approximate range in which N_e is likely to reside for species in the same broad taxonomic categories as the characterized species (2).

of 0.1 per hour was assumed throughout. Cell-division times were taken to be the minima observed in each species, standardized to 20° C; all other details are outlined in *SI Appendix*. In the following, we will refer to the fractional costs of genes (relative to the total cellular energy budget) at the genome, transcript, protein, and cumulative cost levels as s_{DNA}, s_{RNA}, s_{PRO}, and s_c.

For the bacterium *Escherichia coli*, s_c falls in the range of 10^{-7} to 10^{-3} for almost all genes (with absolute costs of 10^3 to 10^8 P), far above the likely minimum values that can be perceived by selection in this large-N_e species (Fig. 3). If such genes were to find themselves in an environment where their functions were no longer useful, inactivating mutations would be strongly selected for. Within eukaryotes, small peaks of lowly expressed (and perhaps misannotated) genes exist with roughly the same absolute costs of *E. coli* genes. However, most eukaryotic genes have total absolute costs exceeding 10^6 P, with substantial fractions in multicellular species falling in the range $>10^8$ P. For many genes in yeast and a substantial fraction in *Caenorhabditis elegans* and *Arabidopsis thaliana*, s_c is sufficiently large for a gene to be opposed by selection if it had few added benefits (e.g., a redundant gene duplicate). The major contribution that pushes s_c past the drift barrier in eukaryotes is the cost of translation. Most values of s_{DNA} and s_{RNA} in multicellular species are below the threshold for efficient selection.

For an additional 31 bacterial and 13 unicellular eukaryotes with annotated genomes, sufficient data exist on cell volumes and cell-division times to estimate lifetime energy requirements using Eqs. **1a–1c**. In the absence of direct information on the single-cell concentrations of gene-specific mRNAs and proteins in this subset of species, we used existing gene annotations and genome sequences to compute species-specific costs of an average gene, relying on the expected average numbers of mRNA and protein molecules per gene extrapolated from the functions in Fig. 2 and the decay rates noted above.

Three general conclusions can be drawn from this extended dataset. First, there is a consistent ranking of $s_{DNA} < s_{RNA} < s_{PRO}$, with a one to two order of magnitude increase from the former to the latter (Fig. 4). Second, average estimates of all three cost measures in bacteria are generally substantially greater than those in eukaryotes, and in most cases are likely large enough to be opposed by selection. For eukaryotes, the chromosome-level costs are generally too low to be detectable by selection, and this is also true in many cases for the average transcription-level costs, confirming the suggestion that gene-sized insertions in large eukaryotes will typically be effectively neutral from a bioenergetic perspective unless they are translated. Third, within both bacteria and eukaryotes, there is a substantial negative scaling of all three levels of cost with cell volume. Although there is almost no overlap in cell volumes between these two groups, it is clear that there is continuity in the scaling of data across groups. These general conclusions still hold for cells growing substantially below maximum rates (*SI Appendix*).

Discussion

A central goal of evolutionary genomics is to understand the mechanisms responsible for the massive expansion in genome size and gene structural complexity from prokaryotes to unicellular eukaryotes to multicellular species, most of which results from the colonization of noncoding DNA. Most exogenous DNA is hazardous in the sense that it increases the ways in which an associated gene can be rendered nonfunctional by mutation (e.g., by altering gene regulation, intron splicing, and/or translation-initiation sites) (2). However, all genes also impose a baseline energetic cost on a cell via the demands at the DNA, RNA, and protein levels (6, 7). These costs are relevant because the long-term preservation of a gene by natural selection requires that its phenotypic benefits exceed the energetic costs to a large enough extent to offset the power of random genetic drift.

