Microbial trait evolution is dominated by frequent and rare pulsed evolution

Yingnan Gao and Martin Wu*
Department of Biology, University of Virginia, Charlottesville, VA 22904, USA;

* Corresponding author: Martin Wu, mw4yv@virginia.edu

Abstract

On the macroevolutionary timescale, does trait evolution proceed gradually or by rapid bursts (pulses) separated by long periods of stasis? Although studies have shown pulsed evolution is prevalent in animals, our knowledge about the tempo and mode of evolution across the tree of life is very limited. This long-standing debate calls for a test in bacteria and archaea, the most ancient and diverse forms of life with unique population genetic properties. Using a likelihood-based framework, we analyzed patterns of microbial genomic trait evolution on a broad macroevolutionary timescale. Here we show that pulsed evolution is both prevalent and predominant in microbes. For the first time, we detected two distinct types of pulsed evolution that are predicted by the punctuated equilibrium and quantum evolution theories. Our findings suggest that major bacterial lineages originated in quick bursts and pulsed evolution is common across the tree of life despite drastically different population genetic properties of animals, plants and microbes.

Introduction

There has been a long-standing debate about the tempo and mode of trait evolution on the macroevolutionary timescale. The gradualism theory states that evolution occurs gradually by small changes that accumulate over a long period of time [1]. The pulsed evolution theory, on the other hand, argues that evolution mostly proceeds in bursts of larger changes (jumps) separated by long periods of stasis [1–3]. Two types of jumps have been proposed in pulsed evolution. Simpson's quantum evolution theory postulates that jumps happen when lineages shift into new adaptive zones and these jumps play an import role in the origination of higher taxa [2], while Eldredge and Gould's later punctuated equilibrium theory focuses exclusively on jumps associated with speciation [1]. Conceptually, these two types of jumps exist side-by-side but differ in their frequencies and magnitudes. Studies of animal fossil records support the punctuated equilibrium theory [4,5], and more recent phylogenetic comparative studies of vertebrate body size [6–8] also provide evidence for quantum evolution. Together, they show that evolution is not solely composed of slow and gradual changes but also instant jumps on the macroevolution timescale.

Analogous studies in bacteria and archaea, the most ancient and diverse forms of life on Earth, are lacking, largely due to the scarcity of fossil records and well-measured quantitative phenotypic traits in microbes. Fortunately, the phenotypic evolution of microbial species can be reconstructed from extant genome sequences. Several genomic features are highly correlated with the microbial life strategy. For example, the GC% of the ribosomal RNA gene is correlated with the optimal growth temperature of bacteria and archaea [9]. According to the genome streamlining theory, genome size, genomic GC% and the nitrogen use in proteins all evolve in response to the nutrient level in the environment [10,11]. These genomic features can be accurately determined from the thousands of complete genomes currently available that represent a broad range of closely and distantly related lineages, making it possible to study the tempo and mode of trait evolution in microbes over a broad spectrum of macroevolution timescales.

Interestingly, long-term experimental evolution has shown evidence of pulsed evolution in *Escherichia coli* cell size [12]. However, on the macroevolutionary timescale, the role of pulsed evolution in microbial trait evolution remains largely unknown. Although it is well known that there are large trait changes between bacterial clades (e.g., the genomic GC% of high GC vs low GC Gram-positives, AT-rich obligate intracellular bacteria vs their free-living relatives)[13,14], it is

unclear whether these large trait changes arose gradually or instantly by jumps during the time the clades diverged from each other. Compared to animals and plants, bacteria and archaea reproduce asexually and have relatively large population sizes, high dispersal rates and short generation times. Another salient feature unique to microbes is that their genomes can often leap by large scale horizontal gene transfer [15], which obviously will have an impact on the tempo and mode of evolution. Given these unique features, a central question is whether the tempo and mode of microbial trait evolution are similar to those of eukaryotes and whether pulsed evolution is a universal theme across the tree of life, and if so, how much does pulsed evolution contribute to microbial trait evolution?

Results

Gradual evolution does not explain microbial genomic trait evolution

We downloaded 10,616 and 263 complete bacterial and archaeal genomes respectively from the NCBI RefSeq database, from which we reconstructed genome trees and selected 6,668 and 247 representative genomes that passed quality control (Methods). For each representative genome, we calculated four genomic traits (genomic GC%, rRNA GC%, genome size and the average nitrogen atoms per residual side chain N-ARSC), all of which showed strong phylogenetic signals (Pagel's $\lambda>0.99$). Although the four genomic traits are significantly correlated, the correlation is very weak (Supplementary Fig. 1). For any trait, the proportion of variation explained by the other traits is less than 13.5% when evaluated using phylogenetically independent contrast (PIC). Therefore, to capture the possible variation in the tempo and mode of evolution, we chose to test all four traits separately. Notably, the PIC distributions of these traits in bacteria drastically deviate from the normal distribution expected by the Brownian motion (BM) model of gradual evolution (Fig. 1 A-D, two-sided Kolmogorov–Smirnov test, $P<0.001$ for all 4 traits). Specifically, all PIC distributions exhibit a strong leptokurtic (heavy-tailed) pattern with a positive excess kurtosis ranging from 5.79 to 13.47, indicating that extremely rapid trait changes occur more frequently than expected by the BM model. For archaea, the deviation of the PIC distribution from the normal expectation is less severe (Supplementary Fig. 2 A-D, two-sided Kolmogorov-Smirnov test, $P<0.001$, $P=0.024$, $P=0.018$ and $P=0.155$ for rRNA GC%, genomic GC%, genome size and N-ARSC, respectively), with the excess kurtosis ranging from 1.47 to 7.58. Although inconsistent with gradual evolution, such a heavy-tailed pattern can be explained by pulsed evolution. Extremely rapid trait changes ($|PIC|>3$) take place more frequently than expected by the normal distribution (0.27%) throughout the bacterial evolutionary history (Supplementary Fig. 3), suggesting repeated episodes of pulsed trait evolution.

Modeling microbial trait evolution

When plotted against the branch length, the trait changes between two sister nodes in the bacteria phylogeny display a "blunderbuss pattern" (Fig. 1 E-H). It starts with a period of stationary fluctuations where trait changes are bounded and the variance does not accumulate over time. Segmented linear regression analysis indicates that this phase of stasis lasts until the branch length reaches 0.001 substitutions/site for rRNA GC% (Supplementary Fig. 4). On longer timescales, the stasis yields to a pattern of increasing divergence over time. The archaeal traits display similar patterns (Supplementary Fig. 2 E-H). This "blunderbuss pattern", first observed in the evolution of vertebrate body-size, is a signature of pulsed evolution [6]. Interestingly, for

rRNA GC%, we observed a second spike in the trait divergence rate at 0.025 substitutions/site (Supplementary Fig. 4), indicating a change of evolution tempo at this point.

To formally test whether pulsed evolution explains the patterns, we model the trait change between two sister nodes using the Levy process [8]. More specifically, we model the trait change as the sum of three independent stochastic variables: pulsed evolution, gradual evolution, and time-independent trait variation. We assume pulsed evolution occurs at a constant rate relative to the molecular divergence and the jump size follows a normal distribution with a mean of zero. As a result, the pulsed evolution is modeled as a compound Poisson process with normal jumps, with parameters λ and σ^2 denoting the frequency (number of expected jumps per lineage per unit branch length) and the magnitude (variance of trait change) of the jumps, respectively. We model gradual evolution using the classic BM model with a single parameter σ_{BM}^2 denoting the rate of the gradual trait change. Meanwhile, we observed trait variation between genomes with identical phylogenetic marker sequence alignments, indicating the presence of time-independent variation in our phylogeny. This variation follows a leptokurtic distribution. Therefore, we model the time-independent variation with the Laplace distribution with one single parameter ε denoting its variance for simplicity and convenience. It should be noted that a jump in the genomic trait may be coupled with an increase in the molecular divergence rate, especially for those traits affecting protein sequences (e.g., genomic GC% and N-ARSC). However, such correlation between the molecular branch length and the trait change will only reduce the signal of pulsed evolution, as the increased branch length provides greater power for gradual evolution to explain the trait variation.

