Relationship between prokaryotic GC contents 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: Many environmental factors are associated with GC content evolution. However, the relationship between GC content and halophilicity has been poorly explored. Higher GC contents have been linked to halophiles because of a few solitary archaeal cases. The only previous statistical study on the relationship ignored the phylogenetic non-independence of species. The habitats of halophilic archaea typically have high-intensity ultraviolet radiation. Avoiding the DNA photoproduct is widely believed as a selective force to increase GC content in archaea living in high saline/ultraviolet-radiation environments. **Results:** Using phylogenetic generalized least squares (PGLS), we examined the relationship between GC content and halophilicity in 1226 bacteria and 181 archaea. Significant positive correlations have been found in bacteria but not in archaea. We suspect that the crowding of most archaeal halophiles in a few phylogenetic lineages dramatically reduces the effective sample size and probably makes the PGLS method less efficient in detecting correlations. The GC contents of different genomic components were explored separately. Bacterial intergenic sequences have GC contents more strongly correlated with halophilicity than protein-coding genes, and 5S rRNA and tRNA genes have GC contents more strongly correlated with halophilicity than 16S and 23S rRNA genes. Using PGLS, we measured the relationship between GC content and photoreactivity, a measure of the vulnerability of DNA sequences under ultraviolet radiation. No significant correlation was observed in the 1226 bacteria or the 181 archaea. **Conclusions:** We proposed that the correlation between bacterial GC content and environmental salinity should be attributed to the GC-biased DNA repairing process and the insertion of GC-rich sequences.

Key words: GC content; halophiles; rRNA genes; expansion segments; phylogenetic generalized least squares.

Background

GC content is the proportion of guanine and cytosine in nucleotide sequences. This fundamental character underlies the stability of double-strand DNA helix and three-dimensional RNA structures. It possibly influences numerous aspects of the performance of genetic information from the 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), to horizontal gene transfer (8, 9). The evolution of GC contents has been found or suggested to be associated with many environmental factors, like growth temperature, nutrient availability, and ultraviolet radiation (10-17).

Diverse high-salinity environments are distributed across Earth, from naturally occurring hypersaline lakes, and deep-sea brines to artificially created saltern. Halophiles have been found in all three domains of life, Archaea, Bacteria, and eukaryotes (18). The first sequenced halophilic archaeon, *Halobacterium* sp. NRC-1, has a main chromosome and two microchromosomes with GC contents of 68%, 58%, and 59%, respectively (19). With more halophilic archaea sequenced, early studies suggested that high GC content is a common feature of extreme halophiles (20-22). However, there is a remarkable exception, the extreme halophilic organism *Haloquadratum walsbyi* has a low GC content of 47.9% (23). Later, Jones and Baxter (24) compared the GC content of 29 halophilic archaea and 2231 other prokaryotes using the Welch Two Sample t-test. They observed a highly significant difference (63.1% ± 1.3% vs. 49.7% ± 0.55%, $P < 2.2 \times 10^{-16}$). However, they did not account for the phylogenetic non-independence among the species. Comparative analysis of phylogenetically non-independent data using conventional statistical methods often gives erroneous results (25-27). The phylogenetic non-independence of halophilic archaea is especially a concern because they are concentrated on a few phylogenetic lineages (18, 28).

For a statistically confident conclusion on the relationship between GC content and halophilicity, we performed phylogenetic generalized least squares (PGLS) regression analyses (29) on 1226 bacterial species (816 non-halophilic and 410 halophilic) and 180 archaeal species (7 non-halophilic and 173 halophilic). Furthermore, the potential mechanisms underlying the correlation between GC content and halophilicity have been explored, and the prevailing hypothesis of avoiding ultraviolet-induced thymidine dimer formation (30) has been re-examined using the present dataset and the phylogenetic comparative methods.

Results

We first constructed a dataset of prokaryotic halophilicity comprising 1226 bacterial species and 181 archaeal species. According to their optimum growth salinity (Sopt), these prokaryotes were divided into four categories: non-halophiles (Sopt < 0.2 M NaCl), slight halophiles (0.2 M ≤ Sopt < 0.85 M NaCl), moderate halophiles (0.85 M ≤ Sopt < 3.4 M NaCl), and extreme halophiles (3.4 M ≤ Sopt < 5.1 M NaCl). For the convenience of PGLS regression analysis, the halophilicity was converted into an ordered categorical variable by designating the four categories as 1, 2, 3, and 4, respectively.

