Journal of Molecular Evolution (2021) 89:165–171
https://doi.org/10.1007/s00239-020-09974-w

COMMENTARY

Revisiting the Relationships Between Genomic G + C Content, RNA Secondary Structures, and Optimal Growth Temperature

Michelle M. Meyer[1]

Received: 16 September 2020 / Accepted: 9 November 2020 / Published online: 20 November 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract

Over twenty years ago Galtier and Lobry published a manuscript entitled "Relationships between Genomic G + C Content, RNA Secondary Structure, and Optimal Growth Temperature" in the *Journal of Molecular Evolution* that showcased the lack of a relationship between genomic G + C content and optimal growth temperature (OGT) in a set of about 200 prokaryotes. Galtier and Lobry also assessed the relationship between RNA secondary structures (rRNA stems, tRNAs) and OGT, and in this case a clear relationship emerged. Increasing structured RNA G + C content (particularly in regions that are double-stranded) correlates with increased OGT. Both of these fundamental relationships have withstood test of many additional sequences and spawned a variety of different applications that include prediction of OGT from rRNA sequence and computational ncRNA identification approaches. In this work, I present the motivation behind Galtier and Lobry's original paper and the larger questions addressed by the work, how these questions have evolved over the last two decades, and the impact of Galtier and Lobry's manuscript in fields beyond these questions.

Keywords rRNA · Genome composition · Thermoadaptation

Motivation of the Original Work

DNA base composition is one of the most fundamental properties of a genome. Chargaff's measurements of base composition in double-stranded DNA (Chargaff 1951) were important for the development and acceptance of Watson and Crick's structural model of DNA (Watson and Crick 1953) long before one could count individual guanine and cytosine residues on a sequencing trace. Organismal genomic G + C content can vary widely from less than 20% to over 75%, yet there is typically less variation between different locations within a given species genome (Bohlin et al. 2010). Over 50 years after the discovery of DNA's structure, understanding what drives variation in genomic G + C content is still very much an open question, despite DNA sequence data from a multitude of biological entities. It is still unclear whether G + C content variation may be generated by neutral processes such as mutational bias or biased gene conversion, or is primarily the result of natural selection. Furthermore, even if such variation is the result of natural selection, is selection acting on the genomic DNA itself, or rather on the molecules (e.g. RNAs and proteins) encoded by the DNA? These questions were ultimately the subject of Galtier and Lobry's paper published in *J. Mol. Evol.* in 1997 entitled 'Relationships between Genomic G + C content, RNA Secondary Structures, and Optimal Growth Temperature' (Galtier and Lobry 1997).

Despite the far-reaching nature of the questions outlined above, Galtier and Lobry sought to test a relatively specific hypothesis in their work. Chargaff is best known for describing the base-composition of double-stranded DNA, in particular that the quantities of adenosine (A) and thymine (T) are equal, and the quantities of guanine (G) and cytosine (C) are equal (Chargaff's first parity rule) (Chargaff 1951). Somewhat surprisingly, this observation also appears to hold true for single-stranded DNA in many cases [termed Chargaff's parity rule 2 or PR2 (Sueoka 1995)], although this rule is not as exact and there are frequently local variations that do not comply. Attributes consistent with PR2 were first described in *Bacillus subtilis* (Rudner et al. 1968a, b; Karkas et al. 1968), but subsequently proved true in a wide variety

Handling editor: **Aaron Goldman**

Michelle M. Meyer
m.meyer@bc.edu

[1] Department of Biology, Boston College, Chestnut Hill, MA 02467, USA

of different genomic sequences (Mitchell and Bridge 2006). Two hypotheses to explain PR2 during the late 1990′s were: (1) this phenomenon is due to mutational bias in the replicating polymerase (Sueoka 1962, 1995); and (2) this property is due to natural selection favoring the formation of self-complementary oligonucleotides within the DNA that might form hairpin structures (Forsdyke 1995). Galtier and Lobry proposed that the second hypothesis would predict that genomic G + C content should increase as organismal optimal growth temperature (OGT or T_{opt}) increases to ensure that DNA hairpin structures would remain stable. Thus, the goal of their study was to determine whether this prediction was supported by a large set of prokaryotic genomes.

