G3, 2022, 12(8), jkac108
https://doi.org/10.1093/g3journal/jkac108
Advance Access Publication Date: 17 May 2022
Investigation

Evolutionary jumps in bacterial GC content

Saurabh Mahajan [1,2,*] Deepa Agashe[1]

[1]National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bengaluru 560065, India,
[2]Atria University, Bengaluru 560024, India

*Corresponding author: Atria University, Bengaluru, Karnataka, India. Email: saurabh.mk@gmail.com

Abstract

Genomic GC (Guanine-Cytosine) content is a fundamental molecular trait linked with many key genomic features such as codon and amino acid use. Across bacteria, GC content is surprisingly diverse and has been studied for many decades; yet its evolution remains incompletely understood. Since it is difficult to observe GC content evolve on laboratory time scales, phylogenetic comparative approaches are instrumental; but this dimension is rarely studied systematically in the case of bacterial GC content. We applied phylogenetic comparative models to analyze GC content evolution in multiple bacterial groups across 2 major bacterial phyla. We find that GC content diversifies via a combination of gradual evolution and evolutionary "jumps." Surprisingly, unlike prior reports that solely focused on reductions in GC, we found a comparable number of jumps with both increased and decreased GC content. Overall, many of the identified jumps occur in lineages beyond the well-studied peculiar examples of endosymbiotic and AT-rich marine bacteria and do not support the predicted role of oxygen dependence. Our analysis of rapid and large shifts in GC content thus identifies new clades and novel contexts to further understand the ecological and evolutionary drivers of this important genomic trait.

Keywords: phylogenetic models; Lévy jumps; ecological drivers

Introduction

GC content refers to the fraction or percentage of GC base pairs in a genome. The GC content of bacterial genomes varies from as low as ∼13% for *Zinderia insecticola* (McCutcheon and Moran 2011) to as high as ∼75% for *Aneromyxobacter dehalogenans* (Thomas *et al.* 2008). Moreover, across bacteria the GC content of four fold degenerate codon sites varies from 5% to 95%, i.e. almost no GC base pairs to only GC base pairs (Muto and Osawa 1987; Hershberg and Petrov 2010). Such differences in GC content profoundly affect critical components of the expression of genomic information, including the usage of different synonymous codons (Knight *et al.* 2001), tRNA pools and tRNA modifying enzymes (Diwan and Agashe 2018), and amino acids (Knight *et al.* 2001; Lightfield *et al.* 2011). Given its fundamental importance for the maintenance and transfer of genetic information, the diversity of GC content and its evolutionary determinants have been investigated for many decades (Sueoka 1961). In general, the GC content of a sequence must be determined by a combination of biases in the mutational process, biases in the fixation process (selection or recombination), and drift. Although these forces may act differently on different sequences within a genome, the GC content of different regions such as intergenic regions, RNA coding genes, and protein coding genes and different codon positions within them are correlated to each other (Muto and Osawa 1987; Zhu *et al.* 2010; Raghavan *et al.* 2012; Brocchieri 2014). Thus, the GC content of a genome can be considered a single trait evolving under a set of common evolutionary pressures.

It is now well accepted that mutations in most bacteria (and also archaea and eukaryotes) are biased toward AT (Hershberg and Petrov 2010; Hildebrand *et al.* 2010) and the actual GC content of most bacteria is typically higher than expected based only on this mutation bias. Thus, on top of the underlying mutation bias, there is almost certainly also a fixation bias such that GC → AT mutations are preferentially removed or AT → GC mutations are favored. This fixation bias could be due to selection for higher GC content (Hershberg and Petrov 2010; Hildebrand *et al.* 2010), or due to a biased recombination process arising from GC-biased gene conversion (Lassalle *et al.* 2015). Although there are differences in the extent of the mutation bias such that mutations in AT rich bacteria are also more biased toward AT (Long *et al.* 2018), it is not clear if differences in the fixation bias contribute to GC content diversity. In addition, a number of ecological factors have been proposed to be correlated with GC content (Agashe and Shankar 2014), e.g. host-association (Moran 2002), aerobiosis (Naya *et al.* 2002; Aslam *et al.* 2019), nitrogen fixation (McEwan *et al.* 1998), and temperature (Musto *et al.* 2004). However, many factors do not show strong correlations with GC content after accounting for the phylogeny or other confounding factors (Marashi and Ghalanbor 2004; Wang *et al.* 2006; Vieira-Silva and Rocha 2008; Aslam *et al.* 2019). Thus, the relationship between ecological factors and GC content diversity is also not clearly understood.

Since change in genome-wide GC content is a slow process, comparative analysis is by far the most informative approach to

Received: February 14, 2022. **Accepted:** April 20, 2022

investigate the evolutionary determinants of GC content. Such studies of GC content diversity across bacteria have provided useful datasets and insights (Hershberg and Petrov 2010; Hildebrand *et al.* 2010; Bobay and Ochman 2017; Long *et al.* 2018). Across the range of GC content observed in bacteria, there are several trends characteristic of different bacterial groups. For instance, most Actinobacteria are GC rich (average >60%), whereas Firmicutes are GC-poor (average ∼40%) [for recent data, see (Lightfield *et al.* 2011; Reichenberger *et al.* 2015)]. Typically, closely related bacteria have similar GC content (Haywood-Farmer and Otto 2003), although there are many well-studied exceptions. For instance, multiple lineages of insect endosymbionts and surface ocean dwelling bacteria have independently evolved exceptionally low GC content compared to their closest relatives (Moran *et al.* 2008; Giovannoni *et al.* 2014). Perhaps the most well-known example of the first kind are bacteria from the genus *Buchnera*, endosymbionts of aphids, whose average GC content is <30% compared to the ∼50% GC content of related Enterobacteria (Lai and Baumann 1992; Moran 1996). Endosymbionts of many other insects also show similar trends of drastically reduced GC content (Moran *et al.* 2008). A well-known marine bacterium with exceptionally low GC content is *Pelagibacter*, an α-proteobacterium with a GC content of ∼30% compared to ∼50–60% of most other α-proteobacteria (Giovannoni *et al.* 2014). Similarly, other marine bacterial lineages are also highly AT-rich (Grzymski and Dussaq 2012; Ghai *et al.* 2013; Giovannoni *et al.* 2014; Luo *et al.* 2017). The extremely low GC contents of endosymbionts or AT-rich marine bacteria are clearly derived from a higher ancestral GC content (that is closer to their respective relatives) by drastic reductions in a relatively short time. The large changes in GC content of these lineages are explained by peculiar biological circumstances such as reduction in overall selection efficiency accompanying endosymbiosis (Moran *et al.* 2008; Wernegreen 2015) or intense selection associated with nutrient-poor surface ocean waters (Giovannoni *et al.* 2014). In contrast, closely related clades do not appear to have undergone such rapid changes in GC content. Thus, the evolution of GC content evolution appears to proceed differently in these special lineages vis-a-vis their relatives. However, it is not clear whether the distinct modes of evolution are a general feature of GC content diversification across bacteria.

Perhaps even more intriguingly, hitherto there are no reports of bacterial lineages with large increases in GC content. It is possible that such lineages exist, but have simply not been identified yet. On the other hand, large changes in GC may occur only in specific biological circumstances that cause reductions in GC content. These alternative scenarios have interesting implications for the diversity of GC content and its evolutionary drivers. If one were to find instances of large increases in GC, it would immediately raise many interesting questions. How frequently do they occur and in what biological circumstances? What are the evolutionary forces behind such changes, and are they similar to those causing GC reductions? Identification of such cases would also broaden the available datasets to better understand the evolution of GC content.

Questions about the generality of the different modes of GC evolution and its direction can be addressed using phylogenetic models of trait evolution (Felsenstein 1985; Pagel and Harvey 1989). These models are regularly used to study the evolution of morphological (Barkman *et al.* 2008; Landis and Schraiber 2017; Baker and Venditti 2019), behavioral (Remeš *et al.* 2015; Hagey *et al.* 2017), and molecular traits (Liedtke *et al.* 2018; Stern and Crandall 2018) of animals or plants. Most simply, trait evolution on a phylogeny is modeled according to a Brownian process, i.e. as random changes accumulating at a constant rate without direction or constraint (Felsenstein 1985); or according to an Ornstein-Uhlenbeck process, i.e. as random changes occurring at a constant rate but with an attraction toward an optimal value (Hansen 1997). Modifications of these simple models capture more realistic evolutionary scenarios where the rate of trait evolution, the optimal value of a trait, or direction of evolution may differ across lineages (Butler and King 2004; O'Meara *et al.* 2006; Beaulieu *et al.* 2012). Another class of models based on the Lévy process capture qualitatively distinct evolutionary scenarios, where trait evolution is discontinuous due to occasional jumps in addition to accumulation of random changes at a constant rate (Duchen *et al.* 2017; Landis and Schraiber 2017). Comparison of the fit of different phylogenetic models and parameter variation across the phylogeny can provide interesting insights into the tempo and mode of trait evolution and their ecological and evolutionary correlates or mechanisms. Unfortunately, very few studies (Haywood-Farmer and Otto 2003; Baidouri *et al.* 2016) have used such approaches to understand bacterial trait evolution. The evolution of the GC content of bacteria was previously analyzed using this approach (Haywood-Farmer and Otto 2003), but before the advent of large datasets and sophisticated trait evolution models. This prior study found that GC content evolution is consistent with a Brownian model of evolution, implying gradual evolution at a constant rate. The discovery of bacterial lineages with rapid changes in GC content highlights the need for an expanded analysis with much larger and comprehensive datasets and new methods. Specifically, several phylogenetic models incorporating large jumps are now available and allow the inference of jumps in trait evolution (Duchen *et al.* 2017). These can be applied to large datasets of bacterial taxa to investigate GC content evolution.

