Aerobiosis Decreases the Genomic GC content in Prokaryotes by Guanine Oxidation

Sidra Aslam, Xin-Ran Lan, and Deng-Ke Niu*

MOE Key Laboratory for Biodiversity Science and Ecological Engineering and Beijing Key Laboratory of Gene Resource and Molecular Development, College of Life Sciences, Beijing Normal University, Beijing 100875, China

*Author for Correspondence: Deng-Ke Niu, College of Life Sciences, Beijing Normal University, Beijing 100875, China. Telephone number: ++86-10-58802064; Email addresses: dkniu@bnu.edu.cn, dengkeniu@hotmail.com

Abstract

Oxidative stress is unavoidably faced by oxygen-consuming organisms. Under this stress, guanine is the most fragile base and so most frequently damaged among the four bases. Replication of DNA containing damaged guanines or incorporation of the damaged guanines into DNA strands would cause G to T mutations at a frequency depending on the efficiency of DNA repairing enzymes and the accuracy of replication enzymes. For this reason, aerobiosis is expected to decrease GC content. However, an opposite pattern of base composition in prokaryotes was reported 15 years ago. Although it has been widely cited, its overlook of the effect of shared ancestry determines the necessity to re-examine the reliability of its results. In the present study, by phylogenetically independent comparisons, we found that aerobic prokaryotes have significantly lower whole-genome GC contents than anaerobic ones. When the GC content is measured only at the 4-fold degenerate sites, the difference between aerobic prokaryotes and anaerobic prokaryotes became larger, being consistent with a mutational force imposed by oxidative stress on the evolution of nucleotide composition.

Key words: oxygen requirement, reactive oxygen species, aerobe, anaerobe, phylogenetically independent, nucleotide composition.

Main Text

Oxygen is an essential environmental factor for most organisms living on earth. Its accumulation is the most significant change in the evolution of biosphere and dramatically influenced the evolutionary trajectory of all the organisms that are exposed to it (Decker and Van Holde 2011). Oxidative metabolism brings not only a large amount of energy to the aerobic organisms, but also an unavoidable byproduct, reactive oxygen species (ROS). The ROS is highly reactive to most of the cellular organic molecules including nucleotides and their polymerized products, DNA and RNA. Among the four bases, the guanine has the lowest oxidation potential and is most susceptible to oxidation (Kanvah, et al. 2010). The direct products of deoxyguanosine oxidation are 8-oxo-7,8-dihydro-guanosine (8-oxoG) and 2,6-diamino-4-hydroxy-5-formamidopyrimidine (Fapy-G). As 8-oxoG has a lower oxidation potential than deoxyguanosine, 8-oxoG is susceptible to be further oxidized into several hyperoxidized products (Delaney, et al. 2012). Replication of DNA containing these damaged deoxyguanosines or incorporation of the damaged nucleotides into DNA would cause G to T mutations, the frequency of which depends on the efficiency of DNA repair enzymes and the accuracy of replication enzymes (Delaney, et al. 2012). No matter the frequency is high or low, the evolutionary direction of oxidatively damaged DNA is expected to decrease the GC content.

However, Naya et al. (2002) observed an entirely opposite pattern, aerobic prokaryotes have higher GC contents than anaerobic prokaryotes, by comparing the GC contents using conventionally statistical analysis. Furthermore, they showed that the pattern was still evident when aerobes and anaerobes were compared within each major phylum of archaea and bacteria. Although their analysis did not specifically account for the effect of shared ancestry, the consistency of the pattern within different phylogenetic lineages has convinced most researchers in this field. In the past 15 years, the study has been cited 115 times (data source:

Google Scholar; access date: June 16, 2007) without fierce doubt. We agree that many factors are shaping the GC content in evolution (Hildebrand, et al. 2010; Rocha and Feil 2010; Agashe and Shankar 2014; Kelkar, et al. 2015; Luo, et al. 2015; Reichenberger, et al. 2015; Seward and Kelly 2016), and the effect of guanine oxidation might be trivial as compared with other factors. But this could only make the difference between aerobes and anaerobes insignificant, rather than lead to a significant difference in the opposite direction. As the GC content displays an obvious phylogenetic signal (Haywood-Farmer and Otto 2003), we suspect that the counterintuitive observation might result from the phylogenetical non-independence of the data (Felsenstein 1985; Whitney and Garland 2010). With the rapid accumulation of sequenced genomes, now it is possible to collect data large enough to carry out a phylogenetically independent comparison.

