NAR Genomics and Bioinformatics, 2024, **6**, lqae088
https://doi.org/10.1093/nargab/lqae088
Advance access publication date: 27 July 2024
Standard Article

Depletion of CpG dinucleotides in bacterial genomes may represent an adaptation to high temperatures

Diego Forni*, Uberto Pozzoli, Alessandra Mozzi, Rachele Cagliani and Manuela Sironi

Scientific Institute IRCCS E. MEDEA, Bioinformatics, 23842 Bosisio Parini, Italy

*To whom correspondence should be addressed. Tel: +39 031877826; Fax: +39 031877499; Email: diego.forni@lanostrafamiglia.it

Abstract

Dinucleotide biases have been widely investigated in the genomes of eukaryotes and viruses, but not in bacteria. We assembled a dataset of bacterial genomes (>15 000), which are representative of the genetic diversity in the kingdom Eubacteria, and we analyzed dinucleotide biases in relation to different traits. We found that TpA dinucleotides are the most depleted and that CpG dinucleotides show the widest dispersion. The abundances of both dinucleotides vary with genomic G + C content and show a very strong phylogenetic signal. After accounting for G + C content and phylogenetic inertia, we analyzed different bacterial lifestyle traits. We found that temperature preferences associate with the abundance of CpG dinucleotides, with thermophiles/hyperthemophiles being particularly depleted. Conversely, the TpA dinucleotide displays a bias that only depends on genomic G + C composition. Using predictions of intrinsic cyclizability we also show that CpG depletion may associate with higher DNA bendability in both thermophiles/hyperthermophiles and mesophiles, and that the former are predicted to have significantly more flexible genomes than the latter. We suggest that higher bendability is advantageous at high temperatures because it facilitates DNA positive supercoiling and that, through modulation of DNA mechanical properties, local or global CpG depletion controls genome organization, most likely not only in bacteria.

Introduction

The genomes of living organisms, as well as those of viruses, have long been known to display important dinucleotide composition biases. Vertebrate genomes are characterized by a depletion of CpG dinucleotides, which is not present in invertebrates and is variably observed in plants and unicellular eukaryotes (1–9). Several viruses also show some degree of CpG depletion, that only partially depends on host associations (8–10). Conversely, TpA dinucleotides are under-represented across the tree of life and in most viruses, particularly in sequences expressed as RNA in the cytoplasm (2,5,7,11,12).

Different hypotheses have been proposed to explain such biases. The depletion of CpG dinucleotides in vertebrates is thought to be at least partially due to methylation and mutational loss of cytosines at CpG sites (1,2,7,13). In viral genomes, another possible pressure driving depletion of CpG dinucleotides is the elusion of cellular sensors that recognize non-self nucleic acids (14–16). In the case of TpA, one of the most supported explanations is the preferential cleavage of UpA dinucleotides by cytosolic RNAses (11).

Recently, we have shown that bacteriophages have a genome dinucleotide composition that tends to resemble their host's genomic CpG content (8). Whereas this seems unlikely to result from specific bacterial defense systems, the underlying reasons are unknown. Indeed, analysis of dinucleotide biases in bacterial genomes have attracted limited attention. A strong CpG under-representation was observed in *Mycoplasma penetrans,* which expresses a CpG-specific DNA methyltransferase (17). However, analyses of diverse bacterial genomes revealed no specific prevalence of CpG methylation by restriction-modification (R-M) systems (18,19). Moreover, such systems are often only transiently associated with their hosts, suggesting that they are unlikely to shape dinucleotide composition in bacterial genomes (20,21). Likewise, other bacterial defense mechanisms based on DNA methylation (e.g. BREX and DISARM) are not known to preferentially target CpG dinucleotides (22,23). Conversely, an early work based on a small number of genomes concluded that CpG dinucleotides in some bacteria are counter-selected because they cause constraint on structural properties of the DNA molecule (24).

Bacteria have circular genomes, which are organized into the nucleoid, a membrane-less compartment where DNA, RNA, and proteins interact to shape chromosome conformation (25). Supercoiling, which results from the torsional strain imposed by DNA-processing enzymes, is a key feature of the spatial architecture of bacterial chromosomes and plays an important role in the regulation of gene expression (26). Supercoiling can be either negative or positive, depending on the direction of the twists with respect to that of the DNA double helix, with both types forming coils and plectonemes (DNA loops in which the double-stranded DNA is wrapped around itself) (25). Different enzymes control DNA topology and supercoiling. Whereas topoisomerases, which dissipate excessive supercoils, exists in all domains of life, bacteria and archaea also encode enzymes that introduce negative (gyrase) and positive (reverse gyrase, RG) supercoiling (27). In particular, RG is restricted to bacteria living at high temperatures (higher

Received: March 18, 2024. **Revised:** June 17, 2024. **Editorial Decision:** July 11, 2024. **Accepted:** July 18, 2024

than 70°C) and is thought to represent an adaptation to avoid DNA damage. By introducing positive supercoling, RG was suggested to limit DNA breathing and to protect DNA from denaturation (28). Recently, the DNA sequence was shown to determine the position of supercoils and DNA intrinsic curvature (29). Also, different dinucleotides were shown to exert different effects on DNA bendability (30).

Herein, we used the full set of reference bacterial genomes to provide a comprehensive view of dinucleotide biases and to test whether bacterial lifestyle traits or DNA mechanical properties influence genome dinucleotide composition.

Materials and methods

Bacterial genomes

The list of bacterial genomes was derived from the BV-BRC site (https://www.bv-brc.org/) by selecting entries with 'good' genome quality and corresponding to representative or reference strains. Only one genome for each species was retained (randomly selected). Genome sequences were downloaded using the ncbi-genome-download python tool (31), querying both refSeq and GenBank databases (Supplementary Table S1).

The presence of reverse gyrase in the genome of thermophilic/hyperthermophilic bacteria was checked by running HMMER v3.1b2 (32). In particular, we retrieved the Hidden markov model of the reverse gyrase family from the NCBI database (TIGR01054), which was used as a query profile against all proteins from the thermophilic/hyperthermophilic bacteria in our dataset using the hmmsearch tool. We considered only hits with an *E*-value $<10^{-100}$; we then aligned these hits and we kept protein sequences that encode both the helicase and the topoisomerase domain. This generated a list of 63 reverse gyrase proteins (Supplementary Table S1). All these proteins were already annotated as reverse gyrase in their corresponding genomes.

Dinucleotide observed/expected ratio

To investigate dinucleotide biases, we calculated the observed/expected ratio (O/E ratio) for all dinucleotides. Specifically, the frequency of each dinucleotide in each genome (i.e. the observed frequency) was divided by the product of the frequencies of the contributing nucleotides (i.e. the expected frequency). For instance, for CpG, we calculated the number of CpG along the genome divided by the number of all possible dinucleotides; this frequency was then divided by the product of C and G frequencies. Dinucleotide composition was calculated using the compseq tool (https://www.bioinformatics.nl/cgi-bin/emboss/compseq), by setting the size of word equal to 2 and using the 'calcfreq' parameter, so that the dinucleotide expected frequencies are calculated from the observed frequency of single bases. Dinucleotides were also counted in the reverse complement of the sequence using the 'reverse' parameter.

