

Eukaryotes evade information storage-replication rate trade-off with endosymbiont assistance leading to larger genomes

Parthasarathi Sahu, Sashikanta Barik, Koushik Ghosh, and Hemachander Subramanian*

Department of Physics, National Institute of Technology Durgapur, India

Abstract

Genome length varies widely among organisms, from compact genomes of prokaryotes to vast and complex genomes of eukaryotes. In this study, we theoretically identify the evolutionary pressures that may have driven this divergence in genome length. We use a parameter-free model to study genome length evolution under selection pressure to minimize replication time and maximize information storage capacity. We show that prokaryotes tend to reduce genome length, constrained by a single replication origin, while eukaryotes expand their genomes by incorporating multiple replication origins. We propose a connection between genome length and cellular energetics, suggesting that endosymbiotic organelles, mitochondria and chloroplasts, evolutionarily regulate the number of replication origins, thereby influencing genome length in eukaryotes. We show that the above two selection pressures also lead to strict equalization of the number of purines and their corresponding base-pairing pyrimidines within a single DNA strand, known as Chargaff's second parity rule, a hitherto unexplained observation in genomes of nearly all known species. This arises from the symmetrization of replicore length, another observation that has been shown to hold across species, which our model reproduces. The model also reproduces other experimentally observed phenomena, such as a general preference for deletions over insertions, and elongation and high variance of genome lengths under reduced selection pressure for replication rate, termed the C-value paradox. We highlight the possibility of regulation of the firing of latent replication origins in response to cues from the extracellular environment leading to the regulation of cell cycle rates in multicellular eukaryotes.

Keywords: *genome size evolution, prokaryote-eukaryote divergence, Chargaff's second parity rule, C-value paradox, endosymbiosis, latent origins, GC-Skew, origins of replication, replichores*

Significance Statement

Understanding the forces shaping genome architecture is a long-standing challenge in evolutionary biology. Our study demonstrates that the balance between replication speed and information storage, constrained by cellular energetics, drives the divergence in genome lengths between prokaryotes and eukaryotes. By quantifying selection pressure as the ratio of replication time to genomic information content, we show that this pressure enforces adaptive constraints, giving rise to features such as symmetric replichores and Chargaff's second parity rule. These insights not only help us resolve an enduring evolutionary puzzle, but also offer a unified framework linking genome organization, cellular specialization, and even potential mechanisms underlying carcinogenesis.

Introduction

Life on Earth began approximately 3.7 billion years ago and evolved from simpler forms to complex and diverse organisms observed today, shaped by various selection pressures [1, 2]. Organisms are broadly classified into two groups: prokaryotes and eukaryotes. Prokaryotes are characterized by simpler structures, whereas eukaryotes are generally more complex and have evolved from prokaryotes, with defining characteristics such as endosymbiotic relationships, nuclear membranes, huge variance in genome size, etc. [3, 4, 5]. Despite emerging earlier in Earth's history, prokaryotes maintain smaller genomes and show less morphological evolution compared to eukaryotes [6, 7, 8]. The constraints limiting the complex morphological evolution of prokaryotes are debated [9, 10, 11].

The tendency of prokaryotes to acquire compact genomes is extensively modeled, with models constructed to include impacts of population size, environmental per-

*Corresponding author: hsubramanian.phy@nitdgp.ac.in

turbations, and selection for metabolic efficiency under nutrient limitation [6, 12, 13, 14]. The evolutionary forces shaping eukaryotic genome length remain comparatively under-explored. In eukaryotes, current frameworks have largely focused on the impact of mutational mechanisms [15, 16] and energetic constraints on genome length [17].

Despite extensive studies, a simple explanation for such a dramatic divergence of genome length between prokaryotes and eukaryotes is lacking. In this study, we use a very simple, parameter-free model that incorporates the influence of two primary evolutionary forces: faster replication and enhanced information storage capacity, to study the evolution of genome length across these two domains of life. We show that the genome lengths of prokaryotes and eukaryotes diverge under the same selection pressure, if we restrict prokaryotes to have a single replication origin, while allowing eukaryotes to have multiple origins. We argue that this differentiation stems from access to the energy supply of mitochondria (or chloroplasts), as evidenced by the observed correlation between the number of mitochondria (or chloroplasts) and the length of the genome in eukaryotes. Surprisingly, the model also reproduces multiple other observations that hold for nearly all species, such as the equality of purines and pyrimidines on a single strand of DNA, called Chagraff’s second parity rule (PR-2), replicore length symmetrization (see below), a preference for deletions over insertion mutations, and a huge variance in the genome lengths seen in eukaryotes, called the C-Value paradox.

The Model

To study the effect of the aforementioned selection pressures on genome length, i.e. faster replication and higher information storage capacity, we utilize a model, where a pool of N identical sequences is evolved over m generations under these selection pressures. The initial sequence pool consists of N identical sequences, containing purines or pyrimidines, homogeneously across the sequence, with all purines on one strand and all pyrimidines on the complementary strand (e.g., 5'-RRRRRRR-3'/3'-YYYYYYY-5'). Each generation involves two major steps: (i) replication of N sequences and mutation of all daughter sequences in the pool and (ii) selection of half of the sequences in the pool based on their ability to satisfy the above two selection pressures. These two steps are repeated m times, and the time-evolution of the sequence length, averaged over all the N sequences, is recorded for every generation, for subsequent analysis.

Mutation: We implement large-scale genomic mutations through random deletions or duplications of regions in the daughter sequence that comprise 5% to 10% of the total genome length (Fig. 2). These large-scale mutations represent well-documented drivers of evolutionary processes [6, 18, 19, 20, 11]. The size and location of these mutations are chosen stochastically. During a mutation involving duplication, the duplicating fragment is randomly cho-

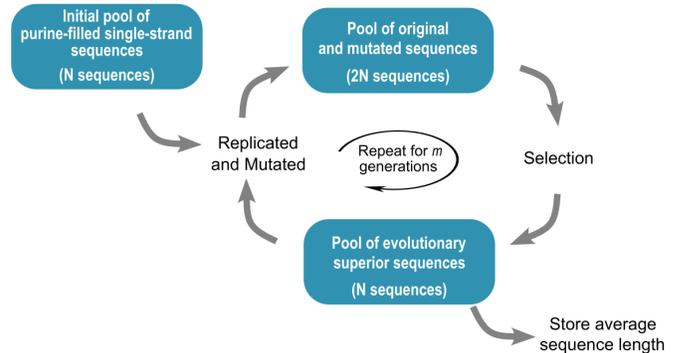


Figure 1: Algorithm of the model. Evaluation of the impact of the two selection pressures, fast replication and high information storage capacity, on the genome length of an organism. An initial pool of N sequences is evolved over m generations that involve two recurring steps: replication and mutation of all the sequences in the pool and applying selection pressure to extract the fittest sequences for the next generation. An initial pool of N identical sequences, composed of all purines or all pyrimidines, are replicated, and the daughter sequences are mutated, producing a pool of $2N$ sequences. Selection acts on this pool, removing N less-fit sequences that do not satisfy the selection pressure adequately. This replication-selection cycle is repeated m times, and the average genome length at every generation is recorded.

sen from either the original strand or the complementary strand. This mechanism ensures that over generations, each strand can have varying amounts of both purines and pyrimidines, even though the initial pool sequences are composed entirely of either purines or pyrimidines. In each generation, every sequence is replicated, and the replicated sequence in the pool undergoes a single mutation, i.e., a duplication or deletion. Following this, the sequence pool is expanded to include N replicated and mutated sequences along with the N unmutated sequences of the previous generation, resulting in a total of $2N$ sequences.

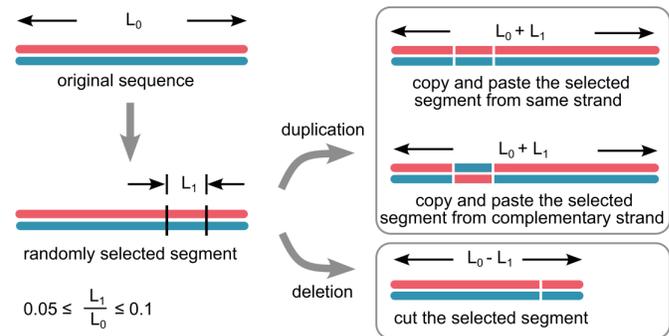


Figure 2: Mutation of a given genome. A DNA double strand, from the initial pool, with a homogeneous distribution of purines on one strand (red) and pyrimidines on its complement (blue). A mutation involves either a deletion or a duplication of a segment of a random length of 5-10% of the total genome length, chosen at a random location of the daughter genome. The mutation results in either a decrease or an increase in the length of the genome by 5-10%. The duplicated fragment can be from either strand, thereby altering the composition of purines/pyrimidines in each strand.