With a dataset including 701 aerobic prokaryotes (species or strains) and 929 anaerobic prokaryotes (species or strains), we manually mapped the character (aerobic or anaerobic) to the phylogenetic tree. For each changing event in oxygen requirement, we got a pair of aerobic and anaerobic species, strains or nodes (fig. 1). The difference of GC content within one pair is phylogenetically independent from the differences within any other pairs. Pairwise comparison of the GC contents of the selected pairs of aerobes and anaerobes can be used to test whether changes in GC content are tightly associated with changes in oxygen requirement. In total, we obtained 87 aerobe-anaerobe pairs, including 81 pairs of bacteria and 6 pairs of archaea.

For a comparative study with that of Naya et al. (2002), we first compared the GC contents of the 701 aerobic genomes and the 929 anaerobic genomes without consideration of their positions on phylogenetic tree. Despite the much larger dataset, we also observed significantly higher GC contents in aerobes than anaerobes (fig. 2A). It seems that increases in sample size unlikely change the result.

By contrast, in the pairwise comparison of the 87 aerobe-anaerobe pairs, we obtained an opposite pattern, the aerobes have significantly lower GC contents than anaerobes (fig. 2B, $P = 0.038$). When the pairwise comparison is limited to the 81 pairs of bacteria, the difference between aerobes and anaerobes becomes insignificant in statistics ($P = 0.072$). In a regression analysis of 488 prokaryotic genomes, Bohlin et al. (2010) did not found significant association between genomic GC content and oxygen requirement when phylogenetic bias is accounted for. It seems that statistical level of the difference depends heavily on the specific samples studied. There are many other factors, like gene conversion, temperature, horizontal gene transfer, and nutrient limitation, that might have increased the GC contents of aerobes or decreased the GC contents of anaerobes in some specific mechanisms unrelated with changes in oxygen requirement (Agashe and Shankar 2014; Lassalle, et al. 2015). If the guanine oxidation is not a very strong mutagenic force, its effect on the evolution of GC content would be easily overwhelmed by other factors. We further proposed that the relationship between oxygen requirement and GC content could be accurately assessed only when oxygen requirement is the sole GC-content-influencing factor differing between the compared lineages. Apparently, distantly related species are more likely to differ in more than one GC-content-influencing factors. As illustrated in fig. 1, besides the oxygen requirement, species 8 and species 9 are assumed to differ in the frequency of GC biased gene conversion, which is suggested as the driver of between-lineage differences in avian GC content (Weber, et al. 2014). The frequent GC biased gene conversion in species 8 might increase the GC content to an amount much larger than that is decreased by guanine oxidation. If so, the aerobic species 8 would have a higher GC content than the anaerobic species 9. For this reason, we measured the divergence time between each pair of lineages using the similarity of 16S rRNA molecules. Two cut-offs, 0.9 and 0.93 were artificially used to define closely related lineages. When all the paired lineages are closely related, the difference between aerobes and anaerobes

became easier to see (fig. 2C). The negative association becomes stronger when higher similarity of 16S rRNA molecules is used to define the close relatedness (fig. 2C). It should be noted that the 16S rRNA similarities of all the 6 pairs of archaea are smaller than 0.9, and so only bacterial genomes are presented in fig. 2C.

Within a pair of genomes, if one contains some newly duplicated transposable elements or sequences recently gained by horizontal gene transfer while the other not, a pairwise comparison of the whole-genome GC content could not accurately reflect the evolutionary forces differentially experienced by the two species. The ideal genomic parts for pairwise comparison are the sequences that have orthologous relationship. For this reason, we compared the GC contents of the orthologous protein-coding genes. The difference between aerobes and anaerobes is larger than that obtained in comparing whole-genome GC contents (Table 1). As a mutational force, the effect of guanine oxidation on the evolution of GC content would be more accurately revealed by analyzing the GC content of selectively neutral sequences or sequences under weak selection. Although the 4-fold degenerate sites might be under selection to maintain specific pattern of codon usage bias (Supek 2016), they are by far the most common candidates for the neutral or weakly selected sequences. At 4-fold degenerate sites, we consistently observed that aerobes have significantly lower GC contents than anaerobes, with the difference in median values much larger than that of whole-genome GC contents and the orthologous protein-coding genes (Table 1). We noticed that the variations of GC contents at 4-fold degenerate sites are larger in both aerobes and anaerobes than the variations of GC contents of whole genomes or orthologous genes. This could probably be attributed to the random noises resulting from the much smaller number of nucleotides counted as 4-fold degenerate sites than in whole genomes or orthologous genes, especially for the genomes that are only partially annotated.