We analyzed the macroevolutionary patterns in bacterial GC content using such phylogenetic models. We found that GC content evolution is better explained by a combination of 2 modes of evolution: gradual diversification and relatively large "jumps." In addition, we identified specific branches that experience such jumps, and analyzed the ecological context in which they occur. We find that large changes in GC content are ubiquitous across bacteria, are not restricted to endosymbionts and marine lineages, and are not consistently related to changes in oxygen dependence. Interestingly, we found a large number of previously unrecognized instances of rapid increase in GC content. The macroevolutionary patterns found here raise further questions and provide interesting datasets to analyse the microevolutionary causes of GC content evolution in bacteria.

Methods

The methods used in this study are summarized in Fig. 1.

Datasets

We started with a phylogeny of ∼22,000 bacterial genomes inferred from universally conserved ribosomal proteins that was downloaded from the Genome Taxonomy Database (release 80) (Parks *et al.* 2018). Since this phylogeny is based on protein sequences, synonymous codon usage bias has not direct influence on its inference. The branch lengths of this phylogeny represent the number of substitutions per site. Due to uneven genome sequencing efforts, clades are unevenly sampled in this phylogeny; e.g. some bacterial species are represented by hundreds of genomes of different strains. These add redundant information about GC content diversity and may bias subsequent analyses.

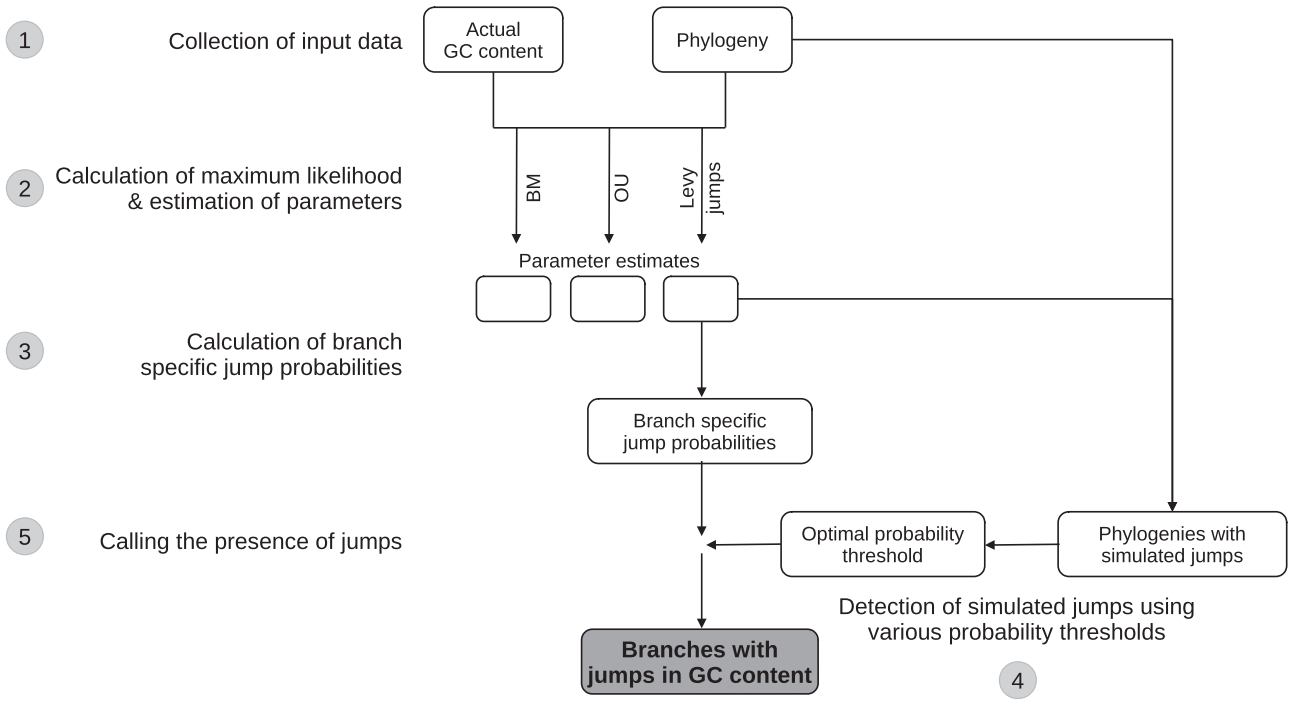

Fig. 1. Summary of methods used in this study. Each major step in the analysis is numbered in the order in which it was performed. The analysis was performed independently for each of 10 order-level clades belonging to 2 bacterial phyla. Step 1: We derived the phylogenies of major bacterial clades and GC content of taxa from the genome taxonomy database (GTDB). Step 2: We obtained the ML and parameter estimates for different phylogenetic models using the phylogenies and the GC content distributions as the input. For the Brownian motion (BM) and Ornstein-Uhlenbeck (OU) models, these were obtained by exact analytical solutions implemented in the *geiger* package in R; while for the Lévy jumps model, these were obtained by expectation-maximization (EM) + markov chain monte carlo (MCMC) sampling method implemented in the *levolution* software. Step 3: For each branch in a phylogeny, we obtained the posterior probability of a jump in GC content using the phylogeny, GC contents, and the best-fit estimates of the parameters of the Lévy jumps model (obtained from step 2). These probabilities were obtained using an empirical Bayes approach implemented in the *levolution* software. Step 4: To calculate posterior probability thresholds to decide the presence or absence of jumps, we first simulated data with GC content jumps. The simulations were performed on the original phylogenies using the best-fit estimates of the parameters of the Lévy jumps model. We then attempted to detect the known jumps in simulated data using various posterior probability thresholds. We chose posterior probability thresholds that led to an optimal choice between precision and recall of the simulated jumps. Step 5: We deemed branches whose posterior probability of experiencing a jump (calculated from actual data in step 3) was greater than the optimal probability threshold (calculated from simulated data in step 4) as those having experienced a jump in GC content.

Therefore, we used a custom script to subsample and retain fewer representatives from densely sampled clades. Briefly, we scanned clades descending from every internal node, and retained only single taxa from clades younger than a specific threshold (0.01 substitutions per site). We chose this threshold in the following way. From the larger phylogeny, we first extracted a smaller clade containing Enterobacteria (which includes many densely sequenced species such as *Escherichia coli* and *Salmonella enterica*). We subsampled this clade with increasing threshold values retaining single taxa from clades younger than this threshold, and picked the threshold where the retained taxa consisted of only one or few strains from each named bacterial species.

To circumvent potential heterogeneity in the macroevolutionary process across distantly related branches and to reduce time required for downstream analyses, we extracted and analyzed subclades roughly at the level of taxonomic orders from the 2 largest bacterial phyla with genomic data: Bacteroidetes and Proteobacteria. Henceforth, the subclades are referred to as "order-level clades" and are specifically referred by the major taxonomic order contained in each of them. Across the 2 phyla, we analyzed 10 order level clades, each containing between ~200 and ~800 taxa.

We obtained genomic GC content data corresponding to all the analyzed genomes from the Genome Taxonomy Database (release 80) (Parks *et al.* 2018).

Comparing the fit of trait evolution models

We compared the likelihood of observed GC content distribution across the phylogeny, under 3 models of trait evolution: the single-rate Brownian (Felsenstein 1985), single-optimum Ornstein-Uhlenbeck ("OU") (Hansen 1997), and the Brownian + stochastic jumps i.e. Lévy jumps (Duchen *et al.* 2017). The Brownian model describes a scenario where the trait value stochastically increases or decreases by a fixed amount per unit time, causing variance to increase at a constant rate (called the Brownian rate σ_0^2). The Ornstein-Uhlenbeck model includes an additional component that "pulls" the trait value toward an optimal value (parameter θ) with a speed that depends on a strength parameter (α) and the difference between the current and optimal value. A Lévy jumps model models a trait that evolves at a constant rate as in a Brownian model, but also experiences additional evolutionary changes as discrete and stochastic events (called "jumps"). In the model formulation used here (Duchen *et al.* 2017), these jumps are assumed to occur according to a Poisson process with frequency λ across the phylogeny. The average magnitude of jumps (i.e. changes in trait value) is modeled as a multiple (α) of the Brownian rate of evolution, although individual jump magnitudes are drawn from a normal distribution (Appendix I of Duchen *et al.* 2017).

We calculated the likelihood of data under the Brownian and OU models and the corresponding best-fit parameter estimates

using the geiger package in R (Harmon *et al.* 2008). For the Lévy jumps model, we obtained the likelihood and the best-fit values of all parameters (except α) using an Expectation Maximization + Markov Chain Monte Carlo procedure implemented in the *levolution* software (Duchen *et al.* 2017). However, this procedure cannot directly estimate the maximum likelihood (ML) value of α, making it necessary to calculate the likelihood and estimate other model parameters independently for a range of α values, and then to choose an α that results in ML. For this purpose, we evaluated α values in the set 0.1, 0.25, 0.5, 1, 2, and 4, i.e. letting the variance contributed by a jump vary between 10 times less and 4 times more than the Brownian rate. From these separate calculations, we chose the α value and corresponding parameter estimates that resulted in the largest ML. In all the clades, ML values peaked in the range of α that were evaluated.