Selection: Following mutations, N sequences are selected from the pool of $2N$ for the next generation based on their ability to satisfy the selection pressure. The selection pressure is quantified through a factor γ , defined as

$$\gamma = \frac{\text{replication time}}{\text{information storage capacity}} \quad (1)$$

To quantify information storage capacity, we utilize the metric of total genome length, since it provides the upper limit for total information storage. Similarly, as a proxy for the replication time for the entire genome, we consider the length of the longest *replichore*, where, replichores are defined as disjoint segments of the genome that replicate independently of each other. This substitution is based on the following considerations: (a) Replichores replicate simultaneously and independently, parallelizing the replication process [21, 22]. Although the firing of multiple replication origins across the genome is generally not synchronized [23, 24], we assume ideal conditions, where origins fire and replication progress across each replichore independently and simultaneously. (b) For simplicity, it is assumed here that the replication speed of a replichore is constant throughout the length of the replichore, although the speed depends on the sequence and the availability of activated nucleotides, in general [25]. This simplifying assumption enables us to use the replichore length as a replacement for the replication time of the whole genome, in our evaluation of γ . Therefore, with the above assumptions, we redefine equation (1) as

$$\gamma = \frac{\text{length of longest replichore}}{\text{length of full genome}} \quad (2)$$

In the selection process, the factor γ is calculated for all the $2N$ (parent and mutated daughter) sequences in the pool, and the N sequences with the lowest values of γ are extracted as the fittest sequences and are carried forward to the next generation.

To identify replichores in a genome, we first locate replication origins and termini by analyzing base composition asymmetry or nucleotide skews, i.e., the excess of G over C and A over T on any given stretch of a single strand, across the genome. This local violation of Chagraff’s second parity rule [26] is regularly utilized to find the replication origins and termini, by equating the locations of peaks and valleys in the skew plot to the replication origins and termini. This is a widely used technique for *in-silico* prediction of replication origins since 1996 [22, 27, 28, 29, 30]. In this study, we have used the purine-pyrimidine (RY) cumulative skew, W , of the sequences [31, 32], where $W_{RY}(n)$ is defined as $W_{RY}(n) = \sum_{i=1}^n (\delta_{S(i),R} - \delta_{S(i),Y})$. Here, S is a genomic sequence of length N bp, composed of four nucleotides, classified into two groups: $R = \{G, A\}$, $Y = \{C, T\}$, and $n = 1 \dots N$. We have taken 1000 purine-filled single strands of length 1024 bp each as the initial sequence pool. These sequences are allowed to mutate, by accrual or deletion of genome fragments, leading to local distortions in the skew in each generation. In order to avoid identifying small-scale skew variations (peaks and

valleys smaller than a certain length scale) as origins or termini, since these are not identified as origins or termini by origin-finding algorithms [32, 33, 34, 35], we concentrate only on large-scale skew variations and ignore small-scale ones. We use wavelet transforms to filter out these origins and termini resulting from small-scale variations, by down-sampling the genome sequence of length N to a length of $N/2^w$, where w is the wavelet level. To ensure that mutating fragments are not smaller than the wavelet compression scale w and thus go unnoticed by the selection pressure, we choose a w such that the smallest mutating fragment is larger than the compression factor; i.e., we impose the condition $0.05N > 2^w$, where N is the full genome length. It should be emphasized that the qualitative divergence of prokaryotic and eukaryotic genome lengths does not depend on the wavelet level used in this down-sampling procedure. Once the origins and termini of the replication are identified, the lengths of the replichores were measured as the distances between neighboring origins and termini, the largest of which is chosen for the calculation of γ . A detailed description of the methodology for identifying replication origins and measuring replichore lengths is provided in the supplementary material.

In our model, we chose the population $N = 1000$, the initial sequence length of 1024 bp (L_0), the number of generations $m = 1000$, and used a 4-level wavelet transformation. We repeated the experiment 100 times to ensure statistical robustness.

Our model also includes an upper threshold for the number of replication origins allowed in a genome (Ori_{max}) to prevent an uncontrolled explosion in genome length. Genomes with a replication origin count greater than Ori_{max} are eliminated during the selection process. For prokaryotes, Ori_{max} is set to 1, while for eukaryotes, Ori_{max} is set to a value much greater than 1 (50 and 100). Although we use parameters such as Ori_{max} , wavelet levels, and mutation size 5% - 10% for computational convenience, the model itself is free of intrinsic parameters, and the observed divergence between the prokaryotic and eukaryotic genome length is completely insensitive to the parameters listed above.

Results

The variation in genome length (in bp) of prokaryotes and eukaryotes over 500 generations is shown in Fig. 3(a) and (b), respectively. The evolutionary minimization of the selection pressure γ , for prokaryotes and eukaryotes, is shown in Fig. 3(c) and (d).

We observe that, in the absence of any restrictions on the number of replication origins, the genome tends to increase in length indefinitely (Fig. 3b). However, taking into account the scarcity of monomer resources and energy required to replicate longer genomes, we restrict genomes to have a maximum of Ori_{max} origins. When Ori_{max} is set high ($\gg 1$), the average genome length is observed to increase over generations. On the other hand, when Ori_{max}

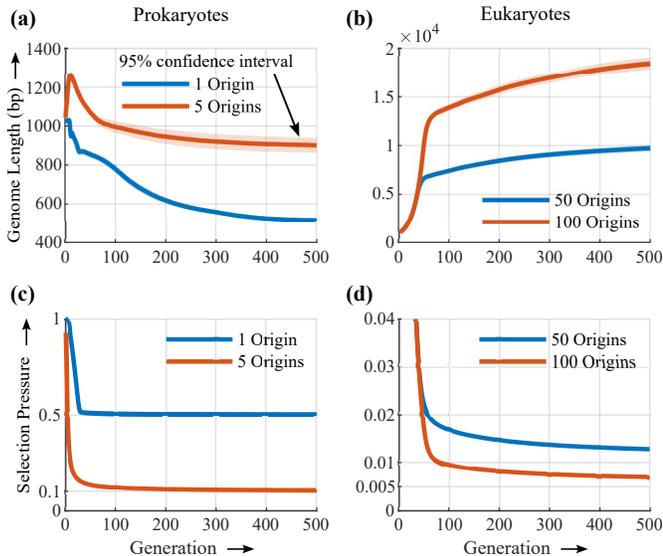


Figure 3: The evolution of genome length over generations. An initial population of 1000 purine-filled single-strand sequences is subjected to mutation and evolved under the selection pressure (γ) to minimize replication time while maximizing information storage capacity. The experiment is repeated 100 times, and the evolution of average genome length over 500 generations is shown, with a 95% confidence interval. (a) When each genome in the pool is constrained to have a single replication origin, mimicking prokaryotic genomes, the average genome length decreases over generations. Also shown is the genome length evolution when a few more origins are allowed, as in the case of archaea. (b) In contrast, when the sequences are allowed to accommodate many more replication origins, mimicking eukaryotic genomes, the average genome length increases. This is because the expansion in genome length of eukaryotes does not come at the expense of replication time. Unlike prokaryotes, eukaryotes can parallelize replication across multiple replichores by replicating them independently and simultaneously, due to the presence of multiple origins, thereby reducing the replication time substantially, while maintaining a large genome length. Due to the constraint of single origin, prokaryotes cannot have more than two replichores, and hence cannot parallelize replication beyond these two, thus restricting their genome length. (c) and (d) Minimization of mutation pressure γ . For both prokaryotes and eukaryotes, γ converges to $1/\text{number of replichores}$. This convergence implies symmetrization of replichorse lengths across a genome. The initial pool has no origins, and hence one replichorse, and the initial value of γ is thus 1.

is restricted to 1, mimicking a prokaryotic genome, the average length of the genome decreases over generations. Despite the applied selection pressure being identical in both scenarios, prokaryotic genomes tend to lose sequence length, while eukaryotes tend to elongate their genomes. This behavior mirrors the evolutionary divergence between the lengths of the prokaryotic and eukaryotic genomes. In the context of the model, the explanation for this pattern of genome evolution is as follows.

Consider a genome with an asymmetric purine-pyrimidine composition, where all nucleotides at the 5'-end of the replication origin are pyrimidines, and those at the 3'-end are purines. The cumulative skew profile of this sequence forms a "V" shape, as illustrated in Fig. 4(a). Prokaryotic genomes with the two replichores emerging from their single origin exhibit such skew profiles. The single replication origin is located at the valley point of the

"V"-shaped skew profile [28, 29, 30, 35]. The skew profile of a eukaryotic genome, on the other hand, has multiple concatenated "V"-shapes, as shown in Fig. 4(b), where, each arm of each "V" corresponds to a replichorse, and the multiple valley points, to multiple origins. Mutations in the sequence alter this skew profile. The selection pressure γ depends on whether the genome undergoes deletion or insertion, and which replichorse arm (shorter or longer) is affected due to the mutation. Fig. 4 shows a few skew profiles of the mutated sequences. Selection acts on the set of genomes with such altered skew profiles and prefers sequences that decrease the replication time and/or increase the information storage.

In the prokaryotic genome (P), mutations involving elongations are not preferred by the selection, since it either increases the longest replichorse length (e.g. P3 and P5), resulting in an increase in overall replication time, or adds an extra valley point to the sequence (e.g. P4 and P6), which is eliminated by the selection pressure, as prokaryotes are restricted to have a single replication origin. However, a mutation that shortens the longest replichorse (e.g., P2) is selected because it decreases the replication time. Unlike prokaryotic genomes, in eukaryotes, our selection algorithm allows for the addition of more origins, and hence more replichores (e.g. E4 and E6). If these new replichores are not the longest among all replichores, mutated sequences containing them will be selected due to their increased information storage capacity and neutral influence on replication time (e.g., E6). Therefore, the genome length of eukaryotes continues to increase over generations through the incorporation of new origins and hence replichores, until selection restricts further increase in the number of origins due to the upper limit Ori_{max} .

