Genome size and nucleotide skews as predictors of bacterial growth rate

Parthasarathi Sahu[1], Sashikanta Barik[1], Koushik Ghosh[1], and Hemachander Subramanian*[1]

[1]Department of Physics, National Institute of Technology Durgapur, India - 713209

Abstract

Bacterial growth rates are constrained by genome replication, yet the role of replication kinetics in bacterial growth rates remains incompletely understood. Here, we examine if genome size, replichore organization, and nucleotide compositional asymmetry are reasonable predictors of bacterial doubling times. In free-living bacteria, both genome size and the length of the longest replichore are found to correlate positively with doubling time, pointing to an influence of replication dynamics on bacterial growth rates. Moreover, fast-growing bacteria are shown to exhibit stronger nucleotide compositional skew. Incorporating skew into the model substantially improves predictive accuracy, suggesting that compositional asymmetry in genomes may facilitate replication fork progression and thereby enhance growth rates. Based on these observations, we speculate that nucleotide skew may play a potential adaptive role in bacterial genome replication. To assess whether the observed association between genome architecture and growth rate reflects an evolutionary signature or a mechanistic link, we reconstructed ancestral states and found that the model fits ancestral traits more strongly, with predictive strength (R^2) decreasing progressively along the evolutionary tree as successive speciations occur. We speculate that this association has been stronger early in bacterial evolution and became subsequently screened as organisms diversified and increased in ecological and physiological complexity.

Keywords: Replication time, Genome evolution, Replichore organization, Fork speed, Nucleotide skew, Asymmetric Cooperativity Model

1 Introduction

The persistence of *compact* genomes in bacteria, despite billions of years of evolution, has been explained by genome streamlining, where natural selection favors the reduction of genome size [1, 2, 3, 4]. For such a reduction to be maintained, it must confer measurable fitness advantages to the organism. Several mechanisms have been proposed: First, smaller genomes require fewer resources for maintenance and expression. Thus, in nutrient-limited environments and in species with large effective population sizes, selection is expected to favor reduced genomes [5, 6, 7]. Second, a reduction in genome size often correlates with smaller cell volume, which increases the surface-to-volume ratio and thereby enhances metabolic efficiency [7, 8]. Third, shorter genomes reduce the time required for replication, which could potentially accelerate cell division and improve competitive fitness. While the first two explanations, nutrient economy and enhanced metabolic efficiency, are widely cited to support streamlining theory, the third explanation has been more controversial. Earlier studies reported little to no correlation between bacterial generation time and genome size, leading to the dismissal of replication time as a significant driver of genome reduction [9, 10, 11, 12].

In this work, we revisit this question. We analyze two sets of bacterial genomes ($n = 367$ and $n = 180$) [11, 13] and their replichore organization to test the influence of genome architecture on the organism's growth rate.

Methods

This study examines how bacterial genome architecture influences replication dynamics and, consequently, organismal growth rates. We correlate growth rates with genome size, replichore size, and nucleotide compositional skew through linear regression models. Two datasets of bacterial growth rates are used: (i) doubling time normalized across reported data, Dataset-I [13], and (ii) minimal doubling time under rapid growth conditions, Dataset-II [11]. Corresponding

*Corresponding author: hemachander@gmail.com

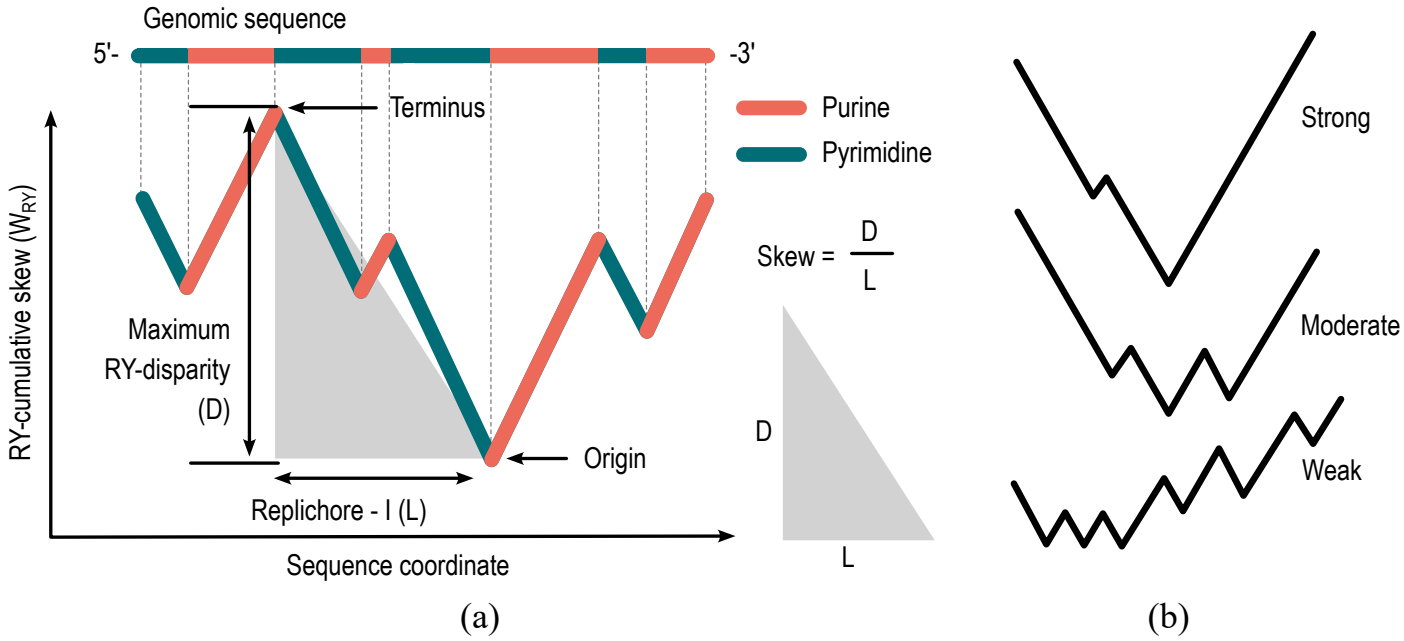


Figure 1: Identification of replication origin and terminus. (a) Cumulative RY-skew (W_{RY}) for a representative genome segment, calculated using Eq. 1. The positions corresponding to the minimum and maximum cumulative RY-skews were used to predict the replication origin and terminus, respectively, thereby dividing the genome into two replichores. The ratio of maximum nucleotide disparity, D, and the corresponding replichore length, L, is defined as "skew", a metric that measures the density of RY asymmetry in a replichore. (b) Cumulative skew profiles for three representative genomes, illustrating variation in skew strength categorized as "strong", "moderate", and "weak".

genomic sequences are retrieved from the NCBI Genome Database, and their accession numbers are provided in Supplementary Data. Only species with both doubling time and genome sequence data available, i.e., 367 in Dataset-I and 180 in Dataset-II, are included in the analysis.

Throughout this study, statistical significance is assessed at $p < 0.05$, and effect sizes are interpreted using the coefficient of determination (R^2), with $R^2 < 0.1$ considered negligible, $0.1 \leq R^2 < 0.2$ weak, and $0.2 \leq R^2 < 0.3$ moderate.

Identification of replication origin and terminus

Nucleotide compositional asymmetries have long been used to infer replication origins and termini [14, 15, 16, 17]. In this study, we employ the purine–pyrimidine skew (RY-skew), which quantifies imbalances between purines ($R = \{A, G\}$) and pyrimidines ($Y = \{C, T\}$) across the genome. The RY-cumulative skew is calculated as:

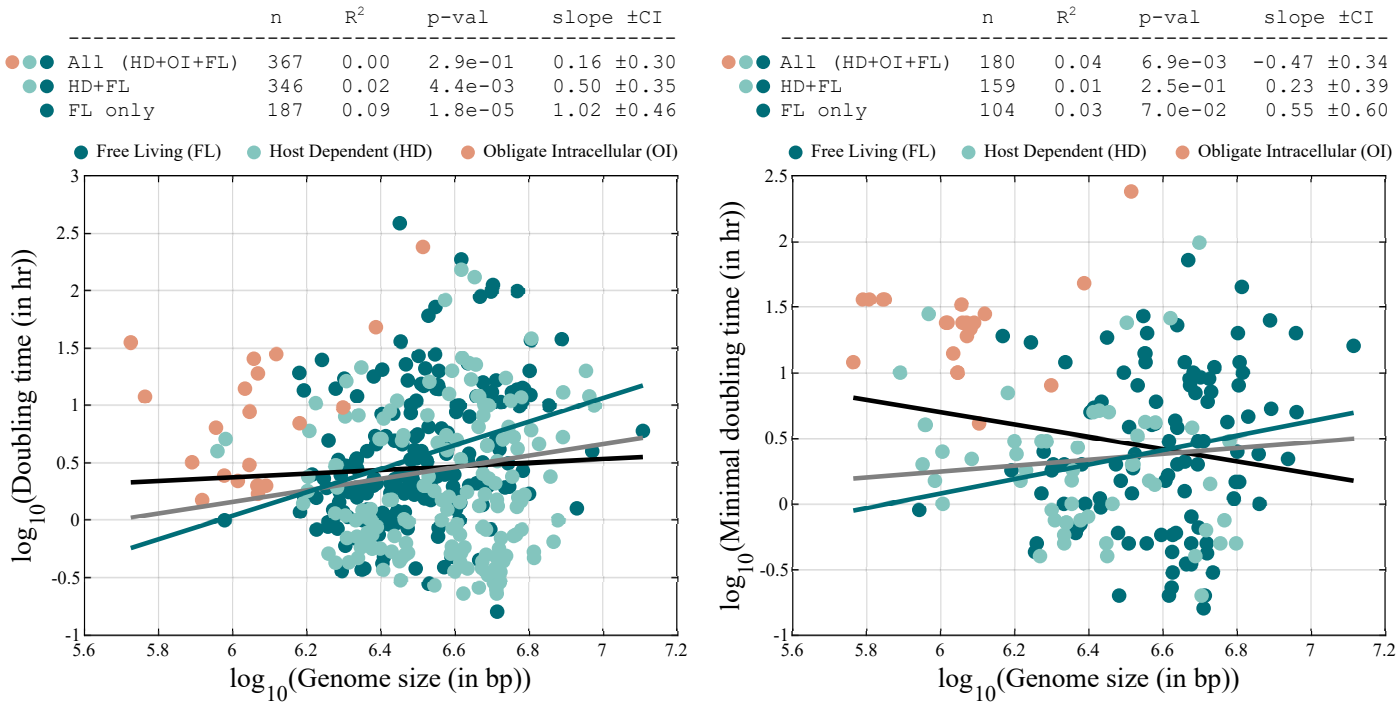
$$W_{RY}(l) = \sum_{i=1}^{l} \left(\delta_{S(i),R} - \delta_{S(i),Y}\right), \tag{1}$$

where S represents a genomic sequence of length G, l is the genomic position index running from 1 to G and δ is the Kronecker delta.

To find the origins and termini of any given sequence, we find the genomic segment that exhibits maximum RY-disparity (Fig. 1). Out of the two extremes of this segment, one falls in a valley position in the cumulative skew diagram, designated as the 'replication origin', and the other falls on a peak and is designated as the 'terminus', thereby partitioning the chromosome into two replichores [14]. A representative cumulative skew profile is shown in Fig. 1. All analyses were implemented in MATLAB, and custom scripts are publicly available (see Supplementary Information).

Results

We investigate whether the size and sequence structure of the genome can predict bacterial doubling time. To this end, we exclusively used genomic sequence characteristics, such as the size and organization of chromosomes and



(a) Model-I on Dataset-I

(a) Model-I on Dataset-II

Figure 2: Relationship between bacterial doubling time and genome size in (a) Dataset-I and (b) Dataset-II. An absence of correlation in Dataset-I and a negligible negative correlation is observed in the full data set. Restricting the analysis to only free-living bacteria reveals a negligible positive correlation in Dataset-I. Regression lines are shown in black for the full dataset, gray for the dataset excluding obligate intracellular species, and green for free-living bacteria only. The same color scheme is followed throughout all figures in this manuscript.

replichores, and their nucleotide compositional skew, to estimate the replication time of the genome and examined whether these characteristics correlate with bacterial doubling time using four regression models.