Identifying evolutionary jumps in GC content

Given the estimated parameter values of a Lévy jumps model, one can also infer the phylogenetic location of jumps using the procedure implemented in the *levolution* software. This is accomplished by scanning multiple combinations of putative jump locations (branches) and then evaluating the posterior probabilities (pp) of one or more jumps occurring on every branch using an empirical Bayes approach (Duchen *et al.* 2017). There are 2 important issues that should be noted here. First, this procedure only allows the calculation of the posterior probability of a branch experiencing *one or more jumps* as defined in the theoretical model (Poisson events that introduce a specified amount of change in the trait value). This implies that one cannot know the exact number of theoretical jumps that are likely to have occurred on that branch. However, multiple theoretical jumps on a branch can be considered empirically equivalent to a single, but larger evolutionary change on the same branch. Therefore, subsequently in this study, we refer to the cumulative evolutionary change occurring on a single branch as a GC content jump. The second issue is that inferring jump locations is not a matter of yes or no, but of choosing an appropriate *pp* threshold above which we can reliably call a branch as having experienced a jump in trait value.

Ideally, all branches that have experienced a jump in trait value must have $pp \approx 1$ and all others, $pp \approx 0$. Thus, a high *pp* threshold should capture most jump locations accurately. In reality, this is not the case. Especially when the magnitude of jumps is small compared to the Brownian component, many branches that have experienced a jump have *pp* values much lower than 1 (Duchen *et al.* 2017). Therefore, choosing a high *pp* threshold can lead to a low "recall" of actual jumps. On the other hand, lowering the *pp* threshold to capture all jumps selects for many branches that have not actually experienced a jump (and therefore rightly received lower pp). This lowers the "precision" of jump inference. Altogether, as the *pp* threshold is varied, there is a negative relationship between precision and recall.

Hypothetically, if the real jumps in any evolutionary history were known, one could choose a *pp* threshold that optimized precision and recall of the inference procedure. Of course, we do not know the location of actual jumps in GC content. Therefore, we determined *pp* thresholds that optimized both precision and recall of jump inference in simulated data with exactly known jumps.

For each order level clade, we decided *pp* thresholds in the following manner.

1) We modified the *ex.jumpsimulator()* function in the *geiger* package (Harmon *et al.* 2008) to simulate 5 independent datasets. In each simulation, GC content evolved according to a Lévy jumps model (described in the previous section) i.e. continuously changing according to a Brownian process with additional changes ("jumps") at branches selected stochastically according to a Poisson process. The parameters for the model: ancestral GC content, Brownian rate (σ_0^2), jump rate (λ), and the average relative jump magnitudes (α), were set to the best-fit estimates obtained from actual GC content data of the respective clades. The location of each simulated jump was recorded by the function used to simulate the datasets.
2) Using the phylogeny of the clade and the simulated GC content of only the tips as inputs, we followed the procedure in *levolution* (explained in the previous section) to estimate the branch-specific pp of jumps for each simulated dataset.
3) Independently for each *pp* threshold in a range of putative *pp* thresholds between 0 and 1, we calculated precision and recall in the following way:

   3.a.We inferred jumps in branches with *pp* higher than the threshold under consideration.
   3.b.We divided the inferred jumps into 2 types: "true jump estimates", when a jump was inferred on a branch with a simulated jump; and "false jump estimates" when a jump was inferred on a branch with no simulated jump.
   3.c.We pooled data across the 5 simulations and calculated the precision and recall as:
   3.c.i.precision = 100 x number of "true jump estimates"/(number of "true jump estimates" + "false jump estimates")
   3.c.ii.recall = 100 x number of "true jump estimates"/(number of simulated jumps).

We chose a *pp* threshold that led to at least 90% precision while trying to achieve maximum recall (Supplementary Fig. 1, Supplementary File 1). The chosen thresholds for different clades resulted in a precision between 91% and 97% and a recall between 2% and 37%. We deemed that all branches of an order-level clade with a posterior probability greater than the chosen threshold had experienced a jump in GC content. In every order-level clade, each such branch was assigned a unique serial number (referred to as "jump index") for reference.

When precision and recall was calculated separately for each simulated dataset instead of the pooled dataset, the chosen thresholds resulted in at least 80% and up to 100% precision in all cases (data summary in Supplementary Table 1, Supplementary File 2). The recall varied considerably across independent simulations: for clades with low overall recall, it varied from 0% to 12% across simulations; whereas for clades with modest recall, it varied from 25% to 42% across simulations.

Analysis of inferred jumps

To analyze the directions and magnitudes of the inferred jumps in GC content, we resorted to an approximate calculation because the procedure in *levolution* cannot estimate the magnitude of jumps occurring on each branch. Therefore, we quantified the impact of each jump by comparing the median GC content of all descendant taxa of the branch affected by the jump with the median GC content of all descendant taxa of the corresponding sister branch. If any descendant branches were also affected by additional nested jumps, we removed the corresponding descendant taxa from the calculation.

To test how often GC content jumps were associated with endosymbiosis (or other forms of host association) or with marine habitats, we searched for primary literature describing the isolation of taxa in clades affected by GC content jumps and their sister clades. We specifically looked for evidence of whether the organism could be cultured independently of a host. If the primary source mentioned that the organism could be isolated and grown independent of the host, we tagged it as: "not host dependent." If the organism was isolated from a host, then we tagged it as: "host associated." In this analysis, we excluded clades represented only with metagenome-assembled genomes, those with large sister clades containing diverse species, and taxa whose phylogenetic placement was unreliable (Supplementary Table 2, Supplementary File 2).

In addition, we also obtained data about the oxygen dependence (anaerobic, facultatively aerobic, aerobic, or obligately aerobic) of taxa from a recent compilation of bacterial phenotypes (Madin *et al.* 2020). We manually assigned oxygen dependence to entire clades (those experiencing GC jumps or sister clades) based on the oxygen dependence of the majority of taxa in each clade.

Results

A Lévy jumps model explains GC content evolution better than a Brownian model

As described in the *Methods* section, we separately analyzed GC evolution in 10 bacterial clades corresponding approximately to major orders from 2 large phyla (Table 1). For each order-level clade, we first visualized the distributions of scaled phylogenetically independent contrasts (PICs) (Felsenstein 1985). The distributions were significantly different compared to normal distributions with excess kurtosis ranging from ~3 to ~17, implying fat-tailed distributions (Supplementary Fig. 2, Supplementary File 1). We also notice few outliers as would be expected for evolutionary jumps indicating larger than expected changes. To quantitatively characterize and compare the possible evolutionary processes that may have led to these distributions, we evaluated and compared the likelihood of GC content distribution under a single-rate Brownian model and a single-optimum Ornstein-Uhlenbeck (OU) model. In all but 2 datasets, we found that the maximum likelihood estimate (MLE) of the constraint parameter (α) in the OU process was ~0 i.e. it just described a Brownian process without constraints. Moreover, in all cases the ML of the Brownian model was equal to the OU model (Table 1). Overall, GC content evolution was not consistent with constraint toward an optimal value. For this reason, we did not test the multi-optima OU models. Further, we found that in all cases, the Lévy jumps model (Duchen *et al.* 2017) explained the data significantly better than the single-rate Brownian model without jumps (Table 1). These results were consistent with our expectation based on the few known lineages with exceptional changes in GC content.

The estimated variance introduced by the Brownian component (σ_0^2), the jump rate (λ), and the total variance introduced by jumps relative to the Brownian component ($\lambda \cdot \alpha$) differed substantially across clades (Supplementary Table 3, Supplementary File 2). The rate (or variance contributed per unit branch length) of the Brownian evolution component varied ~4x, with Flavobacteriales having the lowest and Bacteroidales the highest rate. The estimated jump rate varied >2x from ~2.5 jumps per unit branch length in Bacteroidales to >6 jumps in Flavobacteriales. The total variance introduced by the jumps per unit branch length was between ~2x lower (Bacteroidales) to >3x higher (Rhizobiales) than the Brownian component. Thus, the impact of the baseline (Brownian) rate as well as jumps in GC content evolution, both vary across bacterial orders. The reason for this variation in the frequency of jumps and the relative contributions of jumps to GC content diversity across clades is not very clear.

Identification of branches experiencing GC content jumps

As pointed out earlier, using the procedure in *levolution*, one cannot predict the exact number or magnitude of jumps on each branch, but only estimate the pp of the presence of >0 jumps (as defined in the model). Reliable inference of the phylogenetic location of jump(s) then requires one to choose an appropriate *pp* threshold. Since the jumps in the actual data are not known, we used simulated data to determine appropriate *pp* thresholds that led to optimal precision and recall in the inference of jump

Table 1. Summary statistics of phylogenetic models describing the evolution of GC content in various order-level bacterial clades.

