**Relationship between prokaryotic GC content and environmental salinity**

En-Ze Hu[1,2], Shen Sun[1,2], Deng-Ke Niu[*1,2]

[1]MOE Key Laboratory for Biodiversity Science and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing 100875, China

[2]Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, Beijing 100875, China

*Corresponding authors:

Deng-Ke Niu, dkniu@bnu.edu.cn, dengkeniu@hotmail.com.

**Abstract**

**Background:** The correlation between GC content and halophilicity has received limited attention, despite the numerous environmental factors associated with GC content evolution. While higher GC content has been linked to halophiles in some archaeal cases, it is widely believed that selective pressure from high-intensity ultraviolet radiation in halophilic archaea drives GC content increase, as it prevents DNA photoproduct formation. However, this assumption has not been statistically analyzed in a phylogenetically independent manner prior to our study.

**Results:** Using phylogenetic generalized least squares, we investigated the relationship between GC content and halophilicity in 1226 bacteria and 181 archaea. Our analysis found significant positive correlations in bacteria but not in archaea. Resampling analysis indicates that the absence of significant correlation in archaea may be due to the relatively small sample size. We also observed that the strength of the correlation is negatively influenced by the functional constraint of genomic components. Additionally, we found that halophilic bacteria and archaea do not have lower photoreactivity (a measure of DNA vulnerability to ultraviolet radiation) than the photoreactivity expected from their GC contents.

**Conclusions:** In contrast to previous assumptions, we did not find evidence to support the widespread photoprotection hypothesis or another hypothesis that high GC content in halophiles stabilizes nucleic acid structures. Instead, our findings align with a nonadaptive hypothesis. Halophilic prokaryotes likely evolved high GC content due to frequent GC-biased gene conversion in response to DNA double-strand breaks induced directly or indirectly by high salt concentrations.

Keywords: DNA double-strand breaks; GC content; gene conversion; halophile; photoreactivity; phylogenetic generalized least squares (PGLS).

## Background

GC content refers to the proportion of guanine and cytosine present in nucleotide sequences. This fundamental characteristic stabilizes the double-stranded DNA helix and three-dimensional RNA structures. It can affect various aspects of genetic information performance, including three-dimensional genome organization (1), mutation and recombination rates (2), gene transcription and mRNA translation

rates (3-5), spicing site recognition (6), RNA interference efficiency (7), and horizontal gene transfer (8, 9). Many environmental factors, such as growth temperature, nutrient availability, and ultraviolet (UV) radiation (10-17), have been found or proposed to be associated with the evolution of GC content.

A wide range of high-salinity environments can be found globally, from naturally occurring hypersaline lakes and deep-sea brines to artificially created salterns. Halophiles have been discovered in all three domains of life, including Archaea, Bacteria, and eukaryotes (18). The first sequenced halophilic archaeon, *Halobacterium* sp. NRC-1, comprises a primary chromosome and two microchromosomes with GC content of 68%, 58%, and 59%, respectively (19). Early studies suggested that high GC content is a characteristic feature of extreme halophiles (20-22). However, *Haloquadratum walsbyi*, an extremely halophilic organism, has a low GC content of 47.9% (23). In a later study, Jones and Baxter (24) compared the GC content of 29 halophilic archaea and 2231 other prokaryotes using a Welch two-sample t-test and observed a remarkably significant difference (63.1% ± 1.3% vs. 49.7% ± 0.55%, $P < 2.2 \times 10^{-16}$). However, they did not consider the phylogenetic non-independence among species. Comparative analysis of phylogenetically non-independent data using traditional statistical methods often produces erroneous results (25-27). The phylogenetic non-independence of halophilic archaea is of particular concern because they are concentrated in a few phylogenetic lineages (18, 28).

To confidently determine the relationship between GC content and halophilicity, we conducted phylogenetic generalized least squares (PGLS) regression analyses (29) on 1226 bacterial species (816 non-halophilic and 410 halophilic) and 181 archaeal species (8 non-halophilic and 173 halophilic). Additionally, we explored the potential mechanisms underlying the correlation between GC content and halophilicity and revisited the prevalent hypothesis of avoiding UV-induced thymidine dimer formation (30) using the current dataset and phylogenetic comparative methods.

## Results

Prokaryotes are categorized into four groups based on their optimal range of salt concentration for growth: non-halophiles, slight halophiles, moderate halophiles, and extreme halophiles (31). To obtain a dataset for halophilicity, we utilized the Halodom database (32), Protrait database (33), and Madin et al. prokaryotic trait dataset (34). For PGLS regression analysis, halophilicity was transformed into an

ordered categorical variable that assigned numerical values to the four categories (1, 2, 3, and 4).

