A Positive Correlation between Bacterial GC Content and Growth Temperature

En-Ze Hu[†], Xin-Ran Lan[†], Zhi-Ling Liu, Jie Gao, and Deng-Ke Niu*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering and Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, Beijing, 100875, China

[†]These authors contributed equally to this paper.

*Corresponding author: Deng-Ke Niu, dkniu@bnu.edu.cn, dengkeniu@hotmail.com

Abstract

Because GC pairs are more stable than AT pairs, GC-rich genomes were proposed to be more adapted to high temperatures than AT-rich genomes. Previous studies consistently showed positive correlations between growth temperature and the GC contents of structural RNA genes. However, for the whole genome sequences and the silent sites of the codons in protein-coding genes, the relationship between GC content and growth temperature is in a long-lasting debate. With a dataset much larger than previous studies (681 bacteria and 155 archaea), our phylogenetic comparative analyses showed positive correlations between optimal growth temperature and GC content both in bacterial and archaeal structural RNA genes and in bacterial whole genome sequences, chromosomal sequences, plasmid sequences, core genes, and accessory genes. However, in the 155 archaea, we did not observe a significant positive correlation of optimal growth temperature with whole-genome GC content or GC content at four-fold degenerate sites. We randomly drew 155 samples from the 681 bacteria for 1000 rounds. In most cases (> 95%), the positive correlations between optimal growth temperature and genomic GC contents became statistically nonsignificant ($P > 0.05$). This result suggested that the small sample sizes might account for the lack of positive correlations between growth temperature and genomic GC content in the 155 archaea and the bacterial samples of previous studies. Besides the thermal adaptation hypothesis, we should be open to other intricate explanations for the observed correlations.

Key words: GC content, optimal growth temperature, prokaryotes, evolution, thermophile.

Introduction

As guanine (G) strictly pairs with cytosine (C) and adenine (A) pairs with thymine (T) in DNA double helix, the amount of G is equal to C, and that of A is equal to T in the genomes of any cellular organisms. GC content, i.e., the percentage of G + C, is widely used as a measure of genomic nucleotide composition. It is a highly variable trait ranging from 8% to 75% [1-3]. This genomic trait has been widely studied, and its evolution has been proposed to be associated with numerous mutational and selective forces driven by genetic, metabolic, and ecological factors [4-17]. The high temperature might be the most long-debating one [18-20]. Because G:C pairs have an additional hydrogen bond than A:T pairs, the GC-rich genomes are thermally more stable in high-temperature environments. Bernardi and Bernardi [21] proposed that high GC content is a thermal adaptation of warm-blooded animals.

As prokaryotes have a much wider thermal distribution than plants and animals, bacterial and archaeal genomes are the best materials to test the thermal adaptation hypothesis. An analysis of 764 prokaryotic species, including mesophilic genera and thermophilic genera, did not find a correlation between whole-genome GC content (GC_w) and the optimal growth temperature (Topt) [20]. However, this study found a significant positive correlation between Topt and the GC content of structural RNA (tRNAs and rRNAs). The rationale of these observations is that the secondary structures of tRNAs and rRNAs are more sensitive to high temperatures than the double-strand helix of DNA. In most prokaryotes, protein-coding genes take most of the genome size. Protein structures and functions constrain the GC content evolution at the nonsynonymous sites of the codons. This functional constraint might conceal the hypothetical thermal adaptation. Compared with GC_w, the GC content at the third site of the codons (GC_3) is more desirable to test the thermal adaptation hypothesis. Early solitary cases indicated that GC_3 might be related to growth temperature. For example, the tyrosy1-tRNA synthetase gene isolated from the thermophile *Bacillus stearothermophilus* has a higher GC_3 than the homologous gene in *Escherichia coli*, 68.0% *vs.* 59.4% [22]. The *leuB* gene isolated from the extreme thermophile *Thermus thermophilus* HB8 has an extremely high GC_3, 89.4% [23]. For a general conclusion, Hurst and Merchant [24] examined the relationship between GC_3 and Topt of 29

archaeal species and 72 bacterial species. Unfortunately, they did not find significant correlations between Topt and GC_3 or between Topt and GC_w. At the same time, they also found a significant positive correlation between the GC content of structural RNAs and the Topt in both archaea and bacteria. Their analysis accounted for the effect of shared ancestry, so they provided more robust evidence against the thermal adaptation hypothesis. Soon afterward, Xia et al. [25] showed that the growth at increasing temperature (from 37°C to 45°C) for 14,400 generations did not increase but decreased the genomic GC content of the bacterium *Pasteurella multocida*. Furthermore, Lambros et al. [26] reported a negative correlation between optimal growth temperature and the GC content of protein-coding genes in 550 prokaryotes, despite that the effect of the shared ancestry had not been controlled.

Subsequently, Musto et al. [27] published a debate-provoking study. As many environmental factors likely influence genomic GC content evolution, closely related species are expected to differ in fewer environmental factors than distantly related species. The correlation of GC content with growth temperature is less likely disturbed by other factors when the analysis is limited within closely related species. Therefore, Musto et al. [27] examined the relationship between genomic GC content and Topt with each prokaryotic family. Among the 20 families they studied, the number of families with positive correlations is significantly higher than expected by chance, no matter the effect of the common ancestors was accounted for or not. Meanwhile, they observed a significant positive correlation when considering all independent contrasts from different families together. However, Marashi and Ghalanbor [28] noticed that most of the significant correlations within each family depend heavily on the presence of a few outlier species. Exclusion of only one species would lead to loss of significant correlations in several families. Basak et al. [29] pointed out that the correlation is sensitive to the presence or absence of a few outliers in some families because the sample sizes in these families were too small. Using non-parametric correlation analysis that is not sensitive to the presence of outliers, Musto et al. [30] repeated their analysis and confirmed their previous results. The debate did not end after that. Wang et al. [31] updated the Topt values for some species and found that the positive correlation between Topt and genomic GC content in two families disappeared. Besides, they suggested that the positive correlation between Topt and genomic GC content in the family

Enterobacteriaceae should be explained by the correlation between genome size and optimal temperature. Still, this study did not shake the confidence of Musto et al. [32] on the correlation between Topt and genomic GC content in prokaryotes. Although Musto and coauthors have rebutted all the criticisms, their studies have not convinced later authors of review articles [4, 18, 33]. For example, Agashe and Shankar [4] claimed that "*it seems unlikely that genomic GC content is driven by thermal adaptation*" after reviewing the results of Hurst and Merchant [24] and Xia et al. [25], but without mentioning the debates on Musto et al. [27].

As prokaryotic genomes often have many accessory genes frequently lost and gained, the genome-wide measures of GC content could roughly reflect the shaping effects of environmental factors in evolution. By contrast, the structural RNA genes ubiquitously exist in prokaryotic genomes, and their GC contents are more comparable in large-scale phylogenetic analyses. Similarly, the core genome or strictly defined orthologous genes could also accurately reflect the historical shaping effect of growth temperature on GC content evolution. Ream et al. [34] analyzed the GC contents of two genes (*ldh-a* and *α-actin*) across 51 vertebrate species with adaptation temperatures ranging from −1.86°C to approximately 45°C. They did not find any significant positive correlations between living temperature and GC content, whether the GC content is measured by the entire sequences, the third codon position, or the fourfold degenerate sites. However, Zheng and Wu [35] found a positive correlation between growth temperature and the GC content in the coding regions of four genes across 815 prokaryotic species, including mesophiles, thermophiles, and hyperthermophiles. These four genes shared by all the 815 prokaryotic genomes could be considered strictly defined core genomes.