Clade	Phylum	A. Likelihood under various phylogenetic models				B. Summary parameters		
		ML (single-rate Brownian)	ML (single-optimum OU)	ML (Lévy jumps)	P-value (Lévy jumps vs. BM)	Number of taxa	Ancestral GC	Number of jumps
Cytophagales	Bacteroidetes	−445.252	−445.252	−426.24	5.5e−09	167	40.07	2 (1, 1)
Bacteroidale	Bacteroidetes	−1,886.49	−1,886.49	−1,831.52	1e−16	713	40.1	3 (0, 3)
Flavobacteriales	Bacteroidetes	−1,364.19	−1,364.19	−1,204.14	1e−16	609	39.89	73 (19, 54)
Acetobacterales and related orders	Proteobacteria (α)	−572.371	−572.371	−537.383	6.7e−16	198	62.84	7 (6, 1)
Sphingomonadales	Proteobacteria (α)	−540.688	−540.688	−471.786	1e−16	260	55.84	11 (9, 2)
Rhizobiales	Proteobacteria (α)	−1,040.84	−1,040.84	−863.626	1e−16	538	63.8	23 (17, 6)
Rhodobacterales	Proteobacteria (α)	−1,141.83	−1,141.83	−1,033.02	1e−16	469	63.38	27 (20, 7)
Betaproteobacteriales	Proteobacteria (γ)	−1,859.95	−1,859.949	−1,690.86	1e−16	770	58.12	24 (20, 4)
Enterobacterales	Proteobacteria (γ)	−1,441.81	−1,440.58	−1,314.91	1e−16	602	45.99	18 (7, 11)
Pseudomonadales	Proteobacteria (γ)	−1,485.58	−1,485.572	−1,406.66	1e−16	632	52.02	13 (11, 2)

The table shows 2 sets of data for 10 order-level clades of bacteria: (A) The ML of GC content distributions under 3 phylogenetic models. The MLs of data under the Brownian and OU model are almost identical because the maximum likelihood estimate (MLE) of α, the constraint parameter in OU, was zero in almost all cases, making all best-fit OU models effectively equivalent to the Brownian models. The *P*-value of Lévy jumps model being better fit compared to the Brownian motion model was calculated from a likelihood ratio test (LRT). (B) Some summary parameters. The number of jumps in the last column refers to the number of branches on which the posterior probability of detecting a jump was higher than the chosen threshold for each clade. As described in the methods, we treat each such branch as having experienced a single evolutionary jump in GC content. The numbers in parentheses refer to the number of downward and upward jumps, respectively.

location. Briefly, we simulated GC content evolution according to the best-fit parameters of the Lévy jumps model, then inferred branch-specific pp of jumps from the simulated GC contents of extant taxa, and determined the presence or absence of jumps on any branch according to a *pp* threshold. Varying this threshold and then comparing the inferred jump locations to the locations of simulated jumps allowed us to calculate precision and recall of jump detection under the different thresholds (see *Methods* for details).

In general, using high *pp* thresholds to infer jump locations leads to higher precision but poor recall, whereas using low *pp* thresholds to infer jump locations leads to lower precision but better recall in identifying branches with simulated jumps. However, the precision-recall relations of different order-level clades fell in 2 categories (Supplementary Fig. 1, Supplementary File 1). For 1 set of clades, decreasing *pp* thresholds led to a small decrease in the precision as recall increased to ∼30%, and a large decrease in precision thereafter. Sphingomonadales, Rhizobiales, Rhodobacterales, and Flavobacteriales are examples of this category. For a second set of clades, precision decreased rapidly and recall increased only slightly with decreasing *pp* thresholds. It is not clear why the precision-recall curves are different for these 2 categories, but may have to do with the specific tree topologies or taxon densities. Nevertheless, in the first case, we chose *pp* thresholds of 0.75 which lead to ∼90% precision and ∼30% recall in identification of simulated jumps. For the second case, we chose *pp* thresholds of 0.95 or 0.9 that also lead to ∼90% precision but only ∼2–20% recall. Although the recall appears very poor, larger jumps were detected more frequently as expected (Duchen *et al.* 2017). Simulated jumps with >5% GC content change had at least 40% recall, those with >10% GC content change had at least 60% recall, those with >15% GC content change had at least 80% recall, and finally those with >20% GC content change had almost 100% recall (Supplementary Fig. 3, Supplementary File 1). Therefore, we expect that a majority of the branches experiencing large jumps in the actual data are identified correctly. These large jumps are also biologically more interesting and potentially insightful. We also tested if uncertainty or potential inaccuracies in the topology of the phylogeny could have majorly impacted the inference of branch locations. To do this, we investigated the bootstrap support values of nodes following which GC jumps were detected. A majority (∼67%) of nodes had high bootstrap support of >90% (Supplementary Fig. 4, Supplementary File 1). Only ∼6% nodes showed bootstrap support values <50%. Thus, phylogenetic uncertainty does not majorly impact the inference of jump locations.

For inference of jumps in the actual GC content data, we identified branches with *pp* values greater than these thresholds (defined above) as those experiencing jumps. As examples, the inferred jumps mapped on a phylogeny of Rhizobiales and Acetobacterales are shown in Figs. 2 and 3 (for other clades, see Supplementary Figs. 5–12, Supplementary File 1). Reassuringly, many instances of expected jumps in bacterial lineages with peculiar host-associated lifestyles were captured by this approach. For example, jumps were inferred at the stem branches for the Enterobacterial endosymbiont clades *Buchnera*+*Blochmannia* and *Baumannia* + others (Husník *et al.* 2011), at the base of a clade involving the Flavobacterial endosymbiont *Blattabacterium* (Bandi *et al.* 1995), and at the base of Betaproteobacterial (endo)symbionts *Kinetoplastibacterium* (Alves *et al.* 2013), *Polynucleobacter* (Heckmann and Schmidt 1987), and *Profftella* (Nakabachi *et al.* 2013). In Rhizobiales (Fig. 2), a jump was inferred at the stem

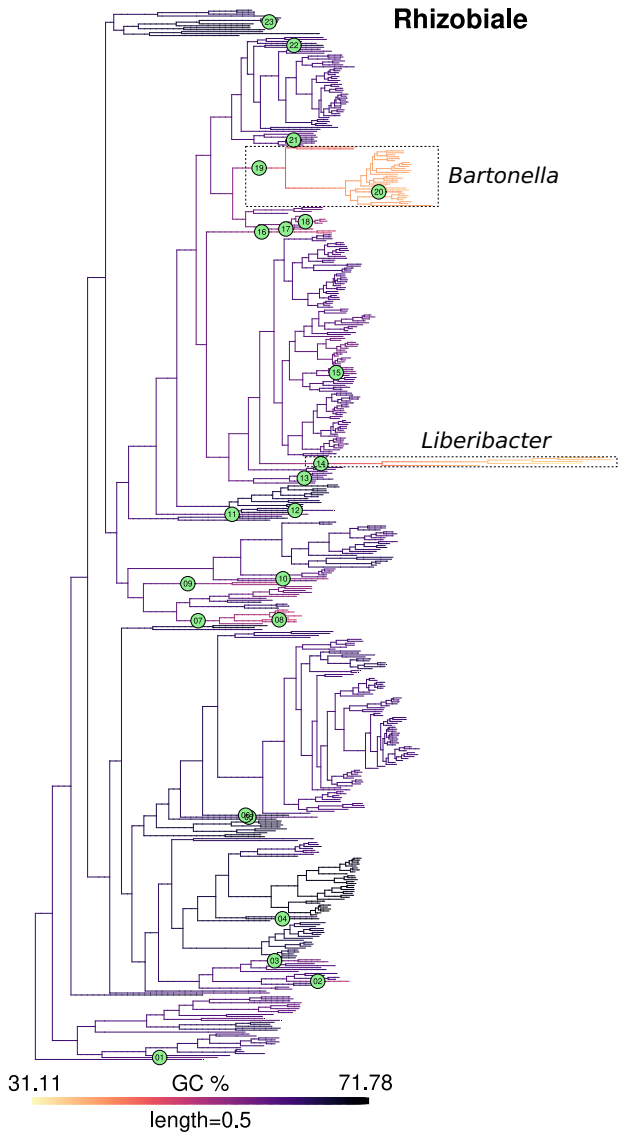

Fig. 2. GC content map and location of inferred jumps in Rhizobiales. GC content was mapped onto a phylogeny of Rhizobiales using the contMap function from R package phytools. This mapping itself is only indicative of trends since it assumes a Brownian model of evolution. Branches with inferred jumps i.e. where the posterior probability of observing jump(s) is greater than the chosen threshold are indexed in filled circles. Two interesting examples of jumps in Rhizobiales are highlighted in dashed boxes, which occur in the stem branches of *Liberibacter* (jump index 14), an obligate plant pathogen and *Bartonella* (jump index 19), an obligate animal pathogen, respectively. Within the genus *Bartonella*, the lineage leading to *B. australis* experienced an upward jump (index 20). Mapping for other clades is shown in Fig. 3 and Supplementary Figs. 5–12.

branch of the genus *Liberibacter*, which includes obligate host-dependent pathogens (Haapalainen 2014).

The total number of jumps inferred in this way ranged between 2 (for Cytophagales) and 73 (for Flavobacteriales) with a median of 15 (Table 1). Flavobacteriales appeared to be an exception since the next largest number of inferred jumps among other clades was 27 (for Rhodobacterales). However, the total number of detected jumps were not related to the ancestral GC content or the number of taxa in the clade (Supplementary Table 3, Supplementary File 2). On the other hand, the fraction of upward jumps in a clade was related to its ancestral GC content. Clades with low ancestral GC content (<50%) experienced proportionally

