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. These findings point to a potential adaptive role of nucleotide skew in bacterial genome replication.

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

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, 9]. 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 [10, 11, 12, 13].

In this work, we revisit this question. We analyze two sets of bacterial genomes ($n = 367$ and $n = 180$) [12, 14] 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 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 [14], and (ii) minimal doubling time under rapid growth conditions, Dataset-II [12]. Corresponding genomic sequences are retrieved from the NCBI Genome Database, with accession numbers provided in Supplementary information, and only species with both doubling time and genome sequence data available, i.e., 180 in the first set and 367 in the second set, are included in the analysis.

Identification of replication origin and terminus

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

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

where S represents a genomic sequence of length G_l, and δ is the Kronecker delta.

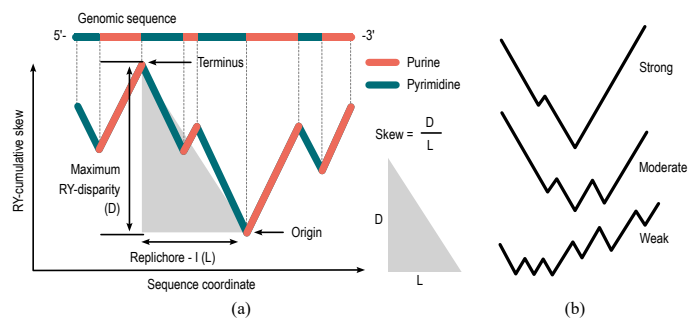


Figure 1: Identification of replication origin and terminus. (a) Cumulative RY-skew for an example genome, calculated using Eq. 1. The positions corresponding to the minimum and maximum cumulative RY-skews were used to predict the replication origin and terminus, respectively, thereby dividing the genome 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".

To find the origins and termini of any given sequence, we rotate the circular genome until the cumulative R-Y disparity (along the y-axis in Fig. 1) is maximum. The valley position

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

in the resulting skew diagram is designated as the 'replication origin', and the peak as the 'terminus', thereby partitioning the chromosome into two replichores. 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 how bacterial genome architecture influences replication time, and whether the replication dynamics exert selective pressure on bacterial doubling times. To this end, we employed linear regression models to correlate bacterial doubling time with genomic sequence characteristics, such as replichore organization, ploidy, and nucleotide skews.

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,

$$T = \alpha\, G_l \tag{2}$$

where T is replication time, G_l is genome size, and α a proportionality constant.

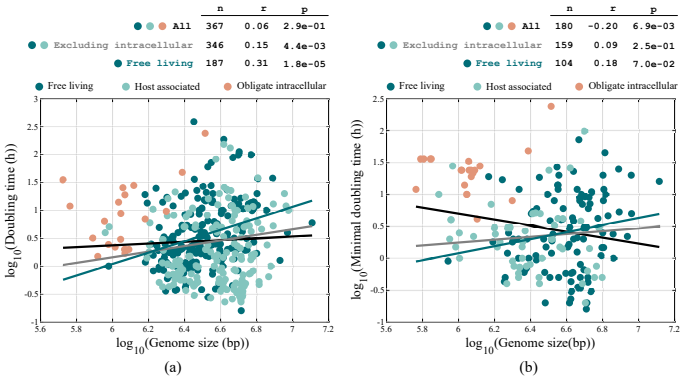


Figure 2: Relationship between bacterial doubling time and genome size in (a) Dataset-I and in (b) Dataset-II. No significant correlation is observed when obligate intracellular species are included. Excluding them yields a very weak correlation, while restricting to free-living bacteria reveals a more pronounced correlation, suggesting that selective pressure to reduce replication time may operate in nutrient-competitive environments but not in host-dependent ones. 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.

In agreement with previous reports [10, 11, 12, 13], we found a negative or no significant correlation between genome size and doubling time when considering the full dataset, $r = 0.06$ for Dataset-I and $r = -0.20$ for Dataset-II (see Fig. 2(a), Table 1 Model-I). Excluding obligate intracellular species that are strictly host-dependent bacteria and hence are largely insulated from nutrient competition [17], we found only a very weak correlation. However, when restricting the analysis to free-living bacteria, excluding facultative host-dependent bacteria, the correlation became more pronounced, $r = 0.31$ for Dataset-I and $r = 0.18$ for Dataset-II, suggesting that genome size reduction in these organisms may indeed be shaped by selective pressure to minimize replication time in *competitive* environments. This modest correlation arises from a simplified model that ignores parallel replication across chromosomes (ploidies) and replichores, thus overestimates replication time relative to biological reality [18, 19]. On excluding ploidies 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 $r = 0.09$ for the entire dataset, and $r = 0.33$ for free-living bacteria in Dataset-I and $r = -0.17$ (entire dataset) and $r = 0.25$ (free-living) in Dataset-II (see Table 1, Model-II).