Using a manually collected dataset of growth temperature and without accounting for the effect of the common ancestors, Sato et al. [36] recently confirmed the results of Galtier and Lobry [20]. It should be noted that the correlation between Topt and the GC content of structural RNA was consistently observed in much more studies than those mentioned above [36-40]. By contrast, as reviewed above, the correlation between Topt and genomic GC content, if it exists, depends heavily on the sample size, the families of prokaryotes, the sequences, and the methods used to detect it.

Benefit from the manually curated dataset of growth temperature from the database TEMPURA

[36], we carried out a comprehensive analysis of the relationship between growth temperature and GC content. The present study covers three indexes of growth temperature (maximal growth temperature [Tmax], Topt, and minimal growth temperature [Tmin]) and a series of GC content indexes, including GC_w, GC content of the protein-coding sequences (GC_p), GC content at fourfold degenerate sites (GC_4), GC content of the genes coding structural RNAs (tRNA, GC_{tRNA}; 5S rRNA, GC_{5S}; 16S rRNA, GC_{16S}; 23S rRNA, GC_{23S}) and GC content of non-coding DNA (GC_{non}, including intergenic sequences and untranslated regions of mRNA that are generally unannotated in prokaryotic genomes). The whole genome, primary chromosome genome sequences, plasmid genomes, core genes, and accessory genes have been examined separately. Our results consistently support a positive correlation between genomic GC content and growth temperature in bacteria.

Results

Strong phylogenetic signals in both GC contents and growth temperatures

A significant force shaping prokaryotic evolution is horizontal gene transfer, making the genealogical relationships among bacteria and archaea exhibit a somewhat network-like structure. If bifurcation is not the phylogeny's dominant pattern, most phylogenetic comparative methods will not be necessary for prokaryotic evolutionary studies. We are unsure how much this impression has influenced the researchers in prokaryotic genomic studies, but many papers did not use any phylogenetic comparative methods. Despite the frequent horizontal gene transfers, careful examination of the prokaryotic phylogeny could see a statistical tree [41-43]. In principle, the necessity of phylogenetic comparative methods depends on the significance of the phylogenetic signal, a measure of the correlation between the evolution of the analyzed trait and the presumed phylogenetic tree. We first measured the phylogenetic signals of the analyzed traits for the 681 bacteria and 155 archaea obtained from the database TEMPURA [36]. As shown in table 1, all the λ values are close to one, which indicates that simple statistical analysis that does not account for common ancestry's effect would lead to inaccurate results [44, 45].

Bacterial but not archaeal genomic GC contents correlated with growth temperatures

We used the phylogenetic generalized least squares (PGLS) regression to examine the relationships between GC contents and growth temperatures. The significant positive and negative slopes of the regressions correspond to significant positive and negative correlations, respectively. The value of the slope represents the phylogenetically corrected rate of change in GC content as growth temperature changes. Four phylogenetic models, the Brownian motion model (BM), the Ornstein-Uhlenbeck model with an ancestral state to be estimated at the root (OUfixedRoot), the Pagel's lambda model (lambda), and the early burst model (EB), have been applied in the analysis. Their results are qualitatively identical and quantitatively similar. As the four models lead to the same conclusion, the trivial differences among their results are unrelated to understanding the relationship between GC content and growth temperature. We present the BM model results in the main text and deposit other models' results as supplementary tables.

Consistent with numerous previous studies, we found positive correlations between the GC contents of structural RNA genes and growth temperature in bacteria and archaea (Table 2). We noticed a rank in the slope values, from Tmax, Topt, to Tmin.

Interestingly, we also found positive correlations of Tmax and Topt with various indexes of genomic GC contents, GC_w, GC_p, GC_4, and GC_{non}, in bacteria (Table 2). Nevertheless, bacterial Tmin is not correlated with three of the four GC content indexes (Table 2). In archaea, none of the three temperature indexes (Tmax, Topt, or Tmin) have any significant correlations with any of the four genomic GC content indexes (Table 2).

If growth temperature could shape GC contents by the stabilities of RNA secondary structures and DNA double helix, a structural RNA or a DNA double helix that is stable at the Tmax or Topt is, of course, stable at the Tmin. This logic makes it reasonable to see that the Tmin has weaker or no significant correlations with GC contents.

The difference in the correlations between bacteria and archaea might be attributed to either unknown intrinsic differences between these two domains or the substantial difference in the sample size, 681 *vs.* 155.

Sample sizes matter

If the lack of significant correlations between genomic DNA and Tmax and Topt in archaea results from the small sample size, the correlations in bacteria will be lost when the sample size of bacteria is reduced to 155. For this reason, we randomly selected 155 bacteria from the 681 bacterial samples for 1000 rounds. The results of the resampling analysis confirmed the idea, the sample sizes matter (Table 3). In > 950 rounds, the genomic GC content indexes (GC_w, GC_p, GC_4, and GC_{non}) are not correlated with Tmax or Topt ($P > 0.05$). This result could also explain the difference between the present study with Hurst and Merchant (2001), which did not find significant correlations between GC_w/GC_3 and Topt by phylogenetic analysis of about 100 prokaryotes. Meanwhile, a few positive correlation cases happen, indicating that significant positive correlations could also be found by chance when the analyzed sample is small.

Besides, the correlations between growth temperature and the GC contents of structural RNA genes might also be lost occasionally when the sample size is severely reduced (Table 3). In the 1000 rounds of resampling, lacking significant correlations happens in 308 (for Tmax) and 473 (for Topt) rounds for 5S rRNA genes, and 12 (for Tmax) and 21 (for Topt) rounds for tRNA genes. However, in the 16S and 23S rRNA genes, positive correlations were consistently observed in all the 1000 rounds of resampling. We suspected that the tens of times more nucleotides in 16S and 23S rRNA than 5S rRNA and tRNAs make the results of 16S and 23S rRNAs less sensitive to small sample sizes.

In statistics, the rule of thumb boundary between small and large samples is $n = 30$. However, the results in Table 3 indicate that $n = 155$ is a too-small sample in the phylogenetic comparative analyses of the relationship between growth temperature and genomic GC content. Because of the common ancestor, two closely related lineages with highly similar growth temperatures and GC contents should be regarded as nearly one effective sample rather than two independent samples. The effective sample size in phylogenetic comparative studies should be much lower than the census number of the analyzed lineages.

Positive correlations observed in genes of both chromosomes and plasmids

Previous studies showed that plasmids have significantly lower GC contents than chromosomes [8, 46,

47]. Therefore, we examined the correlations between growth temperatures and GC contents separately in chromosomes and plasmids. The separations of plasmids and chromosomes are arbitrary. We strictly followed the classifications of chromosomes and plasmids of the NCBI genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/). Among the 681 bacteria and 155 archaea analyzed above, 172 bacteria and 42 archaea have plasmid genomes. The bacterial chromosomes also have GC contents (GC_w, GC_p, GC_4, and GC_{non}) positively correlated with Tmax and Topt (Table 4). Interestingly, the same pattern was also found in the bacterial plasmids (Table 4) in spite that the correlations of Tmax with GC_4 and GC_{non} are just significant at marginal levels ($0.05 < P < 0.1$). All these correlations are not significant in archaea.

In the two previous studies comparing the GC content between plasmids and chromosomes [46, 47], the common ancestor effect was not accounted for. By the way, we performed a phylogenetic paired t-test [48] and confirmed the pattern of lower GC content in plasmids (Table S8).