Like many bioinformaticians, Galtier and Lobry largely compiled existing data for their study (Staley et al. 1984; Dalgaard and Garret 1993; Van de Peer et al. 1994; Sprinzl et al. 1996), and the methods used to determine genomic G + C content (thermal melting curves or buoyant density centrifugation) would be considered quite crude compared to the precision that sequencing provides today. Using this data, Galtier and Lobry found that OGT and genomic G + C content do not display a clear relationship, thus casting doubt on the hypothesis that secondary structures in genomic DNA explain Chargaff's PR2 (Galtier and Lobry 1997). Despite the specific nature of the hypothesis addressed, the two findings for which this paper is most frequently cited are quite general. The first is the lack of relationship between OGT and genomic G + C content. The second is that G + C content in the stems of the 16S and 23S rRNAs, and generally in the 5S rRNA and tRNAs, does correlate with organismal OGT. Both of these trends had previously been established in the context of hyperthermophilic archaea (Dalgaard and Garret 1993). However, the work of Dalgaard and Garret included a small number of organisms (about twenty vs. over one-hundred in Galtier and Lobry), which belonged to a limited phylogenetic distribution with narrow environmental diversity (thermophilic archaea with a few additional model species for comparison). Galtier and Lobry extended the findings of Dalgaard and Garrett across significantly more bacterial species, and in so doing extended the story beyond thermophilic archaea to a much more general phenomenon that attracted significantly more interest.

In the years since the publication of Galtier and Lobry's manuscript, work toward understanding forces at work in genome composition has continued. The debate regarding the relationship between genome composition and thermostability was by no means settled by this work, and satisfying explanations for Chargaff's PR2 and the diversity of G + C observed across diverse genomes remain elusive over 20 years later. The two major findings of Galtier and Lobry have spurred significant further work that encompasses a range of different applications that take advantage of the relationships between OGT, structured RNA G + C content, and genomic G + C content. These include: prediction of organism OGT based on 16S rRNA sequence, separation or enrichment of DNA extracted from microbial communities for a particular sub-populations based on G + C content, and computational methods for structured RNA identification.

Resolving the Relationship Between Genomic G + C Content and Thermoadaptation

The work of Galtier and Lobry provided evidence against adaptation to growth at higher temperature directly impacting genomic G + C content. However, this premise was further assessed using several different genomic subsets or better controlled sets of genomes by many additional studies from a range of authors over the years. Analysis of the three codon positions in coding sequences separately (under the assumption that the third codon position is less likely to be under selection for protein function), showed that GC content of the third codon closely mirrors that of the genome as a whole and does not correlate with OGT (Hurst and Merchant 2001). However, analysis of coding sequence dinucleotide frequencies indicated some OGT correlated changes, suggesting that thermoadaptation could directly impact genome dinucleotide frequencies (Nakashima et al. 2003). Several additional studies have assessed whether better phylogenetically informed sampling (comparing pairs of genomes from within the same class) enable better detection of a correlation between OGT and G + C content (Musto et al. 2004, 2006; Wang et al. 2006). However, findings from such works remain controversial and are not necessarily robust across many bacterial genera. It is clear that many factors such as codon bias (Knight et al. 2001) and changes in protein composition associated with thermoadaptation (Singer and Hickey 2000), may impact genomic G + C content (Hickey and Singer 2004). However, none of these factors yield a clear relationship between genomic G + C content and OGT.