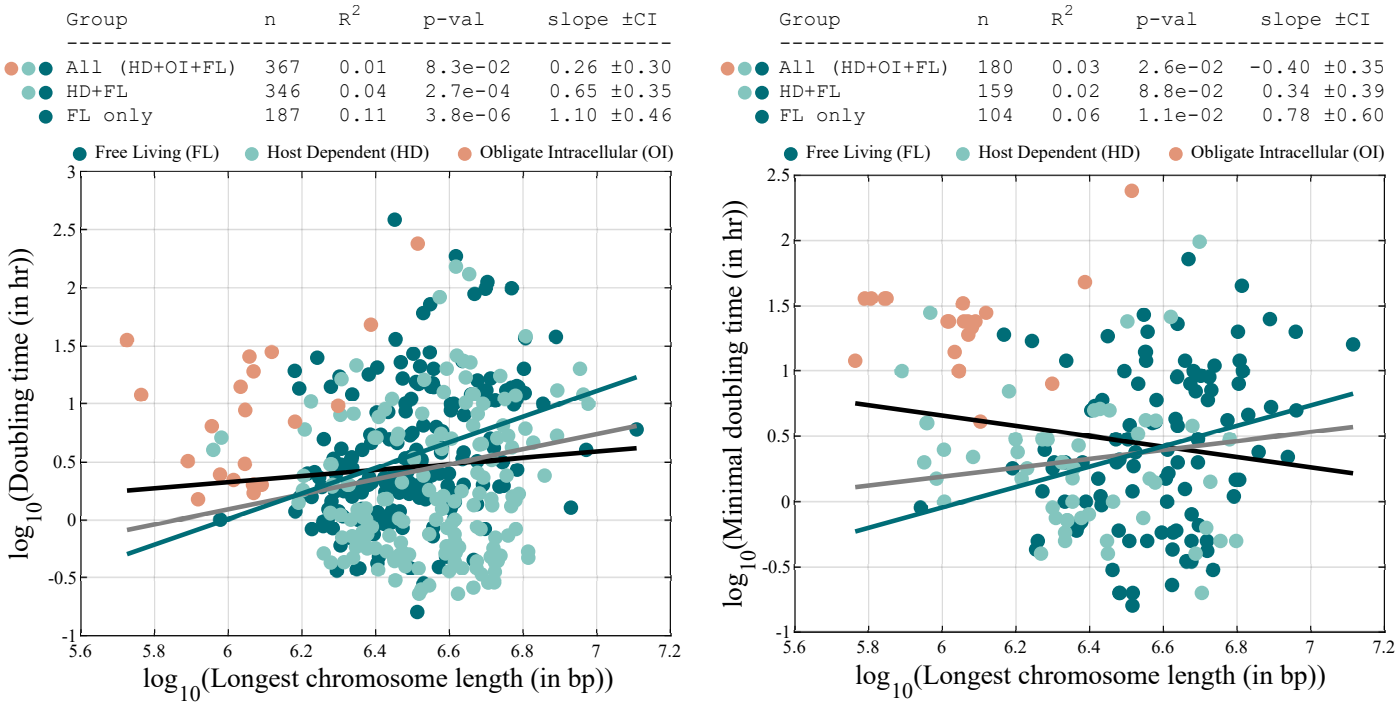
Genome size and growth rate

As a first approximation, we modeled replication time as directly proportional to genome size, assuming a constant replication rate without parallelization. Here, by genome size, we imply all chromosomes present within a single bacterial cell, primary, and secondary chromosomes, if any. In this simplified form,

$$T = b_1 G + b_0 + \epsilon \tag{2}$$

where T is replication time, G is genome size, and b_0 and b_1 are the regression coefficients of the model.

In agreement with previous reports [9, 10, 11, 12], we found a negligible or an absence of correlation between genome size and doubling time when considering the full dataset, $R^2 = 0.00$ for Dataset-I and $R^2 = -0.04$ for Dataset-II, see Fig. 2, Table 1 Model-I. Excluding obligate intracellular species, which are strictly host-dependent bacteria and hence largely insulated from nutrient competition [18], we found an increase in correlation, although it remained insignificant ($p > 0.05$) in Dataset-II. However, when restricting the analysis to free-living bacteria, further excluding facultative host-dependent bacteria, the correlation became more pronounced, $R^2 = 0.09$ in Dataset-I, suggesting that in *competitive* environments, bacterial doubling time is influenced by genome size. This weak correlation arises from a simplified model that ignores parallel replication across chromosomes and replichores, thus overestimates replication time relative to biological reality [19, 20]. On excluding secondary chromosomes and correlating only the size of the *primary* chromosome to the doubling time, we observe a better correlation than that with the total genome size, with $R^2 = 0.01$ for the entire dataset, and $R^2 = 0.11$ for free-living bacteria in Dataset-I and $R^2 = -0.03$ (entire dataset) and $R^2 = 0.06$ (free-living) in Dataset-II, see Fig. 3, Table 1, Model-II.



(a) Model-II on Dataset-I

(b) Model-II on Dataset-II

Figure 3: Relationship between bacterial doubling time and chromosome size in (a) Dataset-I and in (b) Dataset-II. The correlation between doubling time and chromosome size improves compared to that with total genome size. As the secondary chromosomes, ploidies are replicated in parallel [20]; the longest chromosome, typically the primary one, dictates the replication time.

Length of the longest replichore and growth rate

Following the previous analysis, assuming parallel replication across chromosomes and also the replichores, the effective replication time is determined by the longest replichore. Accordingly, we refined the model to

$$T = b_1 R_l + b_0 + \epsilon \tag{3}$$

where R_l is the length of the longest replichore, identified by using purine-pyrimidine compositional skew in the primary chromosome.

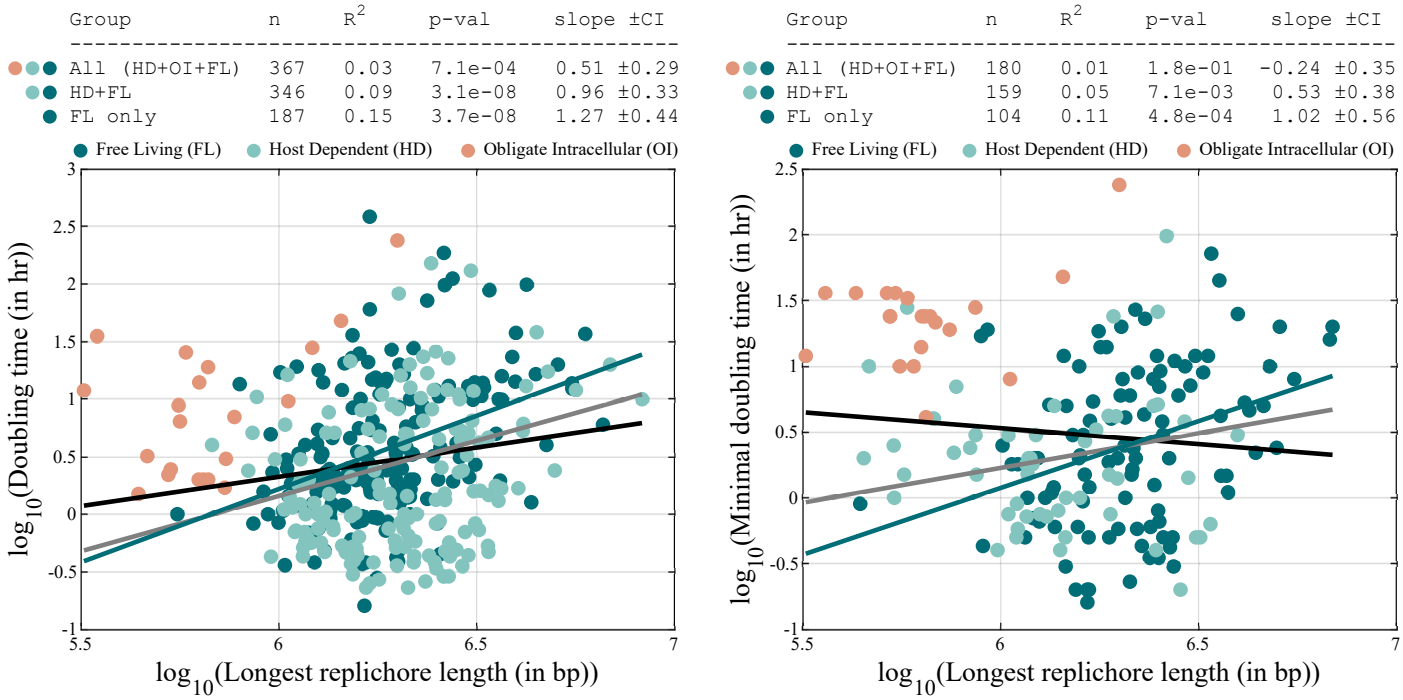
Incorporating replication parallelization improved the model's prediction of doubling time with $R^2 = 0.03$ for all organisms and $R^2 = 0.15$ for free-living bacteria in Dataset-I and $R^2 = -0.01$ for the entire dataset and $R^2 = 0.11$ for free-living bacteria in Dataset-II, see Fig. 4, Table 1, Model-III.

Effect of skew on replication kinetics

Inspired by our previous DNA unzipping simulation [21], which suggested that homogeneous sequences (compositionally skewed sequences) facilitate rapid sequential strand separation, we examined whether such nucleotide compositional asymmetry in bacteria influences their replication dynamics. We define skew as the slope of the genomic segment exhibiting maximum RY-disparity (see Fig. 1). Aligned with previous reports [15, 22], we observed that fast-growing bacteria tend to exhibit stronger purine–pyrimidine skew in their sequence (Fig. 5).

Based on this observation, we hypothesize that nucleotide skew may enhance replication fork kinetics and, consequently, affect bacterial growth rates. To investigate this, we collected reported replication fork speed data for 10 distinct species from 25 different literature sources (see Supplementary Information), and their correlation with nucleotide skew is shown in Fig. 6(b). Due to the limited data size, the co-variation of replication fork speed with nucleotide skew remains inconclusive and requires further analysis.