We re-annotated the representative genomes of the prokaryotic species (identified as reference or representative genomes in the NCBI Genome Database). Considering the strength of functional constraints, we divided prokaryotic genomes into seven genomic components and calculated their GC content: the first and second codon positions of protein-coding sequences ($GC_{1+2}$), the third codon position of protein-coding sequences ($GC_3$), intergenic sequences ($GC_{inter}$), tRNA genes ($GC_{tRNA}$), 5S rRNA genes ($GC_{5S}$), 16S rRNA genes ($GC_{16S}$), and 23S rRNA genes ($GC_{23S}$) (Additional file 1: Table S1-S2). It should be noted that the sample size for rRNA genes might be slightly smaller due to incomplete annotations in some genomes.

**Both genomic GC content and halophilicity have strong phylogenetic signals**

Horizontal gene transfer among prokaryotes makes genealogical relationships somewhat complex. If bifurcation is not the dominant pattern of phylogeny and common ancestors do not significantly influence the trait values of their descendants, conventional statistical methods may be used. To test whether phylogenetic comparative methods are necessary, we calculated Pagel's $\lambda$ values to measure the correlation between the evolution of the analyzed trait and the presumed phylogenetic tree (35). Our analysis revealed that the GC contents of whole genomes and genomic components of both bacteria and archaea in our dataset show strong phylogenetic signals (Additional file 1: Table S3). Similarly, the halophilicity of bacteria and archaea displays significant phylogenetic signals (Additional file 1: Table S3).

Halophilic archaea are naturally exposed to high-intensity UV radiation, and their high GC content was initially thought to be an adaptation that prevents the formation of UV-induced thymine dimers (19, 20, 30, 36). However, further analysis revealed that all four adjacent pyrimidines (TT, TC, CT, CC) are susceptible to photodamage under UV irradiation (37). To quantify photoreactivity (PR), Jones and Baxter (24) developed a formula that takes into account the frequencies of adjacent pyrimidines susceptible to photodamage under UV irradiation (TT, TC, CT, CC) and their intrinsic photoreactivity: $PR = 1.73(qTC) + 1.19(qTT) + 0.61(qCT) + 0.39(qCC)$, where qTC, qTT, qCT, and qCC represent the frequencies of each adjacent pyrimidine. GC content influence DNA photoreactivity since it affects the

frequencies of the four adjacent pyrimidines. Our analysis of the photoreactivity of whole genomes and genomic components of bacteria and archaea in our dataset (Additional file 1: Table S4-S5) also revealed strong phylogenetic signals (Additional file 1: Table S3).

**Whole genome GC content significantly correlates with halophilicity in bacteria but not archaea**

To investigate this relationship, we used halophilicity as the independent variable and $GC_w$ as the dependent variable in PGLS regression analyses. Regression slopes with a plus sign indicate a positive correlation, while those with a minus sign indicate a negative correlation. We used four commonly used models: Brownian motion, Ornstein-Uhlenbeck, Pagel's λ, and Early burst. The best-fit model was determined by selecting the model with the smallest Akaike information criterion (AIC) value. As shown in Table 1, the Pagel's λ model is the best fit for both bacteria and archaea, with considerably smaller AIC values compared to the other models. Under this model, we found a significant positive correlation between the genome-wide GC content and halophilicity in bacteria ($P$ = 0.015) but not archaea ($P$ = 0.908).

Table 1 PGLS regression analyses of whole-genome GC content and halophilicity

| | *n* | Model | AIC | Slope | *P* |
|---|---|---|---|---|---|
| Bacteria | 1226 (816, 35, 348, 27) | BM | 16102 | 0.002 | $1.3\times10^{-7}$ |
| | | OU | 14130 | 0.002 | $1.4\times10^{-7}$ |
| | | Pagel's λ | 7176 | 0.004 | 0.015 |
| | | EB | 16104 | 0.002 | $1.3\times10^{-7}$ |
| Archaea | 181 (8, 0, 69, 104) | BM | 1231.9 | –0.006 | $9.6\times10^{-5}$ |
| | | OU | 1152.6 | –0.006 | 0.001 |
| | | Pagel's λ | 848.8 | $2.6\times10^{-4}$ | 0.908 |
| | | EB | 1233.9 | –0.006 | $9.6\times10^{-5}$ |

$GC_w$, whole-genome GC content; BM, Brownian motion model; OU, Ornstein-Uhlenbeck model; Pagel's

λ, Pagel's λ model; EB, Early burst model; AIC, Akaike information criterion. The sample sizes of the four categories (non-halophiles, slight halophiles, moderate halophiles, and extreme halophiles) are shown in parentheses. A slope significantly > 0 indicates a significant positive correlation between the two traits, and a slope significantly < 0 indicates a significant negative correlation.

Closely related lineage exhibit highly similar GC content and halophilicity due to minimal evolutionary changes since their divergence from a common ancestor. Therefore, such lineages should not be considered as independent samples in phylogenetic comparative studies, and the effective sample size should be well below the census number of the analyzed lineages. We suspect that the lack of a significant correlation between $GC_w$ and halophilicity in archaea might be attributed to the small sample size. To test this hypothesis, we randomly resampled 181 species from the 1226 bacteria 1000 times. Results of PGLS regression analysis indicated that significant positive correlations between $GC_w$ and halophilicity disappeared in most cases (63.7%, $P > 0.05$). If we analyze a bacterial sample as small as the archaeal sample, we might not be able to find a significant correlation between $GC_w$ and halophilicity.