Taken together, our phylogenetically independent comparison provided the evidence for a negative association between oxygen requirement and GC content. This is quite opposite to conventional comparisons that do not account for the effect of shared ancestry, carried out either by previous researchers (2002) or by ourselves (fig. 2A). In the dataset of Naya et al. (2002), the median values of aerobes and anaerobes are 62.0% and 45.0%, respectively. In our dataset, the median values of all aerobes and all anaerobes are 63.4% and 44.6%, respectively. By contrast, in the dataset of most closely related pairs (similarity of 16S rRNA > 0.93), the two median values are 41.5 and 44.2. It is obvious that the aerobic genomes with higher GC contents are oversampled in the conventional comparison. In future, it is interesting to study whether it is just a sampling bias in genome sequencing, or GC-rich aerobic prokaryotes have a higher rate of speciation and a higher species diversity than GC-poor aerobic prokaryotes and anaerobic prokaryotes.

Materials and Methods

From the Genomes Online Database (GOLD, https://gold.jgi.doe.gov/search)(Mukherjee, et al. 2017), we retrieved 3,555 aerobic samples (including 120 archaeal and 3,435 bacterial species or strains) and 2,339 anaerobic samples (including 142 archaeal and 2,197 bacterial species or strains). Among them, the genome sequences of 701 aerobic samples (622 bacteria and 79 archaea) and 929 anaerobic samples (860 bacteria and 69 archaea) were retrieved from the NCBI genome database (ftp://ftp.ncbi.nlm.nih.gov/genomes/). Orthologous genes were identified by the reciprocal best blast hits.

The All-Species Living Tree (Munoz, et al. 2011) was used to locate the phylogenetic positions of the studied prokaryotes. In the cases of polytomy or species/strains that are not present in the All-Species Living Tree, we construct the phylogenetic trees using the software MEGA (Kumar, et al. 2016) with the 16S rRNA.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant numbers 31671321, 31421063, 31371283).

References

Agashe D, Shankar N. 2014. The evolution of bacterial DNA base composition. *J Exp Zool Part B* 322:517-528.

Bohlin J, Snipen L, Hardy SP, Kristoffersen AB, Lagesen K, Dønsvik T, Skjerve E, Ussery DW. 2010. Analysis of intra-genomic GC content homogeneity within prokaryotes. *BMC Genomics* 11:464.

Decker H, Van Holde KE. 2011. Oxygen and the Evolution of Life. Heidelberg: Springer.

Delaney S, Jarem DA, Volle CB, Yennie CJ. 2012. Chemical and biological consequences of oxidatively damaged guanine in DNA. *Free Radical Res* 46:420-441.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat* 125:1-15.

Haywood-Farmer E, Otto SP. 2003. The evolution of genomic base composition in bacteria. *Evolution* 57:1783-1792.

Hildebrand F, Meyer A, Eyre-Walker A. 2010. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet* 6:e1001107.

Kanvah S, Joseph J, Schuster GB, Barnett RN, Cleveland CL, Landman U. 2010. Oxidation of DNA: damage to nucleobases. *Accounts Chem Res* 43:280-287.

Kelkar YD, Phillips DS, Ochman H. 2015. Effects of genic base composition on growth rate in G plus C-rich genomes. *G3 Genes Genom Genet* 5:1247-1252.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* 33:1870-1874.

Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V. 2015. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. *PLoS Genet* 11:20.

Luo HW, Thompson LR, Stingl U, Hughes AL. 2015. Selection maintains low genomic GC content in marine SAR11 lineages. *Mol Biol Evol* 32:2738-2748.

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, Thomas AD, Ali R, Sharma K, Kyrpides NC, et al. 2017. Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res* 45:D446-D456.

Munoz R, Yarza P, Ludwig W, Euzéby J, Amann R, Schleifer K-H, Oliver Glöckner F, Rosselló-Móra R. 2011. Release LTPs104 of the All-Species Living Tree. *Syst Appl Microbiol* 34:169-170.

