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

* Corresponding Author for: 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 unavoidable for oxygen-consuming organisms. Guanine is the most fragile base with regard to oxidative stress and thus the most frequently damaged among the four bases. The incorporation of the damaged guanines into DNA strands and the replication of DNA containing damaged guanines causes G to T mutations at a frequency depending on the efficiency of DNA repair enzymes and the accuracy of replication enzymes. For this reason, aerobiosis is expected to decrease GC content. However, the opposite pattern of base composition in prokaryotes was reported 15 years ago. Although that study has been widely cited, its omission of the effect of shared ancestry requires a re-examination of the reliability of its results. In this study, using phylogenetically independent comparisons, we found that aerobic prokaryotes have significantly lower whole-genome GC contents than anaerobic ones. When the GC content was measured only at 4-fold degenerate sites, the difference between aerobic prokaryotes and anaerobic prokaryotes was greater, which is consistent with the possibility of 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 was the most significant change in the evolution of the 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 aerobic organisms but also an unavoidable byproduct, reactive oxygen species (ROS). ROS are highly reactive with most cellular organic molecules, including nucleotides and their polymerized products, DNA and RNA. Among the four bases, guanine has the lowest oxidation potential and is the 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 further oxidation into several hyperoxidized products (Delaney, et al. 2012). The incorporation of the damaged nucleotides into DNA and the replication of DNA containing these damaged deoxyguanosines can 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). Whether 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, in which aerobic prokaryotes have higher GC contents than anaerobic prokaryotes, by comparing the GC contents using conventional 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 might shape the GC content throughout evolution (Hildebrand, et al. 2010; Rocha and Feil 2010; Raghavan, et al. 2012; Agashe and Shankar 2014; Luo, et al. 2015; Reichenberger, et al. 2015; Ochman and Caro-Quintero 2016; Seward and Kelly 2016; Bobay and Ochman 2017), and the effect of guanine oxidation might be trivial compared with other factors. However, such comparative triviality 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 (Mukherjee, et al. 2017a), it is now possible to collect enough data for 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 change in oxygen requirement, we obtained a pair of aerobic and anaerobic species, strains or nodes (fig. 1). The difference in 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 closely associated with changes in oxygen requirement. In total, we obtained 109 aerobe-anaerobe pairs, including 103 pairs of bacteria and 6 pairs of archaea.

For a study comparable to that of Naya et al. (2002), we first compared the GC contents of the 1044 aerobic genomes and the 952 anaerobic genomes without considering their positions in the phylogenetic tree. Despite the much larger dataset, we also observed significantly higher GC contents in aerobes than anaerobes (fig. 2A). Increases in sample size seem unlikely to change the result.

By contrast, in the pairwise comparison of the 109 aerobe-anaerobe pairs, aerobic prokaryotes have lower, not higher, GC contents than anaerobic ones. The difference is not great, but it is statistically significant (fig. 2B, Wilcoxon signed-rank test, $P = 0.002$). When the pairwise comparison is limited to the 103 pairs of bacteria, the difference between aerobes and anaerobes remains statistically significant (Wilcoxon signed-rank test, $P = 0.003$). However, in a regression analysis of 488 prokaryotic genomes, Bohlin et al. (2010) found no significant association between genomic GC content and oxygen requirement when phylogenetic bias was accounted for. There are many other factors, such as 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 by specific mechanisms unrelated to changes in oxygen requirement (Agashe and Shankar 2014; Lassalle, et al. 2015). If 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 was the sole GC-content-influencing factor differing between the compared lineages. Clearly, distantly related species are more likely to differ in multiple GC-

content-influencing factors. As illustrated in fig. 1, in addition to 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 much greater degree than that by which guanine oxidation decreases it. If so, then 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. An artificial cutoff of 0.93 was used to define closely related lineages. When all the paired lineages are closely related, the difference between aerobes and anaerobes becomes very easy to see (fig. 2C). By contrast, when all the paired lineages are distantly related, there is no significant difference between aerobes and anaerobes (fig. 2D, Wilcoxon signed-rank test, $P = 0.196$). It should be noted that the 16S rRNA similarities of all 6 pairs of archaea are smaller than 0.93, and therefore, only bacterial genomes are presented in fig. 2C.

If one of a pair of genomes contains some newly duplicated transposable elements or sequences recently gained by horizontal gene transfer while the other does not, a pairwise comparison of the whole-genome GC content cannot accurately reflect the evolutionary forces differentially experienced by the two species. The ideal genomic regions for pairwise comparison are sequences with an orthologous relationship. For this reason, we compared the GC contents of 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 patterns of codon usage bias (Supek 2016), they are by far the most common candidates for neutral or weakly selected sequences. At 4-fold degenerate sites, we consistently observed that aerobes have significantly lower GC contents than anaerobes, with a much greater difference in median values than those of whole-genome GC contents and orthologous protein-coding genes (Table 1). We noticed that the variations in GC content at 4-fold degenerate sites are greater in both aerobes and anaerobes than the variations in the GC content of whole genomes or orthologous genes. This increased variation could probably be attributed to the random noise resulting from the much smaller number of nucleotides counted as 4-fold degenerate sites than in whole genomes or orthologous genes, especially for genomes that are only partially annotated.