**Relationship between halophilicity and GC content of genome components**

We also performed PGLS regression analysis on genome components using the four models above, and Pagel's λ model was consistently preferred due to the smallest AIC value. In bacteria, halophilicity is significantly correlated with $GC_3$, $GC_{inter}$, $GC_{tRNA}$, and $GC_{5S}$ ($P = 0.036$, 0.006, 0.001, and $1.50\times10^{-4}$, respectively), and there is a marginal correlation with $GC_{1+2}$ and $GC_{23S}$ ($P = 0.059$ and 0.096, respectively). However, there is no correlation between bacterial halophilicity and $GC_{16S}$ ($P = 0.432$). The significance of the correlation between GC content and halophilicity increases from the first and second codon positions, the third codon position to intergenic sequences (Table 2). Likewise, the GC content of tRNA and 5S rRNA genes is more strongly correlated with halophilicity than the GC content of 16S rRNA and 23S rRNA genes (Table 2).

In archaea, we found no significant correlation between halophilicity and GC content of any genome components ($P > 0.10$ in all cases. Additional file 1: Table S6).

Table 2 Correlations between GC content of genome components and halophilicity in bacteria

| | *n* | Slope | *P* |
|---|---|---|---|
| $GC_{1+2}$ ~ Halophilicity | 1226 | 0.001 | 0.059 |
| $GC_3$ ~ Halophilicity | 1226 | 0.007 | 0.036 |
| $GC_{inter}$ ~ Halophilicity | 1226 | 0.005 | 0.006 |
| $GC_{tRNA}$ ~ Halophilicity | 1226 | 0.001 | 0.001 |
| $GC_{5S}$ ~ Halophilicity | 1109 | 0.004 | $1.50\times10^{-4}$ |
| $GC_{16S}$ ~ Halophilicity | 1083 | $3.1\times10^{-4}$ | 0.432 |
| $GC_{23S}$ ~ Halophilicity | 1038 | $8.6\times10^{-4}$ | 0.096 |

The correlations were examined by PGLS regressions using Pagel's λ model. A slope significantly > 0 indicates a significant positive correlation between the two traits, and a slope significantly < 0 indicates a significant negative correlation. $GC_{1+2}$, GC content of the first and second codon positions of protein-coding sequences; $GC_3$, GC content of the third codon positions of protein-coding sequences; $GC_{inter}$, GC content of intergenic sequences; $GC_{tRNA}$, GC content of tRNA genes; $GC_{5S}$, GC content of 5S rRNA genes; $GC_{16S}$, GC content of 16S rRNA genes; $GC_{23S}$, GC content of 23S rRNA genes.

**Photoreactivity is not positively correlated with halophilicity**

To investigate whether high GC content prevents UV-induced pyrimidine dimers in bacterial genomes (19, 20, 30, 36), we examined the correlation between halophilicity and photoreactivity of different components of bacterial genomes. Our PGLS regression analysis, using halophilicity as an independent variable and photoreactivity as the dependent variable, found no significant correlation between photoreactivity of bacterial whole-genome sequences, protein-coding genes, tRNA genes, 16S rRNA genes, and 23S rRNA genes and halophilicity ($P > 0.100$ for all five cases, as shown in Table 3). However, intergenic sequences and 5S rRNA genes show significant negative correlations with halophilicity (Table 3). We also analyzed archaea and found no significant correlations between halophilicity and photoreactivity ($P > 0.100$ for all cases. Additional file 1: Table S7).

Table 3 Relationship between photoreactivity and halophilicity in bacteria

| | *n* | Slope | *P* |
|---|---|---|---|
| $PR_w$ ~ Halophilicity | 1226 | $-4.9\times10^{-4}$ | 0.220 |
| $PR_p$ ~ Halophilicity | 1226 | $-3.9\times10^{-4}$ | 0.336 |
| $PR_{inter}$ ~ Halophilicity | 1226 | -0.001 | 0.002 |
| $PR_{tRNA}$ ~ Halophilicity | 1226 | $-2.4\times10^{-4}$ | 0.122 |
| $PR_{5S}$ ~ Halophilicity | 1109 | -0.001 | 0.026 |
| $PR_{16S}$ ~ Halophilicity | 1083 | $-1.8\times10^{-4}$ | 0.210 |
| $PR_{23S}$ ~ Halophilicity | 1038 | $-1.4\times10^{-4}$ | 0.349 |

The relationships were examined by PGLS regressions using Pagel's λ model. A slope significantly > 0 indicates a significant positive correlation between the two traits, and a slope significantly < 0 indicates a significant negative correlation. $PR_w$, the photoreactivity of whole-genome sequences; $PR_p$, the photoreactivity of protein-coding genes; $PR_{inter}$, the photoreactivity of intergenic sequences; $PR_{tRNA}$, the photoreactivity of tRNA genes; $PR_{5S}$, the photoreactivity of 5S rRNA genes; $PR_{16S}$, the photoreactivity of 16S rRNA genes; $PR_{23S}$, the photoreactivity of 23S rRNA genes.