The changing tempos revealed by segmented linear regression suggested that one Poisson process may not adequately describe the patterns of pulsed evolution, prompting us to add multiple Poisson processes to our modeling. Therefore, using the framework described above, we tested six different models (Table 1). The BM model delineates gradual evolution. The PE1, PE2 and PE3 models describe pulsed evolution with one, two or three Poisson processes respectively. The PE(n)+BM models represent trait evolution with both pulsed and gradual evolution. Details of these models are provided in the Supplementary Text.

<u>Microbial trait evolution is dominated by frequent and rare pulsed evolution</u>

For the four traits we have examined in bacteria and archaea, the best model is always one with a pulsed evolution component, while the relative support for the BM only model is negligible (all BM Akaike weights < 0.5%, Table 2), indicating that pulsed evolution is present in both bacteria and archaea. The best models fit the PIC distributions much better than the BM does (Fig. 1 and Supplementary Fig. 2). To test the prevalence of pulsed evolution in bacteria, we separately fitted our models on 17 bacterial families that each contained at least 100 genomes. We found that trait evolution in 100%, 94.1%, 58.8% and 35.3% of tested families were best explained by a model with a pulsed evolution component (PE1, PE1+BM, PE2, PE2+BM) for rRNA GC%, genomic GC%, genome size and N-ARSC respectively, indicating that pulsed evolution is prevalent in bacteria (Supplementary Table 1). Our simulation shows that when the number of genomes decreases, the power to detect pulsed evolution in genome size and N-ARSC also decreases (Supplementary Table 2), suggesting that we might have underestimated the prevalence of pulsed evolution in these two traits in the 17 bacterial families. We did not test the prevalence of pulsed evolution in archaea because of insufficient archaeal genomes.

Using parameters of the best models (Table 3), we estimated the relative contribution of each compound Poisson process. The variable ε represents the trait variance in the initial stasis phase (Fig. 1). Its estimated value approximates the intraspecific trait variation between genomes with identical marker gene alignments (i.e., zero branch length) and therefore is used as the baseline. The jumps vary greatly in their frequencies and magnitudes, but can be roughly classified into two types: small and frequent, or large and rare (Table 3). For example, for rRNA GC%, rare jumps (1.96 jumps per lineage per unit branch length) are extremely large in magnitude, as the standard deviation of trait change introduced by one jump is approximately 60 times of $\sqrt{\varepsilon}$, or roughly equivalent to that introduced by gradual evolution under the BM model over a branch length of 0.35 substitution per site, and approximately corresponds to 5.8 °C change in the optimal growth temperature [9]. In comparison, the frequent jumps (118 jumps per lineage per unit branch length) are 60 times more frequent but their sizes are only about 3 times of $\sqrt{\varepsilon}$. Overall, rare jumps predominate in trait evolution as they contribute more than 74% of variation in each trait over the whole phylogeny. Similarly, pulsed evolution also predominates in archaea as the PE1 and PE2 models are the best model in all archaeal traits (Table 2). Due to the limited number of archaeal genomes, we cannot robustly estimate the parameters of each jump process in archaea. To evaluate the effect of the tree topology on our model fitting, we fitted models on the genome tree of the family Enterobacteriaceae (with 748 genomes) made using either FastTree or RAxML. We found that the fitted model parameters are highly similar using these two trees (data not shown).

<u>Rare jumps are correlated with cladogenesis in bacteria</u>

For each branch in the bacterial phylogeny, we calculated the posterior probability of it having at least one jump and mapped rare and super rare jumps onto the phylogeny. We found that jumps occurred throughout the phylogeny (Fig. 2.), again indicating that pulsed evolution is prevalent in bacterial evolution history. Some jumps are associated with known key evolutionary adaptations. For instance, a classic example of adaptation to endosymbiosis occurred within the family Enterobacteriaceae, in the lineage leading to a clade of insect endosymbionts that includes the genera *Buchnera*, *Wigglesworthia* and *Candidatus Blochmannia*. Our model detects large rare jumps at the base of the clade in the genome size and genomic GC% (posterior probability > 0.9, Fig. 2). The recently described order Candidatus Nanopelagicales within the phylum Actinobacteria makes up the most abundant free-living bacteria in freshwater. Nanopelagicales has adapted to live in the nutrient poor environment by streamlining their genomes[16] . Compared to its high GC Gram-positive sister clade, Nanopelagicales has dramatically reduced genome size (~1.4 Mbp) and genomic GC% (~48%). Our modeling indicates that the genomic streamlining process happened not gradually but by jumps. We detected large rare jumps in all genomic traits at the base of the order with extremely high confidence (posterior probability >0.99). Similar patterns have also been observed in the branch leading to the most abundant free-living marine bacteria Pelagibacterales and the intracellular bacteria Rickettsiales and Holosporales. Our model also predicts large rare jumps at higher taxonomic levels such as those at the base branch leading to the α-, β-, γ- and δ-proteobacteria (posterior probability >0.96) and the branch that separates the γ-proteobacteria from the rest of the proteobacteria (posterior probability >0.99). Our results suggest that these key evolutionary adaptations evolved in rapid bursts instead of through slow divergence of species over long periods of time.

Next, we tested whether jumps are correlated with cladogenesis in bacteria by comparing the posterior frequency of jumps to the expected frequency for which we assume no correlation of jumps with cladogenesis (the null hypothesis). For example, if jumps happen significantly more frequently between two congener sister nodes than expected, jumps are considered correlated with the speciation event (cladogenesis at the species level). For frequent jumps, simulations indicated that we lacked the statistical power to reject the null hypothesis at every taxonomic level, and therefore we excluded them from this analysis. For rare and super rare jumps, we tested their correlation with cladogenesis from the species to order levels. We found that rare and super rare jumps occur more frequently than expected for all traits at the genus, family and order levels, except for N-ARSC at the genus level (Table 4). This increase in frequency is significant for rRNA GC%, genomic GC% and genome size at the genus and family levels ($P \leq 0.050$), and for rRNA and genomic GC% at the order level ($P < 0.001$). Interestingly, we found that rare and super rare jumps happen less frequently than expected for all traits at the species level, although it is significant only for ribosomal GC% ($P < 0.001$). Our results suggest that rare and super rare jumps are correlated with cladogenesis at higher taxonomic ranks.

Discussion

Microbes are known for rapid evolution. Why are these genomic traits constrained for millions of years before they diverge? The stasis at the species level can be explained by stabilizing selection that eliminates variants falling outside of a stable niche [17]. Alternatively, it can be maintained by gene flow, as suggested by Futuyma's ephemeral divergence theory [18]. Futuyma proposes that novel adaptive trait variation arises frequently in local populations, but the spatial and temporal mosaic nature of niches prevents such local adaptations from spreading to the entire species because they are wiped out by the gene flow from the prevailing intervening ancestral populations. As a result, trait changes perish and do not accumulate over time, resulting in stationary fluctuations, until speciation interrupts the gene flow. Although reproducing asexually, microbes do exchange genes through homologous recombination and there is evidence that gene flow plays a critical role in bacterial speciation at least under certain conditions [19–21]. Interestingly, the transient trait variation in the initial stasis phase when jumps are absent (the ε term in our model) approximately matches the intraspecific trait variation.

At longer timescales or higher taxonomic levels, trait evolution can be constrained through stabilizing selection exerted by the adaptive zone [22], defined as a set of ecological niches to which a group of species are adapted [2]. This will generate the pattern of phylogenetic conservatism where organisms in a clade tend to have similar traits (synapomorphy) and occupy similar habitats. Accordingly, both genome analyses and ecological studies support that ecological coherence exists at higher taxonomic levels in bacteria [23]. For example, different bacterial clades have their unique set of genes [24] and analysis of thousands of cultured microbial strains showed that strains related at the genus, family or order levels occupy the same habitat more frequently than expected by chance [25].