Positive correlations were observed in both core genes and accessory genes

To correspond to the previous gene-centered studies [35], we examined the correlations in bacterial core genes, i.e., genes present in all the bacteria. With the increase in the number of bacteria, the number of core genes decreases rapidly. With a trade-off between the number of core genes and the number of bacteria, we selected 28 core genes present in 420 genomes, mostly ribosomal protein genes. Significant positive correlations have been found between GC contents (GC_p and GC_4) and growth temperatures, Tmax, and Topt (Table 5).

At the opposite side of the core genes, the accessory genes are present in one or a few bacteria. When we define the accessory genes as the genes present in less than 5% of the analyzed bacterial genomes, on average, each bacterium has 152 accessory genes. Positive correlations were observed between GC contents (GC_p and GC_4) and growth temperatures (Tmax and Topt), although the values of significance are slightly larger than those in core genes (Table 5). Similar patterns were observed when we increased the threshold in defining accessory genes to 10% ($P < 0.05$ for all cases).

By the way, we compared the GC content between bacterial core genes and accessory genes using a phylogenetic paired t-test [48]. Unlike the previous analysis of 36 prokaryotes that did not account

for the effect of common ancestors [49], we did not observe significant differences in GC content between the core genes and the accessory genes (Table S15). We also compared the chromosomal accessory genes and plasmid accessory genes. The accessory genes on chromosomes have significantly higher GC contents than those on plasmids (Table S16).

Qualitative data on growth temperature lead to the same conclusion

In the ProTraits database (http://protraits.irb.hr/) and the IMG database (https://img.jgi.doe.gov/) [50, 51], many prokaryotes lack quantitative measures of growth temperature but are qualitatively classified into four categories: psychrophiles/psychrotrophiles, mesophiles, thermophiles, and hyperthermophiles. We constructed a qualitative dataset of prokaryote growth temperature, including data downloaded from these two datasets and the prokaryotes in the TEMPURA database classified into the four categories referring [36]. By assigning 1, 2, 3, and 4 to the psychrophiles/psychrotrophiles, mesophiles, thermophiles, and hyperthermophiles, respectively, we transformed the qualitative characters into numerical values. Because only some of the genomes have been completely sequenced, we used their GC_w values downloaded directly from the NCBI genome database. PGLS regression revealed a positive correlation between GC_w content and growth temperature in bacteria (slope = 0.457, $P = 0.001$), but not in archaea (slope = –0.582, $P = 0.170$). Although this dataset (4696 bacteria and 279 archaea) is much larger than analyzed above (681 bacteria and 155 archaea), it lost much information during the qualitative classification. All the differences in growth temperature within each category disappear.

We also examined whether the contrast in the temperature category is correlated with the contrast in the GC content between terminal tips of the phylogenetic tree. Referring [6], 273 pairs of bacteria and 41 pairs of archaea were retrieved from the Genome Taxonomy Database [52]. On average, the bacteria with higher ranks in Topt have 1.43% more GC than their paired bacteria with lower ranks. Pairwise comparison showed significantly higher GC contents in the bacteria with higher ranks in growth temperature (Wilcoxon signed rank test, $P = 0.019$, Fig. 1A). Still, no significant differences were observed between paired archaea with different growth temperature ranks (Wilcoxon signed rank test, $P = 0.446$, Fig. 1B).

Evolutionary jumps in GC contents are correlated with growth temperature changes

Recently, Mahajan and Agasheand [3] found that the Lévy jumps model [53] could better explain prokaryotic GC content evolution than the Brownian model. The GC content evolves at a constant rate and sometimes experiences discrete changes, i.e., jumps. Following Mahajan and Agasheand [3], we first confirmed that the Lévy jumps model could better explain the GC_w and Topt in our dataset than the simple Brownian model.

As the Lévy jumps model has not been integrated into the PGLS packages, we cannot replace the BM model with this one. As an alternate, we retrieved the detected jumps in GC_w and examined whether significant changes in Topt accompany them. The phylogenetic locations of jumps were inferred using the *levolution* software [53]. In this procedure, only the posterior probabilities (pp) of the presence of > 0 jumps were estimated, but the exact number or magnitude of jumps on each branch could not be predicted. In practice, the "precision" of jump inference is negatively correlated with the "recall" of actual jumps. By adjusting the threshold of posterior probabilities of the presence of > 0 jumps for a precision > 85%, we obtained the GC_w jumps with 88.5% precision and an acceptable recall of 37.0%. As shown in Fig. 2A, the magnitudes of the GC_w jumps are positively correlated with the changes in Topt (Spearman′s rank correlation, 2-tailed, n = 108, rho = 0.209, P = 0.030). When the precision of jump inference was increased to 96.9%, the recall decreased to 21.5%, and no significant correlation was observed in the smaller sample (Spearman′s rank correlation, 2-tailed, n = 56, rho = 0.195, P = 0.150).

Meanwhile, we detected the evolutionary jumps in Topt using the same model. By adjusting the threshold of posterior probabilities, we inferred the Topt jumps with 95.3% precision and 21.5% recall. A positive correlation was observed between the magnitudes of the jumps in Topt and the changes of GC contents at the positions of Topt jumps (Spearman's rank correlation, 2-tailed, n = 86, rho = 0.280, P = 0.009, Fig. 2B).

Discussion

The GC pairs are thermally more stable than AT pairs in both DNA double helix and structural RNAs. However, this difference is not necessarily a strong enough force to shape the evolution of GC content. As RNA structures are more sensitive to temperature elevation than DNA double helix, the growth temperature is stronger in shaping the GC content evolution of the structural RNA genes than in shaping the genomic GC content evolution. Positive correlations between growth temperature and the GC content of structural RNA genes have been repeatedly observed in various prokaryotic studies [20, 24, 36-40]. However, there was a long debate on the correlation between growth temperature and genomic GC content. Benefit from a new manual-curated dataset of prokaryotic growth temperature [36], we performed a phylogenetic comparative analysis with a much larger sample than previous studies [24, 27]. In 681 bacteria, the genomic GC contents, GC_w, GC_p, GC_4, and GC_{non}, are all positively correlated with growth temperatures, Tmax and Topt. However, in 155 archaea, there are no significant correlations. Then, we resampled 155 bacteria from the 682 bacteria for 1000 rounds. In most cases, the significant positive correlations between genomic GC contents and growth temperatures disappeared. The resampling analysis indicates that the small sample sizes of the previous analyses [24] might lead to the lack of significant correlations. It is easy to increase the sample size several times if accurate phylogenetic relationships [52] are not considered in the analysis. As shown in Table 1, we found that both growth temperatures and GC contents exhibit strong phylogenetic signals. Overlooking the effect of common ancestors would severely affect the accuracy of the results [44].

Our resampling analysis indicates that the lack of significant correlations in archaea might result from the small number of effective samples. We hope to repeat the present study in the future with a larger sample of archaeal genomes. However, it should also be kept in mind that the possibility of no correlation between GC content and growth temperature in archaea has not been convincingly excluded. Some intrinsic differences between bacteria and archaea might produce a sharp difference in the relationship between GC content and growth temperature.

A recent study suggests that sequential amino acid substitutions are involved in the thermal adaptation in the archaeal order *Methanococcales* and revealed arginine as the most favored amino acid [54]. As six GC-rich codons encode the arginine, the thermal adaptation at the proteomic level

would affect the evolution of genomic GC content. Because the 4-fold degenerate sites are free from the evolutionary forces coming from the natural selection acting on protein sequences, our observations of similar correlations of GC_w, GC_p, and GC_4 with growth temperature indicate that the nucleotide composition evolved independently in bacterial adaptation to high temperatures.