In a genome rich in GC, the occurrence of fewer TTs and more CCs could be a random chance. As the formula that quantifies photoreactivity weighs qTT more than qCC (1.19 vs. 0.39), a low photoreactivity in a GC-rich genome could also be attributed to chance. If a low PR is a result of evolutionary adaptation to avoid UV-induced pyrimidine dimers, it should be lower than the expected chance of occurrence. The expected PR of a DNA sequence, E(PR), could be calculated after randomly arranging bases in the sequence (Additional file 1: Table S4-S5). The observed PR is compared with the E(PR) in three datasets (410 bacteria including slight halophiles, moderate halophiles, and extreme halophiles, 375 bacteria including moderate halophiles and extreme halophiles, and 27 bacteria including only extreme halophiles) using phylogenetic paired t-test (38). The results do not show any significant difference in the first two datasets (Table 4). However, in the 27 extremely halophilic bacteria, the PR and E(PR) showed significant differences in the whole-genome sequences, protein-coding genes, intergenic sequences, tRNA genes, 16S rRNA genes, and 23S rRNA genes (Table 4). Upon further examination of

the values, it was observed that the PR values of these sequences are higher than the E(PR) values. The same analysis for archaeal genomes also showed no significant differences ($P > 0.100$ for all cases. Additional file 1: Table S8).

In summary, the results do not support the hypothesis that GC-rich genomes in halophiles evolved to avoid UV-induced pyrimidine dimers.

Table 4 Comparison between observed photoreactivity and the photoreactivity expected by GC content in bacteria

| | *n* | Mean difference | *P* |
|---|---|---|---|
| Slight halophiles, moderate halophiles, and extreme halophiles | | | |
| $PR_w$ vs. $E(PR_w)$ | 410 | 0.011 | 0.989 |
| $PR_p$ vs. $E(PR_p)$ | 410 | 0.011 | 0.990 |
| $PR_{inter}$ vs. $E(PR_{inter})$ | 410 | 0.013 | 0.987 |
| $PR_{tRNA}$ vs. $E(PR_{tRNA})$ | 410 | 0.017 | 0.970 |
| $PR_{5S}$ vs. $E(PR_{5S})$ | 357 | –0.001 | 0.999 |
| $PR_{16S}$ vs. $E(PR_{16S})$ | 297 | 0.003 | 0.997 |
| $PR_{23S}$ vs. $E(PR_{23S})$ | 268 | 0.016 | 0.959 |
| Moderate halophiles and extreme halophiles | | | |
| $PR_w$ vs. $E(PR_w)$ | 375 | 0.011 | 0.989 |
| $PR_p$ vs. $E(PR_p)$ | 375 | 0.011 | 0.990 |
| $PR_{inter}$ vs. $E(PR_{inter})$ | 375 | 0.013 | 0.987 |
| $PR_{tRNA}$ vs. $E(PR_{tRNA})$ | 375 | 0.017 | 0.971 |
| $PR_{5S}$ vs. $E(PR_{5S})$ | 329 | –0.001 | 0.999 |
| $PR_{16S}$ vs. $E(PR_{16S})$ | 276 | 0.003 | 0.997 |
| $PR_{23S}$ vs. $E(PR_{23S})$ | 252 | 0.016 | 0.960 |
| Extreme halophiles | | | |
| $PR_w$ vs. $E(PR_w)$ | 27 | 0.015 | $5.9\times10^{-5}$ |

| | | | |
|---|---|---|---|
| $PR_p$ vs. $E(PR_p)$ | 27 | 0.015 | $1.0\times10^{-4}$ |
| $PR_{inter}$ vs. $E(PR_{inter})$ | 27 | 0.016 | $7.2\times10^{-5}$ |
| $PR_{tRNA}$ vs. $E(PR_{tRNA})$ | 27 | 0.019 | $1.8\times10^{-14}$ |
| $PR_{5S}$ vs. $E(PR_{5S})$ | 25 | –0.003 | 0.307 |
| $PR_{16S}$ vs. $E(PR_{16S})$ | 22 | 0.004 | 0.027 |
| $PR_{23S}$ vs. $E(PR_{23S})$ | 18 | 0.019 | $1.7\times10^{-9}$ |

The expected photoreactivity (PR) of a DNA sequence, E(PR), was obtained by calculating the PR of the sequence after a random base arrangement. The comparisons were performed using a phylogenetic paired t-test (38). The mean difference was the phylogenetically corrected PR – E(PR). $PR_w$, the photoreactivity of whole-genome sequences; $PR_p$, the photoreactivity of protein-coding genes; $PR_{inter}$, the photoreactivity of intergenic sequences; $PR_{tRNA}$, the photoreactivity of tRNA genes; $PR_{5S}$, the photoreactivity of 5S rRNA genes; $PR_{16S}$, the photoreactivity of 16S rRNA genes; $PR_{23S}$, the photoreactivity of 23S rRNA genes.

