Bacterial genomic GC content drifts slowly downward due to a biased mutation rate

Bert Ely*

Department of Biological Sciences, University of South Carolina, Columbia, SC, 29203 USA

*Correspondence to: ely@sc.edu

803-777-2768

Key words: genomic GC content, mutation rate, mutation bias, prokaryotes, Caulobacter,

Abstract

In every kingdom of life, GC->AT transitions are the predominant type of mutation that accumulates over evolutionary time periods. In eukaryotic genomes, this slow loss of GC base pairs is counteracted by biased gene conversion which increases genomic GC content as part of the recombination process. However, this type of biased gene conversion has not been observed in bacterial genomes so we hypothesized that GC->AT transitions cause a reduction of genomic GC content in prokaryotic genomes. To test this hypothesis, we used a phylogenetic approach to analyze triplets of closely related genomes representing a wide range of the bacterial kingdom. The resulting data indicate that genomic GC content is slowly declining in bacterial genomes where GC base pairs comprise 40% or more of the total genome. In contrast, genomes containing less than 40% GC base pairs have fewer opportunities for GC->AT transitions to occur so genomic GC content is relatively stable or actually increasing at a slow rate.

Introduction

Bacterial genomic GC content varies from less than 15% to more than 75% [1]. Furthermore, a wide range of genomic GC content can be observed among the genera within most bacterial phyla. However, genomic GC content usually varies in a narrow range within a single bacterial genus indicating that genomic GC content is relatively constant over evolutionary time scales [1]. In contrast, mutations that occur in bacterial genomes accumulate over time causing changes that have led to the diversification of individual species within a genus. Since mutations that convert GC base pairs to AT base pairs (GC->AT) occur more frequently than any other type of mutation primarily due to cytidine deamination [2], mutations that accumulate over evolutionary time periods could lead to a reduction of genomic GC content in all bacteria. Thus, several hypotheses have been proposed to explain how bacteria are able to maintain genomes with high percentages of GC base pairs despite the mutation pressure [2-6]. For example, the impact of the conversion of GC base pairs to AT base pairs could be counteracted by the process of horizontal gene transfer (HGT) to import GC-rich DNA segments and maintain the overall GC content of the genome. However, this type of HGT would result in patches of imported DNA with higher GC content while the rest of the genome had a continually decreasing GC content. In addition, there would have to be some sort of selection for the import of GC-rich segments into the bacterial genome. In contrast, several analyses have shown that most imported genomic regions obtained via HGT in a variety of bacteria come from closely-related genomes with similar GC content [7-10]. Also, careful analyses of closely related *Caulobacter vibrioides* genomes failed to reveal imported regions with higher than average GC content [11]. Therefore, HGT is not likely to be the primary way that high average GC content is maintained in bacterial genomes.

An alternative hypothesis, biased gene conversion, suggests that AT base pairs are converted to GC base pairs in mismatched regions of DNA that are generated during recombination events. Since this phenomenon has been well established for eukaryotic genomes [12-13], it has been considered a

reasonable hypothesis, and some data have been presented that are consistent with the idea that biased gene conversion has occurred in bacterial genomes [14]. If biased gene conversion did occur in bacteria, then recombination events could help maintain a constant genomic GC content by causing AT-> GC transitions in sufficient quantity to compensate for the excess GC->AT transitions. However, biased gene conversion could only occur during relatively rare recombination events [11] while eukaryotic recombination occurs multiple times during every round of meiosis. Furthermore, we have shown that biased gene conversion did not occur during more than 100 distinct recombination events analyzed in the high GC genomes of *C. vibrioides* [15]. Thus, biased gene conversion does not appear to occur during recombination events in the high GC genomes of this bacterial species.

If recombination and HGT do not impact genomic GC content, then how is high GC content maintained in bacterial genomes? One proposed explanation is that there is some type of selection for higher genomic GC content [2-6]. However, although many possibilities have been ruled out, no one has identified a selective mechanism that would increase genomic GC content. As an alternative, we hypothesized that genomic GC content actually is decreasing due to the GC->AT mutation pressure, and as a result, there would be no need for a separate mechanism to increase genomic GC content. To test this hypothesis, we conducted a series of phylogenetic analyses of mutations that have accumulated over millions of generations in a wide variety of prokaryotic genera. These analyses demonstrated that prokaryotic genomes with high or average levels of GC content are slowly evolving to lower levels of GC content, but the pace of this change is so slow that genomic GC content appears to be constant.