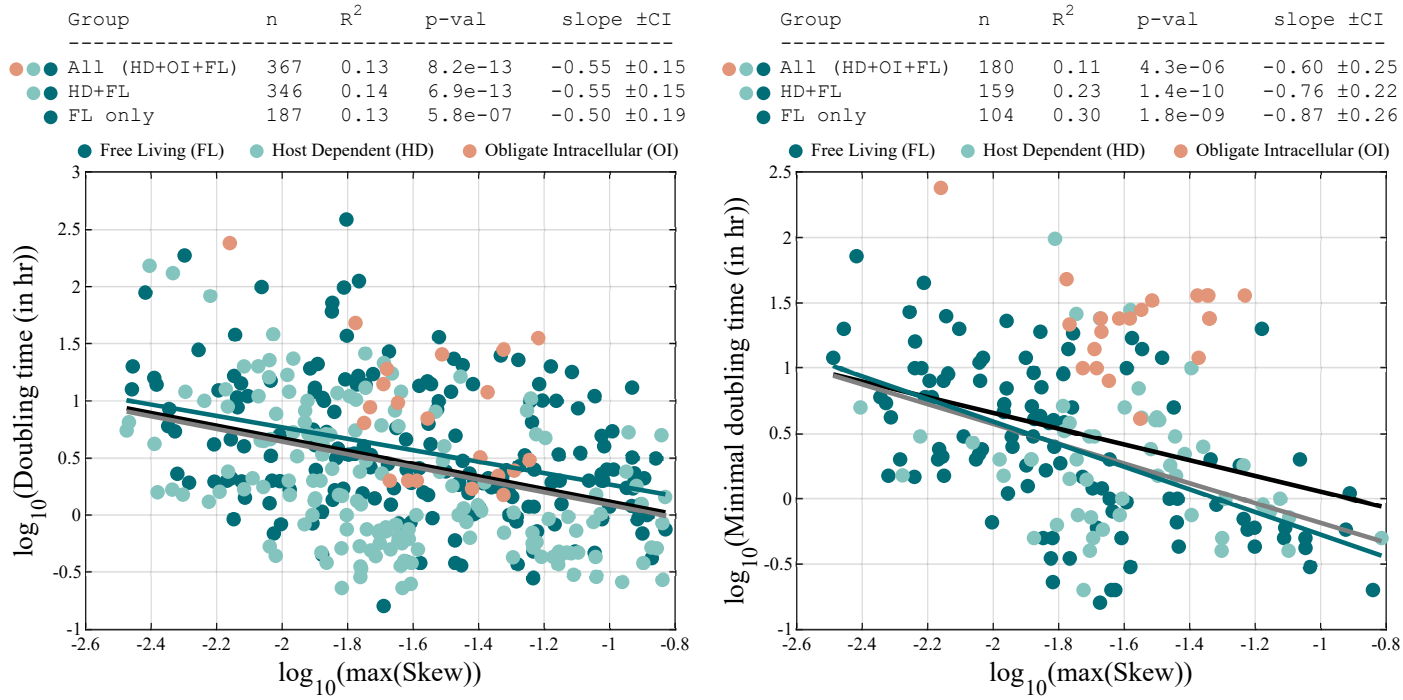
To model the above consideration, we introduce the assumption that the replication fork speed scales with nucleotide skew. The predictor of replication time is now defined as the maximum ratio of replichore length to its corresponding skew (representing fork speed), calculated across both replichores, Model-IV. Under this assumption, replication time



(a) Model-III on Dataset-I

(b) Model-III on Dataset-II

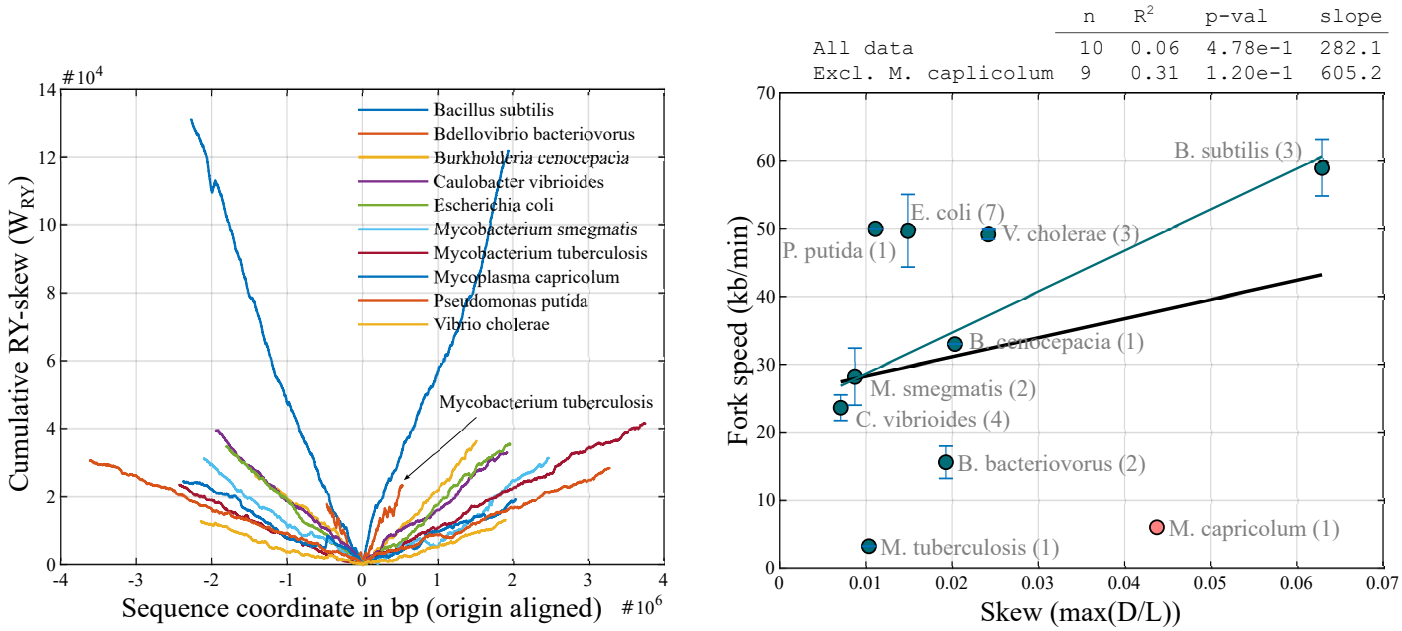
Figure 4: Relationship between bacterial doubling time and the length of the longest replichore in (a) Dataset-I and in (b) Dataset-II. Accounting for replication parallelization leads to a better correlation than when using total genome size, highlighting the importance of replichore architecture in shaping replication dynamics.



(a) Doubling time vs skew in Dataset-I

(b) Doubling time vs skew in Dataset-II

Figure 5: Relationship between bacterial doubling time and maximum purine–pyrimidine skew of the genome in (a) Dataset-I and in (b) Dataset-II. Faster-growing bacteria exhibit stronger compositional skew, suggesting a potential role of skew in replication kinetics.



(a) Skew profile of bacterial chromosome

(b) Replicaiton fork speed vs. nucleotide skew

Figure 6: Correlation between reported replication-fork speed and the maximum purine–pyrimidine skew of the replichores. The principal outlier, *Mycoplasma capricolum*, has a highly reduced genome with very low GC content (¡25%), has lost numerous replication-associated proteins, and lacks *de novo* nucleotide synthesis, relying instead on nucleotide interconversion [23, 24]. Although a moderate correlation is observed when this outlier is excluded, the overall relationship remains inconclusive due to the limited sample size.

can be approximated as

$$T = b_1 \left(\frac{R_i}{S_i} \right)_{\max} + b_0 + \epsilon \tag{4}$$

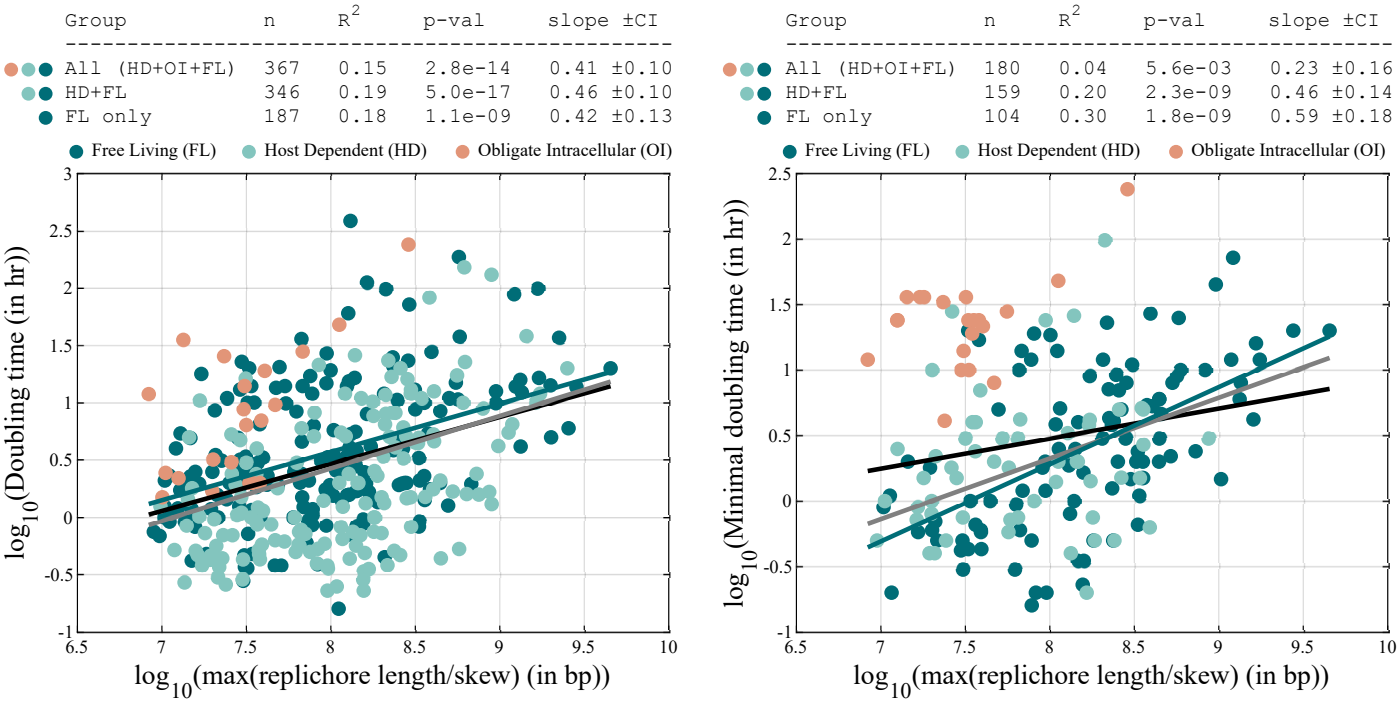
where $(R_i/S_i)_{\max}$ is the maximum of the length-to-skew ratios across the two replichores. Incorporating skew into the regression model improved predictive performance compared with the previous model using replichore size alone, yielding $R^2 = 0.15$ for the total data set and $R^2 = 0.18$ for free-living bacteria in Dataset-I and $R^2 = 0.04$ for the entire dataset and $R^2 = 0.30$ for free-living bacteria in Dataset-II, see Fig. 7 and Table 1, Model-IV.

Test for phylogeny and other confounding factors

We performed additional analyses on Model-IV of Dataset-I to assess the potential influence of confounding factors on the observed correlation. We examined genomic GC content, optimum pH, growth temperature, and optimal temperature using the data provided [13] and found that none of these variables acted as confounders; see the Supplementary Information for the partial correlation analysis. We further observed that although within-clade correlations between the two traits are mostly absent or moderate, correlations between clade-level mean traits, i.e., the geometric mean of the trait values among the extant species within each clade, are noticeably stronger; see Supplementary Information.

Table 2: Comparison of ordinary least squares (OLS) and phylogenetic generalized least squares (PGLS) regression statistics on Dataset-I for the relationship between skew normalized replichore length (predictor) and doubling time (response).

Model-IV Dataset-I	**OLS**		**PGLS** ($\lambda = 0.9$)	
	R^2	Slope $\pm$ SE	R^2	Slope $\pm$ SE
All classes (n = 321)	0.17	0.44 ± 0.05	0.02	0.26 ± 0.10
HD + FL (n = 305)	0.20	0.48 ± 0.06	0.03	0.30 ± 0.11
FL only (n = 166)	0.19	0.44 ± 0.07	0.05	0.42 ± 0.14



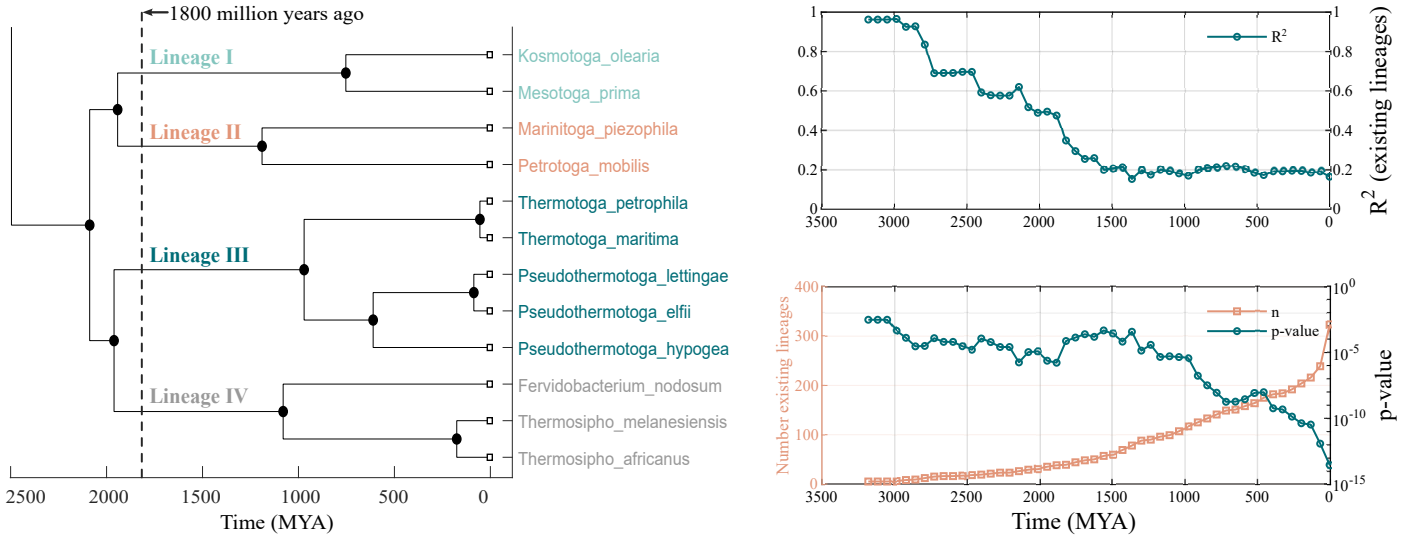
(a) Model-IV on Dataset-I

(b) Model-IV on Dataset-II

Figure 7: Predicted versus observed doubling time under the skew-adjusted model in (a) Dataset-I and in (b) Dataset-II, where replication time is modeled as the ratio of replichore length and the fork speed, with the latter assumed to be proportional to skew. Including nucleotide skew strengthens the predictive power compared to models based solely on genome size or replichore length.