## Discussion

The relationship between environmental salinity and GC content evolution has not been extensively explored compared to other environmental factors. In the early days of genome sequencing, some halophilic archaea were observed to have remarkably GC-rich genomes (19-22). However, the only previous statistical analysis on the correlation between GC content and halophilicity had a methodological flaw as the phylogenetic non-independence of the prokaryotic data was not considered (24). Our PGLS regression analyses confirmed the correlation in 1226 bacteria but not 181 archaea (Table 1). To address the small sample size of archaea, we randomly resampled 181 species from the 1226 bacteria 1000 times and found that in most cases, the correlations between GC content and halophilicity become statistically nonsignificant. This lack of significant correlation in archaea, similar to our previous study on the relationship between GC content and growth temperature (11), is probably due to the small sample size.

The high concentration of NaCl is typically caused by solar-driven water evaporation in the unique habitats of halophilic archaea, which are typically characterized by high-intensity UV radiation. UV penetration is more substantial in saline water (39), and high salt content in water bodies and coastal areas can affect local atmospheric conditions, reducing ozone concentrations and leading to increased UV

radiation (40). Under UV radiation, all four types of adjacent pyrimidines (TT, TC, CT, CC) can potentially undergo photodamage (37). Although high GC content was hypothesized to be an adaptation of the archaea to intense UV radiation (19, 20, 30, 36), the photoreactivity formula (24)[PR = 1.73(qTC) + 1.19(qTT) + 0.61(qCT) + 0.39(qCC)] shows that GC content does not have the strong effect previously assumed in avoiding UV-induced DNA damage. However, photodamage is still a potential driver of GC content evolution, but our analysis in halophilic bacteria or archaea did not find a lower photoreactivity than expected by chance. Therefore, high GC content in halophilic archaea or bacteria should not be interpreted as a photoprotection strategy.

Another possible adaptive effect of high GC content in halophiles is increasing DNA stability under high salt concentrations (41). The low concentrations of cations in the solution can neutralize the negative charge of the phosphate groups on DNA molecules, reducing the repulsive forces between the strands and stabilizing the DNA double helix structure (42, 43). However, when the concentration exceeds a certain threshold, 1M, it can destabilize the DNA double helix structure (44). Some halophilic prokaryotes using the "high-salt-in" strategy need to accumulate molar concentrations of inorganic salts in the cytoplasm, and such salt concentrations exceed the threshold (18, 45). Many halophilic prokaryotes employ the "compatible solute" strategy to cope with osmotic stress by accumulating high concentrations of small organic solutes, such as proline and betaine. These small organic solutes can reduce the stability of the DNA double helix (46-48). The additional hydrogen bonds of G:C base pairs compared to A:T base pairs can help maintain genome stability in high salt concentrations. Therefore, high GC content is also favored in halophilic prokaryotes, whether they accumulate inorganic salts or small organic solutes. However, our genome component observations do not support the DNA double helix stability hypothesis. As DNA double helix structures are more stable than the three-dimensional structures of structural RNA molecules, they are more resistant to destabilizing factors. Therefore, whole-genome GC content correlates less strongly with growth temperature than structural RNA genes (11). However, in this study, we observed that the GC content of 16S and 23S rRNA genes is not significantly correlated with halophilicity (Table 2). Furthermore, the principle of natural selection does not support the DNA double helix stability hypothesis. A slight increase in whole-genome GC content, e.g., from 49.5% to 50.0%, requires thousands of AT to GC mutations, which are unlikely to happen simultaneously. Every single mutation

could only change the GC content of one nucleotide among the millions of nucleotides in a prokaryotic genome and result in a tiny change in the double helix stability that is unlikely to be perceived by natural selection.

Upon comparing the correlations between various genomic components, it was observed that the GC content of the third codon position and intergenic sequences exhibited a stronger correlation with halophilicity than that of the first and second codon positions (Table 2). This result indicates a strong correlation in sequences with weaker functional constraints. It was also noticed that the correlations in tRNA and 5S rRNA genes were stronger than those in 16S rRNA and 23S rRNA genes. The tRNA genes are known for their evolutionary flexibility, as they mutate quickly and are sites of genomic instability (49-51). On the other hand, the ribosomal RNA genes are known for their conservation in evolution. However, 5S rRNA genes exhibit higher intragenomic diversity than 16S or 23S (52). In addition, the average substitution rate per million years for 5S rRNA genes is about half of 16S rRNA genes, 0.5% vs. 1% (53). Therefore, the correlations between halophilicity and GC content among structural RNA genes are also significant in sequences with weaker functional constraints. Researches indicate that high concentrations of salt and organic solutes may damage DNA sequences directly or via oxidative stress, inducing DNA double-strand breaks (DSB) (54-56). Consequently, bacteria living in high salt concentrations should have evolved more efficient repairing systems than nonhalophiles. From bacteria to humans, gene conversion acts as an essential role in repairing DSBs (57). Increasing GC content has been documented as part of the recombination process in gene conversion (58-61). Hence, the correlation between GC content and halophilicity is attributed to this nonadaptive process.