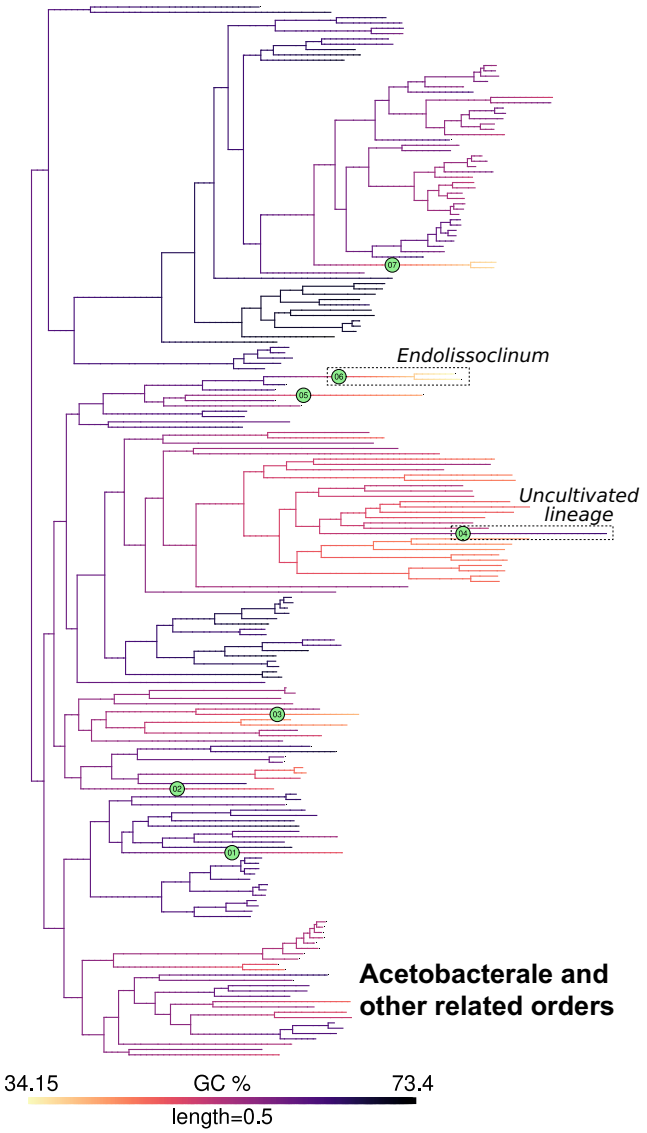

Fig. 3. GC content map and location of inferred jumps in Acetobacterales and related orders. GC content was mapped onto a phylogeny of Acetobacterales and related orders as noted in Fig 2. Branches with inferred jumps i.e. where the posterior probability of observing jump(s) is greater than the chosen threshold are indexed in filled circles. An example of a downward jump in an endosymbiont (*Endolissoclinum*, jump index 6) and an upward jump in an uncultivated bacterial lineage (jump index 4) are highlighted in dashed boxes. Mapping for other clades is shown in Fig. 2 and Supplementary Figs. 5–12.

more upward jumps, whereas clades with high ancestral GC content (>50%) experienced more downward jumps (Fig. 4a).

Magnitude and direction of GC jumps

We estimated jump magnitudes as the difference between median GC contents of taxa affected by jumps, and taxa in the corresponding sister clade. Jumps occurred both in the upward (increasing GC%) and downward (decreased GC%) direction. Although downward jumps ($n = 107$) were more frequent, we also found a comparable number of upward jumps ($n = 85$) (Fig. 4b). In terms of magnitude, downward jumps were bigger ($\Delta GC_{median} = -8.1\%$) than upward jumps ($\Delta GC_{median} = 6.6\%$). Even within the 10% largest jumps in each category, the average magnitude of the downward jumps was larger ($\Delta GC < -19.2\%$) than the upward jumps ($\Delta GC > 13.3\%$). Analyzed another way, among jumps with more than 15% change in GC, there were 18 downward jumps but only 5 upward jumps. Thus, while sudden increases in GC content are not rare, they tend to involve smaller changes in GC content compared to jumps that reduce GC%.

We further analyzed the direction and magnitude of jumps in relation to the estimated ancestral GC content (approximated as the GC content of sister clades that did not experience a jump in GC content). As expected, datasets with more extreme ancestral GC content (either lower or higher) were more likely to experience larger jumps in both directions (Fig. 4c). The pattern was especially striking for endosymbionts with high-GC ancestors, which showed very large downward GC jumps.

Visually, it appears that jumps are concentrated toward the tips i.e. toward more recent branches. However, this could also simply be a result of the number of branches being higher toward the tips in any phylogeny. Indeed, when we compared the distribution of the inferred jumps with jumps randomly placed on the phylogenies (with probability proportional to branch length), we observed that the distributions are not different (Supplementary Fig. 13, Supplementary File 1).

Ecological features associated with inferred GC jumps

Where possible, we extracted information from primary literature about the isolation source of taxa affected by the inferred jumps ("affected") and closely related taxa unaffected by the jumps ("unaffected") (Supplementary Table 4, Supplementary File 2). Because we were interested in understanding the relationship between GC content jumps and *changes* in habitat or life history, we excluded clades where the isolation sources of both sets of taxa could not be reliably inferred. Of the 91 jumps where such data was available, 48 experienced decreased GC content and 43 experienced increased GC content (Fig. 5).

In 7 of these jumps, affected taxa were obligatorily dependent on a host, and only in 4 of these they had switched to obligate host-dependence from a host independent lifestyle (Supplementary Table 4, Supplementary File 2). Thus, only a minor fraction of jumps analyzed here are likely to be caused due to a strict dependence on a host and accompanying changes in evolutionary parameters. In the remaining 84 jumps, the affected taxa could be cultured independently on laboratory media. But in a further 19 of these, affected taxa were associated with hosts (i.e. were isolated from hosts or host-associated material) but unaffected taxa were not; implying a putative switch from free-living to host-association. Of these, 15 experienced a decrease in GC content, whereas 4 experienced an increase in GC content (Fig. 5). Twelve other cases putatively involved an opposite switch (from host-association to no host-association), associated with a GC jump. In this set, 8 affected taxa experienced increased GC content and 4 experienced decreased GC content (Fig. 5). Thus, while about half the analyzed jumps involved a decrease in GC% (48/91 i.e. 53%), decreased GC content is more prevalent in affected taxa that appear to switch from no host-association to host-association (15/19 i.e. 79%) and less prevalent (4/12 i.e. 33%) in affected taxa that appear to switch in the opposite direction ($P < 0.05$ for a Fisher's exact test). Overall, while GC content jumps may sometimes arise due to changes in evolutionary parameters corresponding to such changes in lifestyle, a significant fraction of GC jumps (61/91 i.e. ~67%) did not involve association with (or separation from) hosts.

Independently, in more than half of the analyzed cases (45 of 84, excluding the 7 obligate host-dependent cases), affected taxa were isolated from marine habitats; suggesting that marine environments may impose distinct selection pressures that are

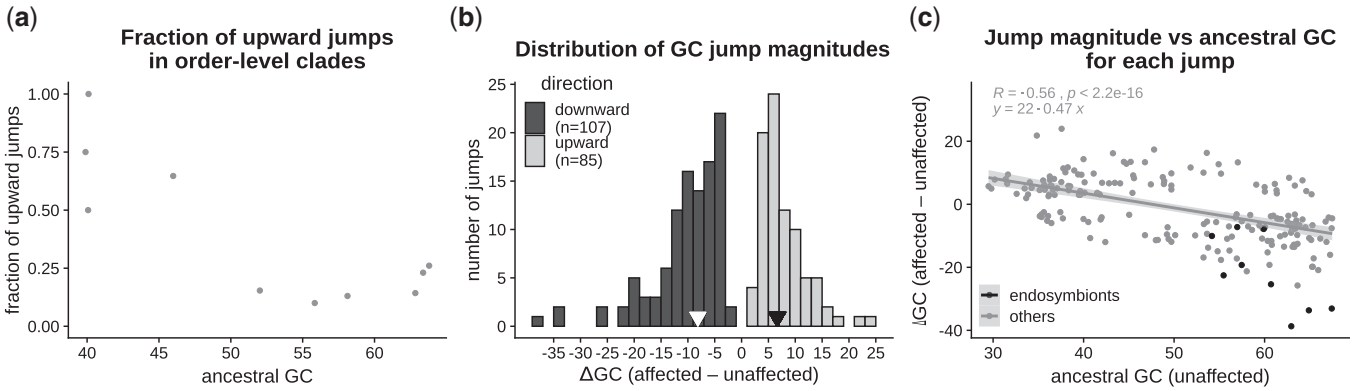

Fig. 4. Direction and magnitude of GC jumps. We estimated the magnitude of each GC content jump as the difference in median GC content of descendant taxa of a branch affected by a jump ("affected") and descendant taxa of a sister branch not affected by a jump ("unaffected"). Jumps involving increased GC content in affected taxa were designated as upward jumps, and those involving decreased GC content of affected taxa were designated as downward jumps. a) The relation between the fraction of total jumps that were upwards and the ancestral GC content of each order-level clade. Ancestral GC content was estimated as a parameter of the Lévy jumps model using the procedure implemented in *levolution*. b) Distribution of jump magnitudes. Black arrows denote the median magnitudes of upward and downward jumps. c) Relation between jump magnitude and the estimated ancestral GC content, with the best-fit regression line (excluding nonendosymbiont clades). Ancestral GC content was estimated as the median GC content of unaffected taxa of the sister clade.

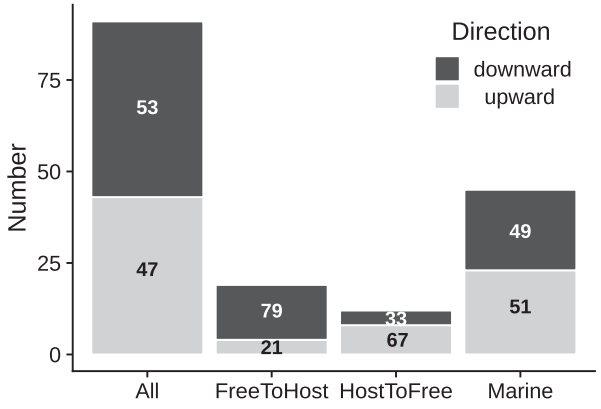

Fig. 5. Proportions of upward and downward jumps across habitat and lifestyle categories. The number of upward and downward jumps are shown across 4 categories of datasets: (1) all datasets that could be analyzed for habitat or lifestyle related changes ($n = 91$) (2) a subset of datasets where the affected taxa (where a GC jump occurred) were host-associated, but related unaffected taxa were free-living ($n = 19$) (3) a subset of datasets where the affected taxa were not associated with hosts, but related unaffected taxa were host-associated ($n = 12$), (4) a subset of datasets where the taxa affected by the jump were isolated from marine habitats ($n = 45$). Numbers in the bars denote the percentage of upward and downward jumps within each category.

especially likely to drive rapid GC shifts. However, in contrast to previous reports of GC reductions from AT rich marine bacteria, in half of the cases (23 of 45) affected taxa isolated from marine habitats showed increased GC content (Fig. 5). Although it was not always possible to discern the exact niche of the involved taxa, at least 12 were isolated from coastal sediments, unlike the AT-rich marine bacteria from nitrogen-poor surface ocean waters. Other taxa affected by large GC jumps were isolated from fresh water, soil, decaying wood, bioreactors, and fermented products (Supplementary Table 4, Supplementary File 2).

