MBE Advance Access published August 22, 2006

UV-targeted dinucleotides are not depleted in light-exposed Prokaryotic genomes

Leonor Palmeira[*], Laurent Guéguen and Jean R. Lobry

Affiliation for all authors:

Laboratoire de Biométrie et Biologie Évolutive (UMR 5558); CNRS;

Univ. Lyon 1, 43 bd 11 nov, 69622, Villeurbanne Cedex, France;

HELIX, Unité de recherche INRIA;

[*] palmeira@biomserv.univ-lyon1.fr; tel: +33 4 72 43 12 84; fax: +33 4 72 43 71 13

Abstract

We have investigated the hypothesis that pyrimidine dinucleotides are avoided in light-exposed genomes, as the result of selective pressure due to high UV exposure. The main damage to DNA produced by UV radiation is known to be the formation of pyrimidine photoproducts: it is estimated that about ten dimers per minute are formed in an *Escherichia coli* chromosome exposed to the ultraviolet light in direct overhead sunlight at sea level. It is also known that on an *Escherichia coli* chromosome exposed to UVb wavelengths (290 to 320 nm), pyrimidine photoproducts are formed in the following proportions: 59% TpT, 7% CpC, and 34% CpT plus TpC. We have analyzed all available complete prokaryotic genomes and the model organism *Prochlorococcus marinus*, and have found that pyrimidine dinucleotides are not systematically avoided. This suggests that prokaryotes must have sufficiently effective protection and repair systems for UV exposure not to affect their dinucleotide composition.

Keywords: G+C content, dinucleotide content, ultraviolet radiation, aerobiosis, *Prochlorococcus marinus*

Introduction

Statistical analysis of the global composition of genomes and its link with environmental or metabolic characteristics has been the focus of considerable interest. It has been established that UV light – at both UVb (290 to 320 nm) and UVa wavelengths (320 to 400 nm) – damages DNA by specific mechanisms. It has been shown that UVb wavelengths are particularly dangerous for DNA and that the damage they most often cause is the formation of cyclobutane pyrimidine dimers by the photo-excitation of adjacent pyrimidines (Setlow, 1966). If one of these dimers is formed on a DNA strand, this leads to a local DNA distortion, which blocks both transcription and replication (Setlow, 1966; Singer and Ames, 1970; Besaratinia et al., 2005). Singer and Ames (1970) have estimated that about ten dimers per minute are formed in an *Escherichia coli* chromosome by the ultraviolet light in direct overhead sunlight at sea level.

The sensitivities of the four pyrimidine dinucleotides to UVb wavelengths differ: experiments on *Escherichia coli* DNA show that TpT photoproducts make up to 59% of the target dimers, CpC up to 7%, and that CpT and TpC share the remaining 34% (Setlow, 1966). It is noteworthy that most dimers involved are T-rich, and so a high G+C content will tend to result in genomes with fewer target dimers (see Figure 1).

Singer and Ames (1970) investigated whether bacteria exposed to higher UV radiation have a higher G+C content, as the result of environmental adaptation. They found a strong link between the genomic **[G+C]** content in bacteria and the amount of UV exposure in their habitat, and they conclude that bacteria exposed to high levels of UV have a higher G+C content than those with less exposure. We wanted to re-assess this hypothesis for the following reasons.

First, the results reported were controversial (Bak et al., 1972). Bak et al. (1972) discuss the UV-exposure experienced by various bacteria, and conclude that some of these species are not as highly exposed as suggested by Singer and Ames (1970). If this is the case, their conclusions would obviously be undermined. They also suggested that the wide variation in the G+C content in bacteria may be mainly attributable to phylogenetic relationships rather than to adaptation to habitat. Since this first controversy, there have been no major follow-up studies, and the question remains open.

Second, the choice of G+C content as an indicator of the impact of UV on pyrimidine dinucleotides is questionable: Fig 1. shows that G+C content is a poor indicator of UV-target content in a genome. Indeed, for a given G+C content, the density of target dinucleotides present in the genome can vary considerably, depending on the degree of aggregation of the pyrimidine dinucleotides (see legend of Fig 1.). Conversely, a given phototarget density can result from many different G+C contents.