No statistically supported conclusion can be drawn for high GC content in halophilic archaea, owing to the limited number of phylogenetic branches diverging in halophilicity. However, almost all the solitary cases, like the class Halobacteria, indicated halophilic archaea have high GC content (11, 18-22). Most halophilic archaea are polyploid and likely use efficient gene conversion to repair their mutations and DSBs (62, 63).

## Conclusions

The commonly observed high GC content in halophilic archaea is widely believed to be an adaptation to

avoid UV-induced pyrimidine dimer formation due to their typically high-intensity UV radiation habitats. By pooling data from three comprehensive databases, a dataset of halophilicity was established, including 1226 bacterial species and 181 archaeal species. Phylogenetic correlation analysis using PGLS demonstrated that genome-wide GC content and halophilicity are significantly correlated in bacteria but not archaea. Further analysis of the correlation between halophilicity and GC content among seven genomic components does not support the widespread photoprotection hypothesis or another hypothesis that the high GC content in halophiles stabilizes nucleic acid structures but instead aligns with a nonadaptive hypothesis. Halophiles evolved high GC content probably due to frequent GC-biased gene conversion, which response to the DNA DSBs induced directly or indirectly by the high salt concentrations.

## Methods

The halophilicity data for bacteria and archaea were obtained from three databases: Halodom (32), Protrait (33), and a prokaryotic trait dataset constructed by Madin et al. (34). Among these, the Halodom database, which was manually curated from published literature, was considered the most reliable. The dataset compiled by Madin et al. through integrating multiple databases was considered the next most reliable. The ProTraits database, which employs semi-supervised text mining methods to extract data from online databases and literature, was considered the least reliable. We consolidated the data from these three sources, prioritizing the Halodom database in case of discrepancies, followed by the dataset of Madin et al. and then the ProTraits database. In total, we obtained the halophilicity of 2871 prokaryotic species.

We retrieved the phylogenetic tree provided by the TimeTree database (64) and obtained the phylogenetic information of 1226 bacterial species (816 non-halophiles and 410 halophiles) and 181 archaeal species (8 non-halophiles and 173 halophiles).

The GC contents of prokaryotic whole genomes were gathered from the NCBI Genome Database (https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt, accessed on 2022-06-25). In cases with multiple sequenced strains, the median GC content of all strains was utilized to represent the $GC_w$ of the respective species.

We downloaded the genomes designated as reference or representative genomes by the NCBI database for each species and re-annotated them using the DFAST software (version 1.2.11) with default parameters (65). We calculated the GC content of different components in the genome, including protein-coding sequences, intergenic regions, tRNA genes, 5S rRNA genes, 16S rRNA genes, and 23S rRNA genes, based on the annotation results from DFAST. For species without a designated reference or representative genome, we annotated all sequenced genomes of that species. We used the median GC content of each component of a species to represent the species data.

The *phytools* package (Version 0.7–70) in R (Version 4.0.3) (66) was used to determine the phylogenetic signals ($\lambda$) and to conduct the phylogenetic paired t-tests. The PGLS regressions were conducted using the *phylolm* package (version 2.6.2) in R (Version 4.0.3) with default parameters (67).

In this study, we did not present the Benjamini-Hochberg (BH)-adjusted *P* values for multiple correlation analyses of the same dataset. In most cases, the *P* and adjusted *P* values were generally on the same side of 0.05, smaller or bigger than 0.05. In a few cases where the significance test results changed because of BH corrections (such as the $GC_3$ ~ Halophilicity in Table 2), we focused on the relative strengths of the correlations rather than the absolute correlation significance values.

**List of abbreviations**

AIC: Akaike information criterion

BH: Benjamini-Hochberg

BM: Brownian motion model

DSB: double-strand breaks

EB: Early burst model

$GC_{1+2}$: GC content of the first and second codon positions of protein-coding sequences

$GC_{16S}$: GC content of 16S rRNA genes

$GC_{23S}$: GC content of 23S rRNA genes

$GC_3$: GC content of the third codon position of protein-coding sequences

$GC_{5S}$: GC content of 5S rRNA genes

$GC_{inter}$: GC content of intergenic sequences

$GC_{tRNA}$: GC content of tRNA genes

$GC_w$: Whole-genome GC content

OU: Ornstein-Uhlenbeck model

PGLS: Phylogenetic generalized least squares

PR: Photoreactivity

$PR_w$: Whole-genome photoreactivity

$PR_p$: Photoreactivity of protein-coding genes

$PR_{inter}$: Photoreactivity of intergenic sequences

$PR_{tRNA}$: Photoreactivity of tRNA genes

$PR_{5S}$: Photoreactivity of 5S rRNA genes

$PR_{16S}$: Photoreactivity of 16S rRNA genes

$PR_{23S}$: Photoreactivity of 23S rRNA genes

UV: Ultraviolet

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional file.

### Competing interests

The authors declare that they have no competing interests.