Bacterial traits

Information on genome length and number of coding sequences (CDS) was derived from BV-BRC annotations. Most other traits were derived from the Bacterial Diversity Metadatabase (BacDive, https://bacdive.dsmz.de/, (33)), which retrieves information from culture collections and primary literature. Specifically, for oxygen tolerance we used BacDive information to create three classes: aerobes ($n = 5276$, which include 'aerobes', 'obligate aerobes', 'facultative anaerobes'), anaerobes ($n = 1468$, which include 'anaerobes' and 'obligate anaerobes'), and microaerophiles (n = 762, which include 'microaerophiles'). For pH preference, bacteria were divided in acidophile ($n = 171$) and alkaliphile ($n = 2529$), as per BacDive classification. Data about motility (yes, $n = 2652$; no, $n = 3958$) and ability of spore formation (yes, $n = 1048$; no, $n = 2276$) were also obtained from BacDive. Data on growing temperatures were instead derived from the GSHC (Genome Sequences: Hot, Cold, and everything in between) database (34) (http://melnikovlab.com/gshc/). The GSHC database periodically retrieves information from public repositories of microorganisms, without manual curation. As a consequence it provides very extensive data with possible minor inaccuracies. As in BacDive, we classified bacteria as psychrophilic (<25°C, $n = 532$), mesophilic (25–39°C, $n = 9104$), thermophilic (40–79°C, $n = 472$) and hyperthermophilic (>80°C, $n = 5$). Because of their small number, hyperthermophilic bacteria were grouped with thermophilic ones. Finally, data on cell shape, cell length and width were derived from BacDive. For rod-shaped bacteria ($n = 3137$), cell volume was calculated using the formula suggested in (35): $V = \pi(W/2)^2(L - W) + 4/3\ \pi(W/2)^3$; where W is the average cell width and L is cell length.

For cell volume, genome length, and number of coding sequences, we divided bacteria in classes based on value distributions and quartiles. Thus, for each trait, we generated four classes: values lower than the first quartile, values between the first and the second quartiles, values between the second and the third quartiles and finally, values higher than the third quartile.

Details regarding all traits are reported in Supplementary Table S1.

Model fitting and ancestral state reconstruction

We modeled the relationship between CpG or TpA O/E ratios and G + C content with cubic smoothing splines using the smooth.spline function in the stats R package (36). The residuals were obtained from the models.

Ancestral state reconstruction was performed by maximum likelihood using the FastAnc function in the phytools R package (37,38), and pyhlogenetic trees were plotted using the contMap and setMap functions.

Phylogenetic tree and ANOVA

The phylogenetic tree was downloaded from the Genome Taxonomy database (GTDB, https://gtdb.ecogenomic.org/, release 2023-04-28, (39)). The tree was built using the sequences of 120 conserved genes. It was pruned to only keep tips corresponding to genomes in our dataset (n = 12049, Supplementary Table S1) using the ape R package (40). Pagel's λ (41)was calculated using the function in the phytools R package.

The pylogenetic ANOVA and post-hoc tests were performed using the phylANOVA function in phytools (37,38,42,43), p-value were calculated with 1000 simulations, and *P*-values for the post-hoc tests were corrected using Holm's method, as suggested.

Intrinsic cyclizability

DNA intrinsic cyclizablity score was predicted by using the DNAcycP python package (44). This package is based on a deep-learning approach from loop-seq data; it takes as input DNA sequences and, using a 50 bp windows approach, it estimates normalized *C* scores per window; for each genome we reported the average value of all windows, as suggested.

Validation dataset

A second bacterial genome dataset was built by retrieving growing temperature information from three different databases. In particular, we used data from the GSHC and BacDive databases, and we also exploited information from the TEMPURA (Database of growth TEMPeratures of Usual and RAre prokaryotes, http://togodb.org/db/tempura) database. We selected all available bacterial strains, we then removed genomes already present in the initial dataset, selected only one genome per species, and kept mesophilic, thermophilic, and hyperthermophilic bacteria. This generated a list of 1732 bacterial genomes (Supplementary Table S2). Genome sequences were downloaded using the ncbi-genome-download tool. To have a sufficiently large and well represented dataset, an equal number of mesophilic and thermophilic/hyperthermophilic bacterial genomes were randomly selected from the starting dataset.

This set of 3464 bacterial genomes was used as the input for the Genome Taxonomy Database Toolkit (GTDB-Tk) (45). GTDB-Tk identified a set of 120 core marker proteins in each genome, which were concatenated and aligned using the MAFFT software (46). The resulting alignment was used as input for the FastTree tool (47) to generate an approximately-maximum-likelihood phylogenetic tree that was used in the pylogenetic ANOVA analyses. Dinucleotide observed/expected ratio, residuals, intrinsic cyclizability, and the presence/absence of the reverse gyrase were obtained as described above.

Results

Dinucleotide biases in bacterial genomes

We assembled a dataset of 15 304 bacterial genomes from the Bacterial and Viral Bioinformatics Resource Center (BV-BRC) (Supplementary Table S1, see Materials and methods). To investigate composition biases, we calculated the observed/expected (O/E) ratio for all dinucleotides, where the expected dinucleotide frequency in a genome is the product of the frequencies of the contributing nucleotides. Ratios lower than 0.78 and higher than 1.23 are generally considered to indicate significant depletion and enrichment (13,48). Across our bacterial genome dataset, TpA dinucleotides were the most depleted, with a median well below 0.78 and the overwhelming majority of genomes having ratios lower than 1. Conversely, CpG dinucleotides were not generally underrepresented, but showed the widest dispersion (Figure 1A).

Previous analyses of genomes from cellular organisms and viruses detected a positive correlation between the O/E ratio of CpG (O/E CpG) and G + C content. A negative correlation with G + C content was instead detected for O/E TpA (7–9,12,49,50). In the bacterial genome dataset, we also detected a strong and positive correlation between O/E CpG and G + C content and a negative one for O/E TpA (Figure 1B). This implies that CpG and TpA dinucleotide abundance cannot be interpreted without taking G + C content into account.

To further investigate how these dinucleotide biases relate to other genomic or lifestyle features, we aimed to devise a measure of dinucleotide representation which accounts for genomic G + C content, as we have previously developed to study viral genomes (8). To this purpose, because the relationships between O/E CpG or O/E TpA and G + C content are not linear, we resolved to cubic smoothing splines to fit the data and to calculate the residuals of the models (Figure\ 1B). Such residuals, hereafter referred to as resCpG and resTpA, indicate how much a given genome deviates in terms of dinucleotide composition from the expected based on G + C content. Residuals can have positive or negative sign, depending whether the dinucleotide is more or less abundant than expected. Because O/E CpG values are more dispersed around the spline than O/E TpA, the distribution of resCpG is wider than that of resTpA (Figure 1C).

Temperature preferences drive CpG dinucleotide biases

To investigate how dinucleotide biases distribute across the bacterial phylogeny, we downloaded a consensus tree of 12271 bacterial genomes from the Genome Taxonomy database. The tree was then pruned to retain only genomes present in our dataset (n = 12049). Using this tree, we reconstructed the maximum likelihood ancestral state of resCpG and resTpA. As expected, clear associations between resCpG and phylogenetic relationships were evident (Figure 2A). This was also the case for resTpA, although the effect was much less evident due to the narrower distribution of resTpA compared to resCpG (Figure 2A). We thus determined whether a phylogenetic signal was detectable in the distribution of resCpG and resTpA by calculating Pagel's λ (41). Estimates of λ resulted equal to 1 for both residuals, indicating a very strong phylogenetic signal.

It follows from these results that exploration of factors driving CpG and TpA abundance requires taking phylogeny into account. We thus applied phylogenetic ANOVA analyses (42) to test whether resCpG and resTpA differ among bacteria having different genomic or lifestyle features. Concerning the latter, we exploited Bac*Dive* (33) to retrieve information about oxygen tolerance, pH preference, motility, ability to form spores, cell shape, cell width, and cell length. These features were selected because they were available for a substantial number of entries corresponding to bacterial genomes in our dataset. In the case of cell shape, because the overwhelming majority of bacteria were rod-shaped ($n = 2672$), we used average cell width and length to calculate cell volume (as described in (35)). Data on temperature preferences were instead obtained from the GSHC database (34). Concerning genomic features, we considered genome length and number of coding sequences, as derived from BV-BRC.