As the frequent gain and loss of plasmids, the plasmid DNAs could be regarded as accessory genomes. Because of the high turnover rates of plasmids and accessory genes in prokaryotic evolution, we could regard them as new immigrants, as opposed to the natives for the chromosomes and core genes. Although the core genes and even the ribosomal RNA genes may occasionally be transferred across different prokaryotic lineages [55, 56], the fitness cost of inter-species replacement of homologous sequences [57] restricts the frequency of the core genes. Genes performing essential informational tasks in the cell are less frequently transferred across lineages [58, 59]. Our phylogenetic correlation analysis showed that positive correlations between GC contents and growth temperatures exist in chromosomes and core genes and exist in plasmids and accessory genes. Also, there is no sharp difference in the correlations between the new immigrants and the natives.

In large-scale analyses of horizontal gene transfer in prokaryotes, GC-content similarity between donor and recipient was found to be the factor, or one of the factors, governing the compatibility of the new immigrants in new hosts [60, 61]. The effect of promoter GC content on the expression of the new immigrants was suggested to be the underlying mechanism governing the compatibility [62]. Here, we suggest that the temperature-associated structural stabilities, including the stability of DNA double helix, the stability of the transient DNA-RNA duplex during transcription, and maybe the stability of the possible secondary structures of mature mRNA [1], might be another nonexclusive factor governing the compatibility. The new immigrants compatible with the host should have GC contents adapted to the host's growth temperature.

A previous serial transfer experiment seems to be contradictory to our results. Increased genomic GC content was not observed in the bacterium *P. multocida* after 14,400 generations of increasing temperature from 37°C to 45°C [25]. Although we observed a positive correlation between genomic GC content and growth temperature, we do not think a small increment in GC content, resulting from either a GC-biased mutator or integration of a GC-rich exogenous sequence, would bring a great

advantage to the host organism. Most likely, it is just a slight advantage. According to the population genetic theory, the slightly beneficial mutants will be efficiently selected only when they are in a large population. The experimental evolution generally involves severe, periodic reductions in population size, and the bottleneck effect dramatically reduces the fixation probability of beneficial mutations [63]. As we see, large-scale statistical analysis has the advantage of revealing slightly beneficial traits.

Musto et al. [27] emphasize that only when closely related species are compared can the growth temperature be the only influencing factor in GC content evolution. Our pairwise comparison of neighboring branches with different ranks of growth temperature (fig. 1) gave the same conclusion as our PGLS analyses. We agree that many factors would influence GC content evolution, and the positive relationship between growth temperature and GC content is a statistically significant result. In the 273 pairs of bacteria, there are 153 pairs where high growth temperature ranks have higher GC contents and 119 pairs with the opposite pattern.

Mahajan and Agasheand [3] and the present study found that the evolutionary rates of GC content and growth temperature have occasional jumps assumed in the Lévy jumps model [53]. As shown in Fig. 2, the jump-ups and jump-downs of GC content are significantly correlated with changes in growth temperature and vice versa. It should be emphasized that not all increases in growth temperature are accompanied by increases in GC content. There are just statistically significant correlations ($P < 0.05$).

We should remark that what we observed are weak correlations between genomic GC content and growth temperature. The slopes of the PGLS regressions are generally between 10^{-3} and 10^{-4}. The bacteria rank higher in growth temperature have just 1.43% more GC (Fig. 1A). Considering the significant difference in the thermoresistence of nucleic acids between *in vivo* and *in vitro* [33], we believe that other cellular components mainly contribute to the thermostability of nucleic acids in thermophiles and hyperthermophiles. The increase of GC content unlikely plays a major role. Moreover, we observed correlations between GC content and growth temperature, which imply rather than prove the causal effects between the two variables. Even if there is a causal effect, we should be open to both the thermal adaptation hypothesis [21] and other intricate explanations. Besides adaptive explanations, nonadaptive processes may be explored in the future. Heat mutagenesis could reduce GC

content by cytosine deamination [64]. In the fission yeast, *Schizosaccharomyces pombe*, there is evidence that high temperature might induce more gene conversions [65], which generally increase the fixation of GC-containing alleles, regardless of their effects on fitness [17]. If the balance between these two competing processes is slightly biased to gene conversion, a weak positive correlation exists between growth temperature and GC content. This paper aims to end the long-standing debate on the relationship between GC content and growth temperature. Only after establishing the positive correlation could the attention of genome biologists be paid to the biological significance of the correlation.

Materials and Methods

We downloaded the prokaryote growth temperatures from the database TEMPURA [36]. This database contains 8,639 manual curated prokaryotes (549 archaea and 8090 bacteria). Using the links to the NCBI Taxonomy database [66] and the taxonomy IDs provided by TEMPURA for each prokaryotic strain, we obtained 1110 prokaryotes whose genome assembly levels were labeled as"complete" from the NCBI database [67]. Among them, we found the phylogenetic information for 682 bacteria and 156 archaea from GTDB (Genome Taxonomy Database) [52]. The sequences of these genomes were downloaded from ftp://ftp.ncbi.nlm.nih.gov/genomes/. To avoid annotation bias resulting from different methods, all the genomes were re-annotated using the DFAST, version 1.2.11, with its default parameters [68]. In total, we obtained the annotations for 681 bacterial genomes and 155 archaeal genomes. The GC contents of these prokaryotes were calculated from their genome sequences.

We also constructed a large dataset according to their growth temperature qualitatively. First, we divided the 836 prokaryotes mentioned above into four categories according to their growth temperature referring to [36]: psychrophiles/psychrotrophiles (Topt < 20°C), mesophiles (20 ≤ Topt < 45°C), thermophiles (45 ≤ Topt < 80°C), and hyperthermophiles (80°C ≤ Topt). Then, we downloaded the lists of prokaryotes labeled with psychrophiles/psychrotrophiles, mesophiles, thermophiles, or hyperthermophiles from the ProTraits database (http://protraits.irb.hr/) and the IMG database (https://img.jgi.doe.gov/) [50, 51]. After discarding the overlapping items, the conflicting items, and

the items lacking phylogenetic information in the GTDB database [52], we obtained a new dataset including 4696 bacteria and 279 archaea (Table S20). The whole-genome GC contents of these prokaryotes were downloaded directly from the NCBI genome database (https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt).

As the contrasts between different pairs of terminal tips of the phylogenetic tree are independent of each other, pairwise comparisons between pairs of terminal tips could control the effect of common ancestors. Referring to reference [6], we wrote a script to select pairs of closely related bacteria with different ranks of growth temperature (psychrophiles/psychrotrophiles, mesophiles, thermophiles, and hyperthermophiles). In cases where two or more neighboring tips with the same rank were used to pair with bacteria with another rank, we used the average value of their GC contents to represent the GC content of their internal node. The script is deposited as supplementary Data S2.

The phylogenetic signals (λ) of both GC contents and growth temperatures were estimated using the *phylosig* function of the R (Version 4.0.3) package *phytools* (Version 0.7-70) [69]. The PGLS regression was performed using the R (Version 4.0.3) package *phylolm* (version 2.6.2) with the default parameters [70].

To avoid false-positive results that might happen in multiple correlation analyses of the same dataset, we controlled the false discovery rate by the Benjamini-Hochberg (BH) procedure using the p.adjust function in R (Version 4.0.3).

Following Mahajan and Agasheand [3], we used the *geiger* package [71] and the *levolution* software [53] to simulate our datasets, estimate the branch-specific posterior probabilities of jumps and infer the phylogenetic location of jumps.

Supplementary Material

Supplementary data are submitted along with the main text.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant number 31671321).

Author Contributions

DKN conceived the study and wrote the manuscript. EZH, XRL, ZLL, and JG performed the data analysis. All authors read, improved, and approved the final manuscript.