In both prokaryotic and eukaryotic organisms, evolutionary selection pressures favor the symmetrization of replichorse lengths. This phenomenon arises because replichores that are shorter than the longest replichorse can undergo elongation through mutational processes, as this enhances the genome's information storage capacity without increasing its replication time, thereby reducing γ . This aligns with the general observation of symmetric replichorse lengths in prokaryotes [36, 37, 38] and the balanced distribution of short sequence motifs in eukaryotes [20, 39]. Collectively, these selective forces drive two key outcomes: (1) the equilibration of replichorse lengths and (2) the addition of one or more new origins, and hence replichores.

The observed genome length reduction in prokaryotes (Fig. 3) is due to a bias of the selection pressure that favors deletions over insertions, although both mutational events are equally probable in our model. As explained in the previous paragraph, selection pressure favors the symmetrization of replichores. When two replichores of the prokaryote are of unequal length, symmetrization requires deletion of the longer replichorse or insertion into the shorter replichorse. Since our choice for the location of these two mutational events is entirely random and, therefore, evenly distributed over the lengths of the genome, the

probability of choosing the shorter replichore for insertion or deletion will always be smaller than the probability of mutations occurring on the longer replichore. Although selection favors insertion into shorter replichore and deletion at longer replichore equally, since the latter is stochastically more favored, deletion occurs more frequently. This computational observation reproduces the strong evolutionary preference observed experimentally for deletions over insertions in prokaryotes [40, 11, 41, 42, 43, 44].

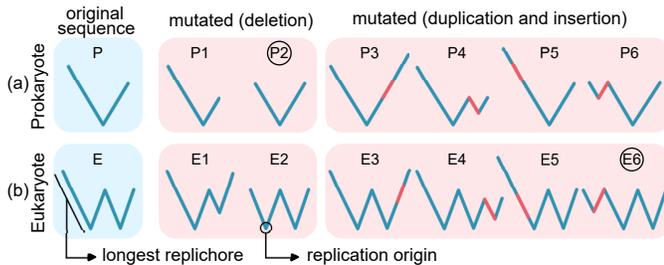


Figure 4: Effect of selection pressure on prokaryotic and eukaryotic genomes. The left panel, in blue background, shows the skew profiles of a prokaryotic and a eukaryotic genome of an intermediate generation, before mutation. The right panel, in red background, shows a few skew profiles of these genomes, after mutation. For eukaryotic genomes, mutations adding an extra replichore originating from a new replication origin (e.g., ‘E6’) are preferred, as they increase information storage by elongating the genome without increasing the replication time (determined by the length of the longest replichore). In prokaryotic genomes, mutations that increase the length of the longest replichore (e.g., ‘P3’ and ‘P5’) or add a new replication origin (e.g., ‘P4’ and ‘P6’) are eliminated by selection pressure. Instead, mutations that shorten the longest replichore length, leading to a symmetrization of replichore lengths (e.g., ‘P2’), are favored, as they optimize the selection pressure ($\gamma_{\text{opt}} = 1/2$) by lowering the replication time.

An initial increase in the average genome length is seen in Fig. 3. This is an artifact of our initial choice of genomes, each of which has been chosen as a homogeneous stretch of purines or pyrimidines. Therefore, these initial sequences have no origin and the length of the entire genome is equal to the length of a replichore, with $\gamma = 1$. The deletions in these sequences do not alter the value of γ , since they affect both the numerator and the denominator of γ equally. However, duplications can reduce γ when the duplicating fragment is from the complementary strand, introducing a new replichore and an origin. The reduction in γ is due to the division of the genome into two replichores, due to the introduction of an origin, thereby reducing the replication time [21]. As a result, within our model, during the early stages of evolution, deletions are not favored by selection, and the average genome length increases.

Symmetrization of replichore lengths leads to Chargaff’s Second Parity Rule

The evolution of genomic sequences in our model begins with a pool of sequences composed entirely of purines (R) on the Watson-strand and complementary pyrimidines (Y) on the Crick-strand. In successive generations, mutation

involving the exchange of strand fragments between complementary strands introduces strand heterogeneity, that is, sequences with interspersed R and Y bases. As demonstrated in the earlier section, evolution under the defined selection pressure γ drives sequences to have replichores of equal length. Half of the replichores, positioned to the left of each origin, become pyrimidine-rich, while the remaining half, to the right of origins, are purine-rich. This symmetry in the cumulative nucleotide skew around replication origins [27] results in global parity in the purine and pyrimidine content throughout the genome, leading to Chargaff’s second parity rule (PR-2).

Chargaff’s first parity rule, identified before the discovery of DNA’s double-stranded structure [45, 46], revealed equal counts of adenine (A) and thymine (T), as well as guanine (G) and cytosine (C) in double-stranded DNA (dsDNA), a pattern now understood as a consequence of Watson-Crick base pairing. In contrast, Chargaff’s second parity rule (PR-2), which extends this symmetry to individual strands of dsDNA, lacks a universally accepted mechanistic explanation [47, 48, 49, 50]. Early hypotheses attributed PR-2 to adaptive intra-strand stem-loop formation, favoring local sequence inversions to achieve the functional benefits of self-complementarity [51, 50]. However, this rationale fails to account for PR-2’s prevalence in non-coding regions, where selective pressures for secondary structures are weak [52]. Alternative theories propose PR-2 as a manifestation of the law of large numbers or an emergent property of entropy maximization in large genomes, where stochastic shuffling of sequences via inversions and transpositions homogenizes base composition over time [53, 54, 55, 56, 57, 58]. However, these mechanisms rely on *no strand-bias* assumptions and do not account for an asymmetry in base substitution frequencies between purine and pyrimidine, i.e. $R \rightarrow Y \neq Y \rightarrow R$ [59, 60]. More importantly, while these theories account for global compositional symmetries, they neither explain nor reproduce local nucleotide compositional skews around replication origins that are prevalent in genomes of nearly all species [27, 28, 29, 30, 61]. The near-universal presence of local violations of PR-2 points to their importance, specifically for replication origin functionality. Stochastic nucleotide shuffling would erase these local PR-2 violations as well, and hence will be counterproductive.

In our framework, although we incorporate inter-strand shuffling, PR-2 emerges not as a passive outcome of random sequence shuffling but as an adaptive response to selection pressure (Fig. 5). Here, selection pressure eliminates any bias in the base substitution frequencies between purines and pyrimidines. Any alteration in the symmetric, V-shaped profile of the cumulative nucleotide skew (see Fig. 4) due to the bias in base substitution frequencies is not tolerated by the selection, since it would adversely affect the replication time, by making the replichore lengths asymmetric. Therefore, selection acts against such biased substitutions, restoring the symmetry of the cumulative skew diagram and that of the replichore lengths, and hence

the global equivalence in the purine-pyrimidine content of a DNA single-strand. The evolutionary trajectory of purine content in our simulations (Fig. 5) illustrates this dynamic, demonstrating a rapid convergence to R/Y parity concurrent with replichore symmetrization. Note that, although the mean R/Y composition equilibrates to 50% in both selective and neutral evolutionary processes (red and blue curves in Fig. 5), the variance in R/Y composition is larger for neutral evolution, when compared with the evolution under selective pressure. This suggests that strong selective pressure leads to strict compliance with PR-2, whereas, non-adaptive evolution allows for deviations from PR-2. If the selection pressure for short replication time is weak or non-existent, as in the case of mitochondria, and for plasmids and viroids, the PR-2 compliance requirement vanishes, according to our model. There is a lack of need to maintain symmetric replichore lengths to minimize replication time in these genomes, since the rate-limiting step for their replication is the replication time of the host genomes, which are much larger. This prediction has been validated experimentally [48, 62]. Chloroplasts' replication is independent of its host cell cycle [63], and it is, therefore, PR-2 compliant.

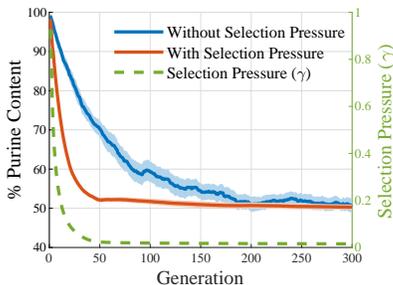


Figure 5: Evolutionary trajectories of base composition over generations. An initial pool of double-stranded DNA, with the Watson strand composed entirely of purines (R) and the Crick strand, of pyrimidines (Y), is evolved through mutations (inversions and inverted transpositions). The variation in purine content of the Watson strand over generations is shown, with shaded regions indicating 95% confidence interval. **Blue curve:** Strand evolution without any selection pressure, exhibiting gradual R/Y parity in DNA single-strand, due to stochastic inter-strand shuffling. This demonstrates PR-2 emerging from stochasticity alone. These sequences do not exhibit any structured nucleotide skew profiles associated with replication origins, seen in most genomes. **Red curve:** Under selection, R/Y parity emerges more rapidly via replichore symmetrization, which is a direct consequence of the selection pressure to minimize replication time by balancing replichore lengths. Selection maintains the nucleotide skew profiles around replication origins, aligning with experimental observations. Selection also preserves PR-2 by eliminating mutational biases for specific nucleotides, removing the necessity for the no-strand bias assumption, used in explanations based on neutral processes.