Both genomic GC contents and halophilicity have strong phylogenetic signals

The widespread horizontal gene transfer among bacteria and archaea makes their genealogical relationships exhibit a somewhat network-like structure. If bifurcation is not the phylogeny's dominant pattern, the common ancestors do not significantly influence the trait values of their descendants. In this case, the trait values of terminal descendants are not strongly related to each other, so that conventional statistical methods could be applied. To examine the necessity of phylogenetic comparative methods, we calculated Pagel's λ values, a phylogenetic signal measuring the correlation between the evolution of the analyzed trait and the presumed phylogenetic tree. Similar to our previous study (11), the whole-genome GC contents (GC_w) of the bacteria and archaea in the present data show strong phylogenetic signals (Table 1). Meanwhile, the Sopts in bacteria and archaea also have significant phylogenetic signals.

Table 1 The phylogenetic signals of genomic GC contents and halophilicity

	Archaea			Bacteria		
	n	Pagel's λ	*P*	*n*	Pagel's λ	*P*
Sopt	180	0.765	4.5×10^{-18}	1226	0.778	1.1×10^{-165}
GC_w	180	0.998	5.5×10^{-97}	1226	0.978	2.2×10^{-274}

GC_w, whole-genome GC contents; Sopt, optimum growth salinity. The λ values range from 0 to 1. When λ = 0, the trait has evolved independently of the phylogeny. By contrast, λ = 1 represents the strongest phylogenetic signal.

Whole-genome GC contents significantly correlated with halophilicity in bacteria but not in archaea

Using Sopt as the independent variable and GC_w as the dependent variable, we performed a PGLS regression analysis to examine the relationship between whole-genome GC contents and halophilicity. The signs of the regression slopes, + and –, indicate positive and negative correlations, respectively. Four widely used models have been applied, the Brownian motion model (BM), Ornstein-Uhlenbeck model (OU), Pagel's lambda model (Pagel's λ), and Early burst model (EB). The minor Akaike information criterion (AIC) value was used to justify the best-fit model. As shown in Table 2, Pagel's λ model is the best-fit model for bacteria and archaea, with much lower AIC values than the other three models. Under this model, a significant positive correlation between whole-genome GC contents and halophilicity was found in bacteria (P = 0.015) but not archaea (*P* = 0.926). The lack of significant correlation in archaea should be cautioned because of the sample size imbalance. The tiny sample for non-halophiles and lack of sample in slight halophiles might introduce a great noise to the analysis.

Table 2 PGLS regression analyses of whole-genome GC content and halophilicity

	n	Model	AIC	Slope	*P*

		BM	16102	0.002	1.30×10^{-7}
Bacteria	1226 (816,	OU	14130	0.002	1.37×10^{-7}
GC_w~Sopt	35, 348, 27)	Pagel's λ	7176	0.004	0.015
		EB	16104	0.002	1.30×10^{-7}
		BM	1222	–0.006	9.97×10^{-5}
Archaea	180 (7, 0,	OU	1143	–0.006	0.001
GC_w~Sopt	69, 104)	Pagel's λ	841	2.1×10^{-4}	0.926
		EB	1224	–0.006	9.97×10^{-5}

GC_w, whole-genome GC contents; Sopt, optimum growth salinity; BM, Brownian motion model; OU, Ornstein-Uhlenbeck model; Pagel's λ, Pagel's lambda 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. The 'Model' represents the corresponding model, and the AIC value is used to judge the model's fit; the model with the smallest AIC value is considered the best model. A slope significantly > 0 indicates a significant positive correlation between the two traits, and a slope significantly < 0 indicates a significant negative correlation.

Correlations of bacterial halophilicity with the GC contents of genome components

We re-annotated the representative genomes (marked as reference or representative genomes in the NCBI Genome Database) of the 1226 bacterial species in our dataset. We then calculated the GC content of the seven genomic components, protein-coding sequences (GC_p), fourfold degenerate sites (GC_4), intergenic sequences (GC_{inter}), tRNA genes (GC_{tRNA}), 5S rRNA genes (GC_{5S}), 16S rRNA genes (GC_{16S}), and 23S rRNA genes (GC_{23S}). The rRNA genes were not thoroughly annotated in all the genomes, so the sample size for the rRNA genes is slightly smaller than other components.