For resTpA, the phylogenetic ANOVA analyses did not detect significance differences among groups for any feature (Table 1). For resCpG, instead, a highly significant association with temperature preferences was detected (Table 1). After post-hoc tests, it was evident that significant pairwise comparisons involved thermophilic/hyperthermophilic bacteria, which had significantly lower resCpG (median well below 0) than mesophiles and psychrophiles (Figure 2B). Indeed, map-

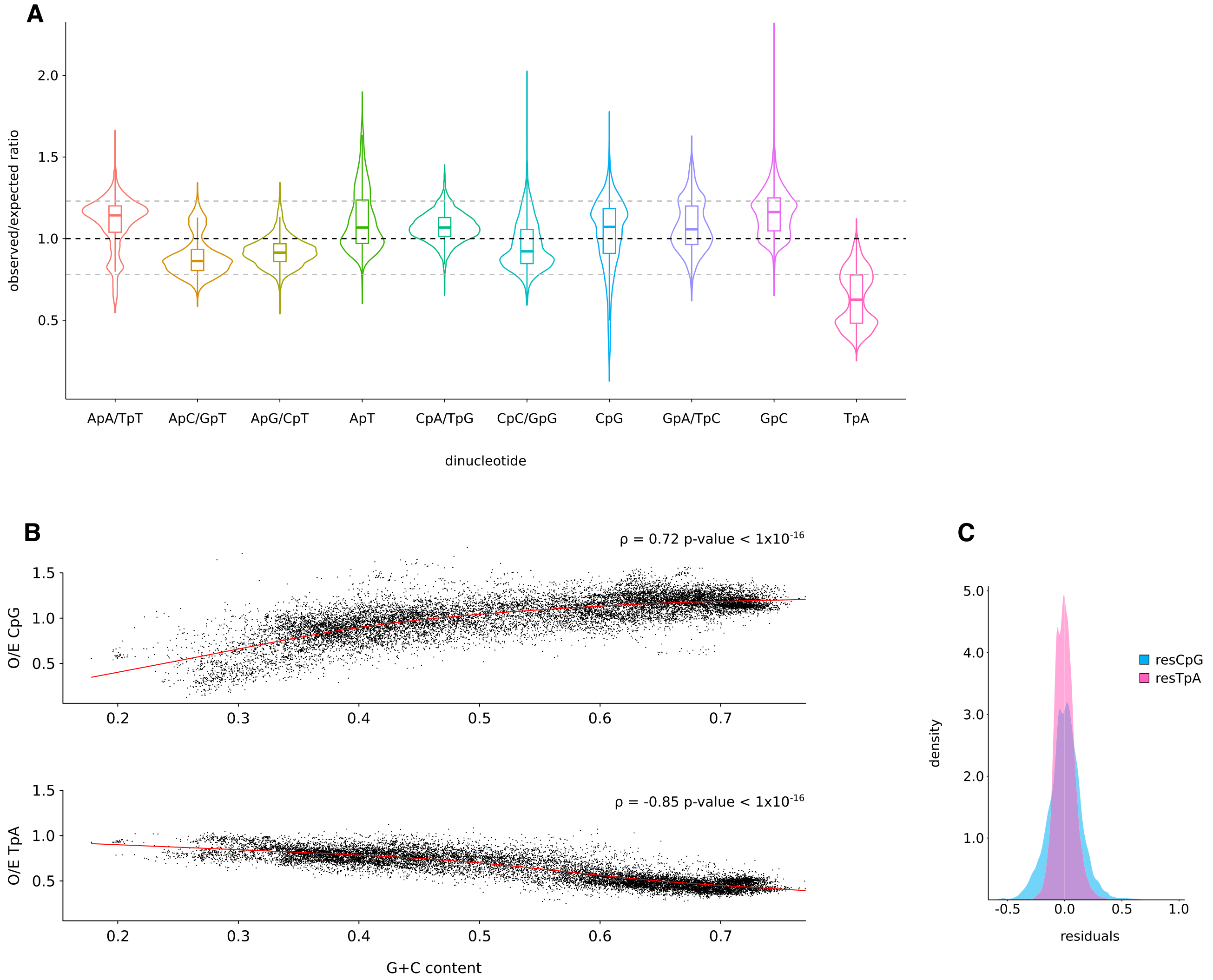

Figure 1. Dinucleotide representation in bacterial genomes. (**A**) Violin plots with boxplots of the observed/expected ratio (O/E) for all dinucleotides. The gray horizontal hatched lines correspond to ratios of 0.78 and 1.23, which are generally considered as significant thresholds for dinucleotide depletion and enrichment, respectively (13,48). (**B**) Observed/expected ratio for CpG (upper panel) and TpA (lower panel) as a function of G + C content. Cubic smoothing splines are shown as solid red lines and the Pearson's correlation coefficient ρ is also reported along with their corresponding *P*-values. (**C**) Distribution of residuals for the O/E CpG and TpA models.

ping of the genomes of thermophilic/hyperthermophilic bacteria on the phylogenetic tree indicated that they populate several clades characterized by very low resCpG. Clearly, though, this is not an all-or-nothing association as thermophiles with non CpG depleted genomes exist and very low resCpG was also observed from a number of bacteria that do not live (or are not known to live) at high temperatures (Figure 2A).

As a comparison, the phylogenetic ANOVA analysis was repeated using O/E CpG instead of resCpG. A significant result was obtained (F = 109.46, P = 0.012), but pairwise comparisons indicated that thermophiles have significantly lower O/E CpG than mesophiles but not than psycrophiles (Supplementary Figure S1). Inspection of genomic G + C content distributions indicated that psychrophilic bacteria have the lowest average content, suggesting that nucleotide composition influences O/E CpG values in this group (Supplementary Figure S1).

Overall, these results indicate that, after accounting for genomic G + C content, temperature preferences influence the abundance of CpG dinucleotides in a subset of bacterial genomes. Conversely, the TpA dinucleotide seems to display a bias that only depends on genomic G + C composition and phylogenetic relationships.

CpG representation affects DNA mechanical properties of bacterial genomes

Several studies have indicated that DNA sequence is a major determinant of DNA flexibility along its central axis (30,51–56). A recent study that measured intrinsic cyclizability (a parameter related to cyclization propensity or bendability) indicated that the CpG dinucleotide has the most negative bendability quotient (30). Interestingly, the genome-wide average cyclizability varies among species and analysis of thermophilic archaea showed that their genomes may be more flexible than expected based on base composition (44,57). We thus hypothesized that CpG depletion in thermophilic and hyperthermophilic bacteria may serve the purpose of increasing the flexibility of their genomic DNA. To test this hypothesis, we used the DNAcycP method (44),