Table 1: The bacterial doubling time is correlated with 4 different predictors. The R^2 and p-values for two datasets under different replication models are presented.

Dataset	Subset	Model I		Model II		Model III		Model IV	
		No parallelization ($T \propto$ genome size)		Parallelization across chromosomes ($T \propto$ chromosome size)		Parallelization across chromosomes & replichores ($T \propto$ longest replichore)		Fork speed depends on RY-skew ($T \propto$ replichore length/skew)	
		R^2	p	R^2	p	R^2	p	R^2	p
Dataset I Total 367 (Madin et al. [13])	All organisms (n = 367)	0.00	3e-01	0.01	8e-02	0.03	7e-04	0.15	3e-14
	Excluding Obligate Intracellular ($n = 346$)	0.02	4e-03	0.04	3e-04	0.09	3e-08	0.19	5e-17
	Free-living ($n = 187$)	0.09	2e-05	0.11	4e-06	0.15	4e-08	0.18	1e-09
Dataset II Total 180 (Vieira et al. [11])	All organisms (n = 180)	-0.04	7e-03	-0.03	3e-02	-0.01	2e-01	0.04	6e-03
	Excluding Obligate Intracellular ($n = 159$)	0.01	3e-01	0.02	9e-02	0.05	7e-03	0.20	2e-09
	Free-living ($n = 104$)	0.03	7e-02	0.06	1e-02	0.11	5e-04	0.30	2e-09



(a) Lineage existing at a time T=1800 MY

(b) Model-IV (on Dataset-I) at different time stamp

Figure 8: Performance of the model over evolutionary time. (a) The ancestral lineages existing at a given evolutionary time slice (shown here at 1800 million years ago, MYA). The predictor and response data for these ancestors are estimated under a Brownian motion model of evolution. (b) Performance of Model-IV is evaluated across evolutionary time, from the present day back to 3200 MYA. For each time slice, the reconstructed ancestral dataset is fitted to the regression model, and the resulting R^2 and p-value of the model are plotted. Going back in time, the decrease in model significance, as indicated by the high p-value, is a result of a decrease in sample size, i.e., the number of clades. The progressive increase in R^2 with deeper evolutionary time suggests that the underlying correlation between traits was stronger in the past and has been progressively obscured by lineage-specific divergence and accumulated stochastic noise over evolutionary history.

To assess the influence of phylogeny on the observed correlations, we assembled divergence-time estimates from TimeTree [25] for 323 of the 367 species in Dataset-I; see Supplementary Information for the phylogenetic tree and Supplementary Data for the Newick file. Two terminals with zero branch lengths were removed to avoid singularities in the phylogenetic covariance matrix C. In addition, we applied a Pagel's λ transformation [26] to the branch lengths to reduce the ill-conditioning of C (initial condition number $\sim 4 \times 10^{10}$). Incorporating phylogenetic structure via Phylogenetic Generalized Least Squares (PGLS) [27] eliminated the correlation (PGLS $R^2 \sim 0.05$) detected under Ordinary Least Squares (OLS) in Model-IV (Table 2).

To better understand this outcome, we quantified phylogenetic signal in both the predictor and the response using Blomberg's K [28], after rescaling the covariance matrix C. We found that the predictor, i.e., skew-normalized replichore length, exhibits a strong phylogenetic signal ($K \approx 1.5$), implying that close relatives are more similar than expected under Brownian motion evolution, whereas the response, doubling time, and its residuals show a weak signal ($K \approx 0.49$). Such data are reported to reduce the reliability and statistical power of PGLS analyses [29, 30]. We further examined the influence of phylogeny on Model-IV by estimating the model's performance in early evolutionary time periods.

To examine ancestral patterns, we estimate trait values for all ancestral lineages (320 internal nodes) in the Dataset-I phylogeny using the observed traits of the 321 extant species. The ancestral states correspond to the set of internal node values that maximize the likelihood of the observed tip data under a Brownian-motion model of trait evolution. These maximum-likelihood (ML) ancestral states can be computed using a closed-form solution in which each internal node value is expressed as a weighted average of the descendant tip values, with weights determined by the variances implied by the branch lengths [31, 32]. We implemented this algorithm to obtain the ML estimates of trait values at all ancestral nodes; the MATLAB script is provided in the Supplementary Data.

The reconstructed ancestral data of the predictor and response at each time slice is then fitted to Model-IV, and the resulting R^2 and p-value are shown as functions of evolutionary time in Fig. 8(b). A consistent increase in R^2 is observed as the evolutionary time deepens, indicating that the association between genome architecture and doubling time in bacteria was stronger on the primitive Earth. The weaker correlations observed in present-day species are likely the result of accumulated lineage-specific divergence and adaptation that progressively obscure the historical signal. The scatter plots of the model fit across 50 time slices are compiled into a video (see Supplementary Information), providing a more intuitive visualization of how the association evolves over deep time.

Discussion

This study examines the relationship between bacterial genome architecture — specifically, genome size, replichore length, and nucleotide compositional skew — and bacterial growth rate. By estimating replication times from these genomic features and comparing them with empirically reported doubling times across a broad set of bacterial species, we present the following three principal outcomes.

First, genome size alone exhibits only a negligible association with doubling time, and this association differs markedly between ecological categories, i.e., obligate intracellular and free-living bacteria show distinct regression slopes and intercepts. Among organisms of comparable genome size, obligate intracellular bacteria generally grow more slowly, likely reflecting the reduced competitive pressures characteristic of their stable host-associated niches. We also observe that the association of genome size is not limited to bacteria; our analysis of archaeal genomes ($n = 153$) also revealed a weak positive correlation ($R^2 = 0.12, p < 1e-05$) between genome size and their doubling time (see Supplementary Information). Second, the length of the longest replichore is a better predictor of growth rate than total genome size, because DNA replication proceeds bidirectionally from the origin, and the slowest-replicating replichore constrains the overall replication time. Third, incorporating nucleotide compositional skew by scaling replichore length with the corresponding skew further improves predictive power. This model outperforms those based solely on genome size, on the longest chromosome, and on the longest replichore (Table 1).

Based on these associations between genome architecture and growth time, we provide some speculative insight on genome evolution.

Adaptive evolution of bacterial genome architecture

Based on the observed association between genome size and doubling time, we speculate that, within a shared ecological setting, bacteria and archaea with smaller genomes may enjoy a replication-time advantage that confers higher fitness, and such an advantage could favor genome reduction. However, because most bacterial genomes are densely packed with protein-coding genes, any reduction in genome size often entails loss of essential functions, including pathways involved in replication fidelity and DNA repair, as observed in Mycoplasma and other minimal-genome bacteria [23, 24]. Thus, bacterial genomes appear to be shaped by opposing pressures to minimize replication time while maintaining genetic functionality.

Beyond overall size, the correlation with the length of the longest replichore in bacteria highlights the importance of genome organization. Because replication proceeds bidirectionally, the slowest-replicating replichore sets the replication time. Based on the observations, we speculate that bacterial genomes may evolve under pressure not only to reduce size but also to balance replichore lengths, minimizing replication delays and enhancing growth efficiency.

Furthermore, to explore whether the correlation arises from closely related taxa contributing redundant information, we reconstructed ancestral states for the dataset and fitted the model to these reconstructed traits. We found that the model fits the ancestral traits more strongly, with predictive strength (R^2) decreasing progressively along the evolutionary tree as successive speciations occur. Based on this observation, we speculate that with the evolution and increasing complexity of organisms, resulting from genetic drift and other selection pressures, the association between growth rate and genome structure may have been screened.

Potential adaptive role of compositional skew

Nucleotide compositional skew has traditionally been interpreted either as a neutral footprint of asymmetric mutational processes—such as cytosine deamination on the lagging strand [33, 34]—or as a consequence of selection for gene orientation on the leading strand [34, 35]. Our results suggest an additional possibility: skew may have functional implications for replication dynamics. We find that fast-growing bacteria tend to exhibit stronger purine–pyrimidine asymmetry, and models incorporating skew better predict doubling times. This raises the hypothesis that compositional skew might facilitate faster replication fork progression and thereby contribute to faster growth. Although we observe a modest correlation between skew and experimentally measured replication fork speeds (Fig. 6b), current data are limited, and larger datasets will be required to investigate their relationship.

The influence of skew on replication fork progression speed can be explained using the "Asymmetric cooperative model", which was originally proposed to explain the evolutionary advantages of unidirectional strand construction during DNA replication and antiparallel orientation of the two strands of dsDNA. An asymmetric kinetic nearest neighbor interaction of base pairs leads to faster strand separation kinetics of a skewed sequence within the above model, thereby facilitating replication and, consequently, becoming evolutionarily advantageous [21, 36]. Wanrooij et al. also

showed that simply swapping the two arms of a highly skewed origin sequence, changing the skew profile of the sequence from a $\vee$-shape to a $\wedge$-shape, abolishes origin functionality [37]. Although this result comes from a non-bacterial system, it demonstrates that changes in compositional asymmetry alone can have mechanistic consequences for replication dynamics. In bacteria, however, any influence of skew on replication fork progression remains speculative and requires direct experimental validation.

Limitations and future directions

Several simplifications were made in this analysis. The classification of bacteria into free-living versus host-dependent was based on a literature survey and may not fully capture ecological diversity. The assumption that replication fork speed scales linearly with skew is also a crude approximation, as zero skew does not imply zero fork speed. More refined models of fork kinetics, potentially incorporating polymerase dynamics and overlapping rounds of replication under fast growth, are needed. Furthermore, while we used RY cumulative skew to identify replication origins, GC or amino-keto (MK) skew may provide stronger predictions in some genomes. Although the primary outcome, that incorporating skew improves predictive power, was robust across RY, GC, and MK definitions, future work should systematically compare these measures. Understanding the mechanistic basis of these differences will be essential for linking compositional biases to DNA replication dynamics and bacterial fitness.

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 additional analysis performed in this work is provided in Supplementary Information.

The scatter plots of the model-IV fit across 50 time slices are compiled into a video and is provided as `TimeEvolutionModel_IV.mp` in the Supplementary Data (GitHub repository, See below).

Data Availability

The data and the MATLAB script used in this analysis are publicly available at the following GitHub repository: `https://github.com/ParthaTbio/Genome_skew_analysis`.

References

[1] Michael Lynch and John S Conery. "The origins of genome complexity". In: *science* 302.5649 (2003), pp. 1401–1404.