Influence of endosymbiont power supply on genome length

According to our model above, prokaryotes tend to minimize their genome length due to the limitation of a single replication origin, whereas eukaryotes tend to acquire

genome content, because of their ability to accommodate multiple replication origins. The rationale for our choice to restrict the number of origins of prokaryotes to one, while allowing the eukaryotes to have many more is to align the model with observations. A deeper reason for this choice lies in the bioenergetic requirement for maintaining multiple replication origins, which is to provide activated monomers and energy supply for the replication machinery *simultaneously* to multiple replicating segments of a large genome. This requirement is met through the endosymbiotic relationship of eukaryotes with mitochondria or chloroplasts [17, 9]. Jordan G. Okie *et al.* have identified a linear relationship between mitochondrial and chloroplast count and cell volume in a number of eukaryotes [64]. Since there is a well-known allometric relationship between cell volume and genome length [65, 66, 67, 68], we have converted the cell volume data of Jordan G. Okie *et al.* to that of genome length and plotted the variation of genome length as a function of mitochondrial/chloroplast count. Fig. 6 shows this linear relationship between the above two variables. This suggests that the length of the eukaryotic genome is in part determined by the availability of the power supply to simultaneously replicate multiple genomic segments, provided by multiple mitochondria or chloroplasts.

This limitation imposed on genome length by cellular energetics is taken into account in our model by limiting the number of origins to a set value, Ori_{max} . In prokaryotes, which lack mitochondria/chloroplasts, the number of origins is generally limited to 1 (or a few), and therefore we take $Ori_{max} = 1$. In eukaryotes, the number of origins can be of the order of tens of thousands, and are limited only by power supply availability, as argued above (6). We therefore set Ori_{max} to 50 or 100, to explore the genome length divergence between prokaryotes and eukaryotes.

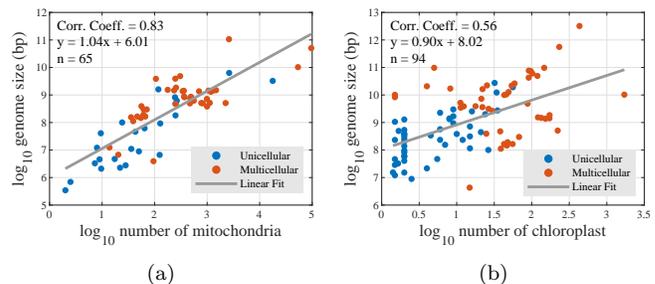


Figure 6: (a) Correlation between the number of mitochondria and genome length in 65 eukaryotic cells. (b) Correlation between the number of chloroplasts and genome length in 94 eukaryotic cells. These plots are produced using data from Jordan G. Okie *et al.* [64]. We estimate the genome length from cell volume using the relation $V = kC^\alpha$ [66], with $k = 3.04 \times 10^{-5} \mu m^3/bp$ and $\alpha = 0.89$. The genome length is positively correlated with mitochondrial or chloroplast count, supporting the argument that the replication of longer genomes of eukaryotes is carried out with the energy provided by the endosymbionts.

Upper and lower bound of genome length

As explained before, the selection pressure symmetrizes the replicore length and adds new origins, and thus replicores, to maximize information storage capacity. After evolving the sequences for a considerable number of generations, the selection produces genomes with nearly equal replicore lengths and the maximum number of origins ($Ori_{max} = 1$ for prokaryotes and $Ori_{max} \gg 1$ for eukaryotes), thus minimizing the selection pressure. The theoretical optimum genome length under these conditions is

$$\begin{aligned} L_{opt} &\approx \text{length of each replicore} \times \text{max no. of replicores} \\ &= \text{length of each replicore} \times 2 Ori_{max} \end{aligned} \quad (3)$$

Therefore the optimum selection pressure is

$$\gamma_{opt} = 1/\text{number of replicores}, \quad (4)$$

The selection pressure can be seen to converge to this optimum in Fig. 3(c) & (d). For prokaryotes, $\gamma_{opt} = \frac{1}{2}$, since a single origin can only support two replicores. Whereas, for eukaryotes, γ_{opt} is 0.01 and 0.005, since, at optimum genome length, they will have ≈ 100 and ≈ 200 replicores, corresponding to 50 and 100 origins, respectively.

Lower bound: The model introduces a wavelet-based resolution limit for distinguishing replication origins. Following a w -level wavelet transform of the genome, pairs of peaks or valleys in the nucleotide skew profile separated by fewer than 2^w nucleotides become indistinguishable, as the transform disregards small-scale variations. This imposes a minimum threshold of 2^w base pairs (bp) on replicore lengths. This threshold may serve as a proxy for the minimal genomic information necessary for cellular viability, reflecting evolutionary constraints on the smallest genome length. Consequently, in our simulations, the smallest genome length achievable for prokaryotes is of the order of 2×2^w nucleotides. Within the model, this lower bound reflects the computational resolution limit for identification of origins, whereas, in the evolutionary dynamics captured by the model, this lower bound reflects the need for sufficient informational complexity for cellular maintenance and replication machinery [69, 70, 71, 72, 73]. Our choice of a specific wavelet level ensures that the genome length does not reduce below a certain viability limit.

Upper bound: One can make an interesting observation from the eq. 4: The optimum length of a eukaryotic genome does not depend on the length of the replicore; only on their number. A genome can therefore increase its own length by increasing the length of each of its replicores evenly, without influencing γ . Such an alteration increases both the replication time and the information storage capacity equally, thereby nullifying its effect on the selection pressure γ . However, our stochastic model cannot make such concerted changes at multiple locations in the genome, and hence cannot alter the length of the genome this way. When the selection pressure is low, evolution, on the other hand, can alter the genome length

by progressively lengthening each of the replicores, by temporarily tolerating an uneven distribution of replicore lengths. This would result in very different genome lengths and corresponding variance in replication times, due to variation in replicore length, even within very similar eukaryotic species, as has been observed abundantly [74, 65, 75, 76]. Therefore, our model does not impose a strict upper limit on the genome length, although it converges to a constant genome length for a fixed number of origins, Ori_{max} . The evolutionarily optimized eukaryotic genome length for a constant Ori_{max} value depends on the average length of the initial genome pool, which can be altered to produce a longer or shorter optimized genome.

Large variance of eukaryotic genome length due to low selection pressure for replication time

In order to verify our statement above that the genome length can vary drastically even within similar species when the selection pressure for replication time is low, we reduce the cost of replication time in our evaluation of selection pressure, by raising it to a power α , where $0 \leq \alpha \leq 1$. The modified expression for selection pressure reads

$$\gamma = \frac{(\text{replication time})^\alpha}{\text{information storage capacity}}. \quad (5)$$

This modification allows for uneven distribution of replicore lengths, and the algorithm tolerates an increase in the replication time by prioritizing information storage capacity, thus enlarging the genome, as explained above. When the importance of replication time reduces, and its cost (α) goes down, the variance in the lengths of the genomes of our initial population increases with generation, as seen from the widening of confidence interval with generation in Fig. 7.

Such computationally observed significant variation in genome length has been documented across species [77, 78, 79] and even within conspecific populations [75, 76, 80, 81], an observation whose evolutionary mechanism is deeply contested. This constitutes the central mystery of the C-value paradox, a set of observations of uncertain evolutionary origins [82, 65]. Prevailing hypotheses posit that persistent upward mutation pressure drives C-value (a measure of genome content) expansion, with species exhibiting slower cellular division rates being more tolerant of random DNA accumulation [74, 83, 84]. This framework is supported by evidence demonstrating strong negative correlations between genome length and both mitotic and meiotic division rates [85, 86]. Our computational model provides an explanation for the above observation, where, low selection pressure for replication time shifts the evolutionary trajectory towards maximizing information storage capacity at the expense of replication time. Apart from increasing the genome content over generations (C-value), this also increases the variance of the genome length, leading to drastically different genome lengths even within conspecific organisms (Fig. 7). Whether the accumulated genome carries information or

not cannot be answered within our model, another mystery of the C-value paradox.

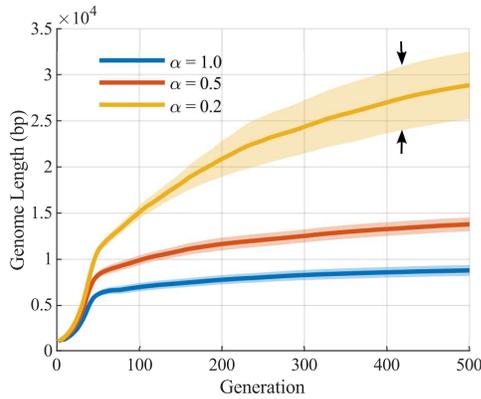


Figure 7: Genome length evolution under reduced selection pressure for replication time. In our model, the replication time is dictated by the length of the longest replicore of the genome. Under low selection pressure for replication time (i.e. low α), replichores expand to enhance information storage capacity while minimally impacting the cost of replication time. This drives progressive genome elongation over generations. This reduced pressure also permits uneven replicore lengths across the genome, as opposed to the optimized symmetrical replicore lengths observed under strong selection. Consequently, genome length variability within a set of related species increases over time, as reflected in the widening confidence intervals for low α . This explains the genome length variability experimentally observed across species and within conspecific organisms. The confidence interval shown here is calculated with 30 samples, for the purpose of visual demonstration. The increase in variance with the reduction in α persists for higher sample numbers.