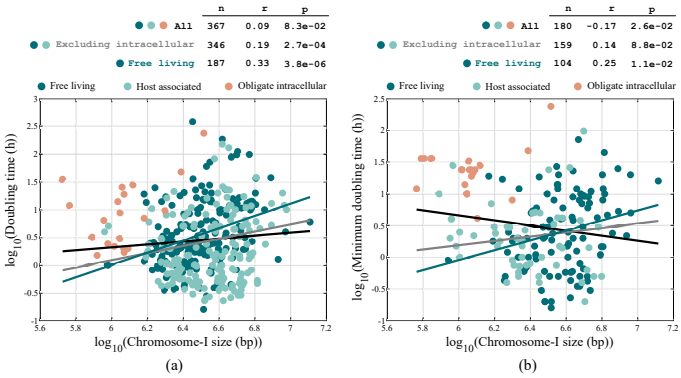


Figure 3: Relationship between bacterial doubling time and chromosome size in (a) Dataset-I and in (b) Dataset-II. The correlation of doubling time with chromosome size improves in comparison to that with total genome size. As the secondary chromosomes, ploidies are replicated in parallel [19]; the longest chromosome, usually the primary one, will dictate the replication time.

Length of the longest replichore and growth rate

Assuming parallelization of replication across chromosomes and replichores, the effective replication time is determined by the longest replichore. Accordingly, we refined the model to

$$T = \beta R_l \tag{3}$$

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

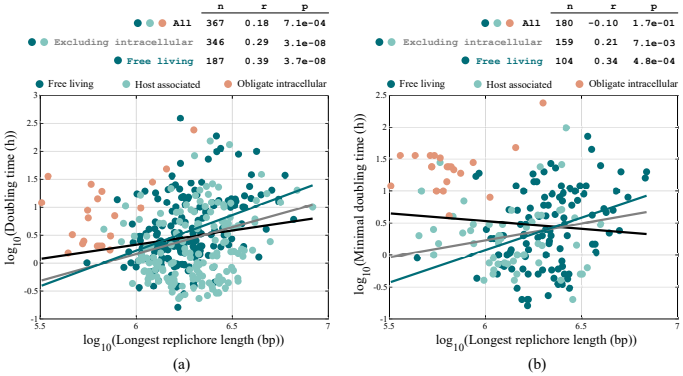


Figure 4: Relationship between bacterial doubling time and the length of the longest replichore in (a) Dataset-I and in (b) Dataset-II. Accounting for replication parallelization leads to a stronger correlation than when using total genome size, highlighting the importance of replichore architecture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

Incorporating replication parallelization improved the correlation with doubling time, $r = 0.18$ for all organisms and $r = 0.39$ for free-living bacteria in Dataset-I and $r = -0.10$ (entire dataset) and $r = 0.34$ (free-living) in Dataset-II (Fig. 4, Table 1, Model-III).

Effect of skew on replication kinetics

Inspired by our previous DNA unzipping simulation [20], 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 [16, 21], we observed that fast-growing bacteria tend to exhibit stronger purine–pyrimidine skew in their sequence (Fig. 5), suggesting that skew may enhance replication fork progression speed and thereby affect bacterial growth rates. For instance, the replication fork progression speed of Mycoplasma capricolum, which shows a weak RY-skew, is just 6 kb/min, which is about 10 times slower than that of E. coli [22].

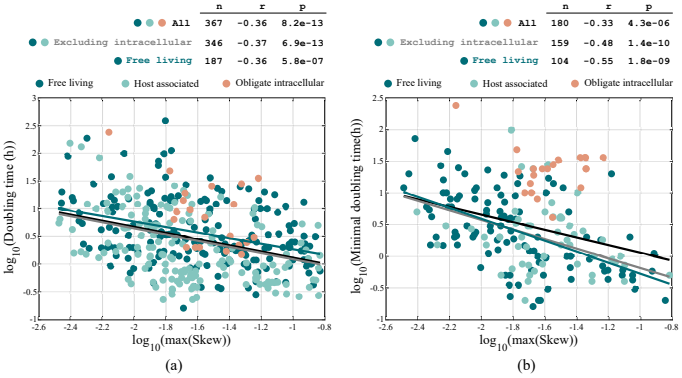


Figure 5: Relationship between bacterial doubling time and maximum purine–pyrimidine skew of the genome in (a) Dataset-I and in (b) Dataset-II. Faster-growing bacteria exhibit stronger compositional skew, suggesting a potential role of skew in facilitating replication fork progression.