### Funding

This work was supported by the National Natural Science Foundation of China (grant number 31671321). The funders had no role in the design of the study or collection, analysis, and interpretation of data or in writing the manuscript.

## Authors' contributions

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

## Acknowledgements

Not applicable.

## References

1. Bouwman BAM, Crosetto N, Bienko M. A GC-centered view of 3D genome organization. Curr Opin Genet Dev. 2023;78:102020.
2. Kiktev DA, Sheng Z, Lobachev KS, Petes TD. GC content elevates mutation and recombination rates in the yeast *Saccharomyces cerevisiae*. Proc Natl Acad Sci USA. 2018;115(30):E7109-E18.
3. Gul IS, Staal J, Hulpiau P, De Keuckelaere E, Kamm K, Deroo T, et al. GC content of early metazoan genes and its impact on gene expression levels in mammalian cell lines. Genome Biol Evol. 2018;10(3):909-17.
4. Sun Y, Tamarit D, Andersson SGE. Switches in genomic GC content drive shifts of optimal codons under sustained selection on synonymous sites. Genome Biol Evol. 2017;9(10):2560-79.
5. Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. High guanine and cytosine content increases mRNA levels in mammalian cells. PLoS Biol. 2006;4(6):e180.
6. Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, et al. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep. 2012;1(5):543-56.
7. Chan CY, Carmack CS, Long DD, Maliyekkel A, Shao Y, Roninson IB, et al. A structural interpretation of the effect of GC-content on efficiency of RNA interference. BMC Bioinformatics. 2009;10(1):S33.
8. 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.
9. 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.

10. Musto H. How many factors influence genomic GC content among prokaryotes? J Mol Evol. 2023;91(1):6-9.

11. Hu E-Z, Lan X-R, Liu Z-L, Gao J, Niu D-K. A positive correlation between GC content and growth temperature in prokaryotes. BMC Genomics. 2022;23(1):110.

12. 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.

13. Luo HW, Thompson LR, Stingl U, Hughes AL. Selection maintains low genomic GC content in marine SAR11 lineages. Mol Biol Evol. 2015;32(10):2738-48.

14. 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.

15. 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.

16. Rudi K. Environmental shaping of ribosomal RNA nucleotide composition. Microb Ecol. 2009;57(3):469-77.

17. Agashe D, Shankar N. The evolution of bacterial DNA base composition. J Exp Zool Part B. 2014;322(7):517-28.

18. Oren A. Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline Syst. 2008;4(1):2.

19. Kennedy SP, Ng WV, Salzberg SL, Hood L, DasSarma S. Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res. 2001;11(10):1641-50.

20. Soppa J. From genomes to function: haloarchaea as model organisms. Microbiology. 2006;152(3):585-90.

21. Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, Deutsch EW, et al. Genome sequence of *Haloarcula marismortui*: a halophilic archaeon from the Dead Sea. Genome Res. 2004;14(11):2221-34.

22. Falb M, Pfeiffer F, Palm P, Rodewald K, Hickmann V, Tittor J, et al. Living with two extremes: conclusions from the genome sequence of *Natronomonas pharaonis*. Genome Res. 2005;15(10):1336-43.

23. Bolhuis H, Palm P, Wende A, Falb M, Rampp M, Rodriguez-Valera F, et al. The genome of the square archaeon *Haloquadratum walsbyi*: life at the limits of water activity. BMC Genomics. 2006;7:169.

24. Jones DL, Baxter BK. Bipyrimidine signatures as a photoprotective genome strategy in G + C-rich halophilic archaea. Life. 2016;6(3):37.

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

26. Whitney KD, Garland T, Jr. Did genetic drift drive increases in genome complexity? PLoS Genet. 2010;6(8):e1001080.

27. Garamszegi LZ, editor. Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice. Berlin: Springer; 2014.

28. Andrei AS, Banciu HL, Oren A. Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. FEMS Microbiol Lett. 2012;330(1):1-9.

29. Grafen A. The phylogenetic regression. Philos Trans R Soc Lond B Biol Sci. 1989;326(1233):119-57.

30. Singer CE, Ames BN. Sunlight ultraviolet and bacterial DNA base ratios. Science. 1970;170(3960):822-6.

31. Kushner DJ. Life in high salt and solute concentrations: halophilic bacteria. In: Kushner DJ, editor. Microbial Life in Extreme Environments. London: Academic Press; 1978. p. 317–68.

32. Loukas A, Kappas I, Abatzopoulos TJ. HaloDom: a new database of halophiles across all life domains. J Biol Res-Thessalon. 2018;25(1):2.

33. 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.

34. Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, et al. A synthesis of bacterial and archaeal phenotypic trait data. Scientific Data. 2020;7(1):170.

35. Pagel M. Inferring the historical patterns of biological evolution. Nature. 1999;401(6756):877-84.

36. Hescox MA, Carlberg DM. Photoreactivation in *Halobacterium cutirubrum*. Can J Microbiol. 1972;18(7):981-5.

37. Matallana-Surget S, Meador JA, Joux F, Douki T. Effect of the GC content of DNA on the distribution of UVB-induced bipyrimidine photoproducts. Photoch Photobio Sci. 2008;7(7):794-801.