Since oxygen dependence has been previously suggested to be associated with higher GC content in bacteria, we also determined changes in oxygen dependence among the clades affected by jumps. Within datasets where such information was available, affected taxa showed no difference in oxygen dependence compared to related unaffected taxa in the majority of cases (44 of 56 comparisons); whereas affected taxa had increased oxygen dependence (predominantly, a change from facultatively aerobic to aerobic) in 7 cases and decreased oxygen dependence (a change from aerobic to facultatively aerobic or anaerobic) in 5 cases (Supplementary Table 5, Supplementary File 2). Interestingly, in 6 of 7 cases where affected taxa had increased oxygen dependence, they had also experienced increased GC content. Overall, in ∼10% of analyzed cases, affected taxa were more dependent on oxygen and experienced increased GC content.

Discussion

Ubiquitous evolutionary jumps in GC content across bacteria

In the study presented here, we analyzed the evolution of GC content across large bacterial datasets (10 orders across 2 bacterial phyla) using phylogenetic models of trait evolution. We found that the diversification of bacterial GC content is more consistent with a mixture of Brownian evolution and ubiquitous evolutionary jumps, rather than pure Brownian evolution. As indicated by the previously reported examples of large evolutionary decreases in the GC content of endosymbionts and AT-rich marine bacteria, evolutionary jumps in bacterial GC content were not entirely unexpected. However, we find that the estimated variance in GC content contributed by such evolutionary jumps was more than the estimated variance contributed by the Brownian component in almost all bacterial clades that we analyzed. Thus, evolution by jumps appears to make a major, but thus far unrecognized, contribution to the diversification of bacterial GC content. Our results are also supported by another recent study that found pulsed evolution to be common in bacterial genome traits, including GC content (Gao and Wu 2021). Our study further describes the frequency, magnitude, and phylogenetic context of the observed jumps in GC content. Specifically, we emphasize 2 novel observations about the characteristics and ecological context of GC content evolution. One, we find a large number of evolutionary jumps that increase GC content, in contrast to previous studies that exclusively report evolutionary reductions in GC content of bacteria. Two, we find evolutionary jumps in GC content

that occur in ecological contexts beyond endosymbiosis and habitation of surface oceans.

Interestingly, we found a comparable number of jumps that increase or decrease GC content. However, upward jumps were not associated with a clearly identifiable set of lifestyles or habitats (such as endosymbiosis). Moreover, we did not find any lineages with large increases comparable to the large reductions in GC content of some of the endosymbionts, and upward jumps were smaller in magnitude than downward jumps on average. These patterns may explain the absence of prior studies recognizing such jumps, though a fifth of the identified upward jumps were moderately large (>10% increase in GC content). Although we have not attempted a detailed analysis of the ecological context or evolutionary causes of upward jumps, the identification of lineages experiencing such jumps presents an opportunity to study them in the future. We found that upward jumps were more common in datasets with lower ancestral GC content, whereas they were less common in datasets with higher ancestral GC content. This is consistent with an evolutionary constraint on the range of observed GC content across bacteria (~25–75%), and indicates that the constraint may also apply to evolutionary jumps.

We must highlight that better fit by a model compared to other competing models does not say anything about the absolute ability of the model to explain the data. A better fitting model among 3 poor models will still be a poor one. Therefore, one has to rely on independent tests of whether model assumptions are satisfied and whether the model offers a good explanation of the data. Unfortunately, such exact "goodness of fit" tests are not available for most macroevolutionary models (Pennell *et al.* 2015). Thus, it is not clear if the Lévy jumps model used here offers an adequate explanation of GC content macroevolution. The modest recall of jump locations simulated according to the best-fit model parameters raises some doubts about the adequacy of the Lévy jumps model. However, the inference of exact locations of jumps is performed separately from the calculation of overall likelihood of data given the Lévy jumps model and estimation of summary parameters such as average jump rate and magnitude. Even when jump locations cannot be efficiently inferred, the average rate and magnitude of jumps can still be estimated accurately (Duchen *et al.* 2017). Moreover, we found that larger jumps in our simulations were recalled with greater frequency, reaching perfect recall for jumps with more than 20% change in GC content. Hence, we suggest that the largest jumps in the evolution of GC content are likely accurately reflected in our analysis.

Our study here considers GC content as a single trait and uses general trait evolution models to reconstruct its evolutionary history. This approach is justified by the correlations between GC content of different genome features such as the 1st, 2nd, and 3rd codon positions, genes with different expression levels, genic and intergenic regions etc. and common evolutionary forces such as mutation biases acting on this trait. However, ancestral reconstruction of complete gene sequences based on branch heterogeneous models may offer an additional source of information for reconstructing the evolutionary history of GC content. This approach benefits from the large number of available sites in genome data, but is challenging due to the computational complexity of nonhomogeneous sequence evolution models. In the future, this approach could be gainfully applied on smaller datasets, perhaps those selected on the basis of the present study.

In addition, we must acknowledge that the inference of jump locations is subject to the specifics of and uncertainties in the underlying phylogenies. For example, the phylogenies used here were derived after de-replication of available genomes, where a few representative taxa among a closely related set were retained. This is true for the derivation of the original datasets (Parks *et al.* 2018) as well as our study (see *Methods*). Such pruning may cause spurious jumps to appear if the retained taxa happen to have different GC content from the closely related taxa not represented in the phylogeny. However, this is unlikely to be true because the representative taxa either had high average nucleotide identity (ANI > 90%) or belonged to the same species as the ones that were removed. Another major source of spurious jumps may be the uncertainties in branch lengths. Specifically, underestimation of branch lengths may lead to trait changes being identified as more exceptional than they are in truth. However, such uncertainties in branch length should not affect our analysis severely since the underlying phylogenies are based on a large number of genes. As discussed earlier, these uncertainties are also less likely to affect the inference of larger jumps in traits. Although Bayesian methods that account for uncertainty in tree topology and branch lengths would be ideal (Huelsenbeck *et al.* 2000), such methods are not available for the jumps model used in our analysis. Another feature of the phylogenies used in our analyses is that branch lengths represent substitution rates rather than time. Consequently, evolutionary time may be underestimated if reduced substitution rates cause spurious jumps to be detected. Conversely, evolutionary time may be overestimated due to accelerated substitutions that may cause some jumps to be missed. More accurate analysis could be obtained by using time trees, but the absence of fossils makes it difficult to reliably date large bacterial phylogenies.

Useful datasets for studying evolutionary factors affecting GC content

Which evolutionary factors lead to GC content diversification is still an unresolved question. Mutational biases correlate with GC content across a diverse set of bacteria (Long *et al.* 2018); thus, changes in mutational biases must contribute to changes in bacterial GC content. However, what causes changes in mutational biases across bacteria is itself not well understood. Deletion of specific DNA replication and repair enzymes alters mutation bias in some bacteria (Dillon *et al.* 2017; Foster *et al.* 2018; Weissman *et al.* 2019) and the natural loss of some repair enzymes is the most likely reason for changes in the mutation bias of endosymbionts (Moran *et al.* 2008; Wernegreen 2015). However, whether such loss or gain contributes to changes in mutation bias and GC content in other lineages has not been investigated so far.

The role of changes in selection or GC biased gene conversion (gBGC) in diversification of GC content is also not clear (Lassalle *et al.* 2015; Bobay and Ochman 2017). Reduced efficiency of overall selection in endosymbiotic bacteria must contribute to reduced selection for GC content, but whether similar changes contribute to other instances of GC change is unclear. Moreover, there are few compelling explanations about what aspects of the biology of organisms could influence these microevolutionary forces leading to GC content jumps. Many environmental factors such as growth temperature, oxygen requirement, and nitrogen availability have been proposed to affect selection on GC content; but none offer convincing evidence after accounting for phylogenetic relatedness in the datasets (Agashe and Shankar 2014). Our analysis of GC jumps also failed to offer strong support for a major role of these environmental factors.

Based on the insights provided from prior studies of endosymbionts and AT-rich marine bacteria, we surmise that jumps in GC

content are more likely to be driven by large changes in one or few different evolutionary factors. In contrast, gradual diversification of GC content (or underlying factors such as mutation bias) across longer time scales may be driven by smaller changes in a number of factors, making it difficult to clearly identify causal relationships. A recent study also proposes that sudden jumps in mutational biases that alter the direction of bias should be generally selectively favored, because such shifts in mutation spectra can allow populations to access under-sampled mutational space (Sane *et al.* 2020). If true, this hypothesis may explain GC jumps involving both increase and decreases in GC content, without invoking specific selection pressures favoring a change in either direction. The bacterial lineages experiencing GC jumps identified here can serve as interesting datasets to test this hypothesis, as well as the role of specific evolutionary factors such as habitat, metabolic requirements, and DNA repair enzymes that may drive GC content changes.

While we think that the jumps in genome GC content are a result of changes in ecological and evolutionary forces acting on GC content per se, it is possible that the observed changes in GC content of a focal clade could have resulted from horizontal gene transfer of a significant number of genes from a host with different GC content. However, we found that the changes in genome GC content during jumps are also reflected in similar changes in median GC content of genes coding for ribosomal proteins (Supplementary Fig. 14, Supplementary File 1) that are unlikely to be horizontally transferred.