Alternative Explanations for Chargaff's Second Parity Rule

Although Galtier and Lobry concluded that ssDNA hairpins are not likely a significant contributor to Chargaff's second parity rule (PR2), during the decades since its original formulation Chargaff's PR2 has largely proven robust as additional sequence data is collected. It applies to most complete genomes (Mitchell and Bridge 2006), although genomes of organelles (Mitchell and Bridge 2006; Nikolaou and Almirantis 2006) and sDNA viruses (Mitchell and Bridge 2006) are notably not compliant. Furthermore, although most complete genomes do follow PR2, there are significant local deviations. In bacterial genomes, the direction of replication and *ori* position significantly impact genome composition (McLean et al. 1998; Nikolaou and Almirantis 2005), sequences that are actively transcribed also tend to display purine loading (Szybalski et al. 1966; Bell and Forsdyke 1999), and exons tend to conform to PR2 more than intronic sequence in eukaryotes (Touchon et al. 2004). Despite such local variations, the rule has been extended from symmetry of mononucleotide frequencies to include symmetry of oligonucleotide frequencies (Qi and Cuticchia 2001; Baisnée et al. 2002; Shporer et al. 2016). The most satisfying explanations for the maintenance of Chargaff's second rule invoke frequent duplication, inversion, and transposition events in the genome (Albrecht-Buehler 2006, 2007; Okamura et al. 2007).

Causes for G + C Content Variability: Neutral Processes or Natural Selection?

The potential causes of diverse genomic G + C content essentially reduce to whether the observed variation is due to neutral processes (Sueoka 1962, 1999) or natural selection. It is easy to imagine how neutral processes may contribute to nucleotide content and several studies have assessed the viability of this option across different species (Zhao et al. 2007; Wu et al. 2012). However, most bacterial polymerases, even those from high G + C content organisms, display a bias toward conversion of G–C pairs into A–T pairs (Lind and Andersson 2008; Hershberg and Petrov 2010; Hildebrand et al. 2010; Wielgoss et al. 2011), although this may not be universally true (Dillon et al. 2015). Increasingly it appears that G + C content in genomes may be the result of a combination of neutral and selection processes that are quite subtle (Reichenberger et al. 2015). In prokaryotes coding sequences tend to be more G + C rich than non-coding regions (Bohlin et al. 2008), coding regions part of the core genome are higher G + C than those of the periphery genome (Bohlin et al. 2017), but modeling studies of substitution rates in the core genome still suggest a universal G–C to A–T mutational bias(Bohlin et al. 2018). Symbiotic bacteria whose genes are under less selective pressure, have both highly reduced and very A + T rich genomes (McCutcheon and Moran 2011) suggesting that lack of selection leads to A + T richness.

An alternative neutral process that has been invoked to explain variation in G + C content is biased gene-conversion. In eukaryotes G–C alleles are more likely to be maintained than A–T alleles during gene conversion events (Mugal et al. 2015). Such events are also proposed to impact bacterial genomes, and a positive correlation is observed between G + C content and evidence of recombination for genes in the core genome (Lassalle et al. 2015). Furthermore, the presence of machinery necessary for non-homologous end joining (NHEJ) is also correlated with increased G–C content (Weissman et al. 2019). The combination of these studies with the observation that increased genomic G + C content may correlate with environmental conditions such as aerobiosis (Naya et al. 2002; Romero et al. 2009), suggests that DNA damage may play a role in prokaryotic genomic G + C content. Thus, the essential question, what causes the strikingly large range of G + C content over diverse prokaryotic genomes, likely has a quite nuanced answer, and remains open even as more, and greater diversity, genomes are available.

rRNA G + C Content and Optimal Growth Temperature

The observation of Galtier and Lobry, that structured non-coding RNAs, and in particular their double-stranded regions, displayed a strong correlation between G + C content and OGT has been widely verified. Additional work shows that the G + C content in rRNAs occurs most noticeably in the regions expected to be base-paired, but also extends to loop regions (although with a small effect size) (Wang et al. 2006). The effect occurs among sequences chosen to control for differences in G + C content due to taxonomy (from the same genera), and cold-adapted organisms (in contrast to just mesophiles and thermophiles) display similar trends in their rRNA (Wang et al. 2006) and tRNA (Dutta and Chaudhuri 2010). Furthermore, the same observation can also be made for other structured RNAs such as the signal recognition particle (SRP) RNA (Miralles 2010). Additionally it has been found that the expression of different copies of the rRNAs with differing G + C composition in the same organism may be tuned to temperature, with higher G + C content rRNAs enabling increased fitness at higher temperatures (Sato et al. 2017; Sato and Kimura 2019).