References

1. Basak S, Mukhopadhyay P, Gupta SK, Ghosh TC. Genomic adaptation of prokaryotic organisms at high temperature. Bioinformation. 2010;4(8):352-6. doi: 10.6026/97320630004352. PubMed PMID: MEDLINE:20975899.
2. Nguyen DT, Wu B, Xiao S, Hao W. Evolution of a record-setting AT-rich genome: indel mutation, recombination, and substitution bias. Genome Biol Evol. 2020;12(12):2344-54. doi: 10.1093/gbe/evaa202.
3. Mahajan S, Agashe D. Evolutionary jumps in bacterial GC content. bioRxiv. 2021:2021.02.16.431469. doi: 10.1101/2021.02.16.431469.
4. Agashe D, Shankar N. The evolution of bacterial DNA base composition. J Exp Zool Part B. 2014;322(7):517-28. doi: 10.1002/jez.b.22565. PubMed PMID: WOS:000343820600005.
5. Wu H, Zhang Z, Hu S, Yu J. On the molecular mechanism of GC content variation among eubacterial genomes. Biol Direct. 2012;7:2. doi: 10.1186/1745-6150-7-2. PubMed PMID: PMC3274465.
6. 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.
7. Glemin S, Clement Y, David J, Ressayre A. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends Genet. 2014;30(7):263-70. doi: 10.1016/j.tig.2014.05.002. PubMed PMID: WOS:000338915700001.
8. Dietel A-K, Merker H, Kaltenpoth M, Kost C. Selective advantages favour high genomic AT-contents in intracellular elements. PLoS Genet. 2019;15(4):e1007778. doi: 10.1371/journal.pgen.1007778. PubMed PMID: WOS:000466866000002.
9. Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. EMBO Rep. 2005;6(12):1208-13. doi: 10.1038/sj.embor.7400538. PubMed PMID: ISI:000233616000020.
10. 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.
11. 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):1380-9. doi: 10.1093/gbe/evv063. PubMed PMID: WOS:000356228800017.
12. Šmarda P, Bureš P, Horová L, Leitch IJ, Mucina L, Pacini E, et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc Natl Acad Sci USA. 2014;111(39):E4096-E102. doi: 10.1073/pnas.1321152111.
13. 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):14504-7. doi: 10.1073/pnas.1205683109.
14. Mann S, Chen YPP. Bacterial genomic G plus C composition-eliciting environmental adaptation. Genomics. 2010;95(1):7-15. doi: 10.1016/j.ygeno.2009.09.002. PubMed PMID: WOS:000273838300002.
15. 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
16. Kogay R, Wolf YI, Koonin EV, Zhaxybayeva O. Selection for reducing energy cost of protein production drives the GC content and amino acid composition bias in gene transfer agents. Mbio. 2020;11(4):13. doi: 10.1128/mBio.01206-20. PubMed PMID: WOS:000572063700003.
17. Lassalle F, Perian 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:e1004941. doi: 10.1371/journal.pgen.1004941. PubMed PMID: WOS:000352081800017.
18. Meyer MM. Revisiting the relationships between genomic G plus C content, RNA secondary structures, and optimal growth temperature. J Mol Evol. 2021;89:165-71. doi: 10.1007/s00239-020-09974-w. PubMed PMID: WOS:000591100500001.
19. Forsdyke DR. Neutralism versus selectionism: Chargaff's second parity rule, revisited. Genetica. 2021;149(2):81-8. doi: 10.1007/s10709-021-00119-5.
20. Galtier N, Lobry JR. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol. 1997;44(6):632-6. PubMed PMID: WOS:A1997XB49700007.
21. Bernardi G, Bernardi G. Compositional constraints and genome evolution. J Mol Evol. 1986;24(1):1-11. doi: 10.1007/BF02099946.
22. Winter G, Koch GLE, Hartley BS, Barker DG. The amino acid sequence of the tyrosy1-tRNA synthetase from *Bacillus stearothermophilus*. Eur J Biochem. 1983;132(2):383-7. doi: 10.1111/j.1432-1033.1983.tb07374.x.
23. Kagawa Y, Nojima H, Nukiwa N, Ishizuka M, Nakajima T, Yasuhara T, et al. High guanine plus cytosine content in the third letter of codons of an extreme thermophile. DNA sequence of the isopropylmalate dehydrogenase of *Thermus thermophilus*. The Journal of biological chemistry. 1984;259(5):2956-60. PubMed PMID: 6321488.
24. Hurst LD, Merchant AR. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc R Soc B. 2001;268(1466):493-7. doi: 10.1098/rspb.2000.1397. PubMed PMID: WOS:000167432800007.
25. Xia X, Wei T, Xie Z, Danchin A. Genomic changes in nucleotide and dinucleotide frequencies in *Pasteurella multocida* cultured under high temperature. Genetics. 2002;161(4):1385-94.
26. Lambros RJ, Mortimer JR, Forsdyke DR. Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles. 2003;7(6):443-50. doi: DOI 10.1007/s00792-003-0353-4. PubMed PMID: ISI:000187157100002.
27. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett. 2004;573(1-3):73-7. doi: 10.1016/j.febslet.2004.07.056. PubMed PMID: WOS:000223697200014.

28. Marashi S-A, Ghalanbor Z. Correlations between genomic GC levels and optimal growth temperatures are not 'robust'. Biochem Biophys Res Commun. 2004;325(2):381-3. doi: 10.1016/j.bbrc.2004.10.051.
29. Basak S, Mandal S, Ghosh TC. Correlations between genomic GC levels and optimal growth temperatures: some comments. Biochem Biophys Res Commun. 2005;327(4):969-70. doi: 10.1016/j.bbrc.2004.12.100.
30. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: A reply to Marashi and Ghalanbor. Biochem Biophys Res Commun. 2005;330(2):357-60. doi: 10.1016/j.bbrc.2005.02.133. PubMed PMID: WOS:000228315800001.
31. 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):681-4. doi: 10.1016/j.bbrc.2006.02.037.
32. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valin F, Bernardi G. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun. 2006;347(1):1-3. doi: 10.1016/j.bbrc.2006.06.054.
33. Grosjean H, Oshima T. How nucleic acids cope with high temperature. Physiology and Biochemistry of Extremophiles: American Society of Microbiology; 2007. p. 39-56.
34. Ream RA, Johns GC, Somero GN. Base compositions of genes encoding alpha-actin and lactate dehydrogenase-A from differently adapted vertebrates show no temperature-adaptive variation in G+C content. Mol Biol Evol. 2003;20(1):105-10. doi: DOI 10.1093/molbev/msg008. PubMed PMID: WOS:000180243700014.
35. Zheng H, Wu H. Gene-centric association analysis for the correlation between the guanine-cytosine content levels and temperature range conditions of prokaryotic species. BMC Bioinformatics. 2010;11:S7. doi: 10.1186/1471-2105-11-s11-s7. PubMed PMID: WOS:000290219200007.
36. Sato Y, Okano K, Kimura H, Honda K. TEMPURA: database of growth TEMPeratures of Usual and RAre Prokaryotes. Microbes Environ. 2020;35(3):ME20074. doi: 10.1264/jsme2.ME20074.
37. Khachane AN, Timmis KN, dos Santos VAPM. Uracil content of 16S rRNA of thermophilic and psychrophilic prokaryotes correlates inversely with their optimal growth temperatures. Nucleic Acids Res. 2005;33(13):4016-22. doi: 10.1093/nar/gki714. PubMed PMID: WOS:000231516900007.
38. Kimura H, Sugihara M, Kato K, Hanada S. Selective phylogenetic analysis targeted at 16S rRNA genes of thermophiles and hyperthermophiles in deep-subsurface geothermal environments. Appl Environ Microbiol. 2006;72(1):21-7. doi: 10.1128/aem.72.1.21-27.2006.
39. Kimura H, Ishibashi J-I, Masuda H, Kato K, Hanada S. Selective phylogenetic analysis targeting 16S rRNA genes of hyperthermophilic archaea in the deep-subsurface hot biosphere. Appl Environ Microbiol. 2007;73(7):2110-7. doi: 10.1128/aem.02800-06.
40. Kimura H, Mori K, Yamanaka T, Ishibashi J-I. Growth temperatures of archaeal communities can be estimated from the guanine-plus-cytosine contents of 16S rRNA gene fragments. Env Microbiol Rep. 2013;5(3):468-74. doi: 10.1111/1758-2229.12035.
41. DeSalle R, Riley M. Should networks supplant tree building? Microorganisms. 2020;8(8). doi: 10.3390/microorganisms8081179. PubMed PMID: WOS:000564624000001.