We also performed PGLS regression analysis using the above four models. As the AIC value of Pagel's lambda model is consistently smaller than the other three models, we only present the regression results provided by Pagel's lambda model (Table 3). To reduce the false-positive results potentially arising from multiple correlation tests on the same dataset, we adjusted the *p* values using the Benjamini-Hochberg (BH) method and denoted them as P_{BH}. GC_p and GC_4 have marginally significant positive correlations with halophilicity (P_{BH} = 0.058). However, GC_4 has a steeper slope than GC_p, indicating a stronger correlation of GC_4 with halophilicity than GC_p. The correlation between GC_{inter} and halophilicity is significant (P_{BH} = 0.017). Among structural RNA genes, GC_{tRNA} and GC_{5S} exhibit significant positive correlations with halophilicity (P_{BH} = 0.004 and 0.001, respectively), while GC_{16S} and GC_{23S} are not correlated with halophilicity (P_{BH} > 0.05 for both cases).

Table 3 PGLS regression analyses of GC contents of genome components and halophilicity

	n	Slope	P	P_{BH}
GC_p~Sopt	1226	0.003	0.042	0.058
GC_4~Sopt	1226	0.008	0.044	0.058
GC_{inter}~Sopt	1226	0.005	0.006	0.017
GC_{tRNA}~Sopt	1226	0.001	0.001	0.004
GC_{5S}-Sopt	1109	0.004	1.50×10^{-4}	0.001
GC_{16S}~Sopt	1083	3.1×10^{-4}	0.431	0.431
GC_{23S}~Sopt	1038	8.6×10^{-4}	0.096	0.110

GC_p, GC contents of protein-coding sequences; GC_4, GC contents of fourfold degenerate sites; GC_{inter}, GC contents of intergenic sequences; GC_{tRNA}, GC contents of tRNA genes; GC_{5S}, GC contents of 5S rRNA genes; GC_{16S}, GC contents of 16S rRNA genes; GC_{23S}, GC contents of 23S rRNA genes; P_{BH}, the corrected P value using the Benjamini-Hochberg method. The AIC value of Pagel's λ model is consistently lower than the other three models (Brownian motion model, Ornstein-Uhlenbeck model, or Early burst model), so the regression results of this model are displayed.

Photoreactivity is not correlated with halophilicity

As halophilic archaea typically suffer high-intensity ultraviolet radiation in their natural habitats, high GC contents are widely believed as an adaptation of the halophilic archaea to avoid the formation of ultraviolet-induced thymine dimers (19, 20, 30, 31). With the discovery that all four adjacent pyrimidines (TT, TC, CT, CC) are susceptible to photodamage under UV irradiation (32), Jones and Baxter (24) established a formula to quantify the photoreactivity (Pg):

Pg = 1.73(qTC) + 1.19(qTT) + 0.61(qCT) + 0.39(qCC).

In this formula, qTC, qTT, qCT, and qCC represent the frequencies of each adjacent pyrimidine. The weight coefficients reflect the different intrinsic photoreactivity of the four adjacent pyrimidines. The GC contents influence the frequencies of the four adjacent pyrimidines in the genome and potentially affect the theoretical photoreactivity of the genome sequence.

To validate this hypothesis, we calculated the Pg for each bacterial genome and performed PGLS regression analysis with halophilicity as the independent variable and the Pg as the dependent variable. No significant correlation was observed (n = 1226, slope = -4.9×10^{-4}, P = 0.220). We also employed the same analysis in archaea. Their genomic photoreactivity is not correlated with halophilicity, either (n = 180, slope = -1.8×10^{-4}, P = 0.717).

Furthermore, we calculated the Pg of bacterial intergenic sequences and found a significant negative correlation between the photoreactivity of intergenic sequences and halophilicity (n = 1226, slope = –0.001, P = 0.002). In a GC-rich genome, fewer TTs and more CCs could happen simply by chance. As qTT is weighted more qCC in the formula quantifying Pg (1.19 *vs.* 0.39), a GC-rich genome might have a low Pg simply by chance. If a low Pg is evolved to avoid the formation of ultraviolet-induced pyrimidine dimers, it should be lower than that expected by chance. The expected Pg of a DNA sequence could be obtained by

calculating the Pg of the sequence after random base arrangement. Then, a relative Pg, defined as the ratio of the observed Pg to the expected Pg, could be used as the signature of evolution against ultraviolet damage. PGLS regression analysis using the relative Pg of bacterial intergenic sequences as the dependent variable and the halophilicity as the independent variable showed that there is no significant correlation between them (n=1226, slope = 4.4×10^{-4}, $P = 0.687$).

In summary, we did not find any evidence supporting the hypothesis that GC-rich genomes evolved in halophiles to avoid the formation of ultraviolet-induced pyrimidine dimers.