Third, the availability of completely sequenced genomes now makes it possible to investigate the impact of UV-exposure on genomes by directly measuring their pyrimidine dinucleotide content. Pyrimidine dinucleotides are the direct targets of UVb wavelengths, and if UV-light has a major impact on genomes, we can expect highly-exposed micro-organisms to display significant depletion of all four pyrimidine dinucleotides (CpC, CpT, TpC and TpT).

Last but not least, recent studies have focused on the relationship between external forces and genomic content (*e.g.*(Naya et al., 2002; Foerstner et al., 2005)). In particular, Naya et al. (2002) have shown that aerobic bacteria have a higher G+C content than anaerobic bacteria. In addition, aerobic bacteria are more likely to be exposed to sunlight, which means that it could be difficult to distinguish between the effects of aerobiosis and those of UV radiation.

Materials & Methods

Systematic study

All 221 bacterial and archae genomes available on the EBI Genome Reviews database were retrieved on June 16, 2005. Out of the 221 complete genomes extracted, two datasets were created: one contained all annotated 'CDS' sequences and will subsequently be referred to as `coding sequences`, and one contained all sequences other than those annotated as 'CDS', 'RRNA' (ribosomal RNA coding regions) or 'TRNA' (transfer RNA coding regions), and will subsequently be referred to as `intergenic sequences`.

We started by a systematic approach, and investigated all available complete bacteria and archae genomes. We then took a closer look at the genomes of three strains of the picocyanobacterium *Prochlorococcus marinus*.

Prochlorococcus marinus as a model organism

Each of the three strains of *Prochlorococcus marinus* we investigated is adapted to a different depth in the water column (Dufresne et al., 2003; Rocap et al., 2003), and therefore exposed to different intensities of ultraviolet radiation. This seemed to make it an ideal model organism for investigating this hypothesis.

Dufresne et al. (2003) have shown that the SS120 strain is adapted to living at a depth of 120 meters. The MIT 9313 strain is adapted to living at a depth of 135 meters, and so both these strains can be considered to be low-light adapted strains (Rocap et al., 2003). The MED4 strain is adapted to living at a depth of 5 meters, and can be considered to be a high-light adapted strain (Dufresne et al., 2003). The residual intensities of 260 nm irradiation (UVb) at various depths in pure water can be estimated from water's absorbance coefficient (Quickenden and Irvin, 1980) as follows: 70% of its original intensity at 5 m depth (MED4 strain), 0.0002% at 120 m depth (SS120 strain), and 0.00007% at 135 m depth (MIT 9313 strain).

The accession numbers and references for these strains are as follows: strain CCMP 1375 / SS120 / SARG (GenBank accession number AE017126) (Dufresne et al., 2003), subsp. pastoris, strain CCMP 1378 / MED4 (GenBank accession number BX548174) and strain MIT 9313 (GenBank accession number BX548175) (Rocap et al., 2003).

Statistical analysis

Our aim was to find out whether pyrimidine dinucleotides are avoided in bacterial genomes. In prokaryote genomes, `coding sequences` can constitute up to 80% of the entire genome, and sometimes even more. These sequences are subjected to strong selective pressure, which makes other effects difficult to detect. Nevertheless, deleting 80% of the data is not a good way to detect small effects. We therefore developed two methods for measuring the over- and under-representation of dinucleotides: one for `coding sequences`, the other for `intergenic sequences`.

The idea was to compute a normalized statistic of the ρ_{xy} statistic (Karlin and Brendel, 1992), in order to make it possible to compare the results for sequences belonging to different species and even to different phyla.

$$\rho_{xy} = \frac{f_{xy}}{f_x \times f_y} \tag{1}$$

where f_{xy}, f_x and f_y are the frequencies in the studied sequence of dinucleotide xy, nucleotide x and nucleotide y respectively.

The normalized statistic is of the following type:

$$z_{score} = \frac{\rho_{xy} - E(\rho_{xy})}{\sqrt{Var(\rho_{xy})}} \tag{2}$$

where $E(\rho_{xy})$ and $Var(\rho_{xy})$ are the expected value and variance of ρ_{xy} according to a given model that describes the sequence. This statistic follows the standard normal distribution. The expected value and the variance can either be computed by simulation, or by analytical calculation if asymptotic results are available.