Habitats and lifestyles of clades experiencing GC jumps

In this study, we attempted a preliminary analysis of the ecological context in which GC content jumps occur. Ideally, one would like to statistically test the association between GC jumps and habitat changes. This requires a complete characterization of all instances of habitat change in the entire dataset, which in turn requires habitat data for all the hundreds of taxa included in this study. Since it was not possible for us to collect this data, we decided to only characterize the clades that experienced GC jumps. But even an analysis of the full set of inferred jumps was precluded by limited data availability. Some lineages were represented only by metagenome-assembled genomes, where lifestyle related information could not be obtained. Other lineages were not represented in systematic collections of microbial phenotypes, and hence we could not use them to analyze the impact of ecological factors. Overall, we could analyze less than half the datasets ($n = 91$ out of 201) for lifestyle or habitat related information, and even fewer ($n = 56$) for oxygen dependence. We hope that in future, new ecological data on some of the interesting lineages with large GC jumps will allow more robust analyses.

Although extreme GC changes in endosymbiotic bacteria are well-studied examples, only a small fraction (~7%) of the GC content jumps in our analysis were attributed to endosymbionts. However, this number is an underestimate for the following reason. In some cases, endosymbiont lineages of independent origins get erroneously lumped together as single clades due to long branch attraction. For example, *Buchnera* and *Blochmannia*, 2 endosymbiont genera with reduced GC content have independent origins (Husník *et al.* 2011), but appear as a single clade in the phylogenies used here. Consequently, our jump inference method detects a single jump (Enterobacterales, jump 15) at the stem of this clade instead of 2 separate jumps. However, such undercounting should have a small effect on the number of jumps involving endosymbionts, because not all jumps with endosymbionts involve multiple endosymbiont lineages.

Beyond endosymbionts, in a further ~25% cases, we found that GC jumps occurred in taxa that had either evolved toward or away from a host association. Changes to GC content in such cases may be explained by changes in evolutionary parameters accompanying changes in lifestyle (e.g. effective population size). In this regard, lineages with upward GC jumps and a putative switch from host-associated to host-independent lifestyle are especially interesting because they could represent a reversion from a low GC, host-associated lifestyle to high GC, host-independent lifestyle. These would represent a changes in the opposite direction to what is observed in endosymbionts. However, we notice that the majority of such lineages are found in orders that are already GC-poor (Flavobacteriales and Cytophagales; GC <40%). Further, the GC changes are relatively small (deltaGC < 6%) and the sister lineages are also not obligately host dependent. Thus, these switches appear to occur in clades that have evolved strong host dependence and associated genomic changes. Regardless, a majority of GC jumps (67%) do not involve a change in lifestyle with respect to host-association. Similarly, in terms of oxygen dependence, affected taxa in a majority of analyzed jumps (80%) did not show a change compared to related unaffected taxa; but in about 10% jumps, affected taxa were more dependent on oxygen (aerobic instead of facultatively aerobic) and had experienced increased GC content. This is consistent with some previous studies that found increased GC content to be associated with increased oxygen dependence (Naya *et al.* 2002; Aslam *et al.* 2019). However, previous studies do not identify specific instances of such associations. The datasets identified here can allow a more detailed investigation of this association and the potential mechanism underlying it.

Separately, about half the analyzed GC jumps occurred in marine lineages and it is possible that streamlining selection in this habitat could be contributing to some of these GC jumps. It was not clear if these lineages were indeed from nitrogen-poor surface waters [as previously reported for AT-rich marine bacteria (Giovannoni *et al.* 2005, 2014; Luo *et al.* 2017)]; but at least 20% were isolated from sediments. Previous studies find that the AT-richness of some bacterial lineages in surface oceans is part of a set of characteristics (genome reduction, smaller intergenic regions, increased coding density, fewer regulatory genes) attributed to streamlining selection due to nutrient limitation (Giovannoni *et al.* 2005, 2014; Grzymski and Dussaq 2012). The specific lineages identified in this study make it possible to assess whether streamlining selection may be relevant to the observed GC changes.

Finally, we find many instances of jumps in GC content of lineages neither related to hosts or marine habitats. As an outstanding example, *Zymomonas* (a genus of free-living, fermenting bacteria) have experienced a large reduction in GC content (~45%) compared to the sister genus *Sphingomonas* (GC content ~55% to 65%). We also identified jumps with >10% reduction in GC content of other putatively free-living bacterial lineages such as *Robiginitomaculum* + *Hellea* (marine), *Aquaspirillum serpens* (aquatic), *Janthinobacterium* sp. B9-8 (soil), *Hirschia* (marine); and jumps with >10% increases in GC content of the lineages *Siphonobacter aqueclarae* (aquatic), *Ferrimonas* (sediment), and *Flavobacterium* CP2B (marine) (Supplementary Table 4, Supplementary File 2). We hope that a detailed analysis of the relevant evolutionary factors in such datasets identified here would lead to further insights into the mechanisms of GC content evolution in bacteria.

Conclusion

We analyzed the diversity of bacterial GC content through a phylogenetic lens and found evolutionary jumps as a predominant mode of diversification of bacterial GC content. We identify these jumps as particularly interesting to study the ecological and evolutionary factors driving GC content evolution. We further surmise that evolutionary jumps—particularly those involving larger changes in GC content—could be driven by changes in ecological or evolutionary factors. However, we did not find strong support for any of the putative ecological factors previously implicated in GC content evolution. Since it will be difficult to experimentally study a large number of bacterial lineages, we suggest that immediate follow-up studies could focus on signatures of selection, drift, and ecological factors that could be gleaned from the available genome sequence data.

Author contributions

SM contributed to the design of methods and analysis, performed the analysis, interpreted results, and wrote the manuscript. DA conceived the study and contributed to the design of methods and analysis, interpretation of results, and writing.

Data availability

The data and code underlying this article are available as Online Supplementary Files and a Github repository at https://github.com/saurabh-mk/manuscript-bacGC. Published sources of data used to derive data in this study are referenced at appropriate places in this article. The repository mentioned above contains the trees containing posterior probability of jumps on each branch that were generated as part of the study. Other data regarding habitats and lifestyles of bacterial taxa compiled as part of this work are available as Supplementary Tables in Supplementary File 2. The repository mentioned above includes code and instructions necessary to reproduce the analyses in this article.

Supplemental material is available at G3 online.

Acknowledgments

The authors thank Gaurav Diwan for contributing to the script for pruning phylogenies, and for useful discussions.

Funding

This work was supported by the Council of Scientific and Industrial Research, India (CSIR) (fellowship "19/372624/SPMF 2013 Award" to SM), and National Centre Biological Sciences (NCBS-TIFR) (RTI 4006 to DA), Department of Atomic Energy, Government of India, Bangalore.

Conflicts of interest

None declared.

Literature cited

Agashe D, Shankar N. The evolution of bacterial DNA base composition. J Exp Zool B Mol Dev Evol. 2014;322(7):517–528. doi:10.1002/jez.b.22565.

Alves JMP, Serrano MG, Maia da Silva F, Voegtly LJ, Matveyev AV, Teixeira MMG, Camargo EP, Buck GA. Genome evolution and phylogenomic analysis of candidatus kinetoplastibacterium, the betaproteobacterial endosymbionts of strigomonas and angomonas. Genome Biol Evol. 2013;5(2):338–350. doi:10.1093/gbe/evt012.

Aslam S, Lan X-R, Zhang B-W, Chen Z-L, Wang L, Niu D-K. Aerobic prokaryotes do not have higher GC contents than anaerobic prokaryotes, but obligate aerobic prokaryotes have. BMC Evol Biol. 2019;19(1):35.doi:10.1186/s12862-019-1365-8.

Baidouri FE, Venditti C, Humphries S. Independent evolution of shape and motility allows evolutionary flexibility in Firmicutes bacteria. Nat Ecol Evol. 2016;1(1):9.doi:10.1038/s41559-016-0009.

Baker J, Venditti C. Rapid change in mammalian eye shape is explained by activity pattern. Curr Biol. 2019;29(6):1082–1088.e3. doi:10.1016/j.cub.2019.02.017.

Bandi C, Sironi M, Damiani G, Magrassi L, Nalepa CA, Laudani U, Sacchi L. The establishment of intracellular symbiosis in an ancestor of cockroaches and termites. Proc Biol Sci. 1995;259(1356):293–299. doi:10.1098/rspb.1995.0043.

Barkman TJ, Bendiksby M, Lim S-H, Salleh KM, Nais J, Madulid D, Schumacher T. Accelerated rates of floral evolution at the upper size limit for flowers. Curr Biol. 2008;18(19):1508–1513. doi:10.1016/j.cub.2008.08.046.

Beaulieu JM, Jhwueng D-C, Boettiger C, O'Meara BC. Modeling stabilizing selection: expanding the Ornstein–Uhlenbeck model of adaptive evolution. Evolution. 2012;66(8):2369–2383. doi:10.1111/j.1558-5646.2012.01619.x.

Bobay L-M, Ochman H. Impact of recombination on the base composition of bacteria and Archaea. Mol Biol Evol. 2017;34(10):2627–2636. doi:10.1093/molbev/msx189.

Brocchieri L. The GC content of bacterial genomes. J Phylogenetics Evol Biol. 2014;2:e108. doi:10.4172/2329–9002.1000e108.

Butler MA, King AA. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. Am Nat. 2004;164(6):683–695. doi:10.1086/426002.

Lai CY, Baumann P. Sequence analysis of a DNA fragment from Buchnera aphidicola (an endosymbiont of aphids) containing genes homologous to dnaG, rpoD, cysE, and secB. Gene. 1992;119(1):113–118. doi:10.1016/0378-1119(92)90074-Y.