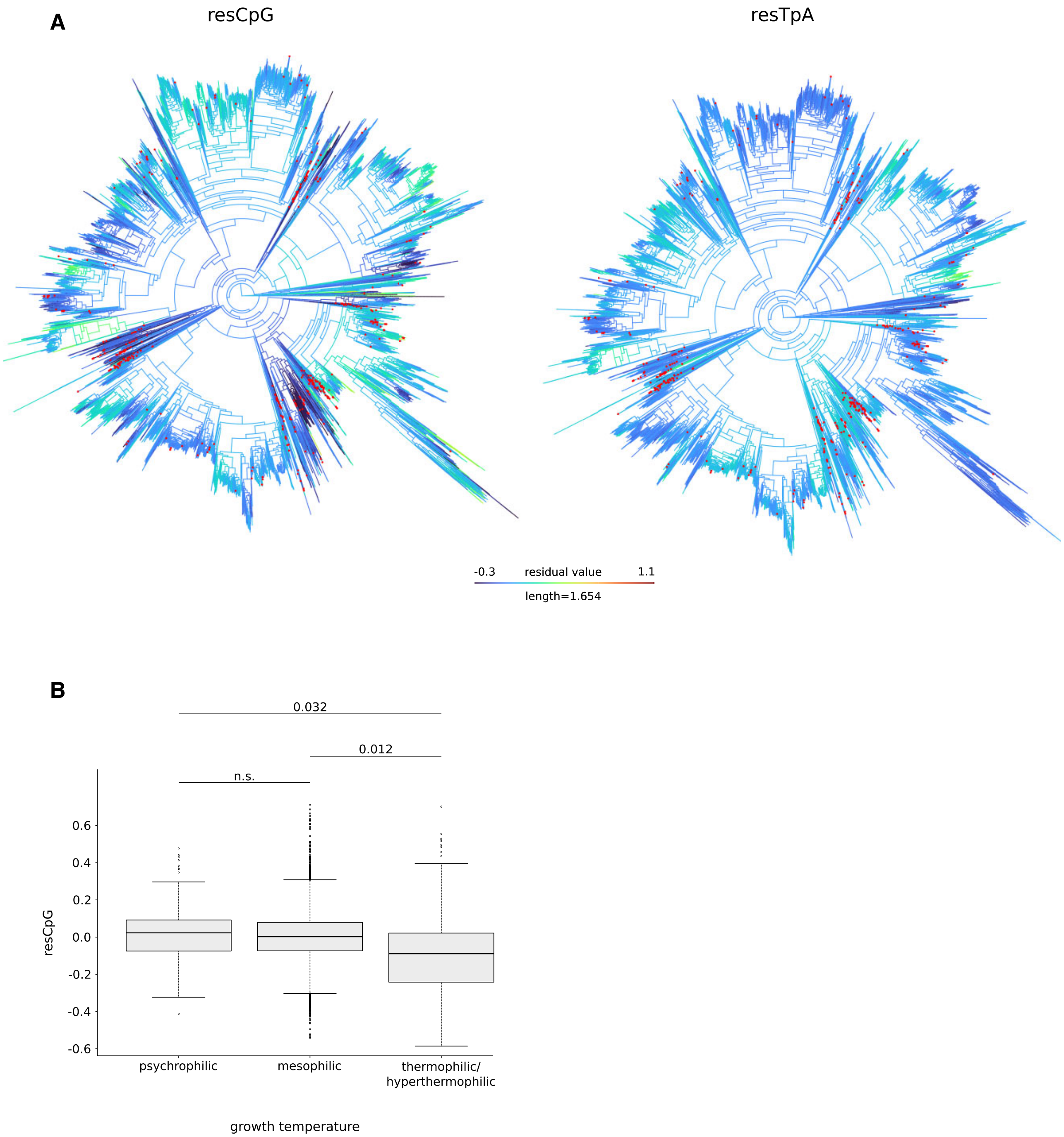

Figure 2. Dinucleotide ancestral state in bacteria. (**A**) Bacterial phylogenetic tree retrieved from the Genome Taxonomy database (see Materials and methods for details) with branches colored by ancestral state reconstruction of resCpG (left) and resTpA (right) values. Red points at tips indicate thermophilic/hyperthermophilic bacteria. (**B**) Boxplot of resCpG values grouped by growing temperature. PhylANOVA pairwise post-hoc tests after Holm's correction are reported above each comparison. Given that a only subset of bacteria from our dataset is present in the Genome Taxonomy database tree, phylogenetic ANOVA analysis was performed on 7228 mesophilic, 451 psychrophilic, and 406 thermophilic/hyperthermophilic genomes.

which is based on a deep learning model trained on DNA flexibility measurements, to calculate the average intrinsic cyclizability of 812 bacterial genomes in our dataset (406 thermophilic/hyperthermophilic and the same number of randomly selected mesophilic genome), which were also included in the genome tree. In line with previous findings, results indicated that thermophilic/hyperthermophilic bacteria have genomes characterized by significantly higher intrinsic cyclizability compared to mesophiles (phylogenetic ANOVA, $F = 55.85$, $P = 0.005$) (Figure 3A). Notably, when we analyzed intrinsic cyclizability and resCpG, negative and significant correlations were detected for both thermophiles/hyperthermophiles and mesophiles (Figure 3B). Overall, these data suggest that CpG dinucleotide biases affect DNA flexibility and that depletion of CpG represents an adaptation to high temperatures.

Table 1. Phylogenetic ANOVA results

	Temperature class	Oxygen tolerance	pH range	Spore formation	Motility	Cell volume[a]	Genome size[a]	Number of CDS[a]
Sample size	Psycrophilic: 451 Mesophilic: 7228 Thermo/ hyper-thermophylic: 477	Aereobe: 4167 Anaerobe: 1096 Microaerophile: 477	Acidophile: 142 Alkaliphile: 2196	Yes: 855 No: 1860	Yes: 2022 No: 2981	<Q1: 665 >Q1 < Q2: 660 >Q2 < Q3: 678 >Q3: 669	<Q1: 3012 >Q1 < Q2: 3012 >Q2 < Q3: 3012 >Q3: 3013	<Q1: 3011 >Q1 < Q2: 3010 >Q2 < Q3: 3013 >Q3: 3015
Phylogenetic ANOVA resCpG	***F* value = 106.59** **Pr(>*F*) = 0.009**	*F* value = 224.38 Pr(>F) = 0.06	*F* value = 10.23 Pr(>F) = 0.20	*F* value = 0.98 Pr(>F) = 0.94	*F* value = 13.54 Pr(>F) = 0.75	*F* value = 3.90 Pr(>F) = 0.59	*F* value = 2.83 Pr(>F) = 0.99	*F* value = 4.43 Pr(>F) = 0.98
Phylogenetic ANOVA resTpA	*F* value = 13.16 Pr(>*F*) = 0.63	*F* value = 31.96 Pr(>*F*) = 0.65	*F* value = 0.10 Pr(>*F*) = 0.91	*F* value = 81.36 Pr(>*F*) = 0.48	*F* value = 5.37 Pr(>*F*) = 0.83	*F* value = 6.54 Pr(>*F*) = 0.47	*F* value = 90.82 Pr(>*F*) = 0.43	*F* value = 52.68 Pr(>*F*) = 0.64

[a] In the case of cell volume, genome size, and number of coding sequences (CDS), we categorized data based on quartile distributions: <Q1, lower than the first quartile; >Q1 < Q2, between the first and second quartiles; >Q2 < Q3, between the second and third quartiles; >Q3, higher than the third quartile.

The presence of reverse gyrase may drive the CpG dinucleotide bias in thermophiles

In the analyses above, we classified thermophilic bacteria as in BacDive (i.e. bacteria with growth temperature higher than 39°C). Thus, the range of temperatures is very wide for species classified as thermophiles in our dataset (40–80°C). It is well known that virtually all hyperthermophilic bacteria, as well as several thermophiles, encode RG (58–61). The positive supercoiling activity of RG requires a temperature of 70°C or higher (62). Because DNA intrinsic curvature is a key structuring factor for positive supercoiling and plectoneme stability (29), we reasoned that CpG under-representation and high intrinsic cyclizability might relate to the presence of RG.

We thus searched (see Materials and methods) for the presence of an RG gene in the genomes of all thermophilic/hyperthermophilic bacteria in our dataset that were also included in the phylogenetic tree. We found that, out of 51 bacteria with temperature preferences equal to or higher than 70°C, 38 encode the RG. Conversely, only 16 genomes out of 415 were predicted to encode RG when bacteria with temperature preferences lower than 70°C were analyzed. Among these, 11 had growing temperatures higher than 65, whereas the others were moderate thermophiles (Supplementary Table S1). This was noted before and the presence of RG in these species is thought to represent an adaptation to short-term exposure to elevated temperatures (63).

We next used phylogenetic ANOVA analyses to investigate resCpG and intrinsic cyclizability in bacteria that encode or do not encode the RG. In line with our hypothesis, we found that the presence of RG associates with significantly lower resCpG and significantly higher intrinsic cyclizability (resCpG phylogenetic ANOVA, $F = 47.46$, $P = 0.014$; cyclizability phylogenetic ANOVA, F = 36.33, $P = 0.027$) (Figure 3C and D). It should however be noted that RG-encoding bacteria live at much higher temperatures than RG-lacking ones (Supplementary Figure S2). Thus, these results are not unexpected and we cannot exclude that the effects on resCpG and intrinsic cyclizability are driven by the temperature rather than by the presence of RG.