Naya H, Romero H, Zavala A, Alvarez B, Musto H. 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J Mol Evol* 55:260-264.

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.

Rocha EPC, Feil EJ. 2010. Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS Genet* 6:e1001104.

Seward EA, Kelly S. 2016. Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biol* 17:226.

Supek F. 2016. The code of silence: widespread associations between synonymous codon biases and gene function. *J Mol Evol* 82:65-73.

Weber C, Boussau B, Romiguier J, Jarvis E, Ellegren H. 2014. Evidence for GC-biased gene conversion as a driver of between-lineage differences in avian base composition. *Genome Biol* 15:549.

Whitney KD, Garland T, Jr. 2010. Did genetic drift drive increases in genome complexity? *PLoS Genet* 6:e1001080.

Table 1. Different GC contents between Closely Related Aerobic and Anaerobic Prokaryotes.

	Aerobes (%)	Anaerobes (%)	*P*
Whole Genome	**41.5** (37.3 - 52.0)	**44.2** (40.4 - 55.3)	< 0.001
Orthologous Genes	**43.5** (39.4 - 54.5)	**49.7** (43.1 - 56.9)	< 0.001
4-Fold Degenerate Sites	**41.3** (27.9 - 72.1)	**54.7** (35.6 - 76.8)	< 0.001

NOTE.—The close relatedness was defined when the similarity of 16S rRNA between paired lineages is higher than 0.93. The same 30 pairs of lineages were used in all the three rows of comparisons. Wilcoxon signed ranks test was used in the comparisons. The median values (highlighted in bold characters) and the interquartile ranges (the first quartile – the third quartile) of GC contents are presented.

Fig. 1. An illustration for the difference between conventional comparison and phylogenetically independent comparison. In conventional comparison, all the aerobes

(including strain 2, species 1, species 2, species 3, species 4, species 7, and species 8) are compared with all the anaerobes (including strain 1, species 5, species 6, and species 9). However, there are only three changing events in oxygen requirement. The first occurred after the divergence of strain 1 from strain 2, the second occurred after the divergence of the common ancestor of species 5 and species 6 from species 4, and the third occurred after the divergence of species 8 and species 9. Only differences of GC content between these three paired lineages are probably associated with the changes in oxygen requirement. Therefore, only these three pairs indicated by the double arrow lines are included in the phylogenetically independent comparison. In this study, we use the median value of species 5 and species 6 to represent the value of node 1.