Interestingly, for the first time, we detected two types of jumps in one dataset: small frequent jumps and large rare jumps. This is possible because the large bacterial dataset spans a wide range of macroevolutionary timescales. For example, the bacterial genome tree in our study has a total branch length of 442.9 substitutions/site. For super rare jumps (e.g., genome size jumps with a rate of 0.17 jump per lineage per unit branch length), it is estimated that there are still 75

events in the entire phylogeny. On the other hand, the resolution of our bacterial genome tree is 5×10^{-5} substitutions/site, meaning that we can detect jumps that occurs as frequently as 20,000 jumps per unit branch length on average. The large difference in the frequency and size of the jumps suggests that they represent different kinds of evolutionary events. Although our modeling does not stipulate the coupling of cladogenesis and pulsed evolution (as in the classical punctuated equilibrium theory), the rate of the frequent jumps in bacteria (115-634 jumps per lineage per unit branch length or 0.06-0.32 jumps per lineage per Myr) approximates the recently estimated bacterial speciation rate (0.03-0.05 speciation per lineage per Myr) where species is defined as having 99% identical 16S rRNAs [26], suggesting that frequent jumps and the speciation events may be correlated. Two features of the rare jumps fit the description of quantum evolution. First, the rare jumps are fairly large in magnitude, most likely resulting from shifting between major adaptive zones. Second, our test shows that rare jumps happen less frequently than expected at the species level but significantly more frequently than expected at higher taxonomic levels (genus, family and order), suggesting there is a correlation between rare jumps and the origination of higher taxa. A key insight from this study is that the major evolutionary adaptations in bacteria and the origination of major bacterial lineages happens in quick bursts (quantum evolution) instead of through slow divergence of species over long periods of time [27].

Microbial genomes are highly dynamic [28,29]. They can change by mutation, gene loss, gene duplication and horizontal gene transfer. Whatever the mechanism, our study suggests that large genome changes happen in episodes of bursts rather than gradually and slowly. These large changes are not due to the simple gain and loss of plasmids as we have excluded plasmids in our study. Chromosomes are in constant exchange with phages, plasmids and other mobile elements and can change by "quantum leaps" in the form of genomic islands [15]. It is worth pointing out that jumps in our model represent trait changes that persist over time, not the processes that drive the changes. The rarity of detected jumps does not mean the evolutionary processes (e.g., selection, population bottleneck) that drive the jumps are rare. It merely means the success rate of such jumps is low. The rarity of jumps can result from adaptation to a large environmental shift that happens infrequently, or it can be manifestation of multiple frequent small jumps occurring in quick succession, which is also rare.

In conclusion, our modeling of phylogenetic comparative data shows that pulsed evolution is both prevalent and dominant in bacteria and archaea genomic traits evolution. The signatures of pulsed evolution detected in this study are consistent with both the punctuated equilibrium and quantum evolution theories. More broadly, our results suggest that pulsed evolution is the rule rather than the exception across the tree of life, despite the drastically different population genetic properties between animals, plants and microbes.

Methods

<u>Bacterial and archaeal phylogeny and genomic traits</u>

We downloaded 10,616 complete bacterial genomes and 263 complete archaeal genomes from the NCBI RefSeq database on September 6, 2018 (Supplementary Table 3). From each genome we identified either 31 bacterial or 104 archaeal protein-coding marker genes using AMPHORA2 [30] with the default options and constructed a bacterial and an archaeal genome tree based on the concatenated and trimmed protein sequence alignment of the marker genes. We reconstructed the archaeal genome tree using RAxML (version 8.2.11) [31] with the option -m

PROTCATLG. Because of its large size, it is impractical to make the bacterial genome tree using RAxML. Instead, we inferred the bacterial genome tree using FastTree (version 2.1.11) [32] with the option -wag -gamma. For better resolution, we re-optimized the branch length of the genome trees with the DNA sequence alignments of the marker genes using RAxML with the option -m GTRGAMMA. We removed genomes with identical alignments, extremely long branches, ambiguous bases or unreliable annotations from the genome trees. For each of the 6,668 bacterial and 247 archaeal genomes remained, we calculated four traits: the ribosomal RNA stem GC% (rRNA GC%), genomic GC%, genome size (excluding plasmids), and the average nitrogen atoms per residual side chain (N-ARSC). We transformed these traits (logit transformation for rRNA GC% and genomic GC%; log transformation for genome size and N-ARSC) to make them comply with the assumption of continuous trait evolution. For conversion from rRNA GC% to the optimal growth temperature, we used the empirical formula determined by Wang *et al* [9]: $T_{optimal}(°C) = 3.75 \times GC_{rRNA}(\%) - 216.27$

Calculating PIC with time-independent variation

Phylogenetically independent contrast (PIC) assumes a Brownian motion (BM) in which trait variance increases linearly with time [33]. However, we observed variation in trait values between genomes that are separated by zero branch length (data not shown). Therefore, we introduce time-independent variation into the BM model and denote its variance with ε. When time-independent variation is normally distributed, the PIC between a pair of sister tips is calculated as

$$PIC = \frac{x_1 - x_2}{\sqrt{(l_1 + l_2)\sigma_{BM}^2 + \varepsilon}}$$

where x_1 and x_2 are the trait values of the tips, l_1 and l_2 are their branch lengths to the parent node, and σ_{BM}^2 is the rate of Brownian motion. The uncertainty of the parent node's trait value is calculated as

$$\varepsilon_0 = \frac{(l_1\sigma_{BM}^2 + 0.5\varepsilon)(l_2\sigma_{BM}^2 + 0.5\varepsilon)}{(l_1 + l_2)\sigma_{BM}^2 + \varepsilon}$$

Testing the pairwise correlation between the four genomic traits

To avoid dependence among extant trait values due to shared ancestry, we selected all 2,003 tip pairs in the bacterial genome tree and calculated their PICs for each trait. Using the PICs, we calculated Pearson correlation coefficient *r* and coefficient of determination R^2 for each trait pair.

Quantifying the frequency of extreme trait changes in bacterial evolution history

We tested whether extremely rapid trait changes happen throughout the bacterial evolutionary history. We calculated the relative distance from the root (last common ancestor of bacteria) to a node *i* as $\hat{d}_{root}(Myr) = \frac{d_{root}}{d_{root}+\overline{d_{tip}}}$ where d_{root} is the branch length of the node *i* to the root, and $\overline{d_{tip}}$ is the average branch length of the node *i* to all its descending tips. It should be noted that a PIC at the node *i* measures the trait difference between its two immediate descending nodes. We binned the PIC based on the relative distance to the root of the node into 7 bins with exponentially distributed boundaries and calculated the frequency of extremely rapid trait changes (|PIC| >= 3) for each bin.

Segmented linear regression of absolute trait change on branch length

To analyze the trend of trait change, we applied segmented linear regression of the absolute trait changes over branch length as described in Uyeda *et al* [6]. For each trait, we log-transformed the absolute trait changes and added a small fixed value (0.001) to obtain approximately normal distribution. To account for uncertainty introduced by ancestral state reconstruction, we adjusted the branch length as described by Felsenstein [33] and log-transformed it as well. When regressing the log-transformed absolute trait changes against the log-transformed adjusted branch lengths, we constrained the slope of the first segment to be zero (to capture the stasis) and allowed the slopes of the remaining regression lines to change at certain breakpoints, but the regression lines had to be continuous (connected). We compared linear regression models with 1 or 2 breakpoints and selected the one with the lowest Akaike Information Criterion (AIC), and used the breakpoints to mark the transitions between different evolution tempos.

Evaluating pulsed evolution in bacteria and archaea

Using maximum likelihood (ML) method, we tested six models of trait evolution (Table 1). For models with more than one compound Poisson processes, we restricted the variances of jump sizes between any two jumping processes to be at least 3-fold different. We fitted the models to trait changes between sister nodes given their branch lengths. For internal nodes, we reconstructed their trait values with Felsenstein's method [33] but took time-independent variation into account. We calculated confidence intervals for model parameters and statistics by bootstrapping with 50 replicates. We selected the best model using AIC.

We estimated two parameters for each compound Poisson process: frequency λ_i and variance of jump sizes σ_i^2, where i is the rank of the Poisson process. For further evaluation of pulsed evolution, we calculated the contribution and the relative jump size of each compound Poisson process in pulsed evolution. The contribution of a Poisson process (as proportion of variance explained, PVE) was calculated by

$$PVE_i = \frac{\lambda_i \sigma_i^2}{\sigma_{BM}^2 + \sum_{j=1}^{n} \lambda_j \sigma_j^2}$$

where i and j are the ranks of Poisson processes, and n is the total number of Poisson processes in the model. The relative jump size was calculated as $\hat{\sigma}_i = \frac{\sigma_i}{\sqrt{\varepsilon}}$. To roughly compare the overall rate of the frequent jumps to the bacterial speciation rate estimated in Myr [26], we calculated the phylogenetically weighted average branch length of all tips to the root in the tree, and then calibrated time assuming the average branch length is equivalent to 3.5 billion years of evolution [34,35].