Naturally, we can propose various models of sequences for calculating the expected value and variance. Each of these models is constructed so as to preserve some of the constraints of the studied sequence for the expected counts. This means that both the expected count and the observed count will share the specified constraints, and the z_{score} statistic will reflect what is over- or under-represented in the studied sequence once the effects of these constraints have been eliminated. Two of these models will be shown here, which means two z_{score} statistics will be presented.

Intergenic sequence analysis

The unconstrained base shuffling model (UBS) describes each sequence as a series of independent draws following the frequencies of the four letters as counted on that sequence. In this model, only the base composition of the analyzed sequence is preserved, and asymptotic results are available (Prum, Rodolphe and de Turckheim, 1995):

$$E(\rho_{xy}) = 1 \tag{3}$$

$$Var(\rho_{xy}) = \frac{(1-f_x)(1-f_y)}{nf_xf_y} \tag{4}$$

The z_{score} statistic computed using this model allows us to answer the following question: is there an anomalously high or low XpY content given the base composition of the studied sequence?

We have used this model on `intergenic sequences`, because we do not know enough about the selective forces involved to be able to develop an appropriate model that would allow us to see what is happening underneath the selective constraints acting on intergenic regions.

Coding sequence analysis

Unlike the intergenic regions, a certain number of constraints in coding regions has been identified. In CDS, we know that there is a bias in codon usage (Grantham et al., 1980), and we therefore expect that the dinucleotides present in the preferred codons will be over-represented in the generic statistic presented in the "Intergenic sequence analysis". We have therefore developed a model that allows us to compute a z_{score} which erases this codon usage bias. This statistic enables us to identify over- and under-representations exist in `coding sequences`, despite the presence of codon usage bias.

In the codon shuffling model (CS), each sequence is described as a series of independent draws of codons following the frequencies of the codons as counted in that sequence. In the CS model, the codon composition of the analyzed sequence is preserved, which means that the base composition of the sequence analyzed is also preserved.

We can show that computing the z_{score} statistic using this model can be reduced to computing the z_{score} statistic on dinucleotides that overlap two codons:

$$z_{score} = \frac{\rho_{XY_{3-1}} - E(\rho_{XY_{3-1}})}{\sqrt{Var(\rho_{XY_{3-1}})}} = \frac{XY_{3-1} - E(XY_{3-1})}{\sqrt{Var(XY_{3-1})}} \tag{5}$$

where $\rho_{XY_{3-1}}$ is the ρ statistic for dinucleotides overlapping two codons, and where XY_{3-1} is the count of dinucleotides that overlap two codons. Asymptotic results are available (Gautier, Gouy and Louail, 1985):

$$E(XY_{3-1}) = \frac{n_1 n_2 - n_3}{n} \tag{6}$$

$$Var(XY_{3-1}) = E(XY_{3-1}) - (E(XY_{3-1}))^2$$

$$+\frac{1}{n(n-1)}[(2n_3(n_1 + n_2 - n_1 n_2 - 1) + n_1 n_2 (n_1 - 1)(n_2 - 1))] \tag{7}$$

where $n1$, $n2$ and $n3$ are the number of codons ending with the letter X, the number of codons starting with the letter Y and the number of codons starting with letter Y and ending with letter X. n is the total number of codons in the sequence respectively.

The z_{score} statistic computed using this model allows us to answer the following question: is there an anomalously high or low XpY content given the codon usage bias of the studied CDS?

Reproducibility

All computations were made using R's seqinR (Charif and Lobry, 2006) package and were conducted using the computation resources available from the IN2P3's Computing Center.

Data and results are available, and can be reproduced online at:

http://biomserv.univ-lyon1.fr/~palmeira/repro/uv.html

Results

Systematic study

No systematic under-representation of CpT, TpC, CpT or TpT dinucleotides was found (see Fig 2). None of the four pyrimidine dinucleotides was globally and significantly over- or under-represented. This clearly does not bear out the initial hypothesis being tested and means that there is no avoidance of these four dinucleotides in prokaryotic genomes, despite the fact that they are major targets for photo-induced damage (Setlow, 1966).