Discussion

Compared with other environmental factors, environmental salinity has been poorly explored in its relationship with GC content evolution. The remarkably GC-rich genomes of a few halophilic archaea were observed in the early genome sequencing era (19-22). The only previous statistical analysis on the relationship between GC contents and halophilicity (24) has a methodological flaw. The phylogenetic non-independence of the prokaryotic data had not been taken into account. Our PGLS regression analyses confirmed the correlation in bacteria but not archaea (Table 2). For the lack of significant correlation in archaea, we proposed that the highly biased distribution of halophiles and non-halophiles on the phylogenetic tree of archaea might make the PGLS less efficient in detecting correlations. In our previous analysis of the relationship between GC content and growth temperature, the small effective sample size had been attributed to the lack of significant correlation in archaea after a re-sampling analysis (11).

The high concentration of MaCl was generally caused by solar-driven water evaporation, which means inevitably intense ultraviolet radiation. The high-intensity ultraviolet radiation typically characterizes the unique habitats of halophilic archaea. It has been shown that ultraviolet penetration is more substantial in saline water (33), and the high salt content in water bodies and coastal areas can affect local atmospheric conditions, reducing ozone concentrations and leading to increased ultraviolet radiation (34). Under ultraviolet radiation, all four types of adjacent pyrimidines (TT, TC, CT, CC) can potentially undergo photodamage (32). Among them, ultraviolet-induced thymidine dimer formation is the first recognized, well-known photodamage. This damage provoked the hypothesis that high GC content is an adaptation of the archaea to the intense ultraviolet radiation of the saline environments (19, 20, 30, 31). The photoreactivity formula (24), Pg = 1.73(qTC) + 1.19(qTT) + 0.61(qCT) + 0.39(qCC), shows that GC contents do not have the strong effect previously assumed in avoiding ultraviolet-induced DNA damages. Although being weak, photodamage is still a potential driver of GC content evolution. In the present study, significant negative correlation between photoreactivity and halophilicity were not found in either bacteria or archaea. Thus, high GC contents in halophilic archaea or bacteria should not be interpreted as a photoprotection strategy.

Another possible adaptive effect of high GC content in halophiles is to increase the stability of DNA molecules under high salt concentrations (35). 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 (36-38). However, when the concentration exceeds a certain threshold, 1M, it will destabilize the DNA double helix structure (39). The halophilic prokaryotes that use the "high-salt-in" strategy need to accumulate molar concentrations of inorganic salts in the cytoplasm, and such salt concentrations exceed the threshold (18, 40). G:C base pairs have additional hydrogen bonds compared to A:T base pairs, which can help maintain the stability of the genome in high salt concentrations. Many halophilic prokaryotes employ another strategy to cope with osmotic stress, the "compatible solute" strategy. They accumulate high concentrations of small organic solutes, such as proline and betaine, to maintain the osmotic balance across their cell membranes. These small organic solutes can reduce the stability of the DNA double helix (41-43). Therefore, high GC contents are also favored in the halophilic prokaryotes with the "compatible solute" strategy.

However, our genome component observation does not support the DNA double helix stability hypothesis. As the DNA double helix structures are more stable than the three-dimensional structures of structural RNA molecules, they are more resistant to high temperatures, so whole-genome GC contents are not as strongly correlated with growth temperature as structural RNA genes (11). In this study, we observed that the GC contents of 16S and 23S rRNA genes are not significantly correlated with halophilicity (Table 3). In addition, the principle of natural selection does not support the double helix stability hypothesis. A slight difference in whole genome GC contents, like 49% *vs.* 50%, required thousands of GC/AT-changing mutations, which is 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 perceived by natural selection.

We noticed that bacterial intergenic sequences have GC contents more strongly correlated with halophilicity than protein-coding sequences, and tRNA and 5S rRNA genes have GC contents more strongly correlated with halophilicity than 16S and 23S rRNA genes (Table 3). These differences might reflect differential retention of GC-content-changing mutations, with most mutations in protein-coding sequences being eliminated because of functional constraints and more mutations retained in intergenic sequences by genetic drift. Accumulating evidence suggests that high salt concentrations and organic solute concentrations could damage DNA sequences directly or via oxidative stress, inducing DNA changes like double-strand breaks (44-46). Furthermore, there is evidence that DNA damage may probably be a fundamental driver of GC content via the GC-biased double-strand break repair process (47).

In animals, rRNA gene size and complexity have been observed to increase in evolution by inserting GC-rich sequences called expansion segments (48-50). A recent computational study revealed the insertion of the expansion segments in the 5S rRNA genes of some disparate bacterial and archaeal genomes (51). Future studies are required to examine whether the insertion of GC-rich expansion segments could account for the significant correlation between the GC contents of 5S rRNA genes and halophilicity and whether it confers the halophilic bacteria some adaptive advantage.