Methods

Genome comparisons

A wide range of representative prokaryotic species were chosen based on the availability of three closely-related and complete genome sequences in the GenBank database. If available, we chose

sets of three closely-related genomic sequences from two genera from each bacterial phylum as well as several from each class of the *Proteobacteria* (Table 1). Each set of three related genomes was aligned with progressiveMauve using standard settings, and the positions of all single nucleotide polymorphisms (SNPs) were exported as a single file that showed which base was present in each of the three genomes at the variable locations [16]. Using a phylogenetic approach, the SNPs that differed between the two more closely related genomes were compared, and the one that matched the third, more distal genome was considered the ancestral allele. Positions that contained three different alleles were omitted from the analysis since the ancestral state could not be determined. Also, genome comparisons that generated fewer than 1000 SNPs with defined ancestral alleles were omitted from further analyses due to the small sample size. Each combination of the ancestral and derived alleles was tallied over the whole genome, and the appropriate combinations were summed to generate total GC->AT transitions, total AT->GC transitions, and the total of the remaining SNPs that resulted in no change in GC content. The number of AT->GC SNPs was subtracted from the number of GC->AT SNPs to determine the net loss of GC base pairs in the genome. (Therefore, a negative number would indicate an increase in genomic GC content). The net change in the number of GC base pairs was divided by the size of the bacterial genome to determine the percentage change in the genomic GC content since the time when the two bacterial strains shared a common ancestor. Also, the net change of genomic GC content was divided by the total number of SNPs analyzed to facilitate comparisons of the change in GC content among the analyzed species.

Results and Discussion

Since our laboratory specializes in C*aulobacter* genetics, we first analyzed closely related *C. vibrioides* genomes that had been recently sequenced [11]. To determine if genomic GC content has remained constant, we identified SNPs by comparing the CB2 and NA1000 genomes and then

determining the direction of mutation by comparison to the more distantly-related CB13 genome [11]. Using this approach, we were able to determine the mutational direction of approximately 23,000 SNPs that had occurred in each genome (Table 2). As expected, we found that GC->AT changes were the most common (49% of all mutations) while AT->GC mutations occurred at a frequency of only 21%. Thus in both strains, there has been a net loss of GC base pairs that would have caused a 0.15% reduction in the genomic GC content during the time period since their most recent common ancestor, which is estimated to have existed 25 million generations ago [11]. To ensure that these calculations were not influenced by the choice of the reference genome, we repeated the analysis using the more distantly-related *C. segnis* TK0057 genome [11, 17] as the reference genome and found a 0.2% reduction of the genomic GC content since CB2 and NA1000 shared a common ancestor (data not shown). Although the magnitude of the genomic GC content reduction is similar in the two analyses, the first estimate is probably more accurate since we were able to determine the ancestral state of 50% more SNPs when the more closely-related CB13 genome was used as the reference. Taken together, these data indicate that the GC content of the *C. vibrioides* NA1000 and CB2 genomes has been slowly drifting downward as a result of the accumulation of GC->AT transitions at a net rate of approximately one fewer GC base pair per 4000 generations.

These results contrast with previous results that showed that AT->GC mutations were the most common types of mutations in *C. vibrioides* [1819]. However, in these previous studies, the number of mutations was both small and limited by selection for a particular phenotype (rifampicin resistance or inactivation of the XylR repressor). Thus, the mutations obtained under the conditions used in these two studies do not appear to be representative of the mutations that accumulate throughout the entire genome over evolutionary time periods.

To determine if slow losses of GC content also had occurred in other types of bacteria, we performed similar analyses using the genomes of 28 bacterial species from 11 different phyla as well as

those of two archaeal genera (Tables 2 and S1). Two simultaneous trends were observed: 1) Greater loss of GC content occurred when the genomic GC content was higher (Fig. 1A); and 2) greater loss of GC content occurred when the genetic distance between the two genomes was greater (represented by the total number of SNPs observed between the two genomes being compared) (Fig. 1B). Both of these trends are consistent with expectations. A larger number of SNPs indicates that a longer time had elapsed since the two strains shared a common ancestor so that there would have been more time for an excess of GC ->AT transitions to accumulate. For example, the six comparisons with the highest numbers of SNPs as a percentage of the genome also were the farthest from the Y axis indicating that they had experienced the greatest changes in GC content (Fig. 1B). Similarly, a higher GC content would provide more opportunities for cytidine deamination to occur. Thus, for genomes with a GC content greater than 40%, there was a trend towards increasing loss of GC content after correcting the total number of mutations relative to the size of the genome (Fig. 1A). Based on these results, we conclude that a slow decrease in genomic GC content is a universal phenomenon for prokaryotic genomes with a genomic GC content that is greater than 40%.

When we analyzed prokaryotic genomes with a genomic GC content that was less than 40%, six of the 16 genome triplets had a surplus of AT->GC transitions resulting an increase in genomic GC content, and the other 10 had a net decrease in genomic GC content that was less than 0.06% (Table 1). Thus genomic GC content seems to be more stable when the percentage of GC base pairs is less than 40% of the total genome. However, in each case, the mutation rates for the GC->AT transitions was higher per GC base pair than the corresponding rate of AT->GC transitions per AT base pair. These data provide further evidence that the conversion of GC base pairs to AT base pairs occurs at a higher rate in all prokaryotic genomes regardless of GC content. Thus, cytidine deamination is probably the most important factor contributing to the higher rate of GC->AT transitions, but the impact of these

transitions on genomic GC content is dependent on the percentage of GC base pairs present in the genome.