The robustness of G + C composition correlation with OGT, has also spurred efforts to more broadly understand what other factors contribute to RNA thermostability. OGT also correlates with a decrease in the prevalence of uracil (U) specifically, although this does not seem to correspond with a replacement of G·U base-pairs with more G–C base-pairs, but rather a decrease in U prevalence across the molecule, including loop regions (Khachane et al. 2005). The structure of a thermophilic ribosome also appears to be more tightly packed than that of a mesophile (Mallik and Kundu 2013), and tRNAs in thermophiles may also display better folding characteristics than those in psychrophiles using in silico models of RNA folding (Dutta and Chaudhuri 2010).

Using rRNA G + C to Predict Optimal Growth Temperature

There are several applications of the observed correlation between OGT and G + C content in functional RNAs. One of these is enrichment of a sampled microbial community for organisms from a specific environment (Kimura et al. 2006). A second application of this correlation is the estimation of OGT, typically based on rRNA sequence (or its composition determined from melt-curves) (Kimura et al. 2010). This approach may be applied to single organisms, or increasingly to confirm the native environment of sequences isolated from metagenomic sequencing (Ragon et al. 2013; Kimura et al. 2013). As the amount of sequence in derived from whole genome shotgun sequencing (WGS) compared with 16S rRNA has shifted in such studies, methods have expanded to include additional features from genomic sequence such as ORF composition, but the composition of the tRNA and rRNA has a significant impact on accuracy even in the context of this addition data (Sauer and Wang 2019), although prediction based on proteomic data alone can also be effective (Li et al. 2019).

Even prior to the development of quantitative regressions to predict OGT based on genomic features, the relationship between rRNA G + C content and OGT was used to speculate about the environment of the last universal common ancestor (LUCA). In an early work Galtier et al. used a Markov model of sequence evolution coupled with maximum likelihood analysis to suggest that the ancestral rRNA contained sequence features consistent with a mesophilic origin (Galtier et al. 1999). However, this finding was rapidly disputed by others using alternative reconstruction techniques (e.g. maximum parsimony), as well as including additional molecules for analysis such as tRNA (Di Giulio 2000), or protein sequences (Di Giulio 2001, 2003). More realistic models based on both protein and rRNA reconstructed sequences indicated the potential for a mesophilic origin followed by divergence and parallel adaptation to higher temperatures followed by subsequent adaptation to more temperature environments (Boussau et al. 2008; Groussin and Gouy 2011). While this question is increasingly tackled by approaches that utilize far more information than what was available 20 years ago to reconstruct entire ancestral gene sets, a clear consensus still has not been reached (Weiss et al. 2016; Akanuma 2017).

Using G + C Content to Identify ncRNA

Another application of the relationship between structured RNA G + C content and organismal OGT coupled with the lack of relationship between G + C content and OGT, is the computational discovery of novel structured RNAs. It is established that stable structures may be formed by many sequences that do not encode functional RNA structures (Rivas and Eddy 2000). However, the premise that in a high A + T genome, structured RNAs should be encoded by regions with higher G + C content so that such molecules retain their stability, is valid. Several different methods for ncRNA identification across a range of different species use some variation of this premise. Deviation from genomic G + C content alone was used to identify ncRNAs within extreme hyperthermophiles *Methanococcus jannaschii* and *Pyrococcus furiosus* (which have modest genomic G + C contents of ~ 30% and ~ 40%, respectively) (Klein et al. 2002), in combination with dinucleotide frequencies to find similar results in *M. jannaschii* (Schattner 2002), or to screen intergenic regions in A + T rich prokaryotic genomes that are further processed by other ncRNA comparative genomic approaches (Meyer et al. 2009; Stav et al. 2019). Other approaches used genome composition as one of many features to identify putative ncRNAs in genomes of mesophiles with less genome composition bias such as *E. coli* (Carter et al. 2001). Finally, several A + T rich eukaryotic genomes have also been screened in a similar manner including *Plasmodium falciparum* (Upadhyay et al. 2005) and *Dictyostelium discoideum* (Larsson et al. 2008). Thus, although any given mRNA may fold into a stable structure, when combined with other information G + C content has proven to be a good screening tool for ncRNA identification in specific situations where the G + C content due to structured RNA stability may rise above the genomic background.