Discussion

The divergence between the lengths of genomes of prokaryotes and eukaryotes has been an enduring enigma, noted since the first systematic C-value measurements carried out in the 1950s. The nearly 100-fold difference in cell sizes between prokaryotes and eukaryotes, and the huge increase in the structural and functional complexity of eukaryotes are generally attributed to this difference. The purpose of this paper is to investigate the evolutionary origin of this divergence. Here, we identify replication time and information storage capacity as the two primary determinants of genome length, with the latter strongly constrained by cellular energetics. The interplay between these two variables dictates the evolutionary trajectory of genome length, for any given species.

We quantify this interplay between replication time and information storage capacity by defining Selection Pressure simply as the ratio of the above two variables. Evolution acts on the genome, attempting to minimize the replication time while maximizing the information storage capacity. By rewriting the selection pressure in terms of the length of the replichores and the total length of the genome, we make the selection pressure computable. We model the evolution of a pool of genomes by introducing stochastic mutations, randomly adding or deleting a frac-

tion of the total genome, and evaluating the effect of the selection pressure on the mutated genomes.

We observe that prokaryotic genomes, modeled here as genomes with a *single* replication origin, lose genome length with evolving generations. This is due to the inability of prokaryotic genomes to parallelize their replication by dividing the genome into multiple segments (replichores) that can replicate independently and simultaneously, because of the restriction on the number of origins. This restriction reduces the genome length for prokaryotes, since only two simultaneously replicating segments are allowed, and any increase in the size of these segments increases the replication time, and is thus evolutionarily disfavored. On the other hand, we observe that eukaryotic genomes tend to increase their genome length indefinitely, without incurring a cost from increased replication time, since the model’s allowance of a large number of origins for eukaryotic genomes allows for massive parallelization of genome replication, by dividing the genome into multiple independently and simultaneously replicating segments. This indefinite expansion of the eukaryotic genome is curtailed only by cellular energetic constraints, the need for activated monomers and energy supply for simultaneously replicating thousands of genomic segments. This constraint is experimentally demonstrated by the linear relationship between the number of endosymbionts and genome length in eukaryotes (Fig. 6). We model this constraint by limiting the number of origins allowed for eukaryotic genomes.

Our model demonstrates that Chargaff’s second parity rule (PR-2), i.e. the symmetry of purine (R) and pyrimidine (Y) frequencies within individual DNA strands, emerges as a direct consequence of selection for replication efficiency. When initialized with purine-filled single-strand sequences, evolution under the defined selection pressures drives them toward parity, yielding strands with equal purine and pyrimidine content. Critically, this symmetry is not a natural outcome of stochastic inter-strand sequence shuffling or thermodynamic entropy maximization, but an adaptive response to selection for balanced replicore lengths. Any compositional bias in a strand would generate asymmetrical replichores, delaying replication and reducing fitness. Thus, PR-2 in our framework reflects an evolutionary optimization: the equalization of R/Y content is a byproduct of selection to harmonize replicore architecture, ensuring efficient bidirectional replication. These results suggest that PR-2 is a signature of replication-driven adaptation, rather than an outcome of neutral processes.

Although not originally intended, surprisingly, our model reproduces several other experimentally observed phenomena, in addition to the prokaryote-eukaryote genome length divergence and reproduction of PR-2 compliance. (a) A general preference for deletion mutations over insertions (Fig. 4a) [40, 11, 41, 42, 43, 44]. (b) A tendency to equalize the lengths of all replichores of the genome, as indicated by the optimized γ value (Fig. 3c,d)

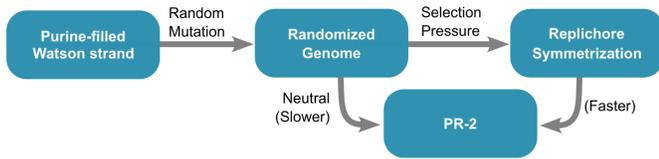


Figure 8: Neutral and adaptive evolutionary trajectories leading to Chargaff’s second parity rule (PR-2). In the neutral process, starting from a purine-filled Watson strand of low fitness (high γ), genomes may reach a state of equal purine (R) and pyrimidine (Y) content on a single strand, i.e. PR-2, solely through a large number of mutations (inversions and inverted transpositions), occurring neutrally without providing any adaptive benefit. The rate of convergence to PR-2 through this route is slower, and the PR-2 compliance is less strict. Moreover, the nucleotide skews that dictate the V-shaped replicore structure are averaged out, nullifying the sequence signature corresponding to replication origins. Alternatively, as demonstrated in this work, selection pressure for rapid replication can *swiftly enforce* PR-2 through replicore length symmetrization, resulting in a *strict* compliance with this parity rule. This adaptive process also retains the sequence signature that determines replication origins.

[20, 36, 37, 38, 39] (c) Correlation between genome length and the number of mitochondria/chloroplasts (Fig. 6) [64, 87] (d) Increase in the variance of genome length resulting from a reduced selection pressure for replication time minimization (C-value paradox) (Fig. 7) [75, 76, 80, 81] (e) Anticorrelation between cell-cycle time and genome length (Fig. 7) [85, 86].

Our model can be tested using *in-vitro* evolution experiments on self-replicating DNA sequences that involve deletions and insertion mutations. Under strong selection pressure for faster replication, the sequences should evolve towards a replicore structure with a central origin of replication and equal and opposite nucleotide skews on the two replicore arms. However, when primed to replicate from the ends, these sequences should evolve toward purine/pyrimidine-filled single strands. When the supply of monomers is adequate, the more fit sequences would exhibit multiple origins. Any imbalance in the lengths of the two replicores of the evolutionarily superior sequences should adversely affect the sequence’s fitness.

Our explanation for the divergence between prokaryotic and eukaryotic genome lengths rests primarily on the number of origins used during DNA replication: using more origins reduces the replication time, all else being equal. Multicellular eukaryotes appear to have invented a new degree of freedom to modulate their cell replication time depending on tissue-level spatial, temporal, energetic, and environmental constraints by employing an appropriate number of origins. As has been amply demonstrated, multicellular organisms do not utilize all available origins of replication to replicate their DNA. Only about 30% of the origins of the human genome are constitutively fired, with the utilization of the rest depending on the tissue/organism-level requirements [88, 89, 90]. This top-down control of replication origin firing partly enables these organisms to create specialized organs with disparate cell-cycle rates, such as human skin and colon,

where rapid cell-cycle rates are crucial, and neurons in the brain, which rarely replicate, presumably to preserve information [91, 92, 93, 94]. An important sequence characteristic that segregates multicellular eukaryotic origins into constitutive, latent, and dormant sets is the magnitude of nucleotide skew at the origin locations [95]. Our model above too uses these nucleotide skews to identify the locations of origins, although the magnitude of the skew is not utilized for determining the efficiency of origin firing, a simplification that will be removed in a later article. We speculate that the local loss of such top-down control on replication origin selection in various tissues leads to rapid replication and, consequently, carcinogenesis [96]. Regulation of the number and firing efficiency of replication origins can modulate genome architecture at evolutionary timescales, while organismal top-down control appears to tame the origins into serving the individual organism at the timescale of the lifetime of that individual.

Statements and Declarations

Competing interests

The authors declare no competing interests.

Acknowledgments

Support for this work was provided by the Science & Engineering Research Board (SERB), Department of Science and Technology (DST), India, through a Core Research Grant with file no. CRG/2020/003555 and a MATRICS grant with file no. MTR/2022/000086.

Supplementary Information

The algorithms and parameters used to generate the plots are provided in supplementary information.

Supplementary Information

Model

We are interested in the investigation of genome length evolution under a selection pressure that attempts to minimize replication time while maximizing information storage capacity. An initial pool of N identical sequences, composed of all purines or all pyrimidines, are replicated, and the daughter sequences are mutated, producing a pool of $2N$ sequences. Selection acts on this pool, removing N less-fit sequences that do not satisfy the selection pressure adequately. This replication-selection cycle is repeated m times, and the average genome length at every generation is recorded.

Algorithms

We followed the following algorithm to mutate each sequence in the pool and then calculate the selection pressure (γ) to extract the fit sequences.

Mutation

We implement large-scale genomic mutations through deletions or duplications of random regions comprising 5% to 10% of the total genome length (Fig. 9). In each generation, every sequence in the pool undergoes a single mutation, i.e., either a duplication or deletion. Following this, the sequence pool is expanded to include the new N mutated sequences along with the N sequences of the previous generation, resulting in a total of $2N$ sequences.