There is a rather good correlation between the XpY content of intergenic sequences and the XpY content of coding sequences, which is strong evidence for general DNA mechanisms common to both coding and intergenic sequences. This shows that in highly constrained CDS sequences our method is able to recover general signals also present in intergenic sequences. This is true not only for the four pyrimidine dinucleotides but for all sixteen dinucleotides, and could be explained by the existence of biased mutational processes acting indifferently on the whole genome and producing genome-wide biases (Chen et al., 2004).

The rather universal over-representation of TpT dinucleotides in all genomes is surprising, even though it was not always statistically significant. Unlike eukaryotic mRNA, where poly-A stretches have a known essential function, there is no evidence for major poly-A or poly-T stretches in bacterial DNA. However, ApA and TpT periodical patterns have been reported in both bacteria and eukaryotes (Tomita, Wada and Kawashima, 1999). This periodicity has been related to DNA coiling and super-coiling and could explain the observed slight over-representation.

Very few outliers have been found for CpC dinucleotide, but those that have been found are the two chromosomes of the fully sequenced *Burkholderia mallei* and *Burkholderia pseudomallei* genomes. These two prokaryotes are both pathogens commonly found in soil and in groundwater and there is no evidence in the literature to suggest that these two strains are exposed to higher levels of UV than the other prokaryotes in the dataset. This feature cannot, therefore, be linked to UV exposure, and may be particular to this genus.

Prochlorococcus marinus as a model organism

No difference was found between the relative abundances of pyrimidine dinucleotides in these three strains (see Fig 3). However, these three ecotypes have been separated long enough to evolve different G+C contents (30.8% for MED4 at 5 m depth ; 36.4% for SS120 at 120 m depth ; 50.8% for MIT9313 at 135 m depth) (Rocap et al., 2003; Dufresne et al., 2003) and at least two of these ecotypes have divergent genomic adaptations (Rocap et al., 2003). These two previous studies show that the genomes of the three strains have diverged, yet this divergence has had no effect on pyrimidine dinucleotide content. There is therefore no evidence of an impact of UV exposure on dinucleotide content.

One possible exception could be the CpC dinucleotide, which seems to be slightly under-represented in the high-light adapted strain, compared to the two low-light adapted strains. For the other three pyrimidine dinucleotides, there is no avoidance of the pyrimidine dinucleotide, and therefore no link between UV exposure and relative pyrimidine dinucleotide abundance. We also note that the variability within one strain can be as great as that between different chromosomes (see Fig 3). This finding is consistent with the lack of any link found between relative dinucleotide abundance and exposure to UV.

Discussion

We have shown that UV exposure has no systematic impact on pyrimidine dinucleotide bias in prokaryotes. This is true not only for all bacteria and archae, but when we looked at strains of *Prochlorococcus marinus*, we once again found no link between UV exposure and pyrimidine dinucleotide abundance. This means that there is no evidence of the avoidance of pyrimidine dinucleotides in micro-organisms exposed to UV.

Prokaryotes have developed mechanisms to repair DNA damage. Our findings show that these systems must be efficient enough to make it unnecessary for pyrimidine dinucleotides to be avoided in their genome. This result is in agreement with recent studies on resistance of marine bacteria to UV radiation, see Agogué et al. (2005) and references therein. From an evolutionary perspective, this is probably due to their inheritance of highly efficient repair systems from ancestral organisms living at the time when there was no ozone layer to filter UV light.

The fact that protection and repair systems in bacteria are efficient enough for the genomic content to have evolved totally independently of UV exposure, tends to support the findings of Naya et al. (2002): UV exposure and aerobiosis are not likely to interfere in their analysis.

Acknowledgments: *This work was funded jointly by the ACI "New Interfaces of Mathematics", the ARC "Integrated Biological Networks" and the ANR "Régularités: Inférence et Statistique" project grants.*

References

Agogué, H., F. Joux, I. Obernosterer, and P. Lebaron. 2005, Resistance of Marine Bacterioneuston to Solar Radiation. Applied and Environmental Microbiology **71**:5282–5289.

Bak, A. L., J. F. Atkins, C. E. Singer, and B. N. Ames. 1972, Evolution of DNA Base Compositions in Microorganisms. Science **175**:1391–1393.