The associations between CpG content, temperature and intrinsic cyclizability are not dependent on genome choice

To validate the associations detected above using a partially independent bacterial genome dataset, we exploited three databases storing information on growth temperatures to obtain a list of 1732 genomes of mesophilic and thermophilic/hyperthermophilic bacteria that were not included in the previous analyses (see Materials and methods). To increase the statistical power, an equal number of mesophilic and thermophilic/hyperthermophilic bacterial genomes in the initial dataset were randomly selected to generate a partially independent set of 3464 genomes. From these genomes, 120 core genes were extracted and aligned to generate a phylogenetic tree. In line with the results reported above, phylogenetic ANOVA analysis indicated that thermophiles/hyperthermophiles have significantly lower resCpG than mesophiles/psychrophiles ($F = 49.34$, $P = 0.041$) (Supplementary Figure S3A).

We next selected all genomes of thermophiles/hyperthermophiles ($n = 144$) and of an equal number of mesophiles to calculate intrinsic cyclizability. As observed in the larger dataset, results indicated that thermophilic/hyperthermophilic bacterial genomes have significantly higher intrinsic cyclizability compared to mesophiles (phylogenetic ANOVA, $F = 29.73$, $P = 0.05$) (Supplementary Figure S3B). For both thermophiles/hyperthermophiles and mesophiles, negative significant correlations between intrinsic cyclizability and resCpG were observed (Supplementary Figure S3C). Finally, evidence of an RG gene was detected in 23 genomes, most of them from bacteria living at temperatures equal or higher than 70°C (Supplementary Table S2, Supplementary Figure S3D). RG-encoding bacteria were found to have more CpG depleted genomes than thermophilic/hyperthermophilic bacteria that do not encode RG (phylogenetic ANOVA, $F = 20.04$, $P = 0.039$). However, as above, the effect of RG presence is impossible to disentangle from that of temperature (Supplementary Figure S3E).

Discussion

The presence of dinucleotide biases, especially the depletion of CpG and TpA dinucleotides, in the genomes of cellular organisms and viruses has been known for years and has spawned interest in understanding its underlying causes (and consequences). Cytosine methylation at CpG sites certainly has a role in determining CpG loss in vertebrate genomes and, most likely, in some vertebrate-infecting dsDNA viruses (1,2,7,13). However, mechanisms other than methylation were shown to contribute to CpG depletion in mammalian genomes and in viruses that infect vertebrates (7–9). Also, some RNA viruses are extremely CpG depleted irrespective of their host range (i.e. whether they infect hosts that encode or do not encode DNA methyltransferases) (9). This latter observation indicates that avoidance of host immune sensing is not a universal

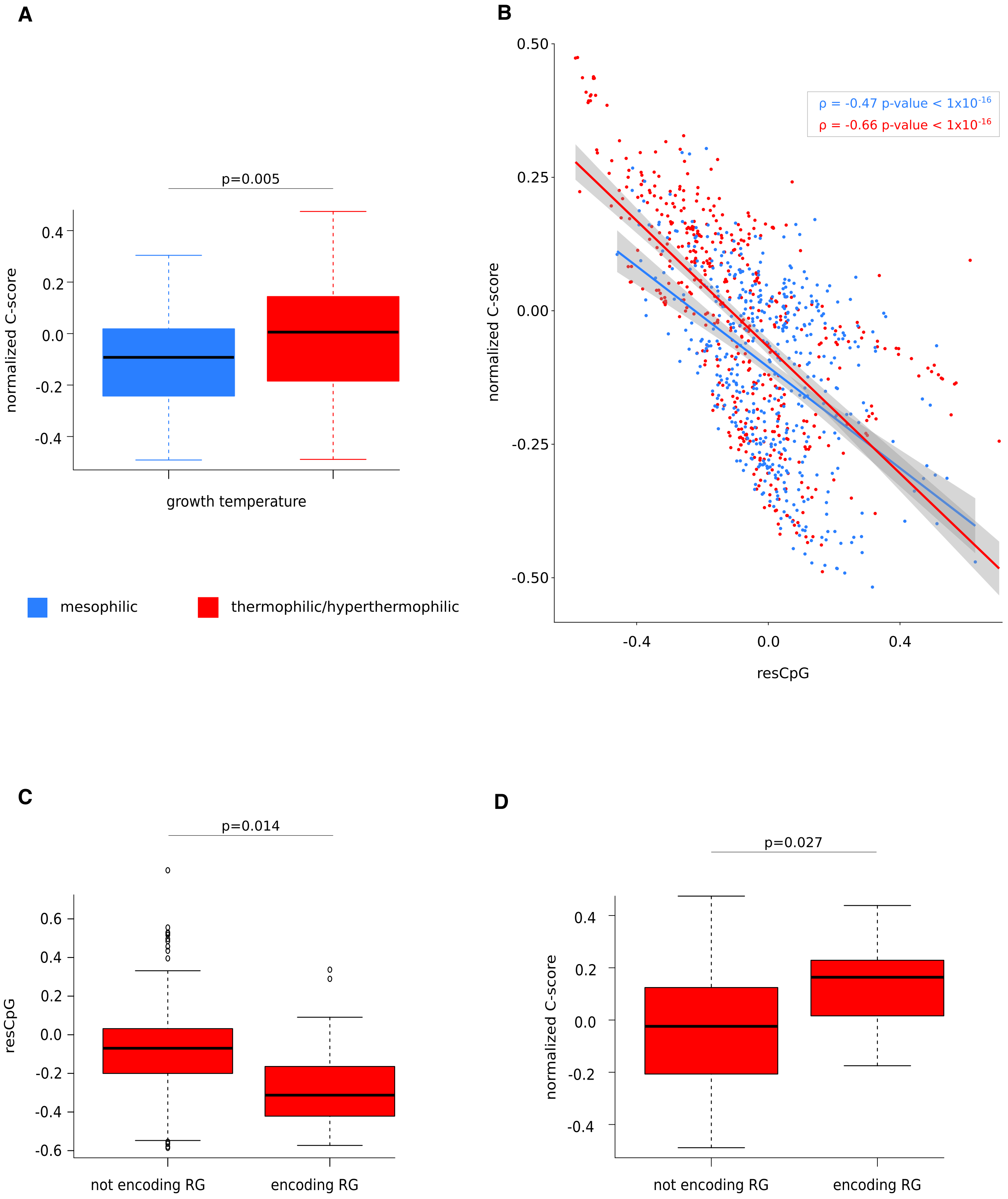

Figure 3. CpG representation and DNA bendability of thermophilic/hyperthermophilic genomes. (**A**) Box plot representation of normalized C-score grouped by growing temperature. Both temperature classes are composed of 406 genomes. Statistical difference among the two groups has been assessed using phylogenetic ANOVA analysis. (**B**) Linear models of normalized C-score as a function of O/E CpG residuals. Dots are colored based on the temperature classes and Pearson's correlation coefficient ρ is also reported. Regression lines are shown with confidence intervals. (**C**) Box plot representation of O/E CpG residuals grouped by thermophilic/hyperthermophilic bacteria that encode or do not encode the reverse gyrase (RG) enzyme. Statistical difference between the two groups was assessed using phylogenetic ANOVA analysis. (**D**) Box plot representation of normalized *C*-score grouped by thermophilic/hyperthermophilic bacteria that encode or do not encode the RG enzyme. Statistical difference between the two groups was assessed using phylogenetic ANOVA analysis.

explanation for dinucleotide biases in viral genomes, either. In fact, the zinc-finger antiviral protein (ZAP) and toll-like receptor 9 (TLR9), which sense CpG-containing non-self nucleic acids, are only encoded by vertebrates (which also posses CpG methylation systems) (14–16). Overall, these observations suggest that some unknown factors contribute to CpG abundance in the genomes of cellular organisms and viruses. Even more mysterious is TpA depletion, which seems to be almost universal and probably related to RNA or DNA stability (11).