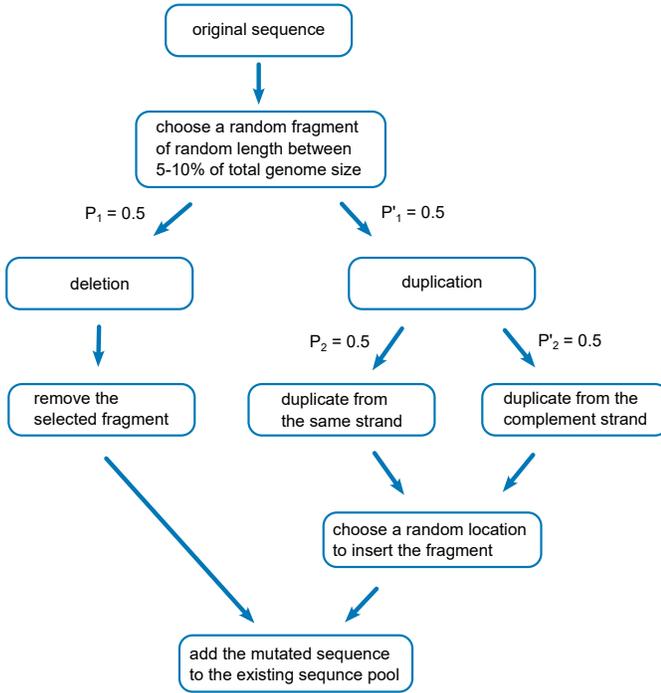


Figure 9: Algorithm for mutation. Here, P_1 and P'_1 are the probability of the sequence undergoing deletion or duplication, respectively. And, P_2 and P'_2 are the probability of the sequence undergoing duplication from the same strand or complementary strand, respectively.

Calculation of selection pressure

Following mutation, N sequences are selected from the pool of $2N$ for the next generation based on a selection pressure aimed at minimizing the factor γ , which is defined as,

$$\gamma = \frac{\text{length of longest replicore}}{\text{full genome length}} \quad (6)$$

A low γ value of a sequence suggests that the sequence is capable of fast replication and high information storage.

We used cumulative skew of the sequences to identify replication origins and termini and hence the replicore

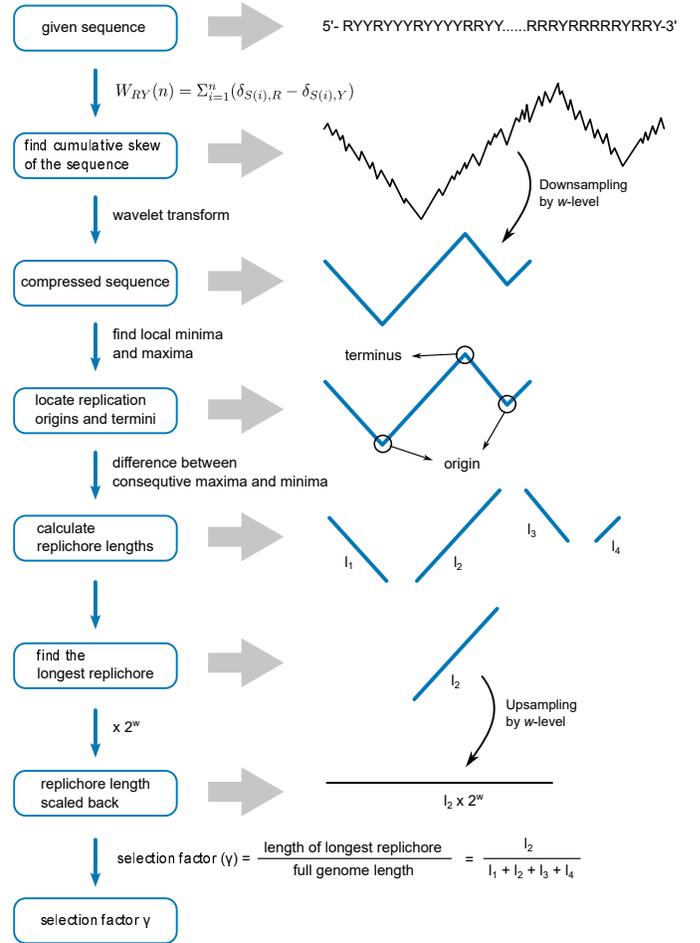


Figure 10: Calculation of selection pressure (γ)

lengths. The purine-pyrimidine (RY) cumulative skew of the sequences, $W_{RY}(n)$, is defined as;

$$W_{RY}(n) = \sum_{i=1}^n (\delta_{S(i),R} - \delta_{S(i),Y}) \quad (7)$$

Here, S is a genomic sequence of length N bp, composed of four nucleotides, classified into two groups: $R = \{G, A\}$, $Y = \{C, T\}$, and $n = 1 \dots N$.

The selection pressure (γ) is calculated for each of the sequences. Thereafter, N sequences with the lowest γ values are selected as evolutionarily superior and are carried forward to the next generation.

Genome Length Evolution Over 1000 Generations

In our simulations, we set the population size to $N = 1000$, the initial genome length to 1024 bp (L_0), and the number of generations to $m = 1000$, employing a four-level wavelet transformation. Each experiment was repeated 100 times to ensure statistical robustness. The genome length stabilized within 500 generations; thus, for clarity, the main article presents results up to 500 generations. The complete genome length evolution over 1000 generations is shown below.

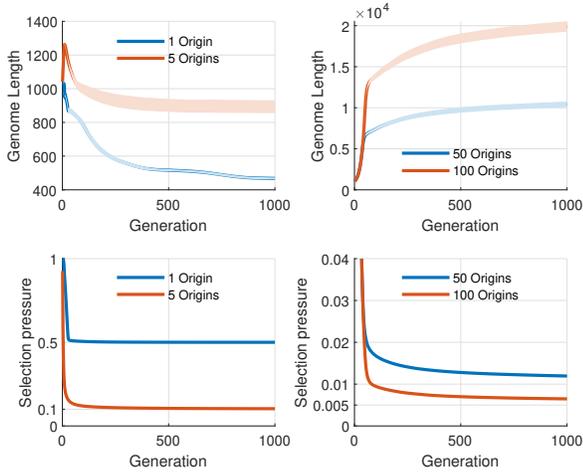


Figure 11: Genome length evolution for 1000 generations

Genome Length and Cell Size Allometry

Genome length has been reported to scale with cell volume following an allometric relationship [65, 67, 68]:

$$V = kC^\alpha, \quad (8)$$

where V represents the cell volume in cubic micrometers, and C denotes genome length in base pairs. We applied this allometric relation to estimate genome lengths from cell volumes in our dataset.

To determine the coefficients k and α , we analyzed the correlation between cell volume and genome length for 60 organisms with available genome data. The estimated coefficients were then used in Eq. 8 to infer genome lengths for the remaining organisms in our dataset.

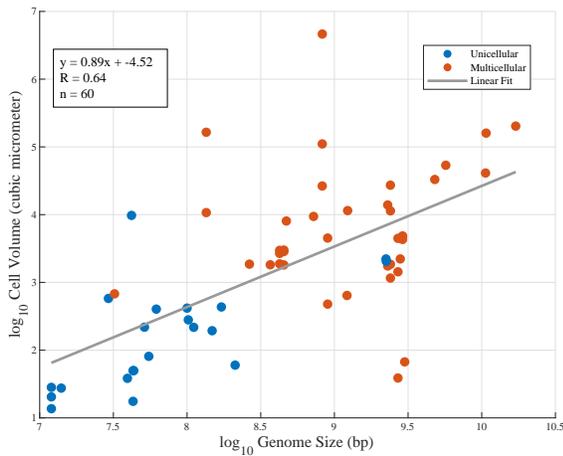


Figure 12: Correlation between genome length and cell volume.

The log-log correlation of genome length and cell volume for these 60 organisms is presented in Fig. 12, yielding estimated coefficients of $k = 3.04 \times 10^{-5} \mu\text{m}^3/\text{bp}$ and

$\alpha = 0.89$. Using these coefficients, we estimated genome lengths for the rest of the organisms from equation (8) and subsequently analyzed their correlation with mitochondria and chloroplast count. The correlation between genome length and mitochondrial count is discussed in the main article.

Parameters

The following parameters were used to generate Fig. (3) of the main article.

1. Length of each sequence in the initial pool, $L_0 = 1024$ bp
2. Number of sequences in the pool, $N = 1000$
3. Wavelet level = 4
4. Minimum length of mutating fragment = 5% of full genome
5. Maximum length of mutating fragment = 10% of full genome
6. Number of generations = 1000

The following parameters were used to generate Fig. (5) of the main article.

1. Length of each sequence in the initial pool, $L_0 = 1024$ bp
2. Number of sequences in the pool, $N = 100$
3. Wavelet level = 4
4. Minimum length of mutating fragment = 5% of full genome
5. Maximum length of mutating fragment = 10% of full genome
6. Maximum number of replication origins $Ori_{max} = 50$
7. Number of generations = 300

The following parameters were used to generate Fig. (7) of the main article.