Taken together, our phylogenetically independent comparison provided evidence for a negative association between oxygen requirement and GC content. This result is quite opposite to conventional comparisons that do not account for the effect of shared ancestry, conducted either by previous researchers (2002) or by ourselves (fig. 2A). In the dataset of Naya et al. (2002), the median values for aerobes and anaerobes are 62.0% and 45.0%, respectively. In our dataset, the median values for all aerobes and all anaerobes are 62.3% and 45.0%, respectively. By contrast, in the dataset of the most closely related pairs (similarity of 16S rRNA > 0.93), the two median values are 41.5% and 44.2%. It is obvious that aerobic genomes with higher GC contents are oversampled in the conventional comparison. In the future, it will be interesting to study whether this oversampling is merely 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. 2017b), 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). We retrieved the GC contents from the Summary section of the homepage of each species or strain in the NCBI Genome database. The genome sequences of the aerobe-anaerobe paired lineages 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 GC contents used in pairwise comparison were calculated from the downloaded genome sequences rather than those retrieved directly from the NCBI Genome database. Although the GC contents from these two sources are not identical, they are highly similar. The regression equation is $y = 1.0012x - 0.0218$ and the R^2 is 0.9972.

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

The data used in fig. 2 are presented in supplementary file S1, Supplementary Material online.

Acknowledgments

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.

Bobay L-M, Ochman H. 2017. Impact of recombination on the base composition of Bacteria and Archaea. *Mol Biol Evol*:10.1093/molbev/msx1189.

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.

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:e1004941.

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, Seshadri R, Varghese NJ, Eloe-Fadrosh EA, Meier-Kolthoff JP, Goker M, Coates RC, Hadjithomas M, Pavlopoulos GA, Paez-Espino D, et al. 2017a. 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. *Nat Biotechnol* advance online publication:10.1038/nbt.3886.

Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, Thomas AD, Ali R, Sharma K, Kyrpides NC, et al. 2017b. 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.

Ochman H, Caro-Quintero A. 2016. Genome size and structure, bacterial. In: Kliman RM, editor. *Encyclopedia of Evolutionary Biology*. Oxford: Academic Press. p. 179-185.

Raghavan R, Kelkar YD, Ochman H. 2012. A selective force favoring increased G+C content in bacterial genes. *Proc Natl Acad Sci USA* 109:14504-14507.

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. Differences in GC Content 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.—Close relationship was defined when the similarity of 16S rRNA between paired lineages was higher than 0.93. The same 30 pairs of lineages were used in all three rows of comparisons. The Wilcoxon signed-rank 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 content are presented.

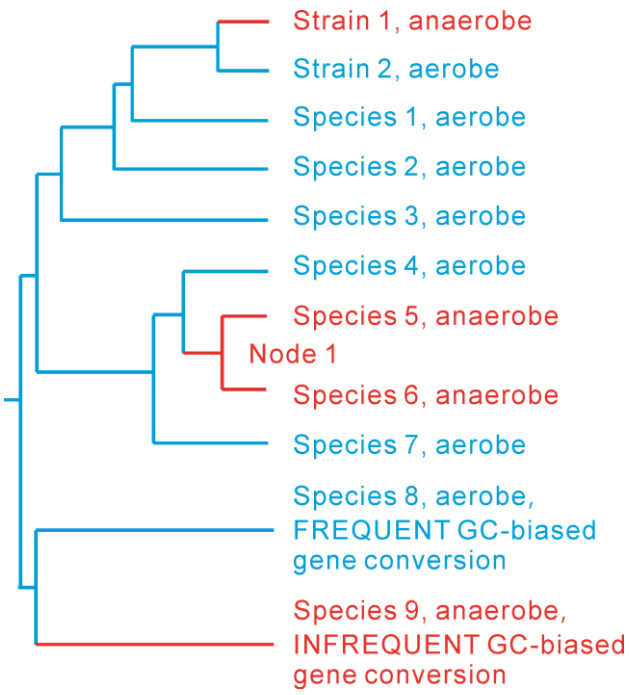

Fig. 1. An illustration of the difference between conventional comparison and phylogenetically independent comparison. In conventional comparison, all the aerobes (including strain 2, species 1, species 2, species 3, species 4, species 7, and species 8) are compared with all the anaerobes (including strain 1, species 5, species 6, and species 9). However, there are only three changes in oxygen requirement, as shown in red color. Only the differences in GC content between the three paired lineages (strain 1 *vs*. strain 2, species 4 *vs*. the common ancestor of species 5 and species 6, and species 8 *vs*. species 9) are likely to be associated with the changes in oxygen requirement. Therefore, only these three pairs should be included in a phylogenetically independent comparison. In this study, we use the average value of species 5 and species 6 to represent the value of node 1.