Besaratinia, A., T. W. Synold, C. Hsiu-Hua, C. Chang, B. Xi, A. D. Riggs, and G. D. Pfeifer. 2005, DNA lesions induced by UV A1 and B radiation in human cells: Comparative analyses in the overall genome and in the p53 tumor suppressor gene. Proceedings of the National Academy of the United States of America **102**:10058–10063.

Charif, D. and J. Lobry. 2006, SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. *In* H. R. U. Bastolla, M. Porto and M. Vendruscolo, eds., Structural approaches to sequence evolution: Molecules, networks, populations, vol. NA of *Biological and Medical Physics, Biomedical Engineering*, NA, Springer Verlag, New York, in press.

Chen, S. L., W. Lee, L. Shapiro, and H. H. McAdams. 2004, Codon usage between genomes is constrained by genome-wide mutational processes. Proceedings of the National Academy of Sciences of the United States of America **101**:3480–3485.

Dufresne, A., M. Salanoubat, F. Partensky, et al. 2003, Genome sequence of the cyanobacterium *Prochlorococcus marinus SS120*, a near minimal oxyphototrophic genome. Proceedings of the National Academy of the United States of America **100**:10020–10025.

Foerstner, K. U., C. von Mering, S. D. Hooper, and P. Bork. 2005, Environments shape the nucleotide composition of genomes. EMBO reports **6**:1208–1213.

Gautier, C., M. Gouy, and S. Louail. 1985, Non-parametric statistics for nucleic acid sequence study. Biochimie **67**:449–453.

Grantham, R., C. Gautier, M. Gouy, R. Mercier, and A. Pavé. 1980, Codon catalog usage and the genome hypothesis. Nucleic Acids Research **8**:r49–r62.

Karlin, S. and V. Brendel. 1992, Chance and Statistical Significance in Protein and DNA Sequence Analysis. Science **257**:39–49.

Naya, H., H. Romero, A. Zavala, B. Alvarez, and H. Musto. 2002, Aerobiosis Increases the Genomic Guanine Plus Cytosine Content (GC%) in Prokaryotes. Journal of Molecular Evolution **55**:260–264.

Prum, B., F. Rodolphe, and E. de Turckheim. 1995, Finding Words woth Unexpected Frequencies in Deoxyribonucleic Acid. Journal of the Royal Statistical Society **57**:205–220.

Quickenden, T. I. and J. A. Irvin. 1980, The ultraviolet absorption spectrumof liquid water. The Journal of Chemical Physics **72**:4416–4428.

Rocap, G., F. Larimer, J. Lamerdin, et al. 2003, Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature .

Setlow, R. B. 1966, Cyclobutane-Type Pyrimidine Dimers in Polynucleotides. Science **153**:379–386.

Singer, C. E. and B. N. Ames. 1970, Sunlight Ultraviolet and Bacterial DNA Base Ratios. Science **170**:822–826.

Tomita, M., M. Wada, and Y. Kawashima. 1999, Apa dinucleotide periodicity in prokaryote, eukaryote, and organelle genomes. Journal of Molecular Evolution **49**:182–192.