1. Length of each sequence in the initial pool, $L_0 = 1024$ bp
2. Number of sequences in the pool, $N = 100$
3. Wavelet level = 4
4. Minimum length of mutating fragment = 5% of full genome
5. Maximum length of mutating fragment = 10% of full genome
6. Maximum number of replication origins $Ori_{max} = 50$
7. Number of generations = 500

References

- [1] T Ryan Gregory. “Understanding natural selection: essential concepts and common misconceptions”. In: *Evolution: Education and outreach* 2 (2009), pp. 156–175.
- [2] Freeman Dyson. *Origins of life*. Cambridge University Press, 1999.
- [3] Radhey S Gupta and G Brian Golding. “The origin of the eukaryotic cell”. In: *Trends in biochemical sciences* 21.5 (1996), pp. 166–171.
- [4] T Martin Embley and William Martin. “Eukaryotic evolution, changes and challenges”. In: *Nature* 440.7084 (2006), pp. 623–630.
- [5] William F Martin, Sriram Garg, and Verena Zimorski. “Endosymbiotic theories for eukaryote origin”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1678 (2015), p. 20140330.
- [6] Itamar Sela, Yuri I Wolf, and Eugene V Koonin. “Theory of prokaryotic genome evolution”. In: *Proceedings of the National Academy of Sciences* 113.41 (2016), pp. 11399–11407.
- [7] Michael Lynch and John S Conery. “The origins of genome complexity”. In: *science* 302.5649 (2003), pp. 1401–1404.
- [8] Eduardo PC Rocha. “The organization of the bacterial genome”. In: *Annual review of genetics* 42.1 (2008), pp. 211–233.
- [9] Nick Lane. “Energetics and genetics across the prokaryote-eukaryote divide”. In: *Biology direct* 6 (2011), pp. 1–31.
- [10] Katsumi Chiyomaru and Kazuhiro Takemoto. “Revisiting the hypothesis of an energetic barrier to genome complexity between eukaryotes and prokaryotes”. In: *Royal Society open science* 7.2 (2020), p. 191859.
- [11] Alex Mira, Howard Ochman, and Nancy A Moran. “Deletional bias and the evolution of bacterial genomes”. In: *TRENDS in Genetics* 17.10 (2001), pp. 589–596.
- [12] Piotr Bentkowski, Cock Van Oosterhout, and Thomas Mock. “A model of genome size evolution for prokaryotes in stable and fluctuating environments”. In: *Genome biology and evolution* 7.8 (2015), pp. 2344–2351.
- [13] Carolina A Martinez-Gutierrez and Frank O Aylward. “Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales”. In: *PLoS Genetics* 18.5 (2022), e1010220.
- [14] Alejandro Rodríguez-Gijón et al. “Linking prokaryotic genome size variation to metabolic potential and environment”. In: *ISME communications* 3.1 (2023), p. 25.
- [15] Stephan Fischer et al. “A model for genome size evolution”. In: *Bulletin of mathematical biology* 76 (2014), pp. 2249–2291.
- [16] Marco Colnaghi, Nick Lane, and Andrew Pomiankowski. “Genome expansion in early eukaryotes drove the transition from lateral gene transfer to meiotic sex”. In: *Elife* 9 (2020), e58873.
- [17] Nick Lane and William Martin. “The energetics of genome complexity”. In: *Nature* 467.7318 (2010), pp. 929–934.
- [18] Patrick Alfred Pierce Moran. “Random processes in genetics”. In: *Mathematical proceedings of the cambridge philosophical society*. Vol. 54. 1. Cambridge University Press. 1958, pp. 60–71.
- [19] Yuri I Wolf et al. “The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages”. In: *Proceedings of the National Academy of Sciences* 106.18 (2009), pp. 7273–7280.
- [20] Guenter Albrecht-Buehler. “Inversions and inverted transpositions as the basis for an almost universal “format” of genome sequences”. In: *Genomics* 90.3 (2007), pp. 297–305.
- [21] Hemachander Subramanian and Robert A Gatenby. “Evolutionary advantage of anti-parallel strand orientation of duplex DNA”. In: *Scientific Reports* 10.1 (2020), p. 9883.
- [22] Parthasarathi Sahu et al. “High Nucleotide Skew Palindromic DNA Sequences Function as Potential Replication Origins due to their Unzipping Propensity”. In: *Journal of Molecular Evolution* (2024), pp. 1–15.
- [23] Dominik Boos and Pedro Ferreira. “Origin firing regulations to control genome replication timing”. In: *Genes* 10.3 (2019), p. 199.
- [24] Prasanta K Patel et al. “DNA replication origins fire stochastically in fission yeast”. In: *Molecular biology of the cell* 17.1 (2006), pp. 308–316.
- [25] Xindan Wang et al. “Replication and segregation of an Escherichia coli chromosome with two replication origins”. In: *Proceedings of the National Academy of Sciences* 108.26 (2011), E243–E250.
- [26] Rivka Rudner, John D Karkas, and Erwin Chargaff. “Separation of B. subtilis DNA into complementary strands. 3. Direct analysis.” In: *Proceedings of the National Academy of Sciences* 60.3 (1968), pp. 921–922.
- [27] Jean R Lobry. “Asymmetric substitution patterns in the two DNA strands of bacteria.” In: *Molecular biology and evolution* 13.5 (1996), pp. 660–665.

- [28] Andrei Grigoriev. “Analyzing genomes with cumulative skew diagrams”. In: *Nucleic acids research* 26.10 (1998), pp. 2286–2290.
- [29] Elisabeth RM Tillier and Richard A Collins. “The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes”. In: *Journal of Molecular Evolution* 50 (2000), pp. 249–257.
- [30] Michael J McLean, Kenneth H Wolfe, and Kevin M Devine. “Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes”. In: *Journal of molecular evolution* 47 (1998), pp. 691–696.
- [31] Ren Zhang and Chun-Ting Zhang. “Identification of replication origins in archaeal genomes based on the Z-curve method”. In: *Archaea* 1.5 (2004), p. 335.
- [32] Natalia V Sernova and Mikhail S Gelfand. “Identification of replication origins in prokaryotic genomes”. In: *Briefings in Bioinformatics* 9.5 (2008), pp. 376–391.
- [33] AC Frank and JR Lobry. “Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes”. In: *Bioinformatics* 16.6 (2000), pp. 560–561.
- [34] Ren Zhang and Chun-Ting Zhang. “Z curves, an intuitive tool for visualizing and analyzing the DNA sequences”. In: *Journal of Biomolecular Structure and Dynamics* 11.4 (1994), pp. 767–782.
- [35] Feng Gao and Chun-Ting Zhang. “Ori-Finder: a web-based system for finding oriC s in unannotated bacterial genomes”. In: *BMC bioinformatics* 9 (2008), pp. 1–6.
- [36] T David Matthews and Stanley Maloy. “Fitness effects of replicore imbalance in *Salmonella enterica*”. In: *Journal of bacteriology* 192.22 (2010), pp. 6086–6088.
- [37] Aaron E Darling, István Miklós, and Mark A Ragan. “Dynamics of genome rearrangement in bacterial populations”. In: *PLoS genetics* 4.7 (2008), e1000128.
- [38] Magnus G Jespersen et al. “Insertion sequence elements and unique symmetrical genomic regions mediate chromosomal inversions in *Streptococcus pyogenes*”. In: *Nucleic Acids Research* 52.21 (2024), pp. 13128–13137.
- [39] Vinayakumar V Prabhu. “Symmetry observations in long nucleotide sequences.” In: *Nucleic acids research* 21.12 (1993), p. 2797.
- [40] Chih-Horng Kuo and Howard Ochman. “Deletional bias across the three domains of life”. In: *Genome biology and evolution* 1 (2009), pp. 145–152.
- [41] Martin S Taylor, Chris P Ponting, and Richard R Copley. “Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes”. In: *Genome research* 14.4 (2004), pp. 555–566.
- [42] Shiheng Tao et al. “Patterns of insertion and deletion in mammalian genomes”. In: *Current Genomics* 8.6 (2007), pp. 370–378.
- [43] Jan O Andersson and Siv GE Andersson. “Pseudogenes, junk DNA, and the dynamics of *Rickettsia* genomes”. In: *Molecular biology and evolution* 18.5 (2001), pp. 829–839.
- [44] T Ryan Gregory. “Insertion–deletion biases and the evolution of genome size”. In: *Gene* 324 (2004), pp. 15–34.
- [45] Erwin Chargaff. “Chemical specificity of nucleic acids and mechanism of their enzymatic degradation”. In: *Experientia* 6.6 (1950), pp. 201–209.
- [46] Erwin Chargaff, Rakoma Lipshitz, and Charlotte Green. “Composition of the desoxyribose nucleic acids of four genera of sea-urchin”. In: *Journal of Biological Chemistry* 195.1 (1952), pp. 155–160.
- [47] Noboru Sueoka. “Intrastrand parity rules of DNA base composition and usage biases of synonymous codons”. In: *Journal of molecular evolution* 40 (1995), pp. 318–325.
- [48] David Mitchell and Robert Bridge. “A test of Chargaff’s second rule”. In: *Biochemical and biophysical research communications* 340.1 (2006), pp. 90–94.
- [49] Pierre-François Baisnée, Steve Hampson, and Pierre Baldi. “Why are complementary DNA strands symmetric?” In: *Bioinformatics* 18.8 (2002), pp. 1021–1033.
- [50] Donald R Forsdyke. “Neutralism versus selectionism: Chargaff’s second parity rule, revisited”. In: *Genetica* 149.2 (2021), pp. 81–88.
- [51] Donald R Forsdyke. “Relative roles of primary sequence and (G+ C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species”. In: *Journal of molecular evolution* 41 (1995), pp. 573–581.
- [52] Shang-Hong Zhang and Ya-Zhi Huang. “Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA”. In: *Bioinformatics* 26.4 (2010), pp. 478–485.
- [53] JR Lobry. “Properties of a general model of DNA evolution under no-strand-bias conditions”. In: *Journal of molecular evolution* 40 (1995), pp. 326–330.
- [54] Guenter Albrecht-Buehler. “Asymptotically increasing compliance of genomes with Chargaff’s second parity rules through inversions and inverted transpositions”. In: *Proceedings of the National Academy of Sciences* 103.47 (2006), pp. 17828–17833.