Conclusions

The major findings of Galtier and Lobry have proven robust nearly 20 years and many additional genomes later. They were not the first to observe the relationship between G + C content of structured RNA and OGT and contrast it with that between genomic G + C and OGT, but they placed this observation into a much larger context than Dalgaard and Garrett (Dalgaard and Garret 1993), and in doing so made the finding accessible to a larger audience and ultimately seeded several other fruitful areas of research. The specific hypothesis that motivated this work has long since been superseded by other explanations, but the root questions remain largely unresolved. Thus, this work remains highly cited today, and will likely continue to be in the future.

Funding This material is based upon work supported by the National Science Foundation (NSF) grant MCB: 1715440 to MMM. The opinions, findings, and conclusions expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

Compliance with Ethical Standards

Conflict of interest The author declare that there is no conflict of interest.

References

Akanuma S (2017) Characterization of reconstructed ancestral proteins suggests a change in temperature of the ancient biosphere. Life. https://doi.org/10.3390/life7030033

Albrecht-Buehler G (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc Natl Acad Sci USA 103:17828–17833. https://doi.org/10.1073/pnas.0605553103

Albrecht-Buehler G (2007) Inversions and inverted transpositions as the basis for an almost universal "format" of genome sequences. Genomics 90:297–305. https://doi.org/10.1016/j.ygeno.2007.05.010

Baisnée P-F, Hampson S, Baldi P (2002) Why are complementary DNA strands symmetric? Bioinformatics 18:1021–1033. https://doi.org/10.1093/bioinformatics/18.8.1021

Bell SJ, Forsdyke DR (1999) Deviations from Chargaff's second parity rule correlate with direction of transcription. J Theor Biol 197:63–76. https://doi.org/10.1006/jtbi.1998.0858

Bohlin J, Skjerve E, Ussery DW (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. PLoS Comput Biol 4:e1000057–e1000059. https://doi.org/10.1371/journal.pcbi.1000057

Bohlin J, Snipen L, Hardy SP et al (2010) Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics 11:464–468. https://doi.org/10.1186/1471-2164-11-464

Bohlin J, Eldholm V, Pettersson JHO et al (2017) The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC Genomics 18:151–211. https://doi.org/10.1186/s12864-017-3543-7

Bohlin J, Eldholm V, Brynildsrud O et al (2018) Modeling of the GC content of the substituted bases in bacterial core genomes. BMC Genomics 19:589–596. https://doi.org/10.1186/s12864-018-4984-3

Boussau B, Blanquart S, Necsulea A et al (2008) Parallel adaptations to high temperatures in the Archaean eon. Nature 456:942–945. https://doi.org/10.1038/nature07393

Carter RJ, Dubchak I, Holbrook SR (2001) A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res 29:3928–3938. https://doi.org/10.1093/nar/29.19.3928

Chargaff E (1951) Structure and function of nucleic acids as cell constituents. Fed Proc 10:654–659

Dalgaard JZ, Garret RA (1993) Chapter 17 Archaeal hyperthermophile genes. In: Kates M, Kushner DJ (eds) The biochemistry of archaea (Archaebacteria). Elsevier, Amsterdam, pp 535–563

Di Giulio M (2000) The universal ancestor lived in a thermophilic or hyperthermophilic environment. J Theor Biol 203:203–213. https://doi.org/10.1006/jtbi.2000.1086

Di Giulio M (2001) The universal ancestor was a thermophile or a hyperthermophile. Gene 281:11–17. https://doi.org/10.1016/s0378-1119(01)00781-8