Testing the effect of sample size on model fitting

To determine the effect of sample size on model fitting, we randomly sampled 2,250, 750, 250, and 100 branches between sister nodes without replacement from the full bacterial phylogeny that contained 6,667 branches between sister nodes. We simulated trait evolution along these branches based on the PE3 model fitted on the bacterial genome size and N-ARSC. We then fitted the six evolution models on the simulated traits, and selected the best model based on AIC. We repeated this procedure 10 times and counted the frequency each model was selected as the best model for each sampling size.

370 Testing the correlation between rare jumps and cladogenesis in bacteria
371 We identified all contrasts between two congener sister nodes. Using a posterior probability
372 threshold of 0.75, we calculated the frequency of having at least one rare jump in these contrasts
373 for each trait (observed frequency). We computed the expected distribution of this frequency
374 through simulations using the estimated model parameters of pulsed evolution under the null
375 hypothesis that there is no correlation between jumps and cladogenesis. By comparing the
376 observed frequency to the expected distribution, we calculated the two-sided P-value of the null
377 hypothesis being true at the species level. We repeated the same statistical test at the genus,
378 family and order levels.

379 **Data availability:** The data that support the findings of this study are publicly available from the
380 NCBI RefSeq database under the accession numbers listed in Supplementary Table 3.
381
382 **Code availability:** The customized R code and scripts used in this study are provided as
383 Supplementary Files.

384

385 **References**

386 1. Eldredge, N. & Gould, S. J. Punctuated equilibria - an alternative to phyletic gradualism.
387 in *Models in Paleobiology* 82–115 (1972).

388 2. Simpson, G. G. *Tempo and mode in evolution*. (Columbia University Press, 1944).

389 3. Stanley, S. M. A theory of evolution above the species level. *Proc. Natl. Acad. Sci.* **72**,
390 646–650 (1975).

391 4. Jackson, J. B. C. & Cheetham, A. H. Phylogeny reconstruction and the tempo of
392 speciation in cheilostome Bryozoa. *Paleobiology* **20**, 407–423 (1994).

393 5. Hunt, G., Bell, M. A. & Travis, M. P. Evolution toward a new adaptive optimum:
394 phenotypic evolution in a fossil stickleback lineage. *Evolution* **62**, 700–710 (2008).

395 6. Uyeda, J. C., Hansen, T. F. & Mcpeek, A. The million-year wait for macroevolutionary
396 bursts. *Proc. Natl. Acad. Sci.* **108**, 15908–15913 (2011).

397 7. Duchen, P. *et al.* Inference of evolutionary jumps in large phylogenies using Levy
398 processes. *Syst. Biol.* **66**, 950–963 (2017).

399 8. Landis, M. J. & Schraiber, J. G. Pulsed evolution shaped modern vertebrate body sizes.
400 *Proc. Natl. Acad. Sci. U. S. A.* **114**, 13224–13229 (2017).

401 9. Wang, H. C., Xia, X. & Hickey, D. Thermal adaptation of the small subunit ribosomal
402 RNA gene: A comparative study. *J. Mol. Evol.* **63**, 120–126 (2006).

403 10. Lauro, F. M. *et al.* The genomic basis of trophic strategy in marine bacteria. *Proc. Natl.*
404 *Acad. Sci.* **106**, 15527–15533 (2009).

405 11. Martinez-Gutierrez, C. A. & Aylward, F. O. Strong purifying selection is associated with
406 genome streamlining in epipelagic Marinimicrobia. *Genome Biol. Evol.* **11**, 2887–2894
407 (2019).

408 12. Elena, S. F., Cooper, V. S. & Lenski, R. E. Punctuated evolution caused by selection of
409 rare beneficial mutations. *Science* **272**, 1802–1804 (1996).

410 13. Andersson, J. O. & Andersson, S. G. E. Genome degradation is an ongoing process in
411 Rickettsia. *Mol. Biol. Evol.* **16**, 1178–1191 (1999).

412 14. Moran, N. A. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc.*
413 *Natl. Acad. Sci. U. S. A.* **93**, 2873–2878 (1996).

414 15. Groisman, E. A. & Ochman, H. Pathogenicity islands: Bacterial evolution in quantum
415 leaps. *Cell* vol. 87 791–794 (1996).

416 16. Neuenschwander, S. M., Ghai, R., Pernthaler, J. & Salcher, M. M. Microdiversification in
417 genome-streamlined ubiquitous freshwater Actinobacteria. *ISME J.* **12**, 185–198 (2018).

418 17. Charlesworth, B., Lande, R. & Slatkin, M. A neo-Darwinian commentary on
419 macroevolution. *Evolution* **36**, 474–498 (1982).

420 18. Futuyma, D. J. Evolutionary constraint and ecological consequences. *Evolution* **64**, 1865–
421 1884 (2010).

422 19. Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial
423 speciation. *Science* **315**, 476–480 (2007).

424 20. Friedman, J., Alm, E. J. & Shapiro, B. J. Sympatric speciation: when is it possible in
425 bacteria? *PLoS One* **8**, e53539 (2013).

426 21. Bobay, L.-M. & Ochman, H. Biological species are universal across life's domains.
427 *Genome Biol. Evol.* **9**, 491–501 (2017).

428 22. Estes, S. & Arnold, S. J. Resolving the paradox of stasis: models with stabilizing selection
429 explain evolutionary divergence on all timescales. *Am. Nat.* **169**, 227–244 (2007).

430 23. Philippot, L. *et al.* The ecological coherence of high bacterial taxonomic ranks. *Nat. Rev.*
431 *Microbiol.* **8**, 523–529 (2010).

432 24. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific
433 marker genes. *Nat. Methods* **9**, 811–814 (2012).

434 25. Von Mering, C. *et al.* Quantitative phylogenetic assessment of microbial communities in
435 diverse environments. *Science* **315**, 1126–1130 (2007).

436 26. Louca, S. *et al.* Bacterial diversification through geological time. *Nat. Ecol. Evol.* **2**,
437 1458–1467 (2018).

438 27. Mayr, E. Speciation and macroevolution. *Evolution* **36**, 1119–1132 (1982).

439 28. Koonin, E. V. & Wolf, Y. I. Genomics of bacteria and archaea: The emerging dynamic
440 view of the prokaryotic world. *Nucleic Acids Res.* **36**, 6688–6719 (2008).

441 29. Ochman, H. & Davalos, L. M. The nature and dynamics of bacterial genomes. *Science*
442 vol. 311 1730–1733 (2006).

443 30. Wu, M. & Scott, A. J. Phylogenomic analysis of bacterial and archaeal sequences with
444 AMPHORA2. *Bioinformatics* **28**, 1033–1034 (2012).

445 31. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
446 large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

447 32. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - approximately maximum-likelihood

trees for large alignments. *PLoS One* **5**, e9490 (2010).

33. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985).

34. Sheridan, P. P., Freeman, K. H. & Brenchley, J. E. Estimated minimal divergence times of the major bacterial and archaeal phyla. *Geomicrobiol. J.* **20**, 1–14 (2003).

35. Betts, H. C. *et al.* Integrated genomic and fossil evidence illuminates life's early evolution and eukaryote origin. *Nat. Ecol. Evol.* **2**, 1556–1562 (2018).

Competing interests: Authors declare no competing interests.