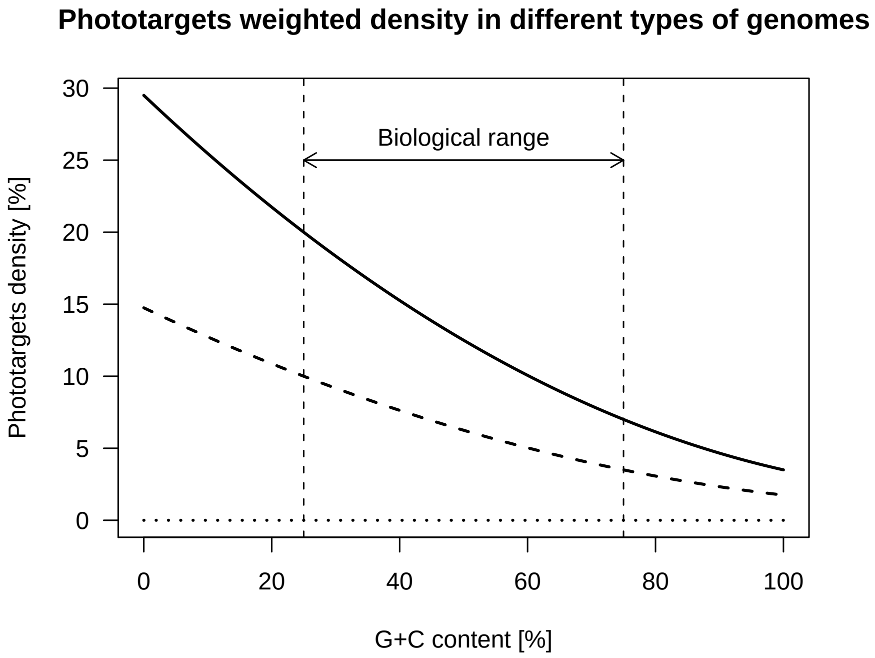

Figure 1: Density of phototargets, weighted by their frequency in the *E. coli* chromosome, and calculated for different G+C contents and for three kinds of random genomes. The weights are as follows: $0.59 * f_{tt} + 0.34 * (f_{tc} + f_{ct}) + 0.07 * f_{cc}$ (where f_{xy} is the frequency of dinucleotide xy in the specified genome). Three models of random genomes are analyzed. In the worst case (solid curve), the genome is the concatenation of a sequence of pyrimidines and a sequence of purines: all pyrimidines are involved in a pyrimidine dinucleotide. In the best case (dotted curve), the genome is an unbroken succession of pyrimidine-purine dinucleotides: no pyrimidine is involved in a pyrimidine dinucleotide. In the "random case" (dashed curve), the frequency of a pyrimidine dinucleotide is the result of chance ($f_{xy} = f_x \times f_y$).