42. Koonin EV. The turbulent network dynamics of microbial evolution and the statistical tree of life. J Mol Evol. 2015;80(5-6):244-50. doi: 10.1007/s00239-015-9679-7. PubMed PMID: WOS:000356532400002.
43. Blais C, Archibald JM. The past, present and future of the tree of life. Curr Biol. 2021;31(7):R314-R21. doi: 10.1016/j.cub.2021.02.052.
44. Felsenstein J. Phylogenies and the comparative method. Am Nat. 1985;125(1):1-15. doi: doi:10.1086/284325.
45. Symonds MRE, Blomberg SP. A primer on phylogenetic generalised least squares. In: Garamszegi LZ, editor. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 105-30.
46. Rocha EPC, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet. 2002;18(6):291-4. PubMed PMID: ISI:000175854700010.
47. Nishida H. Comparative analyses of base compositions, DNA sizes, and dinucleotide frequency profiles in archaeal and bacterial chromosomes and plasmids. Int J Evol Biol. 2012;2012:342482. doi: 10.1155/2012/342482.
48. Lindenfors P, Revell LJ, Nunn CL. Sexual dimorphism in primate aerobic capacity: a phylogenetic test. J Evol Biol. 2010;23(6):1183-94. doi: 10.1111/j.1420-9101.2010.01983.x.
49. Bohlin J, Eldholm V, Pettersson JHO, Brynildsrud O, Snipen L. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC Genomics. 2017;18:151. PubMed PMID: WOS:000394384500004.
50. Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, et al. The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. Nucleic Acids Res. 2020;49(D1):D751-D63. doi: 10.1093/nar/gkaa939.
51. Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Res. 2016;44(21):10074-90. doi: 10.1093/nar/gkw964.
52. Parks DH, Chuvochina M, Chaumeil PA, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38(9):1079-86. doi: 10.1038/s41587-020-0501-8. PubMed PMID: WOS:000529298800004.
53. Duchen P, Leuenberger C, Szilagyi SM, Harmon L, Eastman J, Schweizer M, et al. Inference of evolutionary jumps in large phylogenies using Levy processes. Syst Biol. 2017;66(6):950-63. doi: 10.1093/sysbio/syx028. PubMed PMID: WOS:000412819000006.
54. Lecocq M, Groussin M, Gouy M, Brochier-Armanet C. The molecular determinants of thermoadaptation: *Methanococcales* as a case study. Mol Biol Evol. 2021;38(5):1761-76. doi: 10.1093/molbev/msaa312.
55. Tian R-M, Cai L, Zhang W-P, Cao H-L, Qian P-Y. Rare events of intragenus and intraspecies horizontal transfer of the 16S rRNA gene. Genome Biol Evol. 2015;7(8):2310-20. doi: 10.1093/gbe/evv143. PubMed PMID: WOS:000360819300017.
56. Sato M, Miyazaki K. Phylogenetic network analysis revealed the occurrence of horizontal gene transfer of 16S rRNA in the genus *Enterobacter*. Front Microbiol. 2017;8:10. doi: 10.3389/fmicb.2017.02225. PubMed PMID: WOS:000415285400001.
57. Bershtein S, Serohijos AWR, Bhattacharyya S, Manhart M, Choi J-M, Mu W, et al. Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in

bacteria. PLoS Genet. 2015;11(10):e1005612. doi: 10.1371/journal.pgen.1005612.
58. Kacar B, Garmendia E, Tuncbag N, Andersson DI, Hughes D. Functional constraints on replacing an essential gene with its ancient and modern homologs. mBio. 2017;8(4):e01276-17. doi: 10.1128/mBio.01276-17.
59. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci USA. 1999;96(7):3801-6. doi: 10.1073/pnas.96.7.3801.
60. Popa O, Hazkani-Covo E, Landan G, Martin W, Dagan T. Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes. Genome Res. 2011;21(4):599-609. doi: 10.1101/gr.115592.110.
61. Porse A, Schou TS, Munck C, Ellabaan MMH, Sommer MOA. Biochemical mechanisms determine the functional compatibility of heterologous genes. Nat Commun. 2018;9(1):522. doi: 10.1038/s41467-018-02944-3.
62. Gomes ALC, Johns NI, Yang A, Velez-Cortes F, Smillie CS, Smith MB, et al. Genome and sequence determinants governing the expression of horizontally acquired DNA in bacteria. ISME J. 2020;14(9):2347-57. doi: 10.1038/s41396-020-0696-1. PubMed PMID: WOS:000539137700004.
63. Wahl LM, Gerrish PJ, Saika-Voivod I. Evaluating the impact of population bottlenecks in experimental evolution. Genetics. 2002;162(2):961-71.
64. J W Drake a, Baltz RH. The biochemistry of mutagenesis. Annu Rev Biochem. 1976;45(1):11-37. doi: 10.1146/annurev.bi.45.070176.000303. PubMed PMID: 786147.
65. Brown SD, Audoynaud C, Lorenz A. Intragenic meiotic recombination in *Schizosaccharomyces pombe* is sensitive to environmental temperature changes. Chromosome Res. 2020;28(2):195-207. doi: 10.1007/s10577-020-09632-3.
66. Federhen S. The NCBI Taxonomy database. Nucleic Acids Res. 2012;40(D1):D136-D43. doi: 10.1093/nar/gkr1178.
67. Sayers EW, Beck J, Bolton EE, Bourexis D, Brister JR, Canese K, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2021;49(D1):D10-D7. doi: 10.1093/nar/gkaa892. PubMed PMID: WOS:000608437800002.
68. Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics. 2018;34(6):1037-9. doi: 10.1093/bioinformatics/btx713.
69. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3(2):217-23. doi: 10.1111/j.2041-210X.2011.00169.x.
70. Ho LST, Ane C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst Biol. 2014;63(3):397-408. doi: 10.1093/sysbio/syu005. PubMed PMID: WOS:000334752600009.
71. Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. GEIGER: investigating evolutionary radiations. Bioinformatics. 2008;24(1):129-31. doi: 10.1093/bioinformatics/btm538. PubMed PMID: WOS:000251865000018.

Table 1. The phylogenetic signals of the variables analyzed in this study.