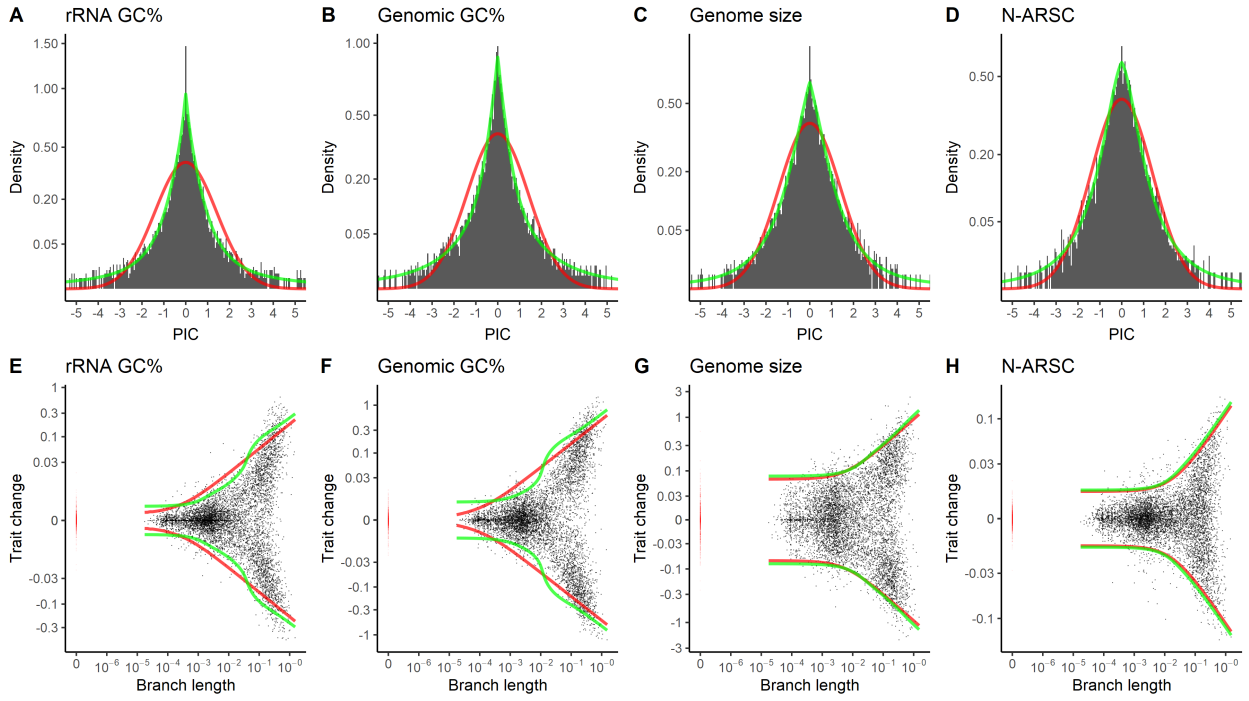


Fig. 1. Pulsed evolution models fit bacterial trait evolution better than the BM model. (A-D) PIC distributions (black bars) deviate significantly from the normal distribution of the BM model (red line). The pulsed evolution models that include two or three Poisson processes (PE2 or PE3, green line) greatly improve the fit to the PIC distributions. Square-root transformation is applied to the y-axis (density) to better show the deviation in the frequency of large PICs. (E-H) Patterns of bacterial trait changes at different branch lengths. Trait changes derived from the bacterial phylogeny are shown in black dots. Trait differences between genomes separated by zero branch length are shown in red dots. The expected 95% confidence intervals (CI) of the models are shown in colored lines (red line for the BM model, green line for the pulsed evolution model). Pseudo-log transformation is applied to the y-axis (trait change) to better show the trend of trait change in short branches.

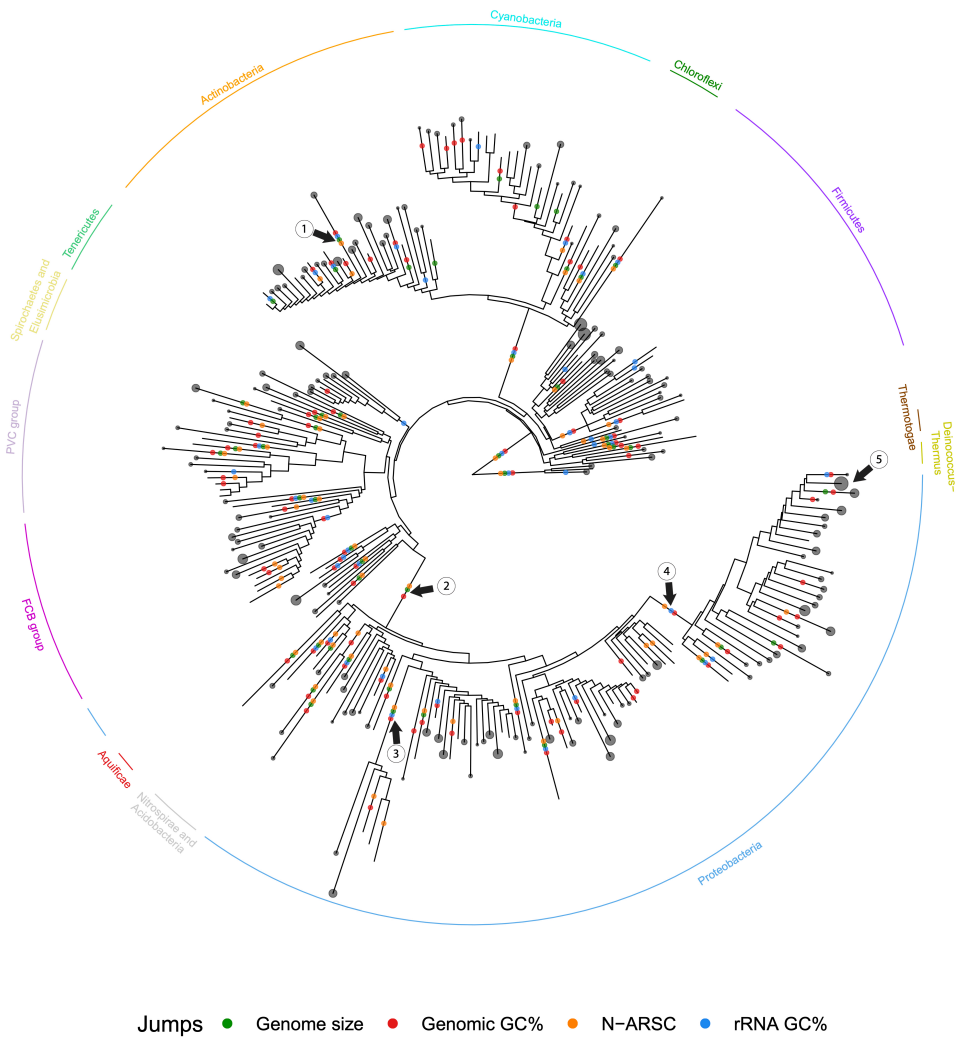


Fig. 2. Rare jumps are widely distributed throughout the bacterial phylogeny. For clarity, clades have been collapsed at the taxonomic rank order. Colored dots are placed on branches where the posterior probability of having at least one rare or super rare jump event is greater than 0.9. Arrows point to branches leading to 1. the order Candidatus Nanopelagicales. 2. the α-, β-, γ- and δ-proteobacteria. 3. the orders Pelagibacterales, Rickettsiales and Holosporales. 4. γ-proteobacteria. 5. the genera *Buchnera*, *Wigglesworthia* and *Candidatus Blochmannia* within the family Enterobacteriaceae.

481 **Table 1. The six trait evolution models tested in this study.**

Models	Free parameters
Brownian motion (BM)	σ_{BM}^2 , ε
One Poisson process (PE1)	λ_1 , σ_1^2 , ε
Two Poisson processes (PE2)	λ_1 , σ_1^2 , λ_2 , σ_2^2 , ε
Three Poisson processes (PE3)	λ_1 , σ_1^2 , λ_2 , σ_2^2 , λ_3 , σ_3^2, ε
One Poisson process and Brownian motion (PE1+BM)	σ_{BM}^2 , λ_1 , σ_1^2 , ε
Two Poisson processes and Brownian motion (PE2+BM)	σ_{BM}^2 , λ_1 , σ_1^2 , λ_2 , σ_2^2 , ε

482

483 **Table 2. List of AIC values for each model fitted for bacterial and archaeal trait evolution.**
484 The AIC values for the best model and models that are not significantly inferior (AIC change <2)
485 in each trait are in bold.