Di Giulio M (2003) The universal ancestor was a thermophile or a hyperthermophile: tests and further evidence. J Theor Biol 221:425–436. https://doi.org/10.1006/jtbi.2003.3197

Dillon MM, Sung W, Lynch M, Cooper VS (2015) The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*. Genetics 200:935–946. https://doi.org/10.1534/genetics.115.176834

Dutta A, Chaudhuri K (2010) Analysis of tRNA composition and folding in psychrophilic, mesophilic and thermophilic genomes: indications for thermal adaptation. FEMS Microbiol Lett 305:100–108. https://doi.org/10.1111/j.1574-6968.2010.01922.x

Forsdyke DR (1995) Relative roles of primary sequence and (G + C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. J Mol Evol 41:573–581. https://doi.org/10.1007/BF00175815

Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44:632–636. https://doi.org/10.1007/pl00006186

Galtier N, Tourasse N, Gouy M (1999) A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221. https://doi.org/10.1126/science.283.5399.220

Groussin M, Gouy M (2011) Adaptation to environmental temperature is a major determinant of molecular evolutionary rates in archaea. Mol Biol Evol 28:2661–2674. https://doi.org/10.1093/molbev/msr098

Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6:e1001115. https://doi.org/10.1371/journal.pgen.1001115

Hickey DA, Singer GAC (2004) Genomic and proteomic adaptations to growth at high temperature. Genome Biol 5:117–127. https://doi.org/10.1186/gb-2004-5-10-117

Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6:e1001107. https://doi.org/10.1371/journal.pgen.1001107

Hurst LD, Merchant AR (2001) High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc Biol Sci 268:493–497. https://doi.org/10.1098/rspb.2000.1397

Karkas JD, Rudner R, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase. Proc Natl Acad Sci USA 60:915–920. https://doi.org/10.1073/pnas.60.3.915

Khachane AN, Timmis KN, dos Santos VAPM (2005) Uracil content of 16S rRNA of thermophilic and psychrophilic prokaryotes correlates inversely with their optimal growth temperatures. Nucleic Acids Res 33:4016–4022. https://doi.org/10.1093/nar/gki714

Kimura H, Sugihara M, Kato K, Hanada S (2006) Selective phylogenetic analysis targeted at 16S rRNA genes of thermophiles and hyperthermophiles in deep-subsurface geothermal environments. AEM 72:21–27. https://doi.org/10.1128/AEM.72.1.21-27.2006

Kimura H, Mori K, Tashiro T et al (2010) Culture-independent estimation of optimal and maximum growth temperatures of archaea in subsurface habitats based on the G+C content in 16S rRNA gene sequences. Geomicrobiol J 27:114–122. https://doi.org/10.1080/01490450903456699

Kimura H, Mori K, Yamanaka T, Ishibashi JI (2013) Growth temperatures of archaeal communities can be estimated from the guanine-plus-cytosine contents of 16S rRNA gene fragments. Environ Microbiol Rep 5:468–474. https://doi.org/10.1111/1758-2229.12035

Klein RJ, Misulovin Z, Eddy SR (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci USA 99:7542–7547. https://doi.org/10.1073/pnas.112063799

Knight RD, Freeland SJ, Landweber LF (2001) A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol 2:RESEARCH0010. https://doi.org/10.1186/gb-2001-2-4-research0010

Larsson P, Hinas A, Ardell DH et al (2008) De novo search for non-coding RNA genes in the AT-rich genome of *Dictyostelium discoideum*: performance of Markov-dependent genome feature scoring. Genome Res 18:888–899. https://doi.org/10.1101/gr.069104.107

Lassalle F, Périan S, Bataillon T et al (2015) GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet 11:e1004941. https://doi.org/10.1371/journal.pgen.1004941

Li G, Rabe KS, Nielsen J, Engqvist MKM (2019) Machine learning applied to predicting microorganism growth temperatures and enzyme catalytic optima. ACS Synth Biol 8:1411–1420. https://doi.org/10.1021/acssynbio.9b00099