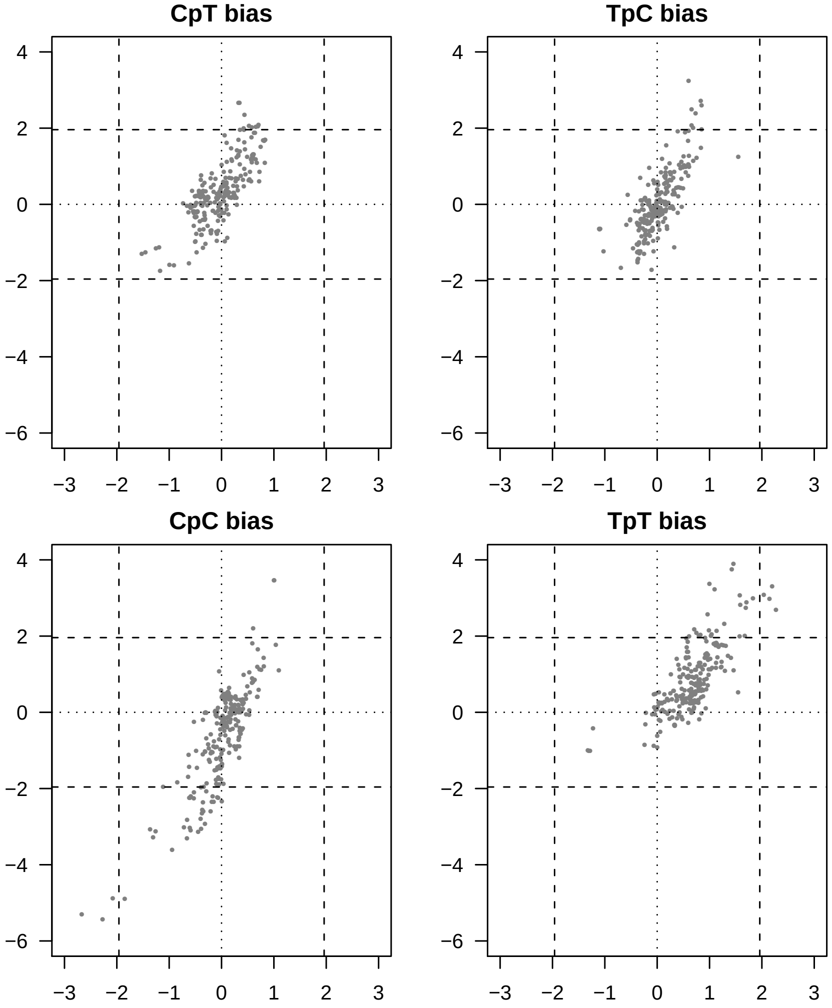

Figure 2: Plot of the mean z_{score} statistics for **intergenic sequences** (x-axis) and for **coding sequences** (y-axis), for each of the four pyrimidine dinucleotides. On each plot, a dot corresponds to the mean of these two statistics in a given prokaryote chromosome. The null x and y axis (dotted lines), and the 5% limits of significance for the standard normal distribution (dashed lines) are plotted as benchmarks. It should be noted note (see Fig 3), that the variability within one chromosome is sometimes as great as that between different chromosomes.

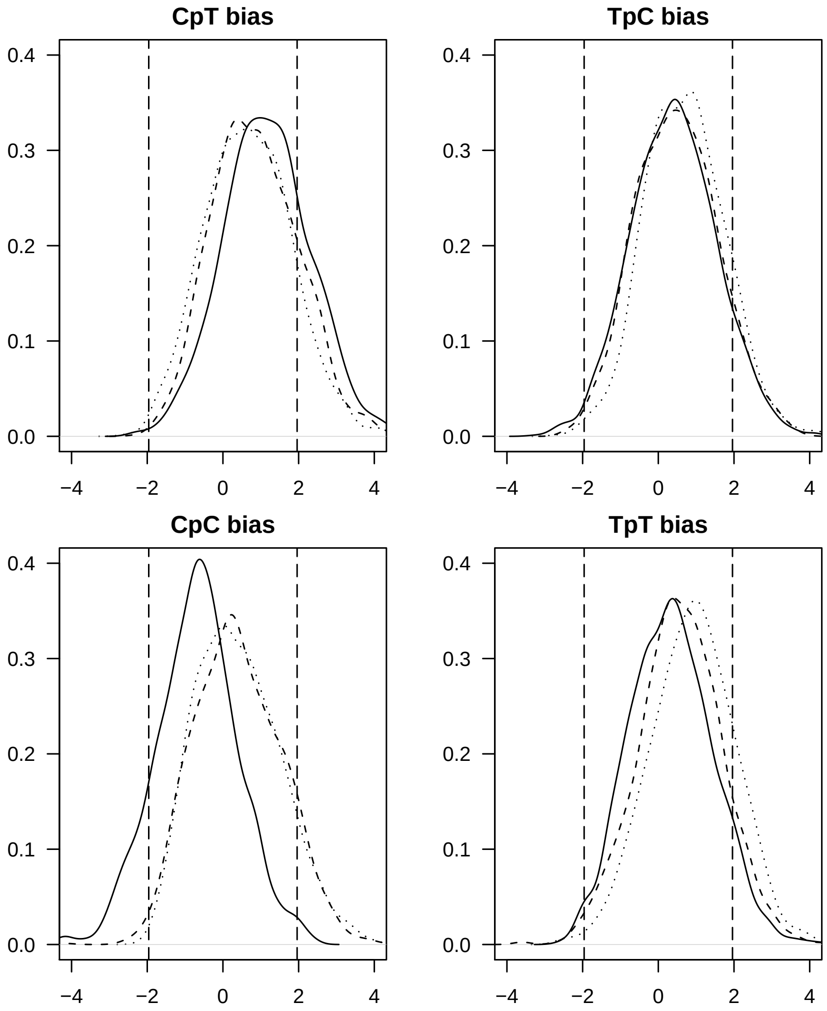

Figure 3: Each figure shows the distributions of the z_{score} in all **coding sequences** corresponding to each of the three strains of *Prochlorococcus marinus*. In each figure, the distribution for the MED4 (a high-light adapted strain) is shown as a solid line; the distribution for the SS120 (a low-light adapted strain) is shown as a dashed line, and the distribution for the MIT 9313 (a low-light adapted strain) is shown as a dotted line. The 5% limits of significance for the standard normal distribution (dashed vertical lines) are plotted as benchmarks.