If the genomic GC content is slowly reduced due to GC->AT mutation pressure, then how did high GC bacterial genomes arise? Our previous analysis of bacterial phyla demonstrated that genomic GC content was fairly constant in most genera, but wide variations occurred within phyla [1]. Thus, during the evolution of bacterial genera, there may have been a period of time when conditions favored divergent genomic evolution patterns, and the mutation pressure and selection that occurred under those conditions might have resulted in the differentiation of genomic GC content among genera within the same phylum. We cannot assume that conditions have not changed since the origin of life on this planet even though bacterial genomes seem to have evolved at a constant rate in recent evolutionary times. We have records of many events that led to mass extinctions of eukaryotic species. Perhaps some of these events or even earlier unknown events greatly impacted bacterial genome evolution as well.

Acknowledgements

I thank Dakota Lee for helping with many of the genome comparisons and Louis Berrios for constructive criticism of the manuscript. This work was funded in part by National Institutes of Health grant GM076277 to BE.

References

1. Lightfield J, Fram NR, Ely B (2011) Across bacterial phyla distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. PLoS ONE 6:e17677. doi10.1371/journal.pone.0017677
2. Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, et al. (2018) Evolutionary determinants of genome-wide nucleotide composition. Nat Ecol Evol 2:237-240. doi:10.1038/s41559-017-0425-y
3. Bohlin J, Eldholm V, Brynildsrud O, Petterson JH-O, Alfsnes K (2018) Modeling of the GC content of the substituted bases in bacterial core genomes. BMC Genomics 19:589. doi:10.1186/s12864-018-4984-3
4. Hershberg R, Petrov DA (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6(9):e1001115. doi:10.1371/journal.pgen.1001115
5. Hildebrand F, Meyer A, Eyre-Walker A (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6(9): e1001107. doi:10.1371/journal.pgen.1001107
6. Hershberg R (2015) Mutation - The engine of evolution: Studying mutation and its role in the evolution of bacteria. Cold Spring Harb Perspect Biol 7(9):a018077. doi:10.1101/cshperspect.a018077
7. Bolotin E, Hershberg R (2017) Horizontally acquired genes are often shared between closely related bacterial species. Frontiers Microbiol 8:1536. doi:10.3389/fmicb.2017.01536
8. Dixit PD, Pang TY, Studier FW, Maslov S (2015) Recombinant transfer in the *E. coli* basic genome. Proc Nat Acad Sci U.S.A. 112:9070-9075 doi:10.1073/pnas.1510839112
9. Fischer S, Klockgether J, Moran Losada P, Chouvarine P, Cramer N, et al. (2016) Intraclonal genome diversity of the major *Pseudomonas aeruginosa* clones C and PA14. Environmental Microbiol Rpts 8:227–234. doi:10.1111/1758-2229.12372

10. Friedman R, Ely B (2012) Codon usage methods for horizontal gene transfer detection generate an abundance of false positive and false negative results. Curr Microbiol 65:639-642. doi:10.1007/s00284-012-0205-52
11. Ely B, Wilson K, Ross K, Ingram D, Lewter T, Herring J, Duncan D, Aikins A, Scott D (2018) Genome evolution observed in wild isolates of *Caulobacter crescentus*. CurrMicrobiol 76: 159-167. doi:10.1007/s00284-018-1606-x
12. Halldorsson BV, Hardarson MT, Kehr B, StyrkarsdottirU, Gylfason A, Thorleifsson G, et al. (2016) The rate of meiotic gene conversion varies by sex and age. Nat Genet 48**:**1377-1384. doi:10.1038/ng.3669
13. Halldorsson BV, Palsson G, Stefansson OA, Jonsson J, Hardarson MT, Eggertsson HP, et al. (2019) Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363:eaau1043. doi:10.1126/science.aau1043
14. Lassalle F, Périan 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(2): e1004941. doi:10.1371/journal.pgen.1004941
15. Ely B. (2020) Recombination and gene loss occur simultaneously during bacterial horizontal gene transfer. PLoS ONE 15(1): e0227987. doi:10.1371/journal.pone.0227987
16. Darling AE, Mau B, Perna NT (2010) progressiveMAUVE: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 5:e11147. doi:10.1371/journal.pone.0011147
17. Patel S, Fletcher B, Scott DC, and Ely B (2015) Genome sequence and phenotypic characterization of *Caulobacter segnis*. Current Microbiology 70: 355-363. doi:10.1007/s00284-014-0726-1.

18. Martins-Pinheiro M, Oliveira AR, Valencia AO, Fernandez-Silva FS, Silva LG, Lopes-Kulishev CO, Italiani VCS, Marques MV, Menck CF, Galhardo RS (2017) Molecular characterization of *Caulobacter crescentus* mutator strains. Gene 626:251-257. doi:10.1016/j.gene.2017.05.038.
19. Valencia AO , Braz VS, Magalhães M, Galhardo RS (2020) Role of error-prone DNA polymerases in spontaneous mutagenesis in *Caulobacter crescentus*. Genet Molec Biol 43:1 e20180283. doi:10.1590/1678-4685-gmb-2018-0283

Table 1. Prokaryotic species used in this study.