[2] Alexis Dufresne, Laurence Garczarek, and Frédéric Partensky. "Accelerated evolution associated with genome reduction in a free-living prokaryote". In: *Genome biology* 6.2 (2005), R14.

[3] Stephen J Giovannoni et al. "Genome streamlining in a cosmopolitan oceanic bacterium". In: *science* 309.5738 (2005), pp. 1242–1245.

[4] Anna K Simonsen. "Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria". In: *The ISME Journal* 16.2 (2022), pp. 423–434.

[5] Brandon K Swan et al. "Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean". In: *Proceedings of the National Academy of Sciences* 110.28 (2013), pp. 11463–11468.

[6] Hongwei Liu et al. "Warmer and drier ecosystems select for smaller bacterial genomes in global soils". In: *Imeta* 2.1 (2023), e70.

[7] Stephen J Giovannoni, J Cameron Thrash, and Ben Temperton. "Implications of streamlining theory for microbial ecology". In: *The ISME journal* 8.8 (2014), pp. 1553–1565.

[8] Niv Sabath et al. "Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation". In: *Genome biology and evolution* 5.5 (2013), pp. 966–977.

[9] 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.

[10] Marie Touchon and Eduardo PC Rocha. "Causes of insertion sequences abundance in prokaryotic genomes". In: *Molecular biology and evolution* 24.4 (2007), pp. 969–981.

[11] Sara Vieira-Silva and Eduardo PC Rocha. "The systemic imprint of growth and its uses in ecological (meta) genomics". In: *PLoS genetics* 6.1 (2010), e1000808.

[12] Mark Westoby et al. "Cell size, genome size, and maximum growth rate are near-independent dimensions of ecological variation across bacteria and archaea". In: *Ecology and Evolution* 11.9 (2021), pp. 3956–3976.

[13] Joshua S Madin et al. "A synthesis of bacterial and archaeal phenotypic trait data". In: *Scientific data* 7.1 (2020), p. 170.

[14] Andrei Grigoriev. "Analyzing genomes with cumulative skew diagrams". In: *Nucleic acids research* 26.10 (1998), pp. 2286–2290.

[15] Peder Worning et al. "Origin of replication in circular prokaryotic chromosomes". In: *Environmental microbiology* 8.2 (2006), pp. 353–361.

[16] James M Freeman et al. "Patterns of genome organization in bacteria". In: *Science* 279.5358 (1998), pp. 1827–1827.

[17] Feng Gao and Chun-Ting Zhang. "DoriC: a database of oriC regions in bacterial genomes". In: *Bioinformatics* 23.14 (2007), pp. 1866–1867.

[18] Lisa Klasson and Siv GE Andersson. "Evolution of minimal-gene-sets in host-dependent bacteria". In: *TRENDS in Microbiology* 12.1 (2004), pp. 37–43.

[19] Caroline Stokke, Torsten Waldminghaus, and Kirsten Skarstad. "Replication patterns and organization of replication forks in Vibrio cholerae". In: *Microbiology* 157.3 (2011), pp. 695–708.

[20] Florian Fournes et al. "Replicate once per cell cycle: replication control of secondary chromosomes". In: *Frontiers in microbiology* 9 (2018), p. 1833.

[21] 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* 92.6 (2024), pp. 761–775.

[22] Kazuharu Arakawa and Masaru Tomita. "Measures of compositional strand bias related to replication machinery and its applications". In: *Current genomics* 13.1 (2012), pp. 4–15.

[23] Angamuthu Kandavelmani and Shanmughavel Piramanayagam. "Comparative genomics of mycoplasma: Insights on genome reduction and identification of potential antibacterial targets". In: *Biomedical and Biotechnology Research Journal (BBRJ)* 3.1 (2019), pp. 9–18.

[24] Alana Mitchell and Lloyd R Finch. "Pathways of nucleotide biosynthesis in Mycoplasma mycoides subsp. mycoides". In: *Journal of Bacteriology* 130.3 (1977), pp. 1047–1054.

[25] Sudhir Kumar et al. "TimeTree 5: an expanded resource for species divergence times". In: *Molecular biology and evolution* 39.8 (2022), msac174.

[26] Mark Pagel. "Inferring the historical patterns of biological evolution". In: *Nature* 401.6756 (1999), pp. 877–884.

[27] Shana R Lavin et al. "Morphometrics of the avian small intestine compared with that of nonflying mammals: a phylogenetic approach". In: *Physiological and biochemical zoology* 81.5 (2008), pp. 526–550.

[28] Simon P Blomberg, Theodore Garland Jr, and Anthony R Ives. "Testing for phylogenetic signal in comparative data: behavioral traits are more labile". In: *Evolution* 57.4 (2003), pp. 717–745.

[29] Liam J Revell. "Phylogenetic signal and linear regression on species data". In: *Methods in Ecology and Evolution* 1.4 (2010), pp. 319–329.

[30] Mark Westoby et al. "Phylogenetically conservative trait correlation: Quantification and interpretation". In: *Journal of Ecology* 111.10 (2023), pp. 2105–2117.

[31] Liam J Revell. "Ancestral state reconstruction of phenotypic characters". In: *Evolutionary Biology* (2025), pp. 1–25.

[32] F James Rohlf. "Comparative methods for the analysis of continuous variables: geometric interpretations". In: *Evolution* 55.11 (2001), pp. 2143–2160.

[33] 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.

[34] Jean R Lobry and Noboru Sueoka. "Asymmetric directional mutation pressures in bacteria". In: *Genome biology* 3.10 (2002), research0058–1.

[35] Eduardo PC Rocha. "Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes?" In: *Trends in microbiology* 10.9 (2002), pp. 393–395.

[36] Hemachander Subramanian and Robert A Gatenby. "Evolutionary advantage of anti-parallel strand orientation of duplex DNA". In: *Scientific Reports* 10.1 (2020), p. 9883.

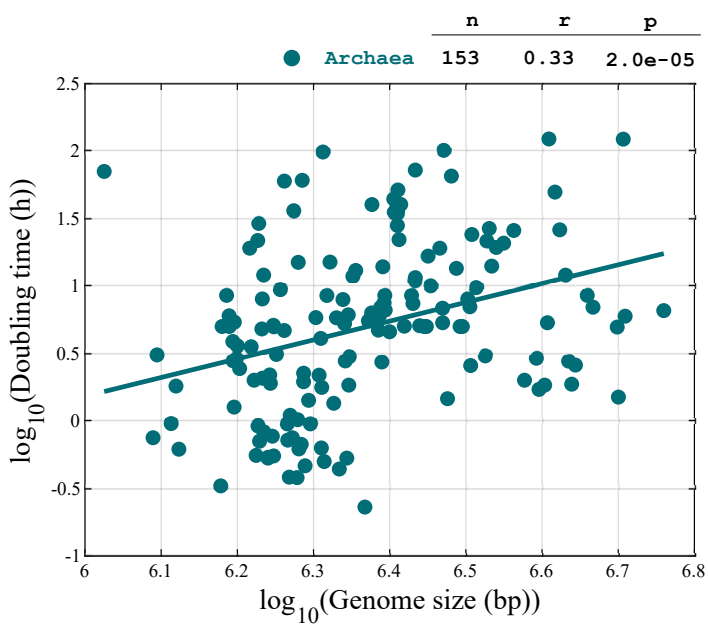
[37] Sjoerd Wanrooij et al. "In vivo mutagenesis reveals that OriL is essential for mitochondrial DNA replication". In: *The EMBO Reports* 13.12 (2012), pp. 1130–1137.

Supplementary Information

This document accompanies the main article "Genome size and nucleotide skews as predictors of bacterial growth rate". The study investigates the influence of genome size, chromosome organization, replichore length, and nucleotide compositional skew on replication kinetics and, consequently, on bacterial growth rates. Additional analyses on Dataset-I that are not included in the main text are presented here.

Archaeal doubling time and genome size

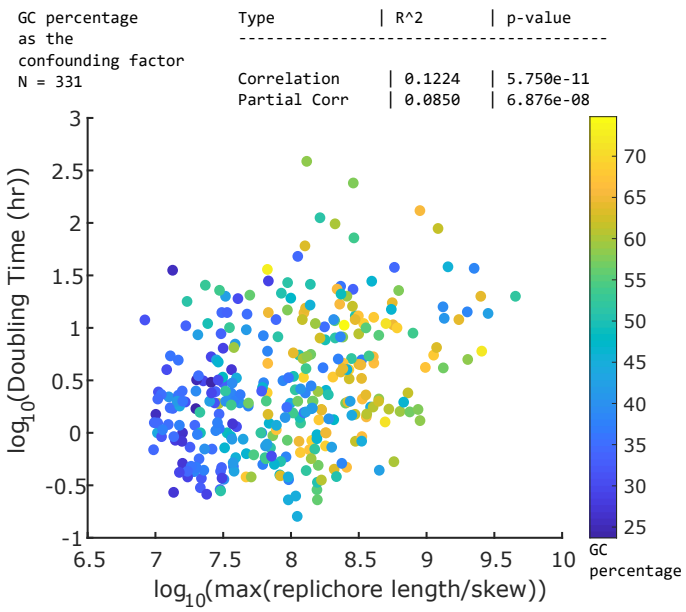
The relationship between archaeal genome size ($n = 153$) and doubling time is shown in Supplementary Fig. S1. A moderate correlation ($R^2 = 0.12$) suggests that genome size contributes to replication kinetics and, consequently, influences growth rate. A stronger correlation is expected when the replichores and their skew are considered rather than the total genome size.



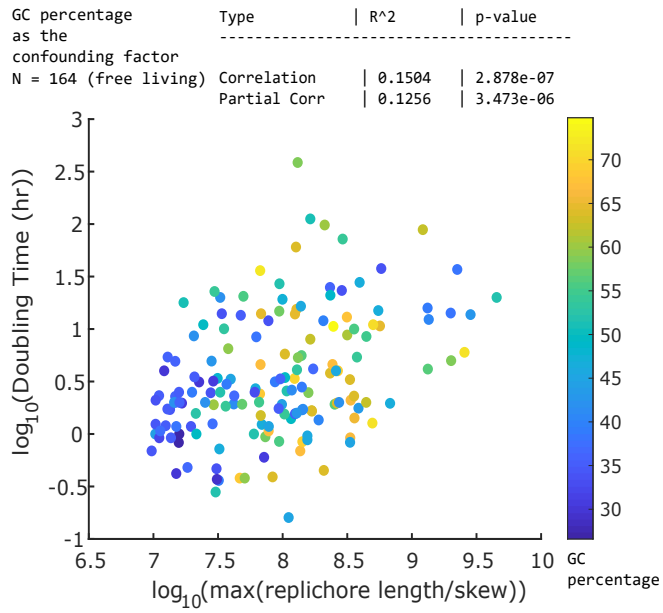
Supplementary Figure S1: Correlation between archaeal doubling time and total genome size. A moderate association is observed even without accounting for replication parallelization across replichores. A stronger correlation is expected when considering the longest replichore rather than the total genome size.

Partial correlation analysis

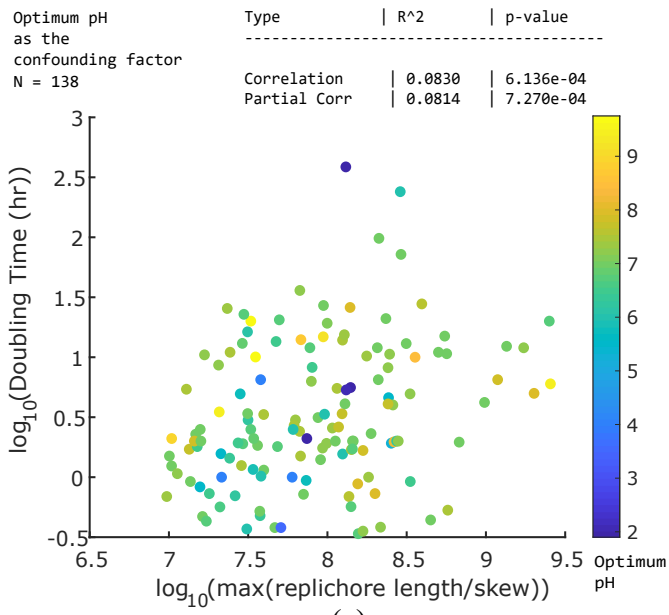
To test whether the observed correlation between the doubling time and the skew normalized replichore length is confounded by other variables, we used the additional data provided in the Dataset-I [13] and performed partial correlation analysis controlling for GC percentage, optimum pH, growth temperature, and optimum temperature, see Supplementary Fig. S2. The analysis is performed on the full data set (Supplementary Fig.S2, left panel) and separately on the free-living ones (Supplementary Fig.S2, right panel). None of these variables shows evidence of confounding the observed association between bacterial doubling time and the skew-normalized replichore length.