486

Domain	Trait	BM	PE1	PE2	PE3	PE1+BM	PE2+BM
Bacteria	Ribosomal RNA GC%	-41650	-43966	**-44810**	-44806	-44390	-44808
	Genomic GC%	-28539	-30921	**-31299**	-31295	-31171	-31297
	Genome size	-15572	-15929	-16075	**-16107**	-16022	-16090
	N-ARSC	-41944	-42123	-42214	**-42217**	-42214	**-42218**
Archaea	Ribosomal RNA GC%	-642	-766	**-802**	-798	-796	-800
	Genomic GC%	-440	-488	**-504**	-500	**-503**	-502
	Genome size	-348	**-359**	-355	-351	-357	-353
	N-ARSC	-1202	**-1212**	-1210	-1206	**-1211**	-1208

487
488

Table 3. Model statistics of pulsed evolution in different bacterial traits.

The 95% confidence interval for each model statistic is listed in the parentheses after the statistic.

Trait	Jump type	Jump rate (lineage^{-1}·unit branch length^{-1})	Relative jumps size	Jump contribution	Time-independent variance
Ribosomal RNA GC%	Rare	1.96 (1.62 – 2.79)	57.8 (50.2 – 63.6)	85.9% (83.8% – 88.3%)	1.81 (1.43 – 2.32) × 10^{-6}
	Frequent	118 (84.8 – 150)	3.0 (2.6 – 3.6)	14.1% (11.7% – 16.2%)	
Genomic GC%	Rare	6.31 (5.42 – 7.34)	34.2 (31.0 – 37.3)	92.7% (91.4% – 93.7%)	1.36 (1.12 – 1.70) × 10^{-5}
	Frequent	167 (101 – 261)	1.9 (1.4 – 2.4)	7.3% (6.3% – 8.6%)	
Genome size	Super rare	0.169 (0.06 – 0.26)	22.1 (19.3 – 35.9)	38.9% (30.3% – 49.1%)	1.40 (1.32 – 1.55) × 10^{-3}
	Rare	5.89 (3.99 – 11.8)	3.7 (2.7 – 5.1)	39.1% (31.0% – 49.9%)	
	Frequent	115 (48.3 – 271)	0.6 (0.4 – 1.0)	22.0% (14.8% – 31.5%)	
N-ARSC	Super rare	0.367 (0.11 – 3.85)	7.8 (6.1 – 11.1)	19.8% (10.4% – 67.3%)	3.54 (3.14 – 3.80) × 10^{-5}
	Rare	7.82 (5.06 – 15.5)	2.8 (2.2 – 3.7)	54.1% (33.6% – 72.1%)	
	Frequent	634 (274 – 1610)	0.2 (0.1 – 0.3)	26.1% (17.1% – 33.8%)	

Table 4. Differences in the percentage of contrasts with at least one rare or super rare jump between those inferred from the empirical data and the expectation from the null hypothesis.

Significant differences are marked with asterisks. P-values and power (β) are listed in parentheses.

Trait	Species	Genus	Family	Order
Ribosomal GC%	-2.3%* (P<0.001, β>0.999)	+9.5%* (P<0.001, β>0.999)	+17.1%* (P<0.001, β =0.985)	+21.3%* (P<0.001, β=0.870)
Genomic GC%	-2.1% (P=0.060, β>0.999)	+8.1%* (P<0.001, β=0.995)	+8.9%* (P<0.001, β=0.610)	+11.5%* (P<0.001, β=0.205)
Genome size	-1.2% (P=0.150, β>0.999)	+3.5%* (P=0.050, β>0.999)	+4.9%* (P<0.001, β>0.999)	+4.1% (P=0.075, β>0.999)
N-ARSC	-0.2% (P=0.630, β>0.999)	-1.8% (P=0.240, β>0.999)	+4.1% (P=0.06, β=0.990)	+3.3% (P=0.290, β=0.770)

Supplementary Text

Probability density of trait change under pulsed evolution

For simplicity, we model pulsed evolution similar to the JN model described in Landis and Schraiber [8]. Specifically, we assume that the pulsed evolution occurs at a constant rate relative to molecular divergence (branch length), causing sudden changes (jumps) in the trait value. We assume that the sizes of these jumps follow a normal distribution with a mean of zero and a fixed variance. As a result, the pulsed evolution is modeled as a compound Poisson process with normal jumps, with parameters λ and σ^2 denoting the frequency and the magnitude (variance) of the jumps, respectively. And between any two tips with branch length Δl in the phylogeny, the trait change Δx_{PE} introduced by pulsed evolution follows the probability density function:

$$\Phi_{PE}(\Delta x_{PE}, \Delta l) = \sum_{n=0}^{\infty} \Phi_{norm}(\Delta x_{PE}, n\sigma^2) \cdot \Phi_{Poisson}(n, \lambda \Delta l)$$

where n denotes the number of jumps occurred over the branch, $\Phi_{norm}(x, \sigma^2)$ denotes the normal probability density function of the random variable x with a mean of 0 and variance σ^2, and $\Phi_{Poisson}(n, \lambda)$ denotes the Poisson probability mass function of n events with expected occurrence of λ. When more than one Poisson processes are involved, the probability density function changes to:

$$\Phi_{PE}(\Delta x_{PE}, \Delta l) = \sum_{n_1=0}^{\infty} \dots \sum_{n_i=0}^{\infty} \left(\Phi_{norm}(\Delta x_{PE}, \sum_{j=1}^{i} n_j \sigma_j^2) \cdot \prod_{j=1}^{i} \Phi_{Poisson}(n_j, \lambda_j \Delta l) \right)$$

where i is the total number of Poisson processes, and j denotes the specific compound Poisson process a variable belongs.

We model the gradual evolution using the classic Brownian motion model with a single parameter σ_{BM}^2 denoting the rate of the gradual trait change. The trait change Δx_{BM} introduced by gradual evolution follows the probability density function:

$$\Phi_{BM}(\Delta x_{BM}, \Delta l) = \Phi_{norm}(\Delta x_{BM}, \sigma_{BM}^2 \Delta l)$$

Meanwhile, we have observed trait variations between genomes with identical marker gene alignment (zero branch length), suggesting that branch length does not explain all the trait variation between tips. Consequently, we introduce the time-independent variation to the model. Because the distribution of the observed time-independent variation is leptokurtic (i.e., heavy-tailed, with positive excess kurtosis), we model it with the Laplace distribution with a mean of 0 and scale parameter b for simplicity and convenience. The probability density function of x following a Laplace distribution is denoted as $\Phi_{Laplace}(x, 2b^2)$, where $2b^2$ is its variance and corresponds to the ε term in our models (i.e., $\varepsilon = 2b^2$). And thus, the time-independent trait change Δx_ε follows the probability density function:

$$\Phi_\varepsilon(\Delta x_\varepsilon, \Delta l) = \Phi_{Laplace}(\Delta x_\varepsilon, \varepsilon)$$

To put these three components together, we first derive the convoluted probability density function of a zero-mean normal distribution and a zero-mean Laplace distribution $\Phi^*(x, \sigma^2, 2b^2)$ as

$$\Phi^*(x, \sigma^2, 2b^2) =$$

$$\frac{1}{2b}\{e^{\frac{\sigma^2}{2b^2}+\frac{x}{b}}[1-\Psi_{norm}(x,-\frac{\sigma^2}{b},\sigma^2)\,]+\ e^{\frac{\sigma^2}{2b^2}-\frac{x}{b}}\cdot\Psi_{norm}(x,-\frac{\sigma^2}{b},\sigma^2)\}$$

where σ^2 is the variance of the normal distribution, $2b^2$ is the variance of the Laplace distribution and $\Psi_{norm}(x,-\frac{\sigma^2}{b},\sigma^2)$ denotes the cumulative distribution function of a normally distributed x with a mean of $-\frac{\sigma^2}{b}$ and variance σ^2.

With the convoluted probability density function, we derive that the trait change Δx between any two tips with branch length Δl in the phylogeny (with one compound Poisson process) follows the probability distribution:

$$\Phi(\Delta x,\Delta l)=\sum_{n=0}^{\infty}\Phi^*(\Delta x,\sigma_{BM}^2\Delta l+n\sigma^2,\varepsilon)\cdot\Phi_{Poisson}(n,\lambda\Delta l)$$

And in the case where more than one compound Poisson processes are involved, the probability density function of the trait change becomes:

$$\Phi(\Delta x,\Delta l)=\sum_{n_1=0}^{\infty}\dots\sum_{n_i=0}^{\infty}\left(\Phi^*(\Delta x,\sigma_{BM}^2\Delta l+\sum_{j=1}^{i}n_j\sigma_j^2\,,\varepsilon)\cdot\prod_{j=1}^{i}\Phi_{Poisson}(n_j,\lambda_j\Delta l)\right)$$

It should be noted that for trait changes that involve internal nodes, the uncertainty of estimated ancestral states (variance) will replace the ε term in the above calculation.