To model this, we introduce the assumption that the replication fork speed scales with nucleotide skew. The predictor of replication time is defined as the maximum ratio of replichore length to its corresponding skew (standing in for fork speed), calculated across both replichores. Under this assumption, replication time can be approximated as

$$T = \gamma \left(\frac{R}{S} \right)_{\max} \tag{4}$$

where $(R/S)_{\max}$ 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 $r = 0.38$ for the total data set and $r = 0.43$ for free-living bacteria in Dataset-I and $r = 0.21$ (entire dataset) and $r = 0.55$ (free-living) in Dataset-II, see Fig. 6 and Table 1, Model-IV.

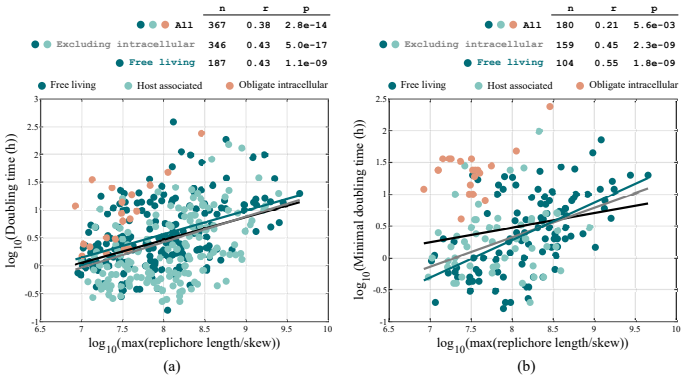


Figure 6: Predicted versus observed doubling time under the skew-adjusted model in (a) Dataset-I and in (b) Dataset-II, where replication time is modeled as the ratio of replichore length and the fork speed, with the latter assumed to be proportional to skew. Including nucleotide skew strengthens the predictive power compared to models based solely on genome size or replichore length.

Discussion

This study examined how bacterial genome architecture influences replication dynamics and, in turn, growth rate. Our analyses unearth three key findings. First, genome size alone is only weakly associated with doubling time, and only among free-living bacteria. Second, accounting for genome organization, particularly the length of the longest replichore, improves this correlation, suggesting that replication time imposes selective constraints on genome architecture. Third, incorporating nucleotide compositional skew further strengthens the predictive power, indicating that RY-asymmetry may play an adaptive role in facilitating replication fork progression.

Selective pressures on genome size and organization

The positive association between genome size and doubling time in free-living bacteria, but not in obligate intracellular species, suggests that selective pressure on replication efficiency is context-dependent. Free-living organisms, which experience strong nutrient competition, may be constrained to reduce genome size in order to minimize replication time [7, 5]. By contrast, host-dependent bacteria may be buffered from such pressures, allowing genome reduction to be shaped primarily by genetic drift, mutational bias, and host–pathogen interactions rather than replication kinetics. However, free-living bacteria also require all the protein-coding sequences for survival in comparison to the obligate host-dependent ones and experience an upward pressure on the genome for information. Consequently, their genomes are shaped by an interplay between downward pressure for speed and upward pressure for information. In addition to these observations on bacteria, our analysis of archaeal genomes ($n = 153$) also revealed a moderate correlation ($r = 0.35, p < 1e-05$) between genome size and doubling time (see Supplementary Data), further supporting the role of genome size in constraining the growth rate.

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. Our results indicate that bacterial genomes may evolve under pressure not only to reduce size but also to balance replichore

Table 1: The bacterial doubling time is correlated with 4 different predictors. The Pearson correlation coefficients (r) and p-values for two datasets under different replication models.