Lind PA, Andersson DI (2008) Whole-genome mutational biases in bacteria. Proc Natl Acad Sci USA 105:17878–17883. https://doi.org/10.1073/pnas.0804445105

Mallik S, Kundu S (2013) A comparison of structural and evolutionary attributes of *Escherichia coli* and *Thermus thermophilus* small ribosomal subunits: signatures of thermal adaptation. PLoS ONE 8:e69898. https://doi.org/10.1371/journal.pone.0069898

McCutcheon JP, Moran NA (2011) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10:13–26. https://doi.org/10.1038/nrmicro2670

McLean MJ, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J Mol Evol 47:691–696. https://doi.org/10.1007/pl00006428

Meyer MM, Ames TD, Smith DP, Weinberg Z, Schwalbach MS, Giovannoni SJ, Breaker RR (2009) Identification of candidate structured RNAs in the marine organism 'Candidatus Pelagibacter ubique'. BMC Genomics 10:268. https://doi.org/10.1186/1471-2164-10-268

Miralles F (2010) Compositional properties and thermal adaptation of SRP-RNA in bacteria and archaea. J Mol Evol 70:181–189. https://doi.org/10.1007/s00239-009-9319-1

Mitchell D, Bridge R (2006) A test of Chargaff's second rule. Biochem Biophys Res Commun 340:90–94. https://doi.org/10.1016/j.bbrc.2005.11.160

Mugal CF, Weber CC, Ellegren H (2015) GC-biased gene conversion links the recombination landscape and demography to genomic base composition. BioEssays 37:1317–1326. https://doi.org/10.1002/bies.201500058

Musto H, Naya H, Zavala A et al (2004) Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett 573:73–77. https://doi.org/10.1016/j.febslet.2004.07.056

Musto H, Naya H, Zavala A et al (2006) Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochem Biophys Res Commun 347:1–3. https://doi.org/10.1016/j.bbrc.2006.06.054

Nakashima H, Fukuchi S, Nishikawa K (2003) Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. J Biochem 133:507–513. https://doi.org/10.1093/jb/mvg067

Naya H, Romero H, Zavala A et al (2002) Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol 55:260–264. https://doi.org/10.1007/s00239-002-2323-3

Nikolaou C, Almirantis Y (2005) A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. Nucleic Acids Res 33:6816–6822. https://doi.org/10.1093/nar/gki988

Nikolaou C, Almirantis Y (2006) Deviations from Chargaff's second parity rule in organellar DNA Insights into the evolution of organellar genomes. Gene 381:34–41. https://doi.org/10.1016/j.gene.2006.06.010

Okamura K, Wei J, Scherer SW (2007) Evolutionary implications of inversions that have caused intra-strand parity in DNA. BMC Genomics 8:160. https://doi.org/10.1186/1471-2164-8-160

Qi D, Cuticchia AJ (2001) Compositional symmetries in complete genomes. Bioinformatics 17:557–559. https://doi.org/10.1093/bioinformatics/17.6.557

Ragon M, Van Driessche AES, García-Ruíz JM et al (2013) Microbial diversity in the deep-subsurface hydrothermal aquifer feeding the giant gypsum crystal-bearing Naica Mine, Mexico. Front Microbiol 4:37. https://doi.org/10.3389/fmicb.2013.00037

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

Rivas E, Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. Bioinformatics 16:583–605. https://doi.org/10.1093/bioinformatics/16.7.583

Romero H, Pereira E, Naya H, Musto H (2009) Oxygen and guanine-cytosine profiles in marine environments. J Mol Evol 69:203–206. https://doi.org/10.1007/s00239-009-9230-9

Rudner R, Karkas JD, Chargaff E (1968a) Separation of *B. subtilis* DNA into complementary strands, I. Biological properties. Proc Natl Acad Sci USA 60:630–635. https://doi.org/10.1073/pnas.60.2.630