<u>Posterior probability of jumps between two sister nodes</u>

We calculate the posterior probability of having n jumps between two sister nodes. Specifically, with one compound Poisson process, for a trait change Δx over branch length Δl, the posterior probability of having n jumps between the two nodes is

$$P(n|\Delta x)=\frac{\Phi^*(\Delta x,\sigma_{BM}^2\Delta l+n\sigma^2,\varepsilon)}{\Phi(\Delta x,\Delta l)}\cdot\Phi_{Poisson}(n,\lambda\Delta l)$$

And in the case where more than one Poisson processes are involved, the posterior probability of having n_1 jumps in one of the Poisson process becomes

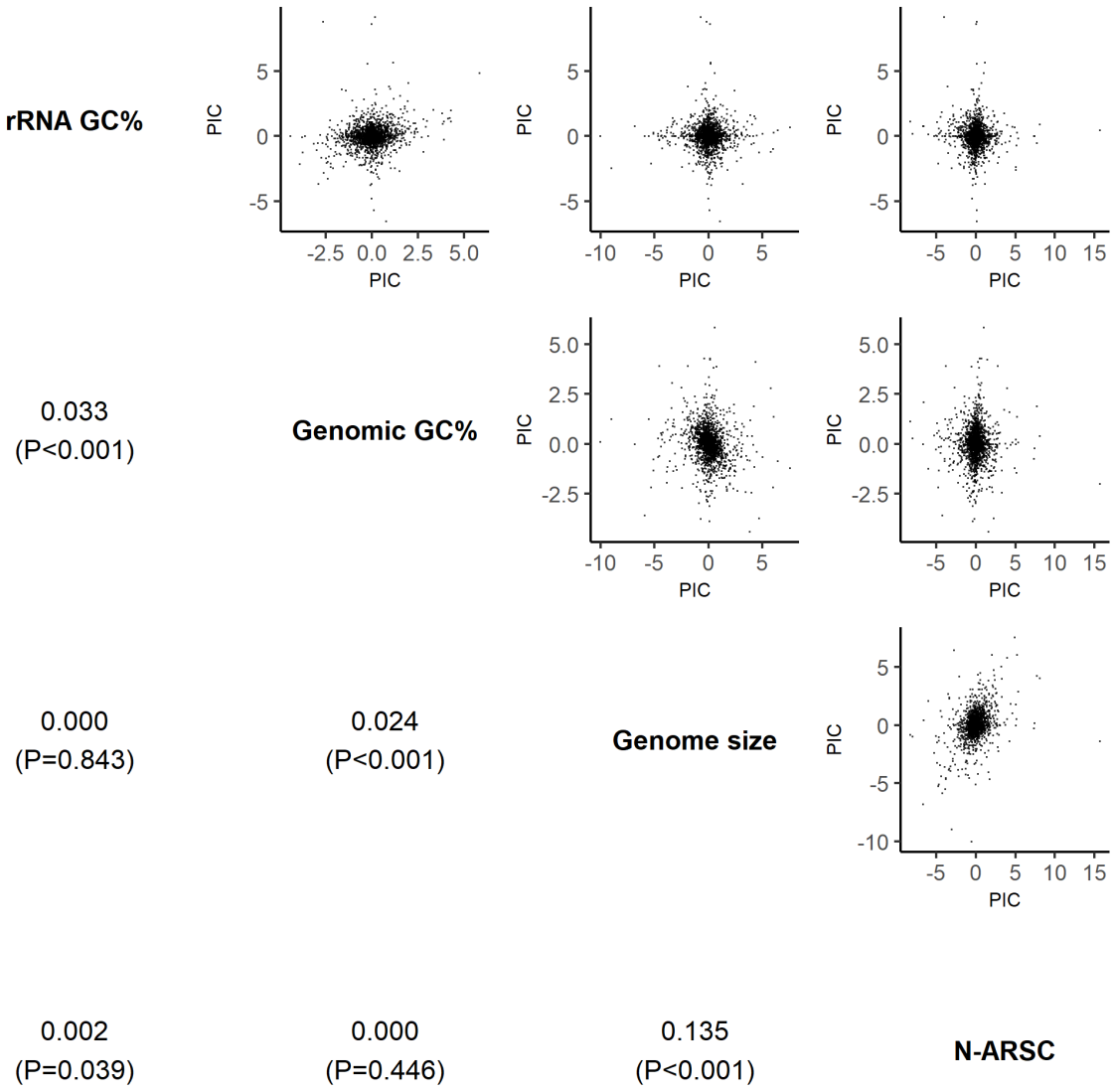
$$P(n_1|\Delta x)=\frac{\Phi_{Poisson}(n_1,\lambda_1\Delta l)}{\Phi(\Delta x,\Delta l)}\cdot$$

$$\sum_{n_2=0}^{\infty}\dots\sum_{n_i=0}^{\infty}\left(\Phi^*(\Delta x,\sigma_{BM}^2\Delta l+n_1\sigma_1^2+\sum_{j=2}^{i}n_j\sigma_j^2\,,\varepsilon)\cdot\prod_{j=2}^{i}\Phi_{Poisson}(n_j,\lambda_j\Delta l)\right)$$

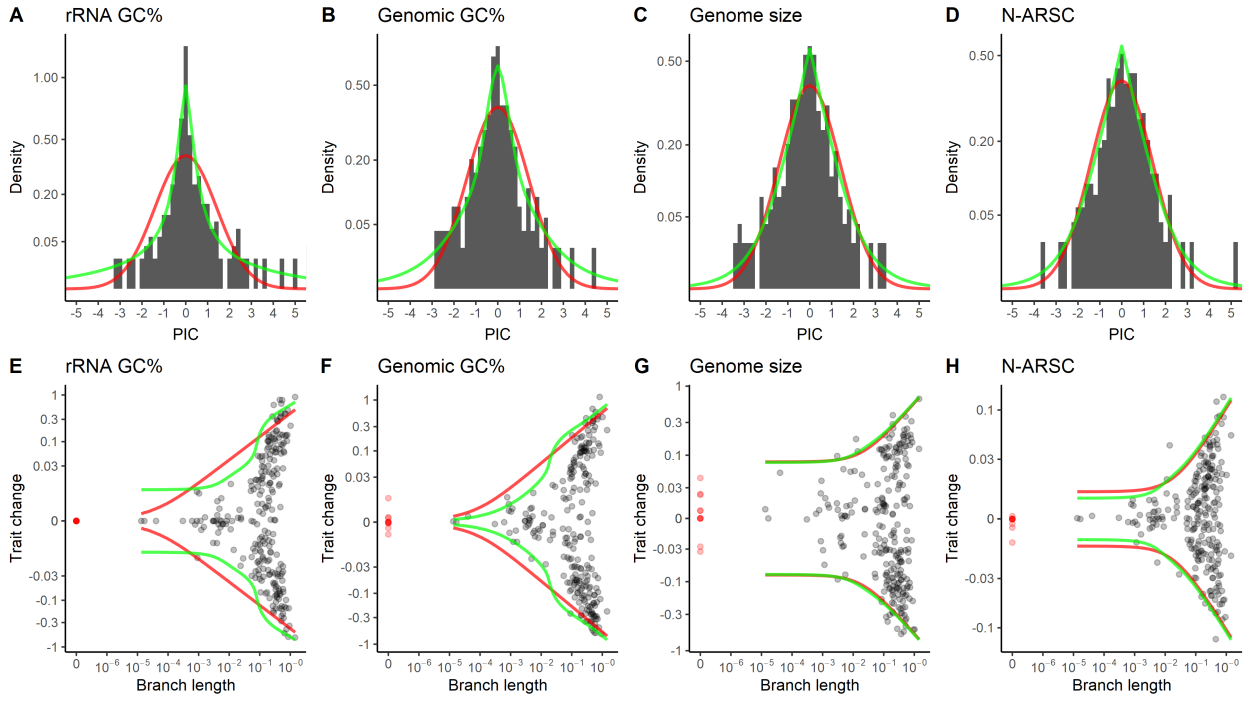
where i is the total number of Poisson processes in the pulsed evolution model.