Dataset	Subset	Model I		Model II		Model III		Model IV	
		No parallelization ($T \propto$ genome size)		Parallelization across ploidy ($T \propto$ chromosome size)		Parallelization across ploidy + replichore ($T \propto$ longest replichore)		Fork speed depends on RY-skew ($T \propto$ replichore length/skew)	
		r	p	r	p	r	p	r	p
Dataset I Total 367 (Madin et al. [14])	All organisms ($n = 367$)	0.06	3e-01	0.09	8e-02	0.18	7e-04	0.38	3e-14
	Excluding Obligate Intracellular ($n = 346$)	0.15	4e-03	0.19	3e-04	0.29	3e-08	0.43	5e-17
	Free-living ($n = 187$)	0.31	2e-05	0.33	4e-06	0.39	4e-08	0.43	1e-09
Dataset II Total 180 (Vieira et al. [12])	All organisms ($n = 180$)	-0.20	7e-03	-0.17	3e-02	-0.10	2e-01	0.21	6e-03
	Excluding Obligate Intracellular ($n = 159$)	0.09	3e-01	0.14	9e-02	0.21	7e-03	0.45	2e-09
	Free-living ($n = 104$)	0.18	7e-02	0.25	1e-02	0.34	5e-04	0.55	2e-09

lengths, minimizing replication delays and enhancing growth efficiency.

Potential adaptive role of compositional skew

Traditionally, nucleotide compositional skew has been interpreted as a neutral byproduct of mutational biases during replication, such as cytosine deamination on the lagging strand, leading to characteristic GC skews [23, 24], or as an adaptive selection bias of coding sequences in the leading strand. However, from the observations we made here, we propose that skew may also have functional significance. We observed that fast-growing bacteria display stronger purine–pyrimidine asymmetry, and our model incorporating skew better predicts doubling time. This raises the possibility that skew is not solely a mutational artifact or a codon preference but may enhance replication fork progression speed and, consequently, bacterial growth. If so, compositional skew may represent an underappreciated adaptive feature of genome evolution.

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 helping to speed up replication, and is therefore adaptive. We request readers to explore our previous articles [20, 25, 26, 27, 28, 29] to understand the mechanistic connection between skews and replication fork speed, and the evolutionary advantage of high nucleotide skews in genomes.

Limitations and future directions

Several simplifications were made in our analysis. The classification of bacteria into free-living versus host-dependent was based on a literature survey and may not fully capture ecological diversity. Our 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. We further observed that GC skew appeared to have a stronger effect than that of the AT skew on replication kinetics. 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 analysis of archaeal data is provided in Supplementary Data.

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. *science*, 302(5649):1401–1404, 2003.

[2] Alexis Dufresne, Laurence Garczarek, and Frédéric Partensky. Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome biology*, 6(2):R14, 2005.

[3] Stephen J Giovannoni, H James Tripp, Scott Givan, Mircea Podar, Kevin L Vergin, Damon Baptista, Lisa Bibbs, Jonathan Eads, Toby H Richardson, Michiel Noordewier, et al. Genome streamlining in a cosmopolitan oceanic bacterium. *science*, 309(5738):1242–1245, 2005.

[4] Anna K Simonsen. Environmental stress leads to genome streamlining in a widely distributed species of soil bacteria. *The ISME Journal*, 16(2):423–434, 2022.

[5] Brandon K Swan, Ben Tupper, Alexander Sczyrba, Federico M Lauro, Manuel Martinez-Garcia, José M González, Haiwei Luo, Jody J Wright, Zachary C Landry, Niels W Hanson, et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences*, 110(28):11463–11468, 2013.

[6] Hongwei Liu, Haiyang Zhang, Jeff Powell, Manuel Delgado-Baquerizo, Juntao Wang, and Brajesh Singh. Warmer and drier ecosystems select for smaller bacterial genomes in global soils. *Imeta*, 2(1):e70, 2023.

[7] Stephen J Giovannoni, J Cameron Thrash, and Ben Temperton. Implications of streamlining theory for microbial ecology. *The ISME journal*, 8(8):1553–1565, 2014.

[8] Niv Sabath, Evandro Ferrada, Aditya Barve, and Andreas Wagner. Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome biology and evolution*, 5(5):966–977, 2013.

[9] John P DeLong, Jordan G Okie, Melanie E Moses, Richard M Sibly, and James H Brown. Shifts in metabolic scaling, production, and efficiency across major evolutionary transitions of life. *Proceedings of the National Academy of Sciences*, 107(29):12941–12945, 2010.

[10] Alex Mira, Howard Ochman, and Nancy A Moran. Deletional bias and the evolution of bacterial genomes. *TRENDS in Genetics*, 17(10):589–596, 2001.

[11] Marie Touchon and Eduardo PC Rocha. Causes of insertion sequences abundance in prokaryotic genomes. *Molecular biology and evolution*, 24(4):969–981, 2007.