Methods

The halophilicity data for bacteria and archaea are retrieved from the Halodom database (52), the Protrait database (53), and the prokaryotic trait dataset constructed by Madin et al. (54). The Halodom database, manually curated from published literature, was considered the most reliable among the three databases. The data set compiled by Madin et al. through integrating multiple databases is considered the next one. The ProTraits database, which employs semi-supervised text mining methods to extract data from multiple online databases and literature, is considered the least reliable. We consolidated the data from these three sources and prioritized the data from 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 (55) and ultimately obtained the phylogenetic information of 1226 bacterial species (816 non-halophiles and 410 halophiles) and 180 archaeal species (7 non-halophiles and 173 halophiles).

The whole-genome GC contents of prokaryotes were obtained from the NCBI Genome Database (https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt, accessed on 2022-06-25). For species with multiple sequenced strains, the median GC content of all strains was taken to represent the whole genome GC content of that 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 (56). According to the annotation results from DFAST, we calculated the GC content of the following components in the genome using Python scripts: protein-coding sequences, fourfold degenerate sites, intergenic regions, tRNA genes, 5S rRNA genes, 16S rRNA genes, and 23S rRNA genes. If a species did not have a designated representative genome, we annotated all sequenced genomes of that species. The median GC content of each component of a species was used to represent the species data.

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. EZH and DKN wrote the manuscript. All authors read, improved, and approved the final manuscript.

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 systems. 2008;4(1):1-13.
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. 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 [Internet]. 2016; 6(3).
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. Hescox MA, Carlberg DM. Photoreactivation in *Halobacterium cutirubrum*. Can J Microbiol. 1972;18(7):981-&.
32. 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.
33. Huovinen PS, Penttilä H, Soimasuo MR. Spectral attenuation of solar ultraviolet radiation in humic lakes in Central Finland. Chemosphere. 2003;51(3):205-14.
34. Stutz J, Ackermann R, Fast JD, Barrie L. Atmospheric reactive chlorine and bromine at the Great Salt Lake, Utah. Geophysical Research Letters. 2002;29(10):18-1--4.
35. 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.
36. Owczarzy R, You Y, Moreira BG, Manthey JA, Huang L, Behlke MA, et al. Effects of sodium ions on DNA duplex oligomers: improved predictions of melting temperatures. Biochemistry. 2004;43(12):3537-54.
37. 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.
38. Singh A, Singh N. Effect of salt concentration on the stability of heterogeneous DNA. Physica A: Statistical Mechanics and its Applications. 2015;419:328-34.
39. Maity A, Singh A, Singh N. Differential stability of DNA based on salt concentration. European Biophysics Journal. 2017;46(1):33-40.
40. Oren A. Life at high salt concentrations, intracellular KCl concentrations, and acidic proteomes. Frontiers in microbiology. 2013;4:315.
41. Rajendrakumar CS, Suryanarayana T, Reddy AR. DNA helix destabilization by proline and betaine: possible role in the salinity tolerance process. FEBS letters. 1997;410(2-3):201-5.
42. 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.
43. Kurz M. Compatible solute influence on nucleic acids: many questions but few answers. Saline systems. 2008;4(1):1-14.
44. Burg MB, Ferraris JD, Dmitrieva NI. Cellular response to hyperosmotic stresses. Physiol Rev. 2007;87(4):1441-74.
45. 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.
46. 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.
47. 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.
48. Biesiada M, Hu Michael Y, Williams LD, Purzycka KJ, Petrov Anton S. rRNA expansion segment 7 in eukaryotes: from signature fold to tentacles. Nucleic Acids Res. 2022;50(18):10717-32.
49. Trahan C, Oeffinger M. The importance of being RNA-est: considering RNA-mediated ribosome plasticity. RNA Biol. 2023;20(1):177-85.
50. Hariharan N, Ghosh S, Palakodeti D. The story of rRNA expansion segments: Finding functionality amidst diversity. WIREs RNA. 2023;14(1):e1732.
51. Stepanov VG, Fox GE. Expansion segments in bacterial and archaeal 5S ribosomal RNAs. RNA. 2021;27(2):133-50.
52. Loukas A, Kappas I, Abatzopoulos TJ. HaloDom: a new database of halophiles across all life domains. J Biol Res-Thessalon. 2018;25(1):2.
53. 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.
54. 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.
55. 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.
56. Tanizawa Y, Fujisawa T, Nakamura Y. DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics. 2018;34(6):1037-9.