GC percentage
as the
confounding factor
N = 331
Type | R^2 | p-value
Correlation | 0.1224 | 5.750e-11
Partial Corr | 0.0850 | 6.876e-08
log10(Doubling Time (hr))
log10(max(replichore length/skew))
GC
percentage

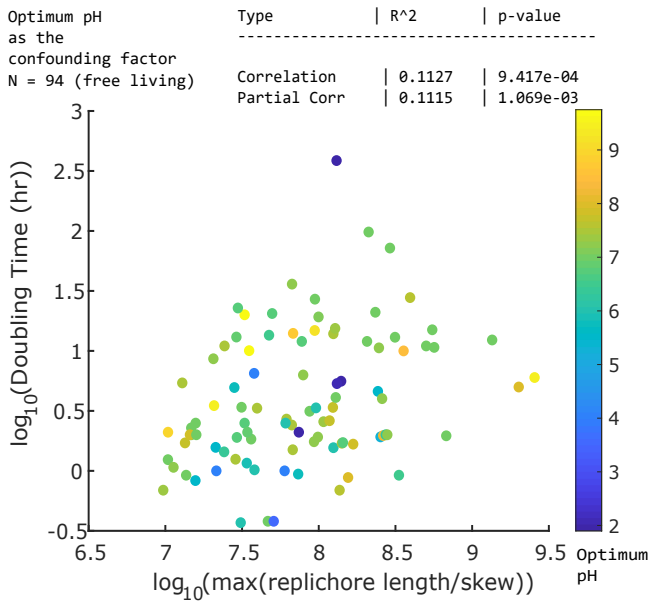
(a)


GC percentage
as the
confounding factor
N = 164 (free living)
Type | R^2 | p-value
Correlation | 0.1504 | 2.878e-07
Partial Corr | 0.1256 | 3.473e-06
log10(Doubling Time (hr))
log10(max(replichore length/skew))
GC
percentage

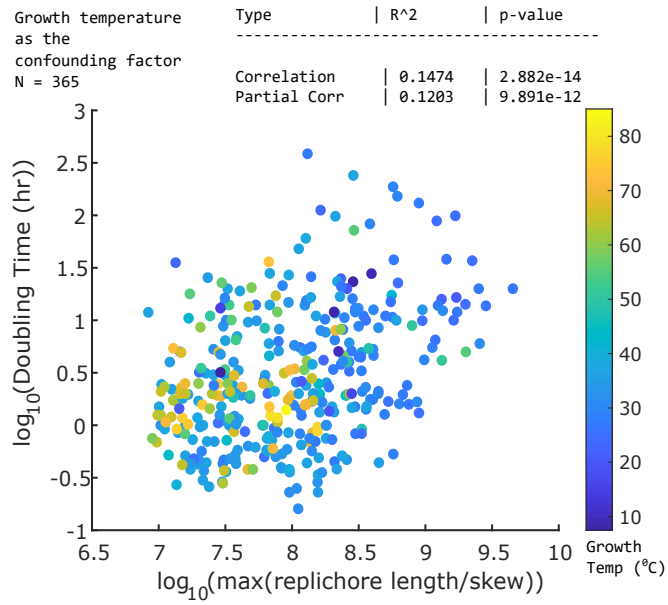
(b)


Optimum pH
as the
confounding factor
N = 138
Type | R^2 | p-value
Correlation | 0.0830 | 6.136e-04
Partial Corr | 0.0814 | 7.270e-04
log10(Doubling Time (hr))
log10(max(replichore length/skew))
Optimum
pH

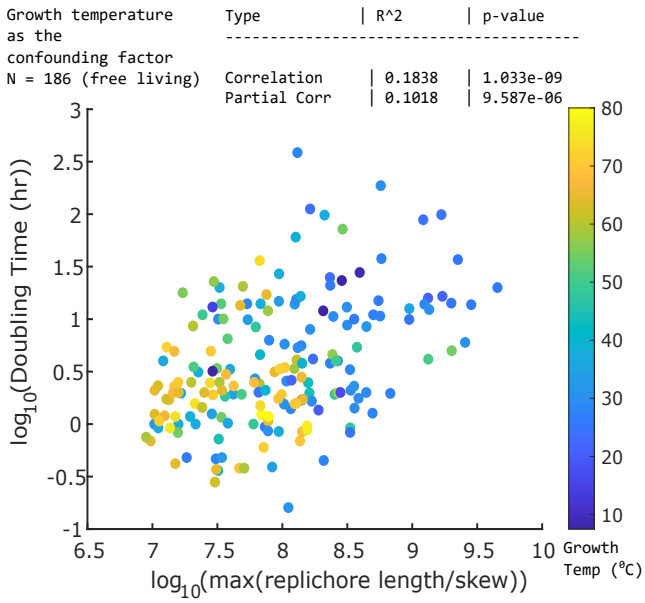
(c)


Optimum pH
as the
confounding factor
N = 94 (free living)
Type | R^2 | p-value
Correlation | 0.1127 | 9.417e-04
Partial Corr | 0.1115 | 1.069e-03
log10(Doubling Time (hr))
log10(max(replichore length/skew))
Optimum
pH

(d)


Growth temperature
as the
confounding factor
N = 365
Type | R^2 | p-value
Correlation | 0.1474 | 2.882e-14
Partial Corr | 0.1203 | 9.891e-12
log10(Doubling Time (hr))
log10(max(replichore length/skew))
Growth
Temp (°C)

(e)


Growth temperature
as the
confounding factor
N = 186 (free living)
Type | R^2 | p-value
Correlation | 0.1838 | 1.033e-09
Partial Corr | 0.1018 | 9.587e-06
log10(Doubling Time (hr))
log10(max(replichore length/skew))
Growth
Temp (°C)

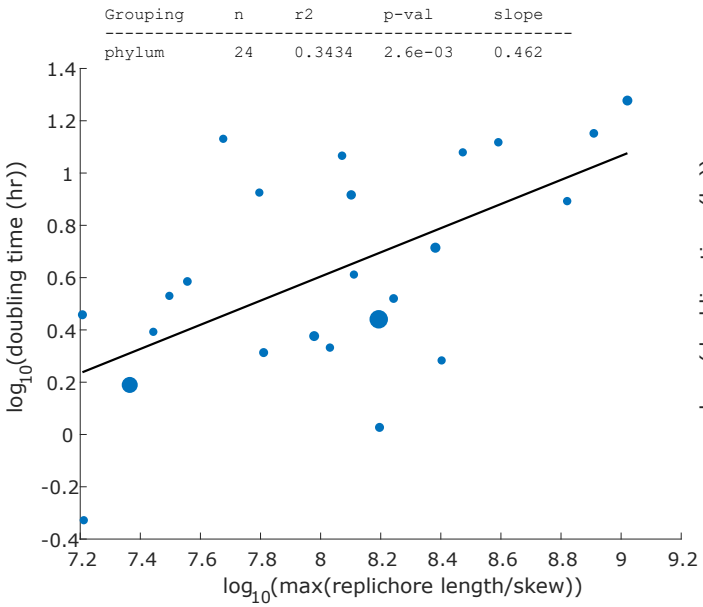
(f)

Intra-clade and inter-clade comparison

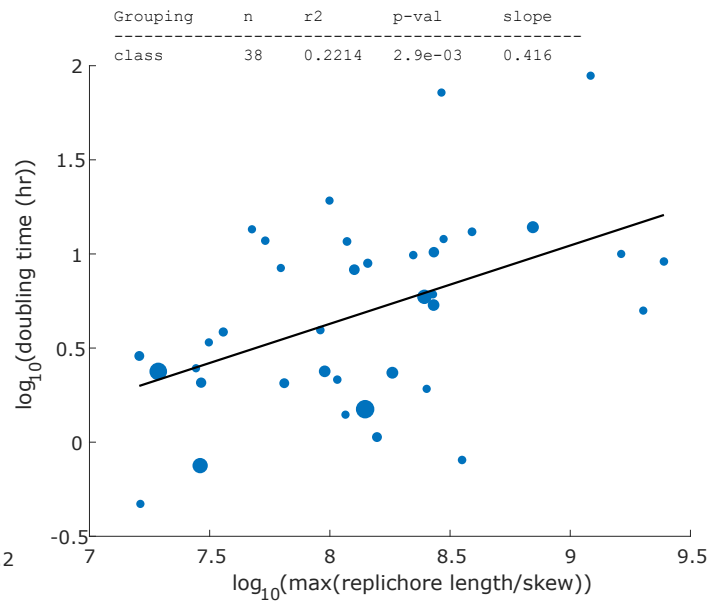
Intra-clade correlation was evaluated among species belonging to the same clade, and the results are presented in Supplementary Fig. S4. To compare across clades, we assigned each clade a representative trait value X_j by computing the geometric mean of the trait values of all extant species within that clade. The resulting inter-clade correlations are shown in Supplementary Fig. S3.

$$X_j = \left(\prod_{i \in C_j} x_i \right)^{\frac{1}{|C_j|}},$$

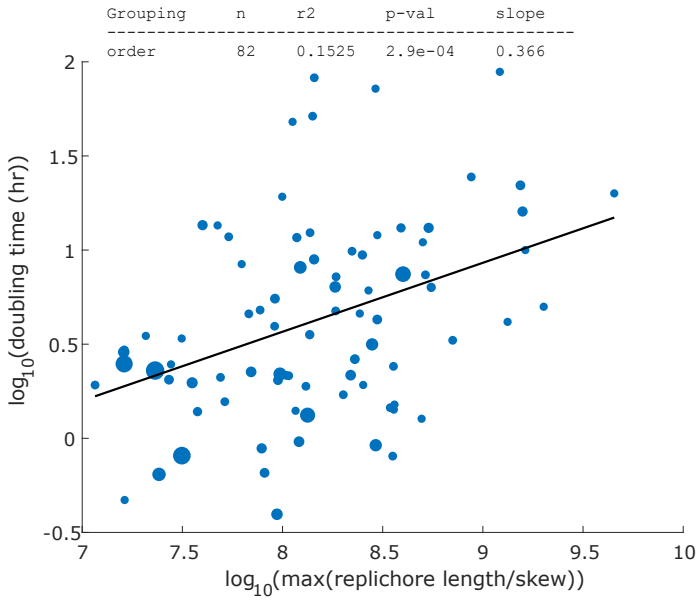
where X_j denotes the trait value assigned to clade j, and C_j is the set of species belonging to clade j.