[12] Sara Vieira-Silva and Eduardo PC Rocha. The systemic imprint of growth and its uses in ecological (meta) genomics. *PLoS genetics*, 6(1):e1000808, 2010.

[13] Mark Westoby, Daniel Aagren Nielsen, Michael R Gillings, Elena Litchman, Joshua S Madin, Ian T Paulsen, and Sasha G Tetu. Cell size, genome size, and maximum growth rate are near-independent dimensions of ecological variation across bacteria and archaea. *Ecology and Evolution*, 11(9):3956–3976, 2021.

[14] Joshua S Madin, Daniel A Nielsen, Maria Brbic, Ross Corkrey, David Danko, Kyle Edwards, Martin KM Engqvist, Noah Fierer, Jemma L Geoghegan, Michael Gillings, et al. A synthesis of bacterial and archaeal phenotypic trait data. *Scientific data*, 7(1):170, 2020.

[15] Andrei Grigoriev. Analyzing genomes with cumulative skew diagrams. *Nucleic acids research*, 26(10):2286–2290, 1998.

[16] Peder Worning, Lars J Jensen, Peter F Hallin, Hans-Henrik Stærfeldt, and David W Ussery. Origin of replication in circular prokaryotic chromosomes. *Environmental microbiology*, 8(2):353–361, 2006.

[17] Lisa Klasson and Siv GE Andersson. Evolution of minimal-genesets in host-dependent bacteria. *TRENDS in Microbiology*, 12(1):37–43, 2004.

[18] Caroline Stokke, Torsten Waldminghaus, and Kirsten Skarstad. Replication patterns and organization of replication forks in vibrio cholerae. *Microbiology*, 157(3):695–708, 2011.

[19] Florian Fournes, Marie-Eve Val, Ole Skovgaard, and Didier Mazel. Replicate once per cell cycle: replication control of secondary chromosomes. *Frontiers in microbiology*, 9:1833, 2018.

[20] Parthasarathi Sahu, Sashikanta Barik, Koushik Ghosh, and Hemachander Subramanian. High nucleotide skew palindromic dna sequences function as potential replication origins due to their unzipping propensity. *Journal of Molecular Evolution*, 92(6):761–775, 2024.

[21] Kazuharu Arakawa and Masaru Tomita. Measures of compositional strand bias related to replication machinery and its applications. *Current genomics*, 13(1):4–15, 2012.

[22] Shintaro Seto and Makoto Miyata. Cell reproduction and morphological changes in mycoplasma capricolum. *Journal of bacteriology*, 180(2):256–264, 1998.

[23] AC Frank and JR Lobry. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, 238(1):65–77, 1999.

[24] Jean R Lobry and Noboru Sueoka. Asymmetric directional mutation pressures in bacteria. *Genome biology*, 3(10):research0058–1, 2002.

[25] Hemachander Subramanian and Robert A Gatenby. Evolutionary advantage of directional symmetry breaking in self-replicating polymers. *Journal of theoretical biology*, 446:128–136, 2018.

[26] Hemachander Subramanian and Robert A Gatenby. Evolutionary advantage of anti-parallel strand orientation of duplex dna. *Scientific Reports*, 10(1):9883, 2020.

[27] Hemachander Subramanian, Joel Brown, and Robert Gatenby. Prebiotic competition and evolution in self-replicating polynucleotides can explain the properties of dna/rna in modern living systems. *BMC Evolutionary Biology*, 20(1):75, 2020.

[28] Hemachander Subramanian. Joint optimization of replicative rate and information storage set the letter size of primordial genetic alphabet. *BioSystems*, 251:105442, 2025.

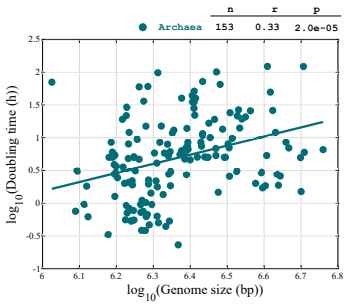
[29] Koushik Ghosh, Parthasarthi Sahu, Shashikanta Barik, and Hemachander Subramanian. Non-enzymatic error correction in self-replicators without extraneous energy supply. *bioRxiv*, pages 2025–06, 2025.

Supplementary Data

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 archaea that is 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 Fig. S1. A moderate correlation ($r = 0.35$) indicates that genome size contributes to replication kinetics and thereby 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 without accounting for replication parallelization across replichores. A stronger correlation is expected when considering the longest replichore rather than the total genome size.