<u>Implementation</u>

The functions and algorithms described above are implemented in R.

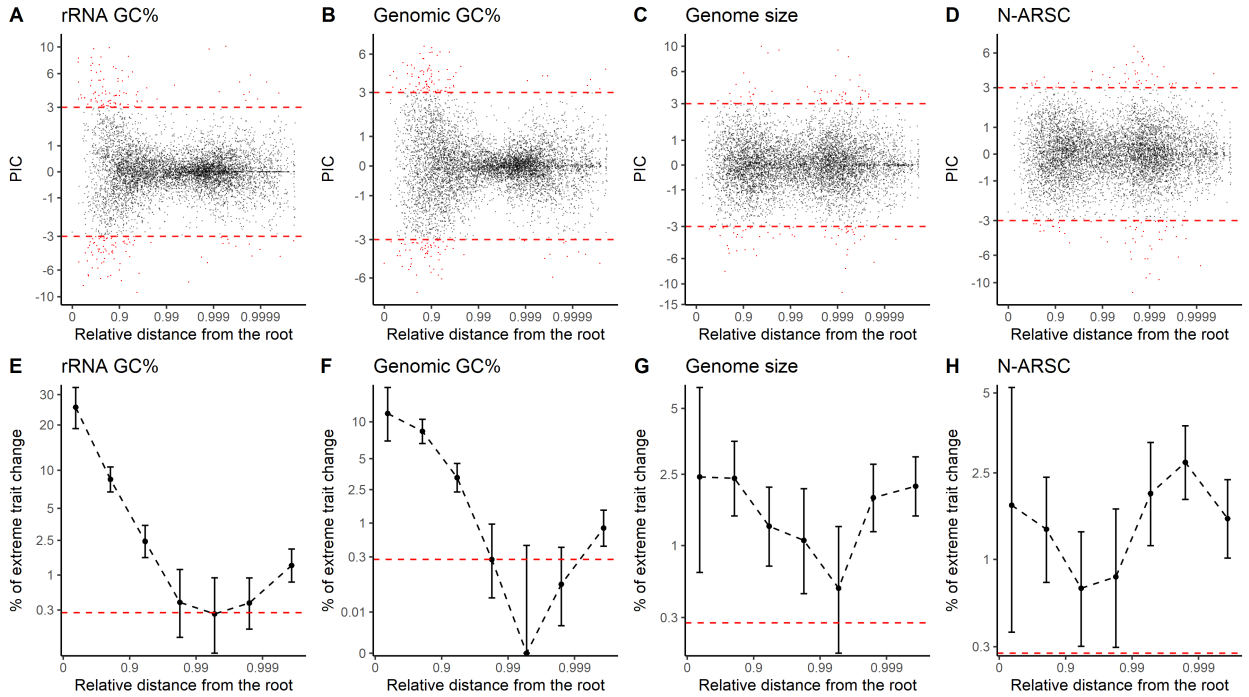


571
572 **Supplementary Figure 1. The pairwise correlation between the four genomic traits**. Upper
573 triangle: the scatter plots of Phylogenetically Independent Contrasts (PICs) between different
574 traits. Lower triangle: the corresponding coefficient of determination R^2 and P value of the
575 correlations.
576



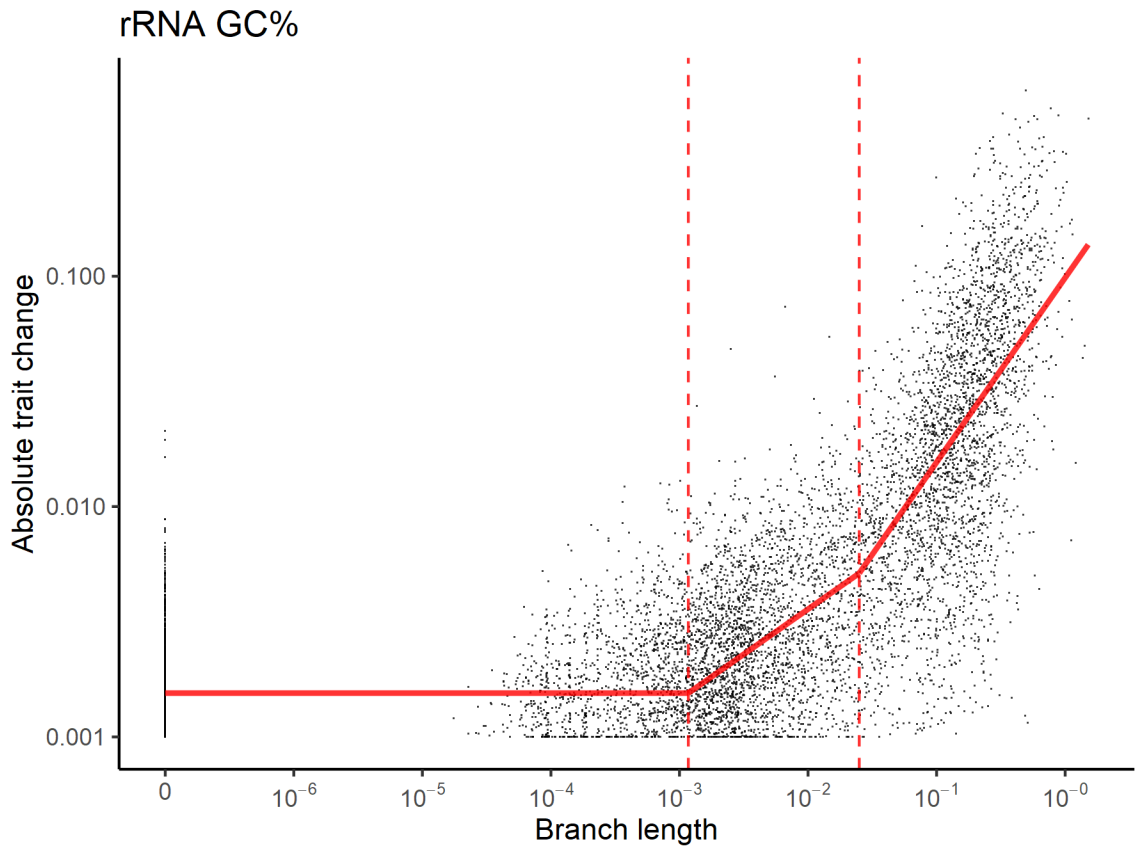
Supplementary Figure 2. Pulsed evolution models fit archaeal trait evolution better than the BM model.

(A-D) Histogram shows that the PIC (black bar) pattern deviates significantly from the expectation of the BM model (red line), while it is better described by the pulsed evolution model with one or two Poisson processes (PE1 or PE2, green line). The y-axis (density) is square-root transformed to better show the deviation in the frequency of large PICs. (E-H) Trait changes derived from the archaeal phylogeny are shown in black dots. Trait differences derived from genomes separated by zero branch length are shown in red dots. The expected 95% CI of BM and pulsed evolution (PE1 or PE2) models are shown by red and green lines, respectively. The y-axis (trait change) is pseudo-log transformed to better show the trend of trait change in short branches.



Supplementary Figure 3. Extremely large PICs occur more frequently than expected by BM throughout the bacterial evolution history.

Distributions of PIC over relative distance from the root are shown for (A) ribosomal RNA GC% (B) genomic GC% (C) genome size and (D) nitrogen use per amino acid N-ARSC. Extremely large PICs (|PIC|>3, outside the red dashed lines) are highlighted in red. Frequencies of the extremely large PIC over relative distance from the root are shown for (E) ribosomal RNA GC% (F) genomic GC% (G) genome size and (H) N-ARSC. The red dashed lines in (E-H) represent the expected frequency of extreme PIC by the BM model. Error bars in (E-H) represents the 95% confidence interval of extremely large PICs' frequency in each bin.



Supplementary Figure 4. Segmented linear regression analysis indicates that the tempo of evolution changes at various points for ribosomal RNA GC%

The distribution of absolute trait change over branch length is shown for ribosomal RNA GC%. The fitted relationship between the mean absolute trait change and branch length is shown in solid red lines, and the time points where the tempo of evolution changes are marked by dashed red lines.