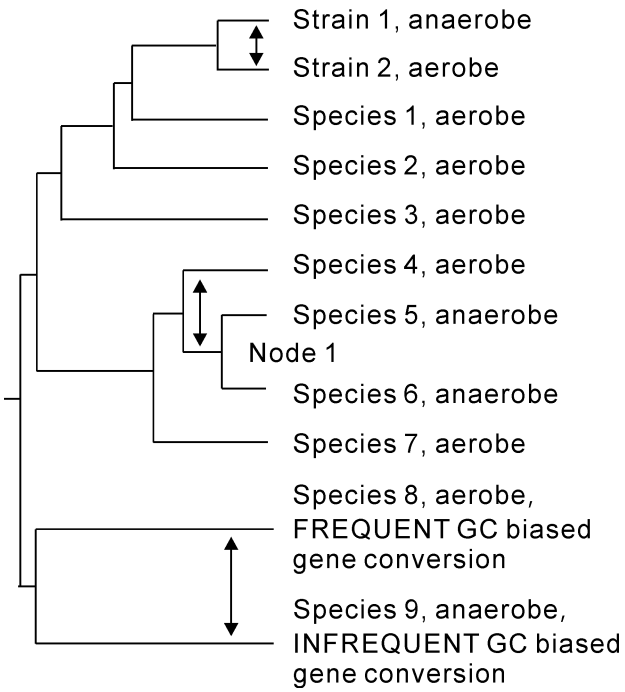
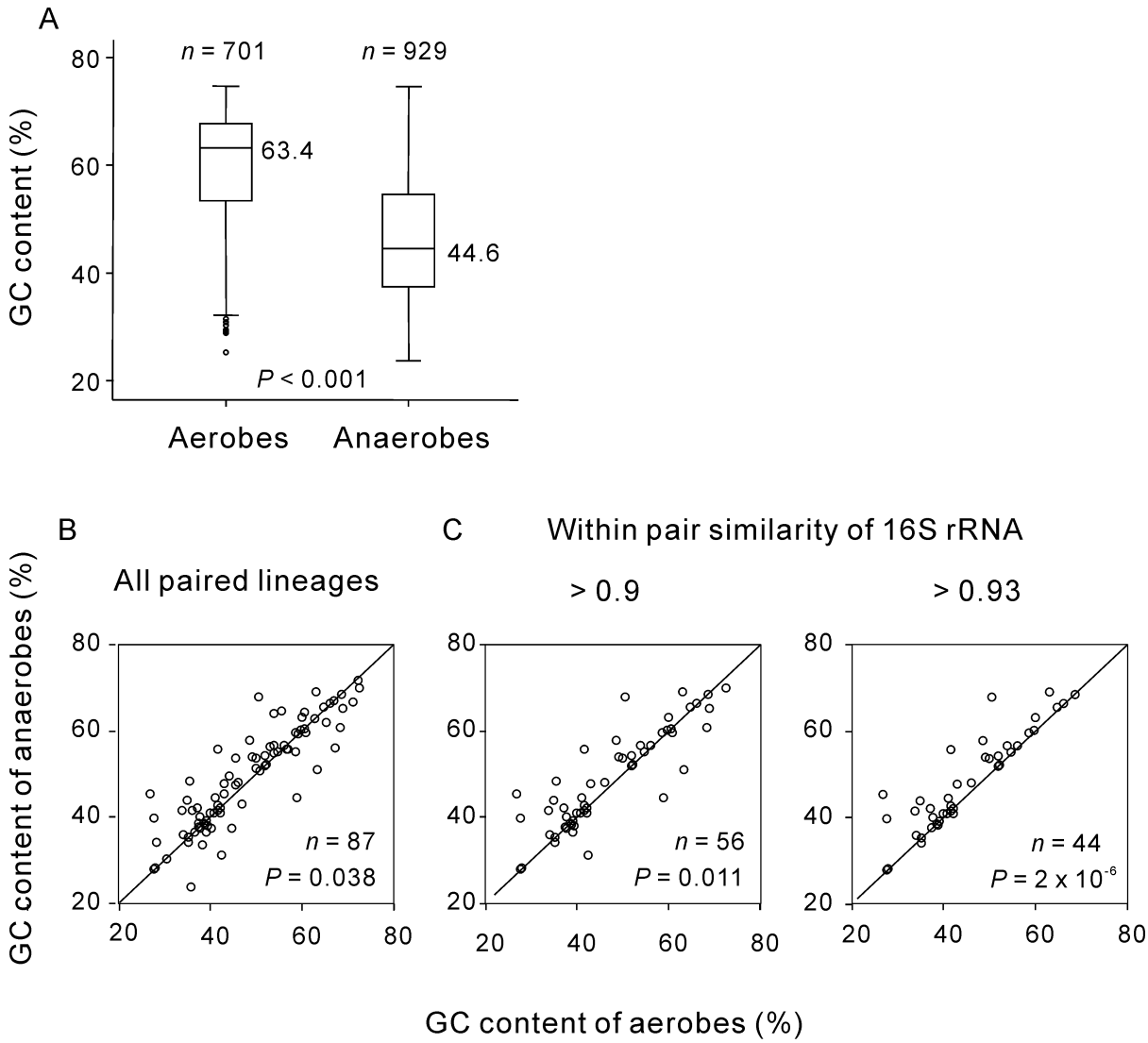


Fig. 2. Comparison of GC content between aerobic and anaerobic prokaryotes. (*A*) Conventional comparison that do not consider the phylogenetic non-independence of the samples indicates that aerobes have higher GC contents than anaerobes (Mann-Whitney *U* test, $P < 0.0001$). (*B*) Pairwise comparison of all the samples we collected gives a result that the GC contents of aerobic prokaryotes are significantly lower than those of anaerobic

prokaryotes. The diagonal line represents cases in which aerobes and its paired anaerobes have the same values of GC content. Points above the line represent cases in which anaerobes have higher GC contents than their paired aerobes while the points below the line indicate cases in which anaerobes have lower GC contents than their paired aerobes. (*C*) After discarding those pairs in which the partners are distantly related, pairwise comparison showed that aerobes have even lower GC contents than anaerobes. The divergence time between each pair of lineages was measured by the similarity of 16S rRNA molecules. The significant values in *B* and *C* were calculated using the Wilcoxon signed ranks test.