Letting s_a be the adaptive advantage of a stretch of DNA, and s_c be the loss of fitness induced by the total energetic costs, the net selective advantage of a DNA segment is

$$s_n = s_a - s_c. \quad [5]$$

Basic evolutionary theory indicates that the absolute value of the net advantage s_n must be greater than $1/N_e$ in a haploid species [and $1/(2N_e)$ in a diploid] to be readily perceived by natural selection. If, for example, the insertion of a segment of DNA provides no immediate phenotypic advantage nor any significant mutational disadvantage, so $s_a \simeq 0$, it will nevertheless be essentially

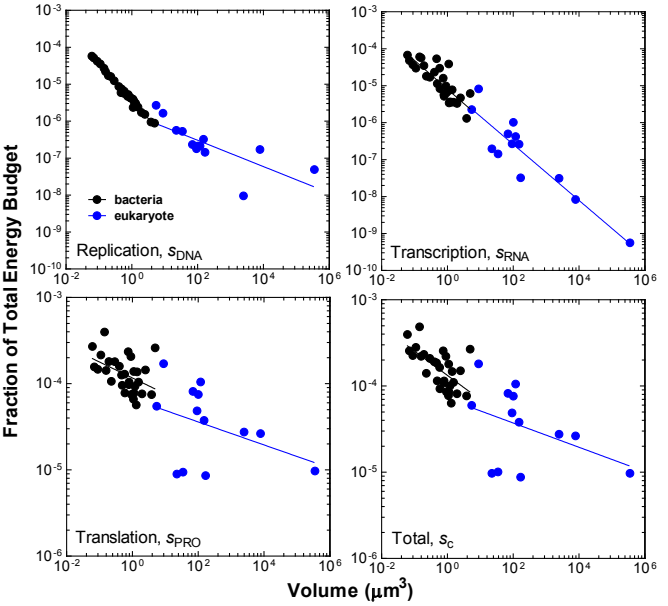

Fig. 4. Fractional costs of average genes in bacteria and unicellular eukaryotes (relative to total cellular energy budgets), subdivided into components at the level of replication, transcription, and translation.

immune to removal from a haploid population by selection, and hence subject to effectively neutral drift and fixation processes, if $s_c < 1/N_e$.

The preceding results indicate that the energetic cost of replicating a DNA segment of even just a few nucleotides (even if nontranscribed) can be sufficient to be perceived by selection in a typical bacterial population with large N_e. In contrast, insertions of even many kilobases often impose a small enough energetic burden relative to the overall requirements of eukaryotic cells to be immune to selection. Although RNA-level costs are frequently greater than those at the DNA level, these are often still not large enough to overcome the power of random genetic drift in eukaryotic cells. This means that many nonfunctional DNAs that are inadvertently, even if specifically, transcribed in eukaryotes (especially in multicellular species) cannot be opposed by selection, a consideration relevant to the debate as to whether transcriptional activity is an indicator of functional significance (see the exchange between refs. 18–20). However, with the cost at the protein level generally being much greater than that at the RNA level, segments of DNA that are translated can sometimes impose a large enough energetic cost to be susceptible to selection, even in multicellular species. This may explain why redundant duplicate genes commonly experience high rates of nonfunctionalization (21).

Lane and Martin (5) have argued that an enhanced ability to generate energy made possible by the origin of the mitochondrion was a prerequisite for the evolution of a vast array of features often associated with complexity in eukaryotic cells, including increased gene number, protein length, protein folds, protein–protein interactions, and regulatory elements. Given the singularity of the mitochondrial colonization event, this idea is not subject to conventional statistical analysis, and we do not provide an evolutionary argument for why the mitochondrion arose.

However, we can now say that in neither complex eukaryotic cells nor in morphologically simpler bacteria does increased cell size induce a condition in which gene addition becomes an increasing energetic burden. In fact, from an evolutionary perspective, the evolution of increased cell size has the opposite effect. Although the absolute cost of a gene does increase with cell size, in terms of the fractional contribution to a cell's energy budget, which ultimately determines whether selection can oppose genome expansion, the cost of an average gene decreases at the DNA, RNA, and protein levels. Even if the relative costs were to remain constant, this cannot compensate for the reduction in effective population sizes in eukaryotes relative to bacteria, which reduces the efficiency of selection. Moreover, because the relative cost of a gene declines with increasing cell-division times (*SI Appendix*), organisms that are more resource-limited experience still weaker selection against inadvertent genome-size expansion.