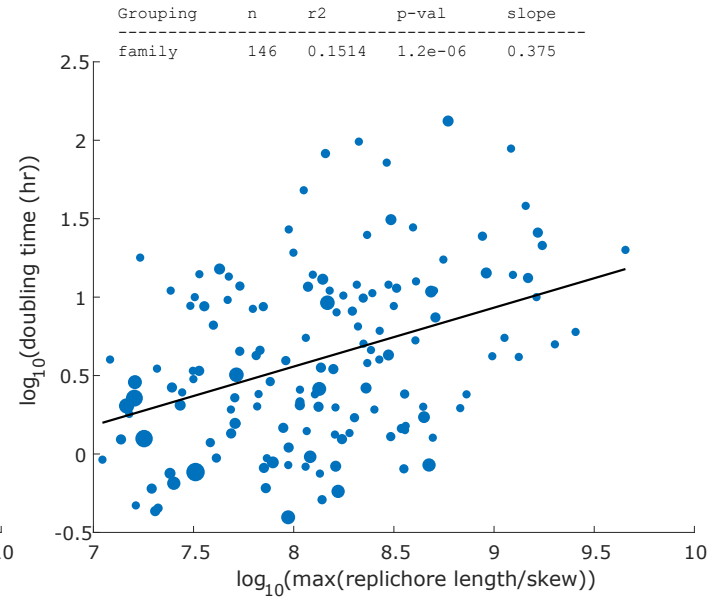
(a) data averged within each phylum



(a) data averged within each class

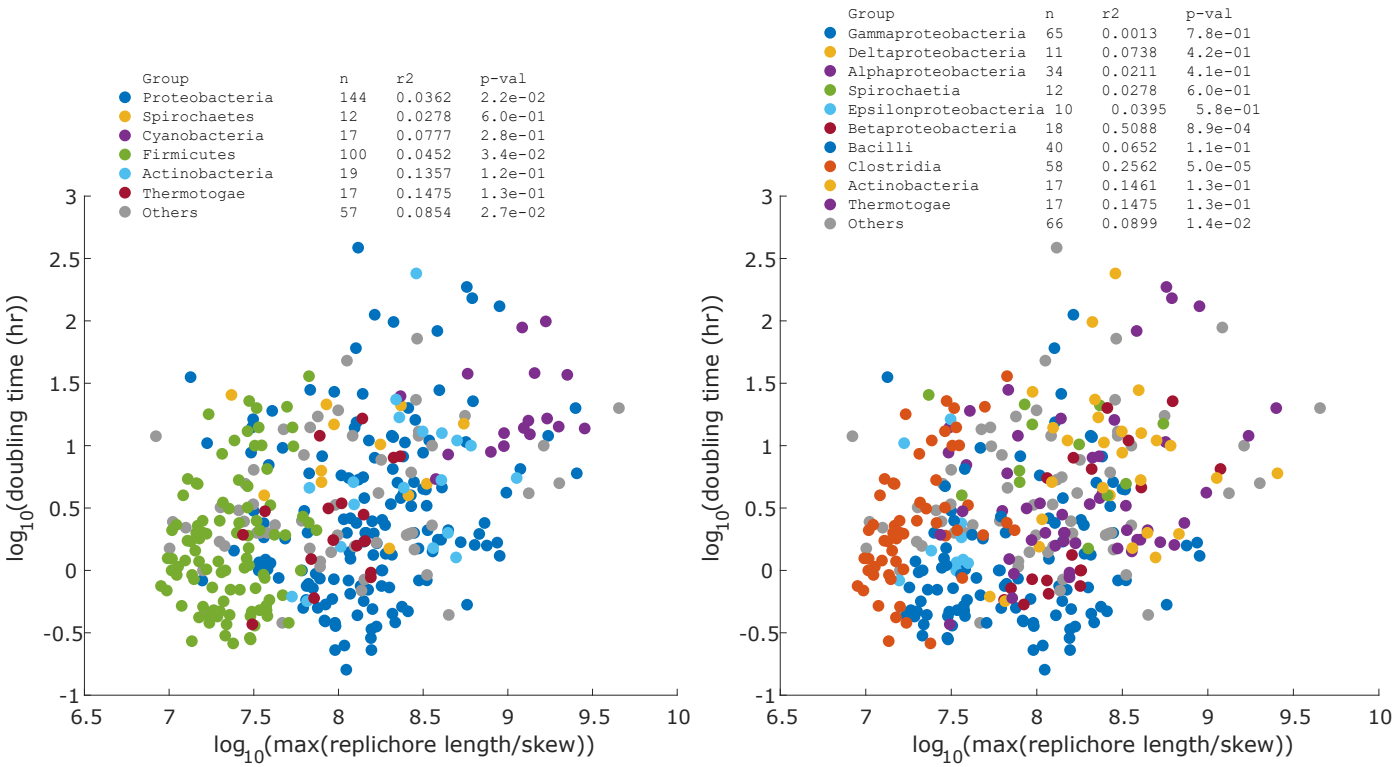


(a) data averged within each order



(a) data averged within each family

Supplementary Figure S3: Among clade correlation. The association between skew-normalized replichore length and growth rate becomes progressively stronger, as reflected by increasing R^2, when examined at broader taxonomic levels (family, order, class, phylum). This pattern motivated us to evaluate the model's fit over deeper evolutionary timescales.



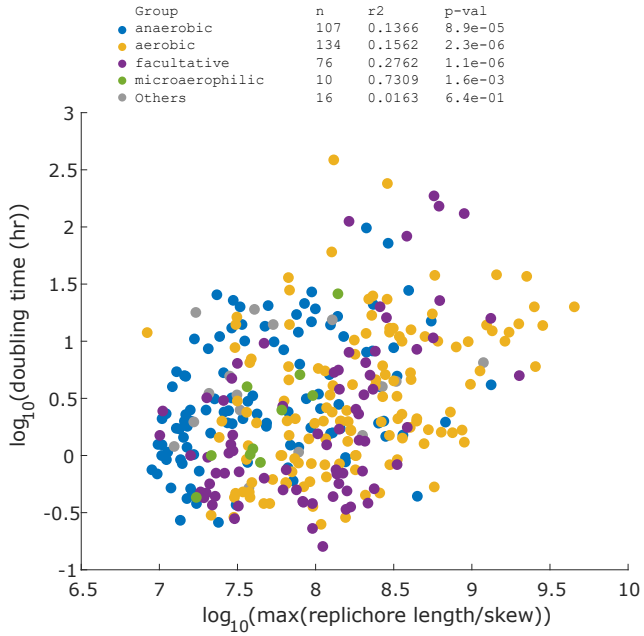
(a) Grouped by Phylum

(b) Grouped by class

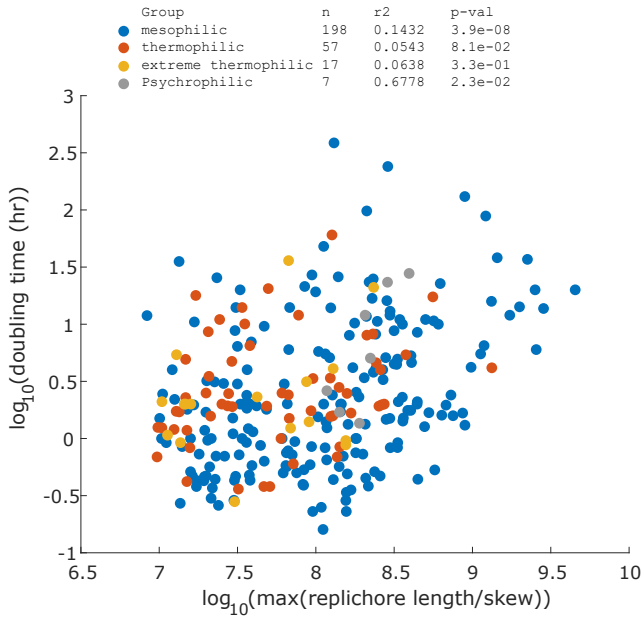
Supplementary Figure S4: Within clade correlation. The association between skew-normalized replichore length and growth rate is not observed to be stronger within individual clades, with the exception of the classes Clostridia and Betaproteobacteria.

Stratification Effects on the Correlation

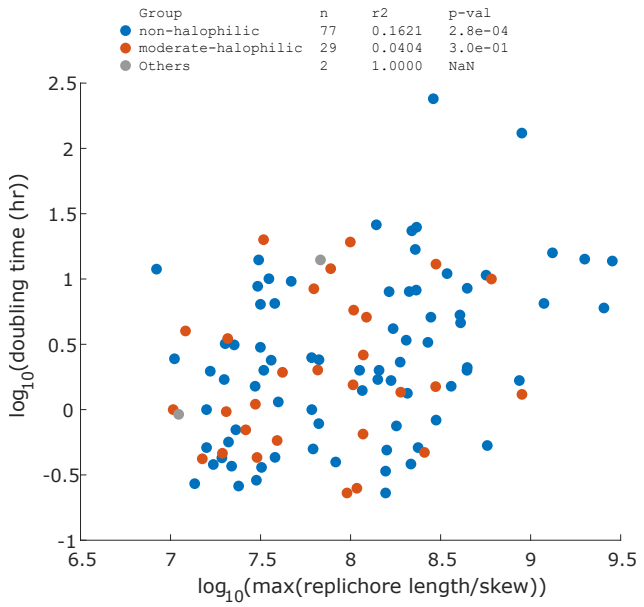
To assess whether the observed correlation could be an artifact of hidden stratification (e.g., a Simpson's paradox–type effect), we examined the relationship within multiple taxonomic and ecological clusters. We analyzed correlations within Families, Phyla (see Supplementary Fig. S4), metabolic categories, salinity ranges, temperature ranges, and Gram stain groups (see Supplementary Fig. S5). None of these clusters showed an opposite (negative) trend. Thus, the positive association in the full dataset is consistent across diverse subgroups and is unlikely to result from phylogenetic or ecological stratification.