38. Lindenfors P, Revell LJ, Nunn CL. Sexual dimorphism in primate aerobic capacity: a phylogenetic test. J Evol Biol. 2010;23(6):1183-94.

39. Arts MT, Robarts RD, Kasai F, Waiser MJ, Tumber VP, Plante AJ, et al. The attenuation of ultraviolet radiation in high dissolved organic carbon waters of wetlands and lakes on the northern Great Plains. Limnol Oceanogr. 2000;45(2):292-9.

40. Stutz J, Ackermann R, Fast JD, Barrie L. Atmospheric reactive chlorine and bromine at the Great Salt Lake, Utah. Geophys Res Lett. 2002;29(10):10.1029/2002GL014812.

41. Litchfield CD. Survival strategies for microorganisms in hypersaline environments and their relevance to life on early Mars. Meteorit Planet Sci. 1998;33(4):813-9.

42. Owczarzy R, Moreira BG, You Y, Behlke MA, Walder JA. Predicting stability of DNA duplexes in solutions containing magnesium and monovalent cations. Biochemistry. 2008;47(19):5336-53.

43. Singh A, Singh N. Effect of salt concentration on the stability of heterogeneous DNA. Physica A. 2015;419:328-34.

44. Maity A, Singh A, Singh N. Differential stability of DNA based on salt concentration. Eur Biophys J. 2017;46(1):33-40.

45. Oren A. Life at high salt concentrations, intracellular KCl concentrations, and acidic proteomes. Front Microbiol. 2013;4:315.

46. Rajendrakumar CS, Suryanarayana T, Reddy AR. DNA helix destabilization by proline and betaine: possible

role in the salinity tolerance process. FEBS Lett. 1997;410(2-3):201-5.

47. Rees WA, Yager TD, Korte J, Von Hippel PH. Betaine can eliminate the base pair composition dependence of DNA melting. Biochemistry. 1993;32(1):137-44.

48. Kurz M. Compatible solute influence on nucleic acids: many questions but few answers. Saline Syst. 2008;4:6.

49. Guimarães AR, Correia I, Sousa I, Oliveira C, Moura G, Bezerra AR, et al. tRNAs as a driving force of genome evolution in yeast. Front Microbiol. 2021;12.

50. Yona AH, Bloom-Ackermann Z, Frumkin I, Hanson-Smith V, Charpak-Amikam Y, Feng Q, et al. tRNA genes rapidly change in evolution to meet novel translational demands. eLife. 2013;2:e01339.

51. Zahra K, Gökçe BA, Wing YN, Gunda D-S, Pabitra N, Jenna G. Large-scale duplication events underpin population-level flexibility in bacterial tRNA gene copy number. bioRxiv. 2022:2022.12.02.516541.

52. Pei A, Li H, Oberdorf WE, Alekseyenko AV, Parsons T, Yang L, et al. Diversity of 5S rRNA genes within individual prokaryotic genomes. FEMS Microbiology Letters. 2012;335(1):11-8.

53. Ochman H, Wilson AC. Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. J Mol Evol. 1987;26(1):74-86.

54. Burg MB, Ferraris JD, Dmitrieva NI. Cellular response to hyperosmotic stresses. Physiol Rev. 2007;87(4):1441-74.

55. Dikilitas M, Collins A, Kocyigit A, EL YAMANI N, Karakas S. DNA damage in potato plants exposed to high level of NaCl stress. Front. Genet. Conference Abstract: ICAW 2015 - 11th International Comet Assay Workshop. doi: 10.3389/conf.fgene.2015.01.00066. Frontiers in Genetics Conference Abstract: ICAW 2015 - 11th International Comet Assay Workshop. 2015:10.3389/conf.fgene.2015.01.00066.

56. Hosseiniyan Khatibi SM, Zununi Vahed F, Sharifi S, Ardalan M, Mohajel Shoja M, Zununi Vahed S. Osmolytes resist against harsh osmolarity: Something old something new. Biochimie. 2019;158:156-64.

57. Santoyo G, Romero D. Gene conversion and concerted evolution in bacterial genomes. FEMS Microbiol Rev.

2005;29(2):169-83.

58. 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.

59. 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.

60. Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, et al. Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol. 2018;2(2):237-40.

61. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285-311.

62. Ludt K, Soppa J. Polyploidy in halophilic archaea: regulation, evolutionary advantages, and gene conversion. Biochem Soc Trans. 2019;47(3):933-44.

63. Naitam MG, Grover M, Kaushik R. Polyploidy in Prokaryotes: Evolutionary Advantages and Strategy for Survival under Extreme Conditions. Haya Saudi J Life Sci. 2021;6:205-12.

64. Kumar S, Suleski M, Craig JM, Kasprowicz AE, Sanderford M, Li M, et al. TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol. 2022;39(8):msac174.

65. Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics. 2018;34(6):1037-9.

66. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3(2):217-23.

67. Ho LST, Ane C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Syst Biol. 2014;63(3):397-408.