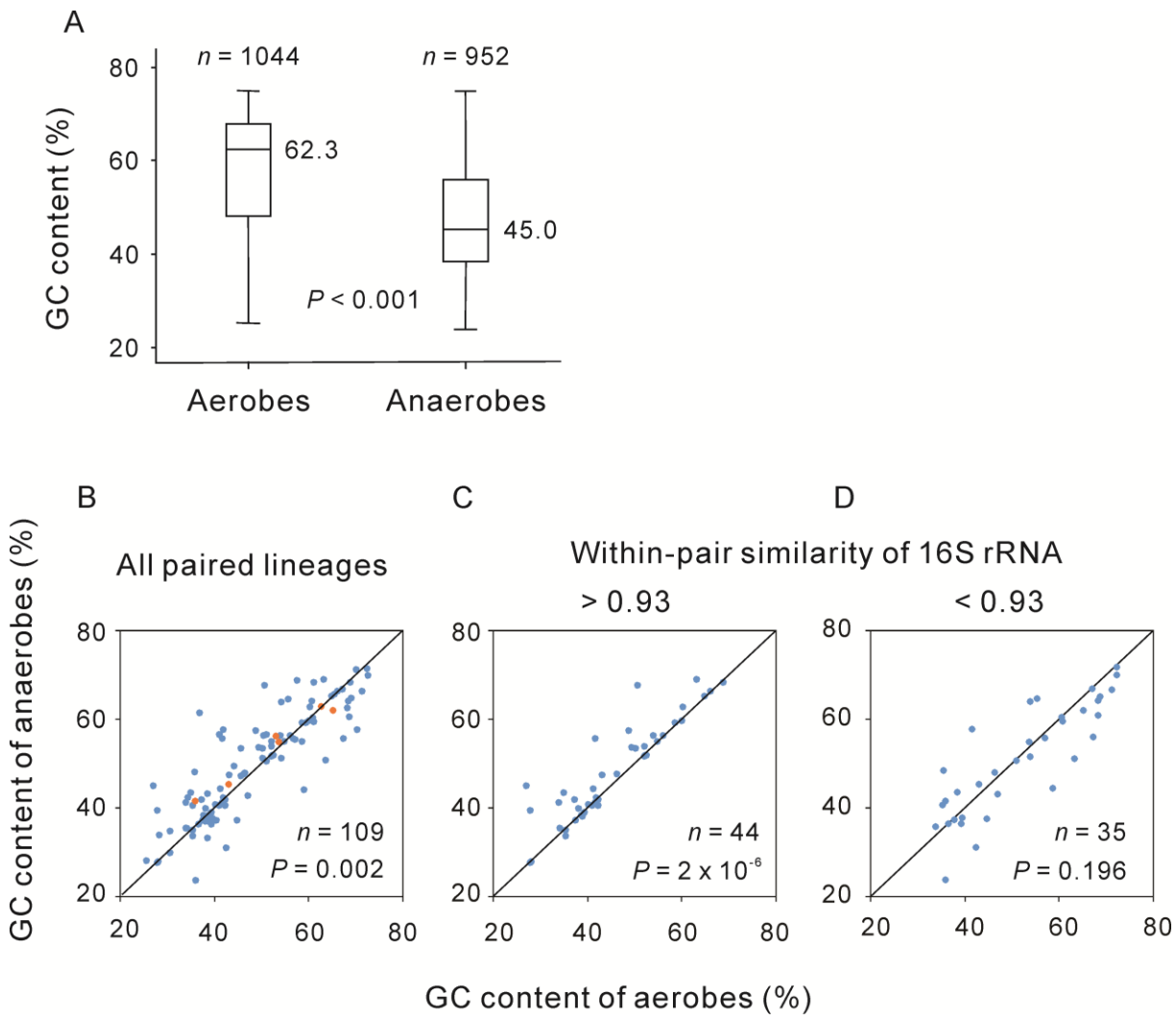

Fig. 2. Comparison of GC content between aerobic and anaerobic prokaryotes. (*A*) Conventional comparison, which does not consider the phylogenetic non-independence of the samples, indicates that aerobes have higher GC contents than anaerobes (Mann-Whitney *U* test, $P = 10^{-88}$). For species with multiple strains that are consistent in oxygen requirement, we used the average value of the strains to represent the GC content of the species. In species including both aerobic strains and anaerobic strains, the average values of GC contents of aerobic strains and anaerobic strains were calculated separately. (*B*) Pairwise comparison of all the samples we collected indicates that the GC contents of aerobic prokaryotes are significantly lower than those of anaerobic prokaryotes. The six pairs of archaea are marked in red. The diagonal line represents cases in which aerobes and their paired anaerobes have the same GC contents. 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*) When the compared aerobes and anaerobes are closely related, the difference in GC content becomes larger. (D) However, when the compared species are distantly related, no significant difference exists between the GC contents of aerobic and anaerobic bacteria. An artificial cutoff of 0.93 in the similarity of 16S rRNA molecules was used to separate close relatedness and distant relatedness. The significance values in *B*, *C* and *D* were calculated using the Wilcoxon signed-rank test.