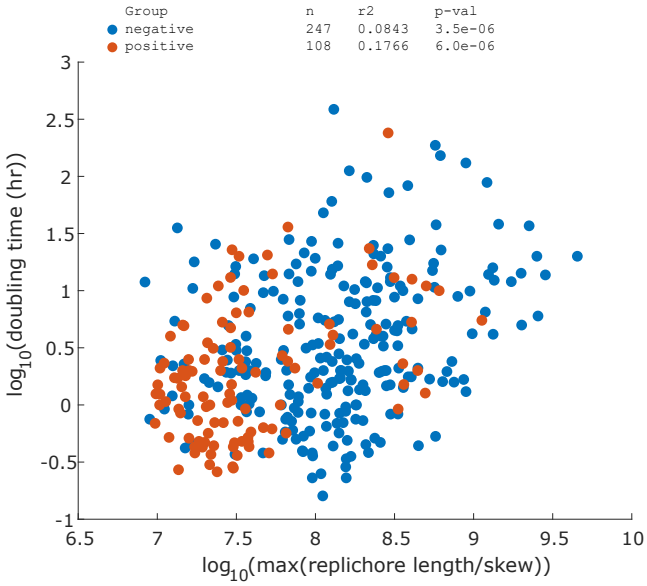
(a) Grouped by metabolism



(b) Grouped by temperature range



(c) Grouped by salinity range

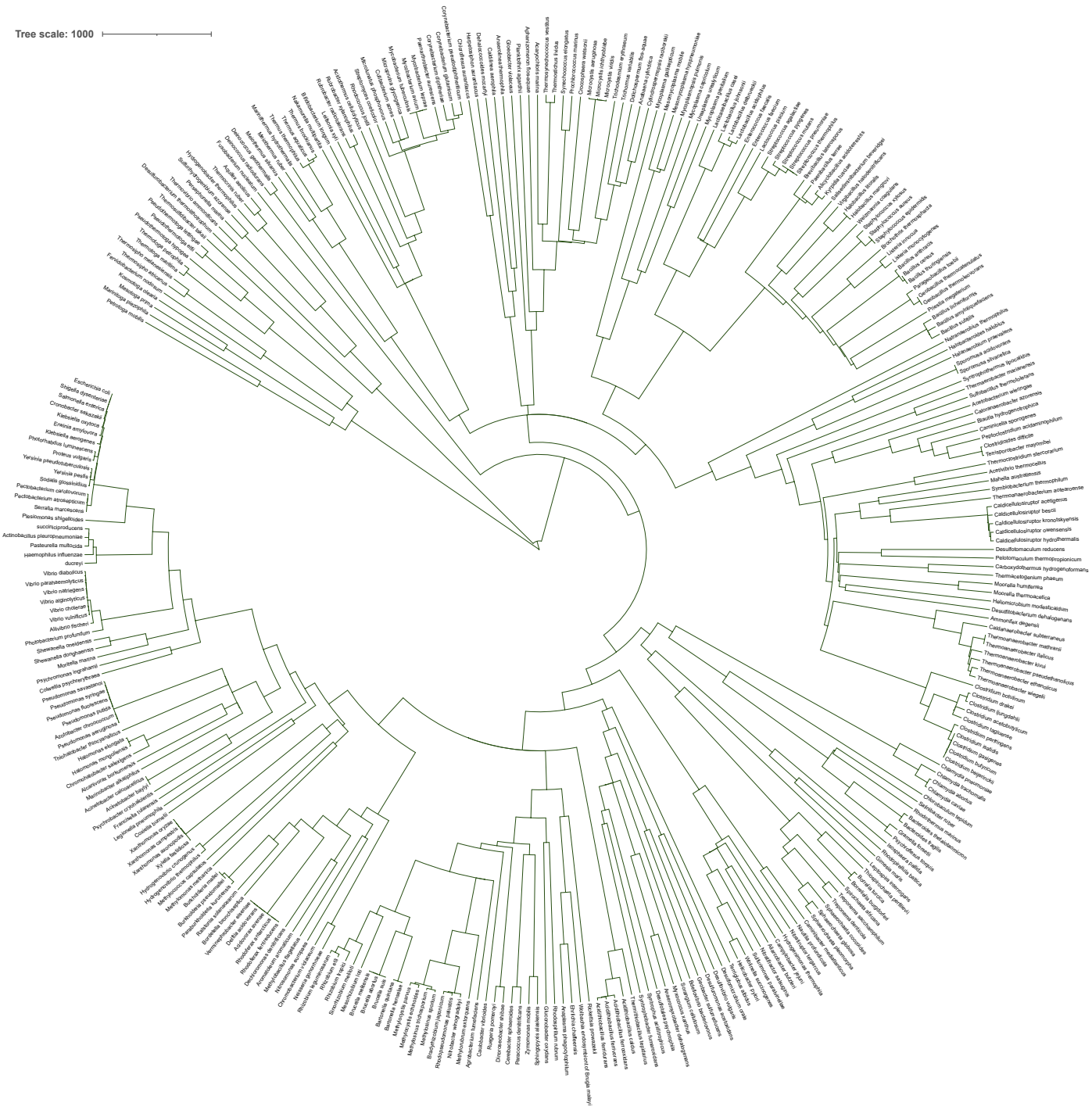


(d) Grouped by gram stain

Supplementary Figure S5: Stratification Effects on the Correlation. Data were clustered by metabolic category, salinity range, temperature range, and Gram stain group. None of these subgroups exhibited an opposite trend to that of the global positive association between skew-normalized replichore length and growth rate. The positive association observed in the full dataset is therefore consistent across diverse ecological and physiological subsets, suggesting that it is unlikely to arise from ecological stratification.

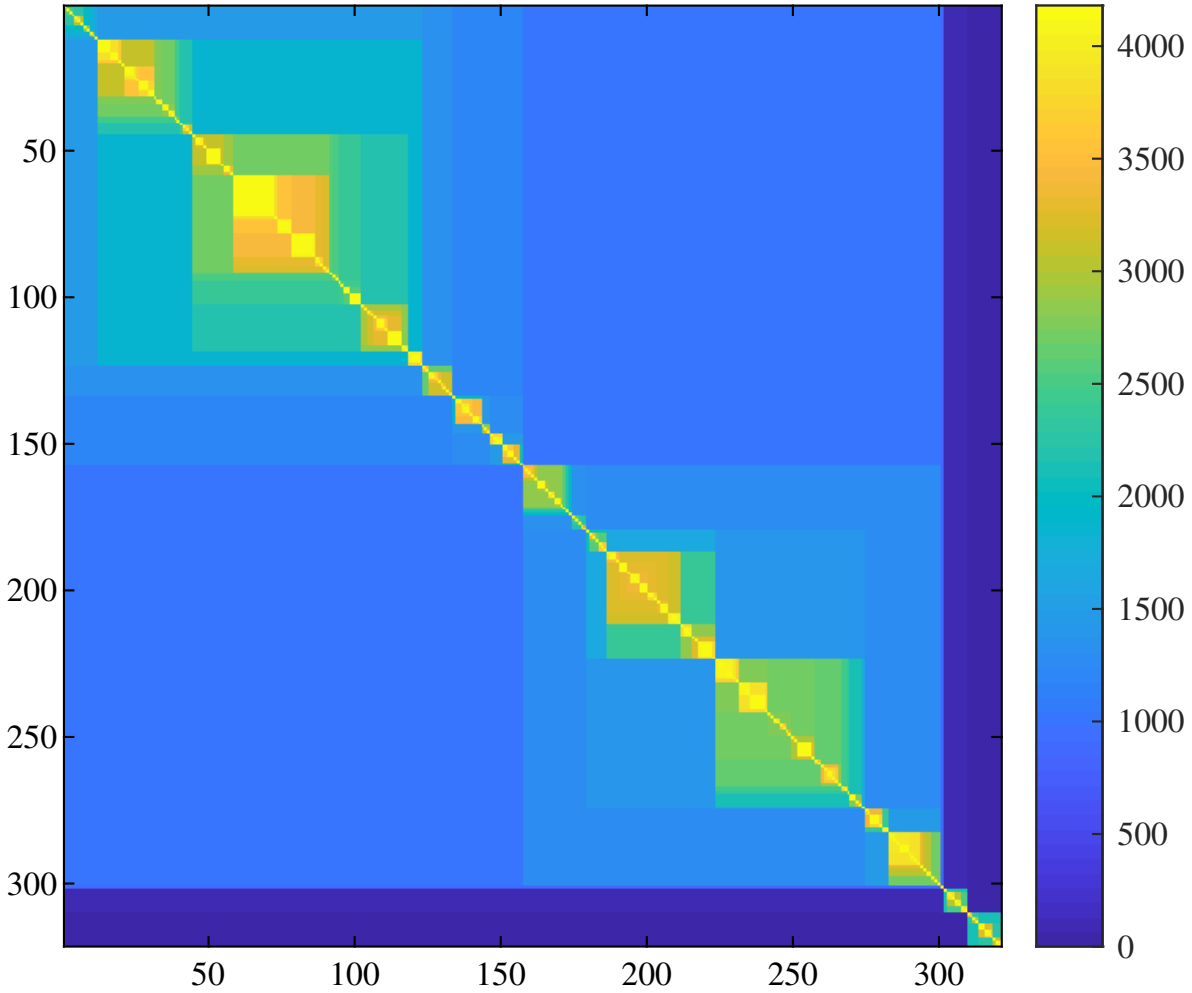
Phylogenetic tree

The phylogenetic time history data of 323 species of Dataset-I imported from TimeTree [25] is shown in Fig. S6. The tree in Newick format is provided in Supplementary Data.



Supplementary Figure S6: Phylogenetic tree with time history data for 323 species of the Dataset-I imported from TimeTree [25] is shown.

The variance-covariance matrix of the phylogenetic tree is presented as a heat map, Supplementary Fig. S7.



Supplementary Figure S7: Variance-Covariance matrix corresponding to the phylogenetic tree, after a branch length transformation using Pagel's $\lambda = 0.9$.

Regression Statistics

The regression coefficients b_0 and b_1 representing the slope and intercept of the regression plots, for all 4 models for both Dataset-I and Dataset-II, are provided here in Supplementary Table S1 and Supplementary Table S2, respectively.

Supplementary Table S1: Ordinary least squares (OLS) regression statistics for the relationship between doubling time (response) and different genomic predictors (Models I–IV) for Dataset-I. Values shown are coefficient of determination (R^2), sample size (n), p-value, and regression slope and intercept with their 95% confidence interval (estimate $\pm$ CI).

Dataset-I	n	R^2	p**-value**	$b_0 \pm$ **SE**	$b_1 \pm$ **SE**
Model I (Predictor: Genome size)					
All	367	0.003	2.9×10^{-1}	-0.58 ± 1.93	0.16 ± 0.30
HD+FL	346	0.023	4.4×10^{-3}	-2.86 ± 2.26	0.50 ± 0.35
FL only	187	0.095	1.8×10^{-5}	-6.10 ± 2.99	1.02 ± 0.46
Model II (Predictor: Longest chromosome)					
All	367	0.008	8.3×10^{-2}	-1.26 ± 1.95	0.26 ± 0.30
HD+FL	346	0.038	2.7×10^{-4}	-3.80 ± 2.26	0.65 ± 0.35
FL only	187	0.109	3.8×10^{-6}	-6.61 ± 2.97	1.10 ± 0.46
Model III (Predictor: Longest replichore)					
All	367	0.031	7.1×10^{-4}	-2.73 ± 1.83	0.51 ± 0.29
HD+FL	346	0.085	3.1×10^{-8}	-5.61 ± 2.10	0.96 ± 0.33
FL only	187	0.152	3.7×10^{-8}	-7.42 ± 2.74	1.27 ± 0.44
Model IV (Predictor: Skew-normalized replichore)					
All	367	0.147	2.8×10^{-14}	-2.81 ± 0.81	0.41 ± 0.10
HD+FL	346	0.185	5.0×10^{-17}	-3.23 ± 0.82	0.46 ± 0.10
FL only	187	0.183	1.1×10^{-9}	-2.80 ± 1.03	0.42 ± 0.13

Supplementary Table S2: Ordinary least squares (OLS) regression statistics for the relationship between doubling time (response) and different genomic predictors (Models I–IV) for Dataset-II. Values shown are coefficient of determination (R^2), sample size (n), p-value, and regression slope and intercept with their 95% confidence interval (estimate $\pm$ CI).

Dataset-II	n	R^2	p**-value**	$b_0 \pm$ **SE**	$b_1 \pm$ **SE**
Model I (Predictor: Genome size)					
All	180	0.040	6.9×10^{-3}	3.51 ± 2.19	-0.47 ± 0.34
HD+FL	159	0.008	2.5×10^{-1}	-1.10 ± 2.52	0.23 ± 0.39
FL only	104	0.032	7.0×10^{-2}	-3.23 ± 3.94	0.55 ± 0.60
Model II (Predictor: Longest chromosome)					
All	180	0.028	2.6×10^{-2}	3.04 ± 2.25	-0.40 ± 0.35
HD+FL	159	0.018	8.8×10^{-2}	-1.85 ± 2.55	0.34 ± 0.39
FL only	104	0.062	1.1×10^{-2}	-4.73 ± 3.92	0.78 ± 0.60
Model III (Predictor: Longest replichore)					
All	180	0.010	1.8×10^{-1}	1.98 ± 2.18	-0.24 ± 0.35
HD+FL	159	0.045	7.1×10^{-3}	-2.94 ± 2.39	0.53 ± 0.38
FL only	104	0.113	4.8×10^{-4}	-6.02 ± 3.53	1.02 ± 0.56
Model IV (Predictor: Skew-normalized replichore)					
All	180	0.042	5.6×10^{-3}	-1.36 ± 1.30	0.23 ± 0.16
HD+FL	159	0.204	2.3×10^{-9}	-3.38 ± 1.17	0.46 ± 0.14
FL only	104	0.300	1.8×10^{-9}	-4.43 ± 1.46	0.59 ± 0.18

Testing for potential outliers

Comparison of ordinary least squares (OLS) regression and OLS refined statistics on Dataset-I for the relationship between skew-normalized replichore length (predictor) and doubling time (response) is shown in Supplementary Table S3. "OLS-refined" denotes OLS refitted after excluding influential data points with Cook's distance $D_i > 4/(n-k-1)$, where n is the sample size and k is the number of independent variables ($k = 1$ in this analysis). The consistency in R^2 and p-value suggests the robustness of the OLS, indicating that the regression is not biased by a single or a few data

points.

Supplementary Table S3: Comparison of ordinary least squares (OLS) regression and OLS-refined statistics on Dataset-I for the relationship between skew-normalized replichore length (predictor) and doubling time (response).

Model-IV Dataset-I	**OLS**			**OLS refined**		
	n	R^2	p**-value**	n	R^2	p**-value**
All classes	367	0.15	3e-14	353	0.15	3e-14
HD + FL	346	0.19	5e-17	332	0.19	9e-17
FL only	187	0.18	1e-09	182	0.17	1e-08