Rudner R, Karkas JD, Chargaff E (1968b) Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci USA 60:921–922. https://doi.org/10.1073/pnas.60.3.921

Sato Y, Kimura H (2019) Temperature-dependent expression of different guanine-plus-cytosine content 16S rRNA genes in *Haloarcula* strains of the class *Halobacteria*. Antonie Van Leeuwenhoek 112:187–201. https://doi.org/10.1007/s10482-018-1144-3

Sato Y, Fujiwara T, Kimura H (2017) Expression and function of different guanine-plus-cytosine content 16S rRNA genes in *Haloarcula hispanica* at different temperatures. Front Microbiol 8:482. https://doi.org/10.3389/fmicb.2017.00482

Sauer DB, Wang D-N (2019) Predicting the optimal growth temperatures of prokaryotes using only genome derived features. Bioinformatics 35:3224–3231. https://doi.org/10.1093/bioinformatics/btz059

Schattner P (2002) Searching for RNA genes using base-composition statistics. Nucleic Acids Res 30:2076–2082. https://doi.org/10.1093/nar/30.9.2076

Shporer S, Chor B, Rosset S, Horn D (2016) Inversion symmetry of DNA k-mer counts: validity and deviations. BMC Genomics 17:696–713. https://doi.org/10.1186/s12864-016-3012-8

Singer GA, Hickey DA (2000) Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. Mol Biol Evol 17:1581–1588. https://doi.org/10.1093/oxfordjournals.molbev.a026257

Sprinzl M, Steegborn C, Hübel F, Steinberg S (1996) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 24:68–72. https://doi.org/10.1093/nar/24.1.68

Staley JT, Bryant MP, Pfenning N, Holt J (1984) Bergey's manual of systematic bacteriology. The Williams & Wilkins Co, Baltimore

Stav S, Atilho RM, Mirihana Arachchilage G, Nguyen G, Higgs G, Breaker RR (2019) Genome-wide discovery of structured noncoding RNAs in bacteria. BMC Microbiol 19:66. https://doi.org/10.1186/s12866-019-1433-7

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592. https://doi.org/10.1073/pnas.48.4.582

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

Sueoka N (1999) Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. J Mol Evol 49:49–62. https://doi.org/10.1007/pl00006534

Szybalski W, Kubinski H, Sheldrick P (1966) Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. Cold Spring Harb Symp Quant Biol 31:123–127. https://doi.org/10.1101/sqb.1966.031.01.019

Touchon M, Arneodo A, d'Aubenton-Carafa Y, Thermes C (2004) Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. Nucleic Acids Res 32:4969–4978. https://doi.org/10.1093/nar/gkh823

Upadhyay R, Bawankar P, Malhotra D, Patankar S (2005) A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of *Plasmodium falciparum*. Mol Biochem Parasitol 144:149–158. https://doi.org/10.1016/j.molbiopara.2005.08.012

Van de Peer Y, Van den Broeck I, De Rijk P, De Wachter R (1994) Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res 22:3488–3494. https://doi.org/10.1093/nar/22.17.3488

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

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171:737–738. https://doi.org/10.1038/171737a0

Weiss MC, Sousa FL, Mrnjavac N et al (2016) The physiology and habitat of the last universal common ancestor. Nat Microbiol 1:16116–16118. https://doi.org/10.1038/nmicrobiol.2016.116

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

Wielgoss S, Barrick JE, Tenaillon O et al (2011) Mutation rate inferred from synonymous substitutions in a long-term evolution experiment with *Escherichia coli*. G3 (Bethesda) 1:183–186. https://doi.org/10.1534/g3.111.000406

Wu H, Zhang Z, Hu S, Yu J (2012) On the molecular mechanism of GC content variation among eubacterial genomes. Biol Direct 7:2–16. https://doi.org/10.1186/1745-6150-7-2

Zhao X, Zhang Z, Yan J, Yu J (2007) GC content variability of eubacteria is governed by the pol III alpha subunit. Biochem Biophys Res Commun 356:20–25. https://doi.org/10.1016/j.bbrc.2007.02.109