- [55] Andrew Hart, Servet Martínez, and Felipe Olmos. “A Gibbs approach to Chargaff’s second parity rule”. In: *Journal of Statistical Physics* 146 (2012), pp. 408–422.
- [56] Piero Fariselli et al. “DNA sequence symmetries from randomness: the origin of the Chargaff’s second parity rule”. In: *Briefings in bioinformatics* 22.2 (2021), pp. 2172–2181.
- [57] Bakhyt T Matkarimov and Murat K Saparbaev. “Chargaff’s second parity rule lies at the origin of additive genetic interactions in quantitative traits to make omnigenic selection possible”. In: *PeerJ* 11 (2023), e16671.
- [58] Patrick Pflughaupt and Aleksandr B Sahakyan. “Generalised interrelations among mutation rates drive the genomic compliance of Chargaff’s second parity rule”. In: *Nucleic Acids Research* 51.14 (2023), pp. 7409–7423.
- [59] Jean-Pierre Vartanian, Michel Henry, and Simon Wain-Hobson. “Hypermutagenic PCR involving all four transitions and a sizeable proportion of transversions”. In: *Nucleic acids research* 24.14 (1996), pp. 2627–2631.
- [60] AC Frank and JR Lobry. “Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms”. In: *Gene* 238.1 (1999), pp. 65–77.
- [61] Boris Bartholdy et al. “Allele-specific analysis of DNA replication origins in mammalian cells”. In: *Nature communications* 6.1 (2015), p. 7051.
- [62] Christoforos Nikolaou and Yannis Almirantis. “Deviations from Chargaff’s second parity rule in organellar DNA: Insights into the evolution of organellar genomes”. In: *Gene* 381 (2006), pp. 34–41.
- [63] Yukihiro Kabeya and Shin-ya Miyagishima. “Chloroplast DNA replication is regulated by the redox state independently of chloroplast division in *Chlamydomonas reinhardtii*”. In: *Plant physiology* 161.4 (2013), pp. 2102–2112.
- [64] Jordan G Okie, Val H Smith, and Mercedes Martin-Cereceda. “Major evolutionary transitions of life, metabolic scaling and the number and size of mitochondria and chloroplasts”. In: *Proceedings of the Royal Society B: Biological Sciences* 283.1831 (2016), p. 20160611.
- [65] T Ryan Gregory. “Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma”. In: *Biological reviews* 76.1 (2001), pp. 65–101.
- [66] T. Ryan Gregory, Paul Hebert, and Ju Kolasa. “Evolutionary Implications of the Relationship between Genome Size and Body Size in Flatworms and Copepods”. In: *Heredity* 84 (Pt 2) (Mar. 2000), pp. 201–8. DOI: 10.1046/j.1365-2540.2000.00661.x.
- [67] Ettore Olmo. “Nucleotype and cell size in vertebrates: a review.” In: *Basic and applied histochemistry* 27.4 (1983), pp. 227–256.
- [68] MANFREDI ROMANINI MG. “The DNA nuclear content and the evolution of vertebrates”. In: *Cytotaxonomy and vertebrate evolution* (1973), pp. 39–81.
- [69] Alexander V Markov, Valery A Anisimov, and Andrey V Korotayev. “Relationship between genome size and organismal complexity in the lineage leading from prokaryotes to mammals”. In: *Paleontological Journal* 44 (2010), pp. 363–373.
- [70] José E González-Pastor, José L San Millán, and Felipe Moreno. “The smallest known gene.” In: *Nature* 369.6478 (1994), pp. 281–281.
- [71] John I Glass et al. “Essential genes of a minimal bacterium”. In: *Proceedings of the National Academy of Sciences* 103.2 (2006), pp. 425–430.
- [72] Claire M Fraser et al. “The minimal gene complement of *Mycoplasma genitalium*”. In: *Science* 270.5235 (1995), pp. 397–404.
- [73] Rosario Gil et al. “Determination of the core of a minimal bacterial gene set”. In: *Microbiology and Molecular Biology Reviews* 68.3 (2004), pp. 518–537.
- [74] Mark Pagel and Rufus A Johnstone. “Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox”. In: *Proceedings of the Royal Society of London. Series B: Biological Sciences* 249.1325 (1992), pp. 119–124.
- [75] Judith K Greenlee, KS Rai, and Alton D Floyd. “Intraspecific variation in nuclear DNA content in *Collinsia verna* Nutt.(Scrophulariaceae)”. In: *Heredity* 52.2 (1984), pp. 235–242.
- [76] Petr Šmarda and Petr Bureš. “Understanding intraspecific variation in genome size in plants.” In: (2010).
- [77] Mike D Bennett and Ilia J Leitch. “Nuclear DNA amounts in angiosperms: targets, trends and tomorrow”. In: *Annals of botany* 107.3 (2011), pp. 467–590.
- [78] Ilia J Leitch and Andrew R Leitch. “Genome size diversity and evolution in land plants”. In: *Plant genome diversity Volume 2: Physical structure, behaviour and evolution of plant genomes*. Springer, 2012, pp. 307–322.
- [79] T Ryan Gregory. *The evolution of the genome*. Elsevier, 2011.
- [80] MA Mowforth and JP Grime. “Intra-population variation in nuclear DNA amount, cell size and growth rate in *Poa annua* L.” In: *Functional Ecology* (1989), pp. 289–295.

- [81] Samuel F Lockwood and John W Bickham. “Genome size in Beaufort Sea coastal assemblages of Arctic ciscoes”. In: *Transactions of the American Fisheries Society* 121.1 (1992), pp. 13–20.
- [82] Susumu Ohno. “So much” junk” DNA in our genome. In” Evolution of Genetic Systems”. In: *Brookhaven symposium in biology*. Vol. 23. 1972, pp. 366–370.
- [83] Michael David Bennett. “Nuclear DNA content and minimum generation time in herbaceous plants”. In: *Proceedings of the Royal Society of London. Series B. Biological Sciences* 181.1063 (1972), pp. 109–135.
- [84] Dmitri A Petrov. “Mutational equilibrium model of genome size evolution”. In: *Theoretical population biology* 61.4 (2002), pp. 531–544.
- [85] Irena Šimová and Tomáš Herben. “Geometrical constraints in the scaling relationships between genome size, cell size and cell cycle length in herbaceous plants”. In: *Proceedings of the Royal Society B: Biological Sciences* 279.1730 (2012), pp. 867–875.
- [86] Dennis Francis, M Stuart Davies, and Peter W Barlow. “A strong nucleotypic effect on the cell cycle regardless of ploidy level”. In: *Annals of Botany* 101.6 (2008), pp. 747–757.
- [87] Peyman Fahimi, Chérif F. Matta, and Jordan Okie. “Are Size and Mitochondrial Power of Cells Inter-determined?” In: *Journal of Theoretical Biology* 572 (July 2023), p. 111565. DOI: 10.1016/j.jtbi.2023.111565.
- [88] Lilas Courtot, Jean-Sébastien Hoffmann, and Valérie Bergoglio. “The protective role of dormant origins in response to replicative stress”. In: *International journal of molecular sciences* 19.11 (2018), p. 3569.
- [89] RA Scalfani and TM2292467 Holzen. “Cell cycle regulation of DNA replication”. In: *Annu. Rev. Genet.* 41.1 (2007), pp. 237–280.
- [90] Antoine Aze and Domenico Maiorano. “Recent advances in understanding DNA replication: cell type-specific adaptation of the DNA replication program”. In: *F1000Research* 7 (2018).
- [91] Ichiro Hiratani et al. “Global reorganization of replication domains during embryonic stem cell differentiation”. In: *PLoS biology* 6.10 (2008), e245.
- [92] Tyrone Ryba et al. “Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types”. In: *Genome research* 20.6 (2010), pp. 761–770.
- [93] David M Gilbert. “Making sense of eukaryotic DNA replication origins”. In: *Science* 294.5540 (2001), pp. 96–100.
- [94] Nicholas Rhind and David M Gilbert. “DNA replication timing”. In: *Cold Spring Harbor perspectives in biology* 5.8 (2013), a010132.
- [95] Guillaume Guilbaud et al. “Determination of human DNA replication origin position and efficiency reveals principles of initiation zone organisation”. In: *Nucleic Acids Research* 50.13 (2022), pp. 7436–7450.
- [96] Haiqing Fu et al. “Dynamics of replication origin over-activation”. In: *Nature Communications* 12.1 (2021), p. 3448.