Compared to viruses, the analysis of CpG and TpA dinucleotides in bacterial genomes has attracted limited attention and only a few genomes have been investigated in this respect (17,18,24,64,65). We thus assembled a large dataset of bacterial genomes, which are representative of the genetic diversity in the kingdom Eubacteria, with the aim of analyzing dinucleotide biases and their underlying drivers. We found that the representation of both CpG and TpA dinucleotides is strongly dependent on genomic G + C content. This seems to be a universal feature of biological sequences, as it is observed across all domains of life and in viruses (with either RNA or DNA genomes) (7–9,49,66). The reasons for the existence of such relationships are presently unclear and cytosine methylation cannot be the only factor involved because invertebrates have poorly methylated genomes and methylation does not occur in RNA viruses. In bacteria, DNA methylation is extremely common and is primarily associated with R-M systems, but CpG and TpA dinucleotides do not represent preferential targets (19). Moreover, it is estimated that, because of the high rate of horizontal transfer of defense genes, only 4% of R-M systems are found in the core genomes of prokaryotic species (20,21). As a consequence R-M systems are unlikely to shape the genomic representation of dinucleotides or to explain their relationship with overall G + C content.

Irrespective of the underlying causes, the relationships between G + C content and TpA or CpG representation implies that dinucleotide biases must be parsed together with genomic base composition. We thus used the residuals of the fitted models to obtain a measure that can be analyzed against different traits. Such residuals are not absolute measures of dinucleotide abundance (e.g. resTpA equal to 0 does not indicate that the representation of TpA dinucleotides corresponds to the expected based on nucleotide frequency). Conversely, resCpG and resTpA depend on the G + C content and on the distributions of values that are fitted by the models. As expected, we obtained a strong phylogenetic signal for both the residuals, meaning that closely related bacterial species have similar dinucleotide representation. We thus used phylogenetic ANOVA analyses to investigate whether genomic or lifestyle features are related to dinucleotide biases.

In addition to having a wide distribution of genomic G + C content, bacteria are extremely diverse in terms of genome size and coding capacity, but also with respect to lifestyle and environmental distribution. Indeed, several species of bacteria (and archaea) are well known for their ability to thrive in extreme environmental conditions, including very high and very low temperatures, hypersaline ecosystems, and acidic habitats (67). We reasoned that investigation of the relationship between dinucleotide biases and different genomic or lifestyle features might provide insight into the role of such biases for bacterial adaptation and explain the underlying drivers. The choice of traits to analyze was conditioned on their availability for a substantial number of strains. This clearly represents a limitation of this study, as a number of traits were available for a limited number of species (e.g. nutrition type, optimum salt concentration), or were too multidimensional (e.g. metabolite utilization) or both (e.g. fatty acid profile). Moreover, different traits were recorded for different numbers of species and we cannot exclude that we failed to detect associations with some traits because of a lack of statistical power. Finally, and most importantly, the biological complexity of these traits is unrecognized in our categorization. For instance, many bacterial species can live at different temperatures and there is no single designation of ‘preferred temperature’ (it may be the temperature at which the bacterium grows the fastest or the one at which it is most often sampled). Similar considerations apply to pH preferences, as several bacteria can live within a relatively wide pH range (68). Likewise, cell volume and shape can vary widely depending on growth conditions and media, and display ample variability for different strains in the same species (69–72). Moreover, our classification is necessarily limited to present knowledge, whereas new bacterial lifestyle traits are constantly described. Whereas future studies will be required to extend analysis to a wider range of traits and species, we were able to identify a correlation between temperature preferences and CpG depletion. Conversely, we did not detect any association between TpA dinucleotide abundance and bacterial traits. One possible reason is that the relatively narrow distribution of resTpA affects statistical power. Alternatively, the narrow distribution itself might indicated that G + C content is the major driver of TpA aboundance in bacterial genomes, irrespective of other features. Clearly, it is also possible that bacterial traits other than the ones analyzed here associate with the TpA dinucleotide bias.

Early investigations based on a handful of prokaryotic genomes found CpG under-representation in thermophilic bacteria and archaea, and proposed that this effect was related to some advantage in terms of DNA structure, such as supercoiling or chromatin packaging (64,65). Here, we used a recently developed tool, DNAcycP, to demonstrate that indeed CpG depletion is correlated with increased DNA bendability, which in turn may associate with positive supercoiling (57). DNAcycP was trained on data obtained with the loop-seq method, which uses a sequencing-based approach to allow a high-throughput measurement of DNA mechanical properties (57,73). This underscores another limitation of our approach, as loop-seq data refer to relatively short, linear sequences of relaxed DNA with cyclizability properties measured at room temperature. Conversely, DNA in living bacteria is not relaxed, as negative and positive supercoilings are pervasive (74,75). It is thus unsure whether the predictions are fully applicable to bacterial genomes in the living context. However, loop-seq data and DNAcycP predictions were shown to describe relevant features of DNA organization in eukaryotic cells, such as transcription factor binding sites and nucleosome positioning, suggesting that they at least partially reflect cyclizability as it occurs *in vivo* (44,73). With respect to temperature, its effects on the mechanical properties of DNA were only partially explored (76–78). At present, it is difficult to determine to which degree loop-seq data and DNAcycP predictions can be influenced by temperature. Most likely, experimental data will be required to address this point.

These caveats notwithstanding, our data are consistent with the observation that dinucleotide frequencies and their respective pairwise distances play a major role in determining DNA mechanical properties. Indeed, using loop-seq data, Basu and coworkers showed that, while the overall G + C content has little effect on cyclizability, the CpG dinucleotide has the most negative bendability quotient (30). Our data show that thermophiles, with generally CpG depleted genomes, may have more flexible genomes than mesophiles, at least as assessed through DNAcycP. Notably, though, we found that resCpG and cyclizability are negatively correlated in both thermophiles and mesophiles, suggesting that the results are not secondary to an unknown effect exerted by temperature preferences. Interestingly, both DNAcycP and another DNA cyclizability prediction tool (CycPred) indicated that two thermophilic archaea have very flexible genomes (44,57), suggesting that DNA flexibility is an adaption of prokaryotes living at elevated temperatures.

The reason why higher DNA flexibility might be favorable for life at high temperatures is not fully clear. Virtually all hyperthermophilic bacteria and archaea encode RG, which is able to induce positive DNA supercoiling (58–61,79). The enzyme is necessary for growth at elevated temperatures (80,81) and positively supercoiled DNA molecules show high stability at temperatures as high as 90°C (82). Recently, Kim and coworkers showed that, in a positive supercoiling regime, DNA curvature determines the formation of plectonemes (29). It is thus possible that increased bendability facilitates the formation of positively supercoiled structures by RG. We should also add that RGs and gyrases have different properties and mechanisms of action (61). Based on structural considerations, it was proposed that the introduction of positive supercoiling by RG requires overall bending of the DNA (62), suggesting that enzymatic activity is facilitated by flexible genomes. These observations are in line with our data showing that thermophiles that encode RG have more CpG-depleted and may have more flexible genomes than those that do not. However, because RG has enzymatic activity at 70°C or higher, it is typical of hyperthermophiles or thermophiles living at very high temperatures. As a consequence, our analyses cannot disentangle the effect of RG presence from that of temperature. It remains however true that, whatever the underlying factor, bacterial adaptation to very high temperatures entails CpG depleted genomes, possibly characterized by high bendability.

It should be noted that our data show that some non-thermophilic bacteria also have CpG-depleted genomes. Likewise, positive DNA supercoiling is not a prerogative of bacteria living at high temperatures. Indeed, a recent analysis of Escherichia coli, a mesophile, indicated that positive and negative supercoiling contributes to the organization of the bacterial genome (74). The authors suggested that positive supercoiling has a role in genome packaging and in buffering the effects of negative supercoiling. It is thus possible that local or global CpG depletion affects the supercoiling and organization of bacterial genomes through modulation of DNA mechanical properties, even in bacteria that do not live at high temperatures. This hypothesis may also be translatable to eukaryotic organisms and viruses, and it may help explain CpG deficiency in genomes that are not subject to cytosine methylation. In this respect, it is also worth mentioning that experimental data showed that CpG methylation stiffens DNA in all sequence contexts (30). Thus, CpG dinucleotides might be particularly counter-selected in organisms that methylate them, as observed in vertebrates. This might explain the existence of a selective pressure in mammalian genomes against CpG, but not directly resulting from cytosine mutation (7). Experimental approaches will be necessary to test these hypotheses.