Taken together, our observations suggest that an energetic boost associated with the emergence of the mitochondrion was not a precondition for eukaryotic genome expansion. First, the absence of a dichotomous break in the relationship between lifetime cellular ATP requirements and cell volume between bacteria and eukaryotes (Fig. 1*A*) is inconsistent with the idea that cells with greater internal complexity impose greater energy demands. Second, two of the central costs of a gene, the steady-state numbers of mRNA and protein molecules maintained, scale sublinearly with cell volume (Fig. 2), and again in a continuous fashion within and between bacteria and eukaryotes. Third, within bacteria alone, although larger cells have higher energetic requirements per cell lifetime (Fig. 1*A*), species with larger cell sizes have reduced cell-division times, implying a higher efficiency of energy conversion despite having larger genome sizes (Fig. 1*B*). Thus, population-genetic arguments based on both the mutational-hazard hypothesis (2) and on the observed features of cellular energetics lead to the conclusion that passive increases in genome size are expected to naturally arise in organisms with increased cell sizes (which, by correlation, have reduced effective population sizes). This supports the view that variation in the power of random genetic drift has played a central role in the historical diversification of genome and possibly cellular architecture across the tree of life.

Materials and Methods

This work is based on data derived from the literature and various existing databases, the details of which are outlined in *SI Appendix*. Briefly, estimates of cell growth and maintenance requirements were derived from chemostat studies involving growth of microbes on defined media with known rates of conversion into ATP. Cell division rates were derived from an extensive literature survey. Expressions for the various costs of a gene were obtained from the basic biochemistry and molecular-biological literature on the underlying processes, and these were converted into absolute values for particular species and genes using genomic, transcriptomic, and proteomic databases.

ACKNOWLEDGMENTS. Support was provided by Multidisciplinary University Research Initiative Awards W911NF-09-1-0444 and W911NF-14-1-0411, the US Army Research Office, National Institutes of Health Award R01-GM036827, and National Science Foundation Award MCB-1050161. This material is also based upon work supported by National Science Foundation Grant CNS-0521433.

1. Lynch M (2007a) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA* 104(Suppl 1):8597–8604.
2. Lynch M (2007b) *The Origins of Genome Architecture* (Sinauer, Sunderland, MA).
3. Lukeš J, Archibald JM, Keeling PJ, Doolittle WF, Gray MW (2011) How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life* 63(7):528–537.
4. Booth A, Doolittle WF (2015) Eukaryogenesis, how special really? *Proc Natl Acad Sci USA* 112(33):10278–10285.
5. Lane N, Martin W (2010) The energetics of genome complexity. *Nature* 467(7318): 929–934.
6. Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli and Bacillus subtilis*. *Proc Natl Acad Sci USA* 99(6): 3695–3700.
7. Wagner A (2005) Energy constraints on the evolution of gene expression. *Mol Biol Evol* 22(6):1365–1374.
8. Kimura M (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ Press, Cambridge, UK).
9. Bauchop T, Elsden SR (1960) The growth of micro-organisms in relation to their energy supply. *J Gen Microbiol* 23:457–469.
10. Pirt SJ (1982) Maintenance energy: A general model for energy-limited and energy-sufficient growth. *Arch Microbiol* 133(4):300–302.
11. Tempest DW, Neijssel OM (1984) The status of YATP and maintenance energy as biologically interpretable phenomena. *Annu Rev Microbiol* 38:459–486.
12. Russell JB, Cook GM (1995) Energetics of bacterial growth: Balance of anabolic and catabolic reactions. *Microbiol Rev* 59(1):48–62.
13. Glazier DS (2015) Body-mass scaling of metabolic rate: What are the relative roles of cellular versus systemic effects? *Biology (Basel)* 4(1):187–199.
14. Dillon SC, Dorman CJ (2010) Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat Rev Microbiol* 8(3):185–195.
15. Lynch M (2006) Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* 60:327–349.
16. Lynch M, Bobay LM, Catania F, Gout JF, Rho M (2011) The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genomics Hum Genet* 12:347–366.
17. Wang M, Kurland CG, Caetano-Anollés G (2011) Reductive evolution of proteomes and protein structures. *Proc Natl Acad Sci USA* 108(29):11954–11958.
18. Doolittle WF (2013) Is junk DNA bunk? A critique of ENCODE. *Proc Natl Acad Sci USA* 110(14):5294–5300.
19. Graur D, et al. (2013) On the immortality of television sets: "Function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biol Evol* 5(3): 578–590.
20. Kellis M, et al. (2014) Defining functional DNA elements in the human genome. *Proc Natl Acad Sci USA* 111(17):6131–6138.
21. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290(5494):1151–1155.