	Bacteria			Archaea		
Traits	*n*	Pagel′s λ	*P*	*n*	Pagel′s λ	*P*
Tmax	681	0.957	3.5×10^{-178}	155	1.000	1.5×10^{-72}
Topt,	681	0.950	1.1×10^{-196}	155	0.988	5.7×10^{-70}
Tmin	681	0.933	6.6×10^{-152}	155	0.966	1.9×10^{-53}
GC_w	681	1.000	2.6×10^{-294}	155	1.000	5.8×10^{-60}
GC_p	681	1.000	4.7×10^{-292}	155	1.000	2.4×10^{-59}
GC_4	681	1.000	4.8×10^{-238}	155	1.000	5.1×10^{-53}
GC_{non}	681	1.000	8.6×10^{-305}	155	1.000	2.0×10^{-65}
GC_{tRNA}	681	0.998	6.0×10^{-275}	155	1.000	2.1×10^{-91}
GC_{5S}	646	1.000	7.0×10^{-178}	130	1.000	9.1×10^{-51}
GC_{16S}	681	0.999	8.7×10^{-250}	155	0.996	3.7×10^{-86}
GC_{23S}	681	1.000	2.1×10^{-245}	155	1.000	6.8×10^{-83}

Tmax, Topt, and Tmin represent maximal, optimal, and minimal growth temperature, respectively; GC_w, GC_p, GC_4, GC_{tRNA}, GC_{5S}, GC_{16S}, GC_{23S}, and GC_{non} represent the GC contents of the whole genome, the protein-coding sequences, the fourfold degenerate sites, the genes coding tRNAs, the genes coding 5S rRNA, the genes coding 16S rRNA, the genes coding 23S rRNA, and the non-coding DNA (including intergenic sequences and untranslated regions of mRNA), respectively. The phylogenetic signals of the chromosomal genes, the plasmid genes, the core genes, and the accessory genes are also very close to one and deposited in supplementary Tables S1-S4.

644 Table 2. PGLS regression of GC contents and growth temperatures.

	Bacteria			Archaea		
	Slope	P	P_{BH}	Slope	P	P_{BH}
GC_w-Tmax	7.1×10^{-4}	7.1×10^{-4}	9.4×10^{-4}	6.6×10^{-4}	0.115	0.153
GC_w-Topt	5.7×10^{-4}	0.009	0.011	3.3×10^{-4}	0.377	0.503
GC_w-Tmin	2.8×10^{-4}	0.156	0.226	5.2×10^{-4}	0.126	0.168
GC_p-Tmax	6.6×10^{-4}	0.002	0.002	5.6×10^{-4}	0.183	0.209
GC_p-Topt	5.3×10^{-4}	0.015	0.016	2.4×10^{-4}	0.522	0.597
GC_p-Tmin	2.5×10^{-4}	0.202	0.231	4.6×10^{-4}	0.180	0.205
GC_4-Tmax	0.001	0.003	0.003	9.9×10^{-4}	0.321	0.321
GC_4-Topt	0.001	0.016	0.016	2.2×10^{-4}	0.806	0.806
GC_4-Tmin	5.5×10^{-4}	0.239	0.239	6.9×10^{-4}	0.393	0.393
GC_{non}-Tmax	8.0×10^{-4}	1.7×10^{-4}	2.8×10^{-4}	9.1×10^{-4}	0.025	0.041
GC_{non}-Topt	6.4×10^{-4}	0.004	0.006	6.4×10^{-4}	0.080	0.129
GC_{non}-Tmin	2.7×10^{-4}	0.170	0.226	6.5×10^{-4}	0.048	0.077
GC_{tRNA}-Tmax	4.1×10^{-4}	2.2×10^{-16}	5.9×10^{-16}	7.1×10^{-4}	1.8×10^{-11}	7.2×10^{-11}
GC_{tRNA}-Topt	3.9×10^{-4}	2.6×10^{-14}	6.9×10^{-14}	5.0×10^{-4}	2.5×10^{-7}	6.7×10^{-7}
GC_{tRNA}-Tmin	1.5×10^{-4}	9.1×10^{-4}	0.002	4.2×10^{-4}	1.8×10^{-6}	4.7×10^{-6}
GC_{5S}-Tmax	5.5×10^{-4}	1.2×10^{-6}	2.4×10^{-6}	0.001	1.9×10^{-5}	3.9×10^{-5}
GC_{5S}-Topt	4.4×10^{-4}	1.4×10^{-4}	2.9×10^{-4}	8.9×10^{-4}	1.6×10^{-4}	3.2×10^{-4}
GC_{5S}-Tmin	3.5×10^{-4}	0.001	0.002	6.1×10^{-4}	0.005	0.010
GC_{16S}-Tmax	5.4×10^{-4}	2.2×10^{-16}	5.9×10^{-16}	8.2×10^{-4}	3.9×10^{-11}	1.0×10^{-10}

GC_{16S}-Topt	5.2×10^{-4}	2.2×10^{-16}	8.8×10^{-16}	7.2×10^{-4}	1.1×10^{-10}	4.5×10^{-10}
GC_{16S}-Tmin	4.6×10^{-4}	2.2×10^{-16}	8.8×10^{-16}	5.5×10^{-4}	8.5×10^{-8}	3.4×10^{-7}
GC_{23S}-Tmax	6.6×10^{-4}	2.2×10^{-16}	5.9×10^{-16}	0.001	2.2×10^{-16}	1.8×10^{-15}
GC_{23S}-Topt	6.5×10^{-4}	2.2×10^{-16}	8.8×10^{-16}	0.001	1.2×10^{-14}	9.5×10^{-14}
GC_{23S}-Tmin	4.9×10^{-4}	2.2×10^{-16}	8.8×10^{-16}	8.3×10^{-4}	8.0×10^{-11}	6.4×10^{-10}

GC contents were the dependent variables, and growth temperatures were the independent variables. The results in this table were obtained using the Brownian motion model. Similar results obtained from three other models are deposited in supplementary Table S5-S7. P_{BH}, Benjamini-Hochberg adjusted P value. Please see Table 1 for the meanings of the other abbreviations.

Table 3. The appearance of correlations in 1000 rounds of resampling analyses.

	Significantly Negative ($P < 0.05$)	Not Significant ($P > 0.05$)	Significantly Positive ($P < 0.05$)
GC_w-Tmax	0	974	26
GC_w-Topt	0	991	9
GC_p-Tmax	0	976	24
GC_p-Topt	0	993	7
GC_4-Tmax	0	962	38
GC_4-Topt	0	992	8
GC_{non}-Tmax	0	974	26
GC_{non}-Topt	0	992	8
GC_{tRNA}-Tmax	0	12	988
GC_{tRNA}-Topt	0	21	979
GC_{5S}-Tmax	0	308	692
GC_{5S}-Topt	0	473	527
GC_{16S}-Tmax	0	0	1000
GC_{16S}-Topt	0	0	1000
GC_{23S}-Tmax	0	0	1000
GC_{23S}-Topt	0	0	1000

In each round of resampling, 155 samples were randomly drawn from the 681 bacteria. PGLS regression analyses were performed for each round. GC contents were the dependent variables, and growth temperatures were the independent variables. The results in this table were obtained using the Brownian motion model. Please see Table 1 for the meanings of the other abbreviations. The datasets for each round of resampling are deposited in Supplementary Data S1.

Table 4. PGLS regression of GC contents and growth temperatures in chromosomes and plasmids.

	Plasmid			Chromosome		
	Slope	P	P_{BH}	Slope	P	P_{BH}
GC_w-Tmax	0.001	0.009	0.043	9.6×10^{-4}	0.029	0.043
GC_w-Topt	0.001	0.005	0.031	9.6×10^{-4}	0.023	0.031
GC_p-Tmax	0.001	0.016	0.043	9.1×10^{-4}	0.038	0.046
GC_p-Topt	0.001	0.010	0.031	9.2×10^{-4}	0.031	0.034
GC_4-Tmax	0.002	0.072	0.072	0.002	0.027	0.043
GC_4-Topt	0.002	0.044	0.044	0.002	0.017	0.031
GC_{non}-Tmax	8.3×10^{-4}	0.055	0.060	0.001	0.021	0.043
GC_{non}-Topt	9.3×10^{-4}	0.025	0.031	0.001	0.021	0.031

GC contents were the dependent variables, and growth temperatures were the independent variables. The results in this table were obtained using the Brownian motion model. Similar results obtained from three other models are deposited in supplementary Table S9-S11. P_{BH}, Benjamini-Hochberg adjusted P value. Please see Table 1 for the meanings of the other abbreviations.