Data availability

The list of bacterial species and their associated traits analyzed in this study is available in Supplementary Tables S1 and S2.

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

Italian Ministry of Health (‘Ricerca Corrente’ to M.S.); A.P.C. funded by Bibliosan.

Conflict of interest statement

None declared.

References

1. Bird,A.P. and Taggart,M.H. (1980) Variable patterns of total DNA and rDNA methylation in animals. *Nucleic Acids Res.*, **8**, 1485–1497.
2. Burge,C., Campbell,A.M. and Karlin,S. (1992) Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Nat. Acad. Sci. USA*, **89**, 1358–1362.
3. Gentles,A.J. and Karlin,S. (2001) Genome-scale compositional comparisons in eukaryotes. *Genome Res.*, **11**, 540–546.
4. Gonçalves-Carneiro,D., Takata,M.A., Ong,H., Shilton,A. and Bieniasz,P.D. (2021) Origin and evolution of the zinc finger antiviral protein. *PLoS Pathog.*, **17**, e1009545.
5. Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: A genomic signature. *Trends Genet.*, **11**, 283–290.
6. Provataris,P., Meusemann,K., Niehuis,O., Grath,S. and Misof,B. (2018) Signatures of DNA methylation across insects suggest reduced DNA methylation levels in Holometabola. *Genome Biol. Evolut.*, **10**, 1185–1197.
7. Simmonds,P., Xia,W., Baillie,J.K. and McKinnon,K. (2013) Modelling mutational and selection pressures on dinucleotides in eukaryotic phyla–selection against CpG and UpA in cytoplasmically expressed RNA and in RNA viruses. *BMC Genomics*, **14**, 610–610.
8. Forni,D., Pozzoli,U., Cagliani,R. and Sironi,M. (2024) Dinucleotide biases in the genomes of prokaryotic and eukaryotic dsDNA viruses and their hosts. *Mol. Ecol.*, **33**, e17287.
9. Forni,D., Pozzoli,U., Cagliani,R., Clerici,M. and Sironi,M. (2023) Dinucleotide biases in RNA viruses that infect vertebrates or invertebrates. *Microbiol. Spectr.*, **11**, e0252923.
10. Giallonardo,F.D., Schlub,T.E., Shi,M. and Holmes,E.C. (2017) Dinucleotide Composition in Animal RNA Viruses Is Shaped More by Virus Family than by Host Species. *J. Virol.*, **91**, e02381-16.
11. Beutler,E., Gelbart,T., Han,J.H., Koziol,J.A. and Beutler,B. (1989) Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. *Proc. Nat. Acad. Sci. U.S.A.*, **86**, 192–196.
12. Simmen,M.W. (2008) Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. *Genomics*, **92**, 33–40.
13. Karlin,S. and Mrázek,J. (1997) Compositional differences within and between eukaryotic genomes. *Proc. Nat. Acad. Sci. U.S.A.*, **94**, 10227–10232.

14. Bowie,A.G. and Unterholzner,L. (2008) Viral evasion and subversion of pattern-recognition receptor signalling. *Nat. Rev. Immunology*, **8**, 911–922.
15. Luo,X., Wang,X., Gao,Y., Zhu,J., Liu,S., Gao,G. and Gao,P. (2020) Molecular mechanism of RNA recognition by zinc-finger antiviral protein. *Cell Rep.*, **30**, 46–52.
16. Takata,M.A., Gonçalves-Carneiro,D., Zang,T.M., Soll,S.J., York,A., Blanco-Melo,D. and Bieniasz,P.D. (2017) CG dinucleotide suppression enables antiviral defence targeting non-self RNA. *Nature*, **550**, 124–127.
17. Wojciechowski,M., Czapinska,H. and Bochtler,M. (2013) CpG underrepresentation and the bacterial CpG-specific DNA methyltransferase M.MpeI. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 105–110.
18. Wang,Y., Rocha,E.P.C., Leung,F.C.C. and Danchin,A. (2004) Cytosine methylation is not the major factor inducing CpG dinucleotide Deficiency in bacterial genomes. *J. Mol. Evol.*, **58**, 692–700.
19. Blow,M.J., Clark,T.A., Daum,C.G., Deutschbauer,A.M., Fomenkov,A., Fries,R., Froula,J., Kang,D.D., Malmstrom,R.R., Morgan,R.D., *et al.* (2016) The epigenomic landscape of prokaryotes. *PLoS Genet.*, **12**, e1005854.
20. Oliveira,P.H., Touchon,M. and Rocha,E.P.C. (2014) The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res.*, **42**, 10618–10631.
21. Bernheim,A. and Sorek,R. (2020) The pan-immune system of bacteria: antiviral defence as a community resource. *Nat. Rev. Micro.*, **18**, 113–119.
22. Goldfarb,T., Sberro,H., Weinstock,E., Cohen,O., Doron,S., Charpak-Amikam,Y., Afik,S., Ofir,G. and Sorek,R. (2015) BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J.*, **34**, 169–183.
23. Ofir,G., Melamed,S., Sberro,H., Mukamel,Z., Silverman,S., Yaakov,G., Doron,S. and Sorek,R. (2017) DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nat. Microbiol.*, **3**, 90–98.
24. Wang,Y. and Leung,F.C.C. (2004) DNA structure constraint is probably a fundamental factor inducing CpG deficiency in bacteria. *Bioinformatics*, **20**, 3336–3345.
25. Dame,R.T., Rashid,F.-Z.M. and Grainger,D.C. (2020) Chromosome organization in bacteria: mechanistic insights into genome structure and function. *Nat. Rev. Genet.*, **21**, 227–242.
26. Le,T.B.K., Imakaev,M.V., Mirny,L.A. and Laub,M.T. (2013) High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science*, **342**, 731–734.
27. Duprey,A. and Groisman,E.A. (2021) The regulation of DNA supercoiling across evolution. *Protein Sci.*, **30**, 2042–2056.
28. Garnier,F., Couturier,M., Débat,H. and Nadal,M. (2021) Archaea: a gold mine for topoisomerase diversity. *Front. Microbiol.*, **12**, 661411.
29. Kim,S.H., Ganji,M., Kim,E., Van Der Torre,J., Abbondanzieri,E. and Dekker,C. (2018) DNA sequence encodes the position of DNA supercoils. *eLife*, **7**, e36557.
30. Basu,A., Bobrovnikov,D.G., Cieza,B., Arcon,J.P., Qureshi,Z., Orozco,M. and Ha,T. (2022) Deciphering the mechanical code of the genome and epigenome. *Nat. Struct. Mol. Biol.*, **29**, 1178–1187.
31. Blin,K. (2023) ncbi-genome-download. https://doi.org/10.5281/ZENODO.8192432.
32. Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.
33. Reimer,L.C., Sardà Carbasse,J., Koblitz,J., Ebeling,C., Podstawka,A. and Overmann,J. (2022) Bac *Dive* in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Res.*, **50**, D741–D746.
34. Helena-Bueno,K., Brown,C.R. and Melnikov,S. (2021) Gosha: A database of organisms with defined optimal growth temperatures Evolutionary Biology. bioRxiv doi: https://doi.org/10.1101/2021.12.21.473645, 30 May 2023, preprint: not peer reviewed.
35. Touchon,M., Bernheim,A. and Rocha,E.P.C. (2016) Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME J.*, **10**, 2744–2754.
36. Wilkinson,G.N. and Rogers,C.E. (1973) Symbolic description of factorial models for analysis of variance. *Appl. Stat.*, **22**, 392.
37. Revell,L.J. (2012) phytools: an R package for phylogenetic comparative biology (and other things): *phytools:Rpackage*. *Methods Ecol. Evol.*, **3**, 217–223.
38. Revell,L.J. (2023) phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ*, **12**, e16505.
39. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumeil,P.-A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.*, **50**, D785–D794.
40. Paradis,E. and Schliep,K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.
41. Pagel,M. (1999) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.
42. Garland,T., Dickerman,A.W., Janis,C.M. and Jones,J.A. (1993) Phylogenetic Analysis of Covariance by Computer Simulation. *Syst. Biol.*, **42**, 265–292.
43. Harmon,L.J., Weir,J.T., Brock,C.D., Glor,R.E. and Challenger,W. (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.
44. Li,K., Carroll,M., Vafabakhsh,R., Wang,X.A. and Wang,J.-P. (2022) DNAcycP: a deep learning tool for DNA cyclizability prediction. *Nucleic Acids Res.*, **50**, 3142–3154.
45. Chaumeil,P.-A., Mussig,A.J., Hugenholtz,P. and Parks,D.H. (2022) GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, **38**, 5315–5316.
46. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
47. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
48. Karlin,S., Doerfler,W. and Cardon,L.R. (1994) Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J. Virol.*, **68**, 2889–2897.
49. Duret,L. and Arndt,P.F. (2008) The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet.*, **4**, e1000071.
50. Odon,V., Fiddaman,S.R., Smith,A.L. and Simmonds,P. (2022) Comparison of CpG- and UpA-mediated restriction of RNA virus replication in mammalian and avian cells and investigation of potential ZAP-mediated shaping of host transcriptome compositions. *RNA*, **28**, 1089–1109.
51. Yoo,J., Park,S., Maffeo,C., Ha,T. and Aksimentiev,A. (2021) DNA sequence and methylation prescribe the inside-out conformational dynamics and bending energetics of DNA minicircles. *Nucleic Acids Res.*, **49**, 11459–11475.
52. Hagerman,P.J. (1986) Sequence-directed curvature of DNA. *Nature*, **321**, 449–450.
53. Koo,H.-S., Wu,H.-M. and Crothers,D.M. (1986) DNA bending at adenine · thymine tracts. *Nature*, **320**, 501–506.
54. Olson,W.K., Gorin,A.A., Lu,X.-J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 11163–11168.
55. Rosanio,G., Widom,J. and Uhlenbeck,O.C. (2015) In vitro selection of DNA s with an increased propensity to form small circles. *Biopolymers*, **103**, 303–320.