Dillon MM, Sung W, Sebra R, Lynch M, Cooper VS. Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in vibrio cholerae and vibrio fischeri. Mol Biol Evol. 2017;34(1):93–109. doi:10.1093/molbev/msw224.

Diwan GD, Agashe D. Wobbling forth and drifting back: the evolutionary history and impact of bacterial tRNA modifications. Mol Biol Evol. 2018;35(8):2046–2059. doi:10.1093/molbev/msy110.

Duchen P, Leuenberger C, Szilágyi SM, Harmon L, Eastman J, Schweizer M, Wegmann D. Inference of evolutionary jumps in large phylogenies using Lévy processes. Syst Biol. 2017;66(6):950–963. doi:10.1093/sysbio/syx028.

Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985;125(1):1–15.

Foster PL, Niccum BA, Popodi E, Townes JP, Lee H, MohammedIsmail W, Tang H. Determinants of base-pair substitution patterns revealed by whole-genome sequencing of DNA mismatch repair defective *Escherichia coli*. Genetics. 2018;209(4):1029–1042. doi:10.1534/genetics.118.301237.

Gao Y, Wu M. Microbial genomic trait evolution is dominated by frequent and rare pulsed evolution. 2021. bioRxiv 2021.04.19.440498. doi:10.1101/2021.04.19.440498.

Ghai R, Mizuno CM, Picazo A, Camacho A, Rodriguez-Valera F. Metagenomics uncovers a new group of low GC and ultra-small

marine Actinobacteria. Sci Rep. 2013;3:2471.doi:10.1038/srep02471.
Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, *et al.* Genome streamlining in a cosmopolitan oceanic bacterium. Science. 2005;309(5738):1242–1245. doi:10.1126/science.1114057.
Giovannoni SJ, Thrash JC, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8(8):1553–1565. doi:10.1038/ismej.2014.60.
Grzymski JJ, Dussaq AM. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. 2012;6(1):71–80. doi:10.1038/ismej.2011.72.
Haapalainen M. Biology and epidemics of *Candidatus* Liberibacter species, psyllid-transmitted plant-pathogenic bacteria. Ann Appl Biol. 2014;165(2):172–198. doi:10.1111/aab.12149.
Hagey TJ, Uyeda JC, Crandell KE, Cheney JA, Autumn K, Harmon LJ. Tempo and mode of performance evolution across multiple independent origins of adhesive toe pads in lizards. Evolution. 2017;71(10):2344–2358. doi:10.1111/evo.13318.
Hansen TF. Stabilizing selection and the comparative analysis of adaptation. Evolution. 1997;51(5):1341–1351. doi:10.1111/j.1558-5646.1997.tb01457.x.
Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. GEIGER: investigating evolutionary radiations. Bioinformatics. 2008;24(1):129–131.
Haywood-Farmer E, Otto SP. The evolution of genomic base composition in bacteria. Evoltion. 2003;57(8):1783–1792. doi:10.1111/j.0014-3820.2003.tb00586.x.
Heckmann K, Schmidt HJ. *Polynucleobacter necessarius* gen. nov., sp. nov., an obligately endosymbiotic bacterium living in the cytoplasm of *Euplotes aediculatus*. Int J Syst Bacteriol. 1987;37(4):456–457. doi:10.1099/00207713-37-4-456.
Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 2010;6(9):e1001115.doi:10.1371/journal.pgen.1001115.
Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010;6(9):e1001107.doi:10.1371/journal.pgen.1001107.
Huelsenbeck JP, Rannala B, Masly JP. Accommodating phylogenetic uncertainty in evolutionary studies. Science. 2000;288(5475):2349–2350. doi:10.1126/science.288.5475.2349.
Husník F, Chrudimský T, Hypša V. Multiple origins of endosymbiosis within the Enterobacteriaceae (γ-Proteobacteria): convergence of complex phylogenetic approaches. BMC Biol. 2011;9:87.doi:10.1186/1741–7007-9–87.
Knight RD, Freeland SJ, Landweber LF. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Gnome Biol. 2001;2(4):research0010.
Landis MJ, Schraiber JG. Pulsed evolution shaped modern vertebrate body sizes. Proc Natl Acad Sci USA. 2017;114(50):13224–13229. doi:10.1073/pnas.1710920114.
Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet. 2015;11(2):e1004941.doi:10.1371/journal.pgen.1004941.
Liedtke HC, Gower DJ, Wilkinson M, Gomez-Mestre I. Macroevolutionary shift in the size of amphibian genomes and the role of life history and climate. Nat Ecol Evol. 2018;2(11):1792–1799. doi:10.1038/s41559-018–0674-4.
Lightfield J, Fram NR, Ely B. Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. PLoS One. 2011;6(3):e17677.doi:10.1371/journal.pone.0017677.
Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, Patterson C, Gregory C, Strauss C, Stone C, *et al.* Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol. 2018;2(2):237–240. doi:10.1038/s41559-017-0425-y.
Luo H, Huang Y, Stepanauskas R, Tang J. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nat Microbiol. 2017;2:17091.doi:10.1038/nmicrobiol.2017.91.
Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, Engqvist MKM, Fierer N, Geoghegan JL, Gillings M, *et al.* A synthesis of bacterial and archaeal phenotypic trait data. Sci. Data 2020;7:170.
Marashi S-A, Ghalanbor Z. Correlations between genomic GC levels and optimal growth temperatures are not 'robust'. Biochem Biophys Res Commun. 2004;325(2):381–383. doi:10.1016/j.bbrc.2004.10.051.
McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbol. 2011;10(1):13–26. doi:10.1038/nrmicro2670.
McEwan CE, Gatherer D, McEwan NR. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas. 1998;128(2):173–178.
Moran NA. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci USA. 1996;93(7):2873–2878.
Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. Cell. 2002;108(5):583–586. doi:10.1016/S0092-8674(02)00665-7.
Moran NA, McCutcheon JP, Nakabachi A. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet. 2008;42:165–190. doi:10.1146/annurev.genet.41.110306.130119.
Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett. 2004;573(1–3):73–77. doi:10.1016/j.febslet.2004.07.056.
Muto A, Osawa S. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA. 1987;84(1):166–169. doi:10.1073/pnas.84.1.166.
Nakabachi A, Ueoka R, Oshima K, Teta R, Mangoni A, Gurgui M, Oldham NJ, van Echten-Deckert G, Okamura K, Yamamoto K, et al. Defensive bacteriome symbiont with a drastically reduced genome. Curr Biol. 2013;23(15):1478–1484. doi:10.1016/j.cub.2013.06.027.
Naya H, Romero H, Zavala A, Alvarez B, Musto H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol. 2002;55(3):260–264. doi:10.1007/s00239-002-2323-3.
O'Meara BC, Ané C, Sanderson MJ, Wainwright PC. Testing for different rates of continuous trait evolution using likelihood. Evolution. 2006;60(5):922–933. doi:10.1111/j.0014-3820.2006.tb01171.x.
Pagel MD, Harvey PH. Comparative methods for examining adaptation depend on evolutionary models. Folia Primatol (Basel). 1989;53(1–4):203–220. doi:10.1159/000156417.
Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36(10):996–1004.

Pennell MW, FitzJohn RG, Cornwell WK, Harmon LJ. Model adequacy and the macroevolution of angiosperm functional traits. Am Nat. 2015;186(2):E33–E50. doi:10.1086/682022.

Raghavan R, Kelkar YD, Ochman H. A selective force favoring increased G+C content in bacterial genes. Proc Natl Acad Sci USA. 2012;109(36):14504–14507. doi:10.1073/pnas.1205683109.

Reichenberger ER, Rosen G, Hershberg U, Hershberg R. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. Genome Biol Evol. 2015;7(5):1380–1389. doi:10.1093/gbe/evv063.

Remeš V, Freckleton RP, Tökölyi J, Liker A, Székely T. The evolution of parental cooperation in birds. Proc Natl Acad Sci USA. 2015;112(44):13603–13608. doi:10.1073/pnas.1512599112.

Sane M, Diwan GD, Bhat BA, Wahl LM, Agashe D. Shifts in mutation spectra enhance access to beneficial mutations. bioRxiv. 2020;2020. doi:10.1101/2020.09.05.284158.

Stern DB, Crandall KA. The evolution of gene expression underlying vision loss in cave animals. Mol Biol Evol. 2018;35(8):2005–2014. doi:10.1093/molbev/msy106.

Sueoka N. Variation and heterogeneity of base composition of deoxyribonucleic acids: A compilation of old and new data. Journal of Molecular Biology. 1961;3(1):31–IN15.

Thomas SH, Wagner RD, Arakaki AK, Skolnick J, Kirby JR, Shimkets LJ, Sanford RA, Löffler FE. The mosaic genome of Anaeromyxobacter dehalogenans strain 2CP-C suggests an aerobic common ancestor to the delta-proteobacteria. PLoS One. 2008;3(5):e2103.

Vieira-Silva S, Rocha EPC. An assessment of the impacts of molecular oxygen on the evolution of proteomes. Mol Biol Evol. 2008;25(9):1931–1942. doi:10.1093/molbev/msn142.

Wang H-C, Susko E, Roger AJ. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. Biochem Biophys Res Commun. 2006;342(3):681–684. doi:10.1016/j.bbrc.2006.02.037.

Weissman JL, Fagan WF, Johnson PLF. Linking high GC content to the repair of double strand breaks in prokaryotic genomes. PLoS Genet. 2019;15(11):e1008493.doi:10.1371/journal.pgen.1008493.

Wernegreen JJ. Endosymbiont evolution: predictions from theory and surprises from genomes. Ann N Y Acad Sci. 2015;1360:16–35. doi:10.1111/nyas.12740.

Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic Acids Res. 2010;38(12):e132. doi:10.1093/nar/gkq275.

Communicating editor: A. Wong