Table 5. PGLS analysis of GC contents and growth temperatures in core genes and accessory genes

	Core Genes			Accessory Genes		
	Slope	P	P_{BH}	Slope	P	P_{BH}
GC_p-Tmax	7.6×10^{-4}	9.6×10^{-4}	0.002	9.0×10^{-4}	0.001	0.002
GC_p-Topt	6.4×10^{-4}	0.007	0.025	6.3×10^{-4}	0.026	0.030
GC_4-Tmax	0.002	6.3×10^{-4}	0.002	0.002	0.003	0.003
GC_4-Topt	0.002	0.004	0.025	0.002	0.019	0.030

GC contents were the dependent variables, and growth temperatures were the independent variables. The results in this table were obtained using the Brownian motion model. Similar results obtained from three other models are deposited in supplementary Table S12-S14. P_{BH}, Benjamini-Hochberg adjusted P value. Please see Table 1 for the meanings of the other abbreviations.

Figure legend

Figure 1. Pairwise comparison of the GC contents between closely related prokaryotes with different growth temperature ranges. Both bacteria (A) and archaea (B) were classified into four ranks according to their growth temperature, from low to high: psychrophiles/psychrotrophiles, mesophiles, thermophiles, and hyperthermophiles. The diagonal line represents cases in which prokaryotes with different ranks have the same GC contents. Points above the line (153 pairs of bacteria and 17 pairs of archaea) represent cases in which prokaryotes with higher ranks have higher GC contents than their paired relatives, while points below the line (119 pairs of bacteria and 24 pairs of archaea) indicate the reverse. The *p* values were calculated using two-tailed Wilcoxon signed-rank tests. The exact values of the GC contents are present in supplementary Table S17.

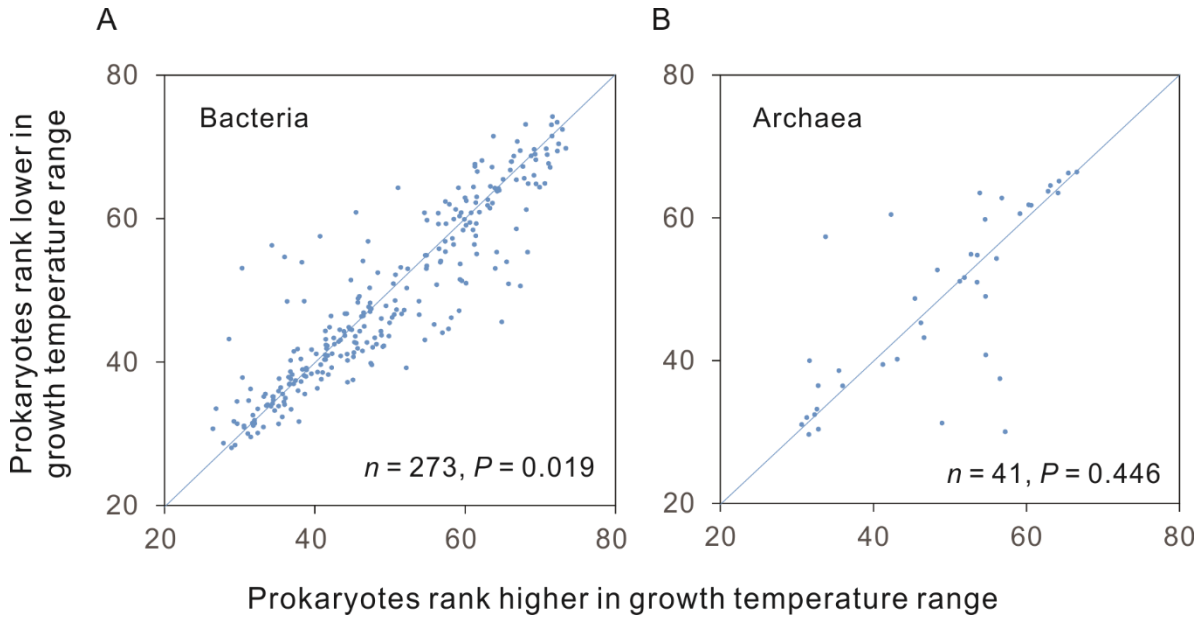
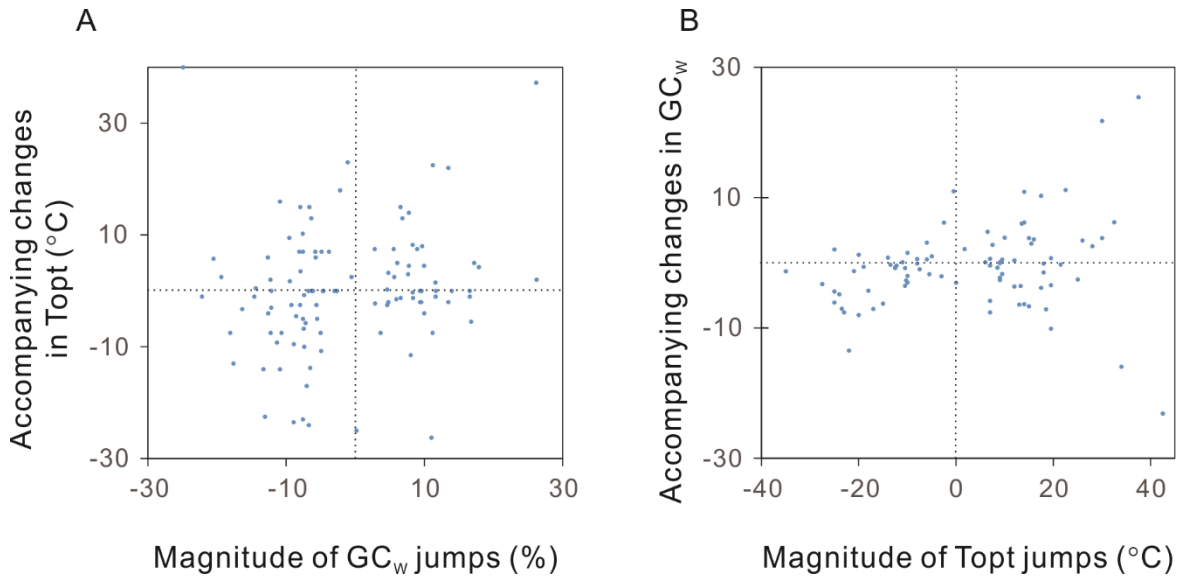
Figure 2. Positive correlations between the sudden changes in GC content and growth temperature of bacteria. Following Mahajan and Agasheand [3], the evolutionary jumps of GC_w (whole-genome GC content) and Topt (optimal growth temperature) in the bacterial phylogenetic tree were detected using the Lévy jumps model [53]. (A) the magnitude of the GC_w jumps are significantly correlated with the accompanied changes in Topt (Spearman′s rank correlation, 2-tailed, n = 108, rho = 0.209, P = 0.030). (B) the magnitude of the Topt jumps is significantly correlated with the accompanying change in GC_w (Spearman's rank correlation, 2-tailed, n = 86, rho = 0.280, P = 0.009). The exact values shown in this figure are present in supplementary Tables S18-S19.