56. Geggier,S. and Vologodskii,A. (2010) Sequence dependence of DNA bending rigidity. *Proc. Natl. Acad. Sci. U.S.A.*, **107**, 15421–15426.
57. Back,G. and Walther,D. (2023) Predictions of DNA mechanical properties at a genomic scale reveal potentially new functional roles of DNA flexibility. *NAR genom. bioinform.*, **5**, lqad097.
58. Heine,M. and Chandra,S.B.C. (2009) The linkage between reverse gyrase and hyperthermophiles: a review of their invariable association. *J. Microbiol.*, **47**, 229–234.
59. Brochier-Armanet,C. and Forterre,P. (2006) Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. *Archaea*, **2**, 83–93.
60. Forterre,P. (2002) A hot story from comparative genomics: reverse gyrase is the only hyperthermophile-specific protein. *Trends Genet.*, **18**, 236–237.
61. Lulchev,P. and Klostermeier,D. (2014) Reverse gyrase—recent advances and current mechanistic understanding of positive DNA supercoiling. *Nucleic Acids Res.*, **42**, 8200–8213.
62. Ogawa,T., Yogo,K., Furuike,S., Sutoh,K., Kikuchi,A. and Kinosita,K. (2015) Direct observation of DNA overwinding by reverse gyrase. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 7495–7500.
63. Catchpole,R.J. and Forterre,P. (2019) The evolution of reverse gyrase suggests a nonhyperthermophilic last universal common ancestor. *Mol. Biol. Evol.*, **36**, 2737–2747.
64. Karlin,S., Ladunga,I. and Blaisdell,B.E. (1994) Heterogeneity of genomes: measures and values. *Proc. Nat. Acad. Sci. U.S.A.*, **91**, 12837–12841.
65. Karlin,S., Mrázek,J. and Campbell,A.M. (1997) Compositional biases of bacterial genomes and evolutionary implications. *J. Bacteriol.*, **179**, 3899–3913.
66. Duret,L. and Galtier,N. (2000) The covariation between TpA deficiency, CpG deficiency, and G+C content of human isochores is due to a mathematical artifact. *Mol. Biol. Evol.*, **17**, 1620–1625.
67. Shu,W.-S. and Huang,L.-N. (2022) Microbial diversity in extreme environments. *Nat. Rev. Micro.*, **20**, 219–235.
68. Jin,Q. and Kirk,M.F. (2018) pH as a primary control in environmental microbiology: 1. Thermodynamic perspective. *Front. Environ. Sci.*, **6**, 21.
69. Si,F., Li,D., Cox,S.E., Sauls,J.T., Azizi,O., Sou,C., Schwartz,A.B., Erickstad,M.J., Jun,Y., Li,X., *et al.* (2017) Invariance of initiation mass and predictability of cell size in *Escherichia coli*. *Curr. Biol.*, **27**, 1278–1287.
70. Harris,L.K. and Theriot,J.A. (2018) Surface area to volume ratio: a natural variable for bacterial morphogenesis. *Trends Microbiol.*, **26**, 815–832.
71. Gallet,R., Violle,C., Fromin,N., Jabbour-Zahab,R., Enquist,B.J. and Lenormand,T. (2017) The evolution of bacterial cell size: the internal diffusion-constraint hypothesis. *ISME J.*, **11**, 1559–1568.
72. Ojkic,N., Serbanescu,D. and Banerjee,S. (2022) Antibiotic resistance via bacterial cell shape-shifting. *mBio*, **13**, e00659–e22.
73. Basu,A., Bobrovnikov,D.G., Qureshi,Z., Kayikcioglu,T., Ngo,T.T.M., Ranjan,A., Eustermann,S., Cieza,B., Morgan,M.T., Hejna,M., *et al.* (2021) Measuring DNA mechanics on the genome scale. *Nature*, **589**, 462–467.
74. Fu,Z., Guo,M.S., Zhou,W. and Xiao,J. (2024) Differential roles of positive and negative supercoiling in organizing the *E. coli* genome. *Nucleic Acids Res.*, **52**, 724–737.
75. Guo,M.S., Kawamura,R., Littlehale,M.L., Marko,J.F. and Laub,M.T. (2021) High-resolution, genome-wide mapping of positive supercoiling in chromosomes. *eLife*, **10**, e67236.
76. Zhang,Y., He,L. and Li,S. (2023) Temperature dependence of DNA elasticity: An all-atom molecular dynamics simulation study. *J. Chem. Phys.*, **158**, 094902.
77. Dohnalová,H., Matoušková,E. and Lankaš,F. (2024) Temperature-dependent elasticity of DNA, RNA, and hybrid double helices. *Biophys. J.*, **123**, 572–583.
78. Driessen,R.P.C., Sitters,G., Laurens,N., Moolenaar,G.F., Wuite,G.J.L., Goosen,N. and Dame,R.T.. (2014) Effect of temperature on the intrinsic flexibility of DNA and its interaction with architectural proteins. *Biochemistry*, **53**, 6430–6438.
79. Kikuchi,A. and Asai,K. (1984) Reverse gyrase—a topoisomerase which introduces positive superhelical turns into DNA. *Nature*, **309**, 677–681.
80. Atomi,H., Matsumi,R. and Imanaka,T. (2004) Reverse gyrase is not a prerequisite for hyperthermophilic life. *J. Bacteriol.*, **186**, 4829–4833.
81. Lipscomb,G.L., Hahn,E.M., Crowley,A.T. and Adams,M.W.W. (2017) Reverse gyrase is essential for microbial growth at 95°C. *Extremophiles*, **21**, 603–608.
82. Bettotti,P., Visone,V., Lunelli,L., Perugino,G., Ciaramella,M. and Valenti,A. (2018) Structure and properties of DNA molecules over the full range of biologically relevant supercoiling states. *Sci. Rep.*, **8**, 6163.

Received: March 18, 2024. **Revised:** June 17, 2024. **Editorial Decision:** July 11, 2024. **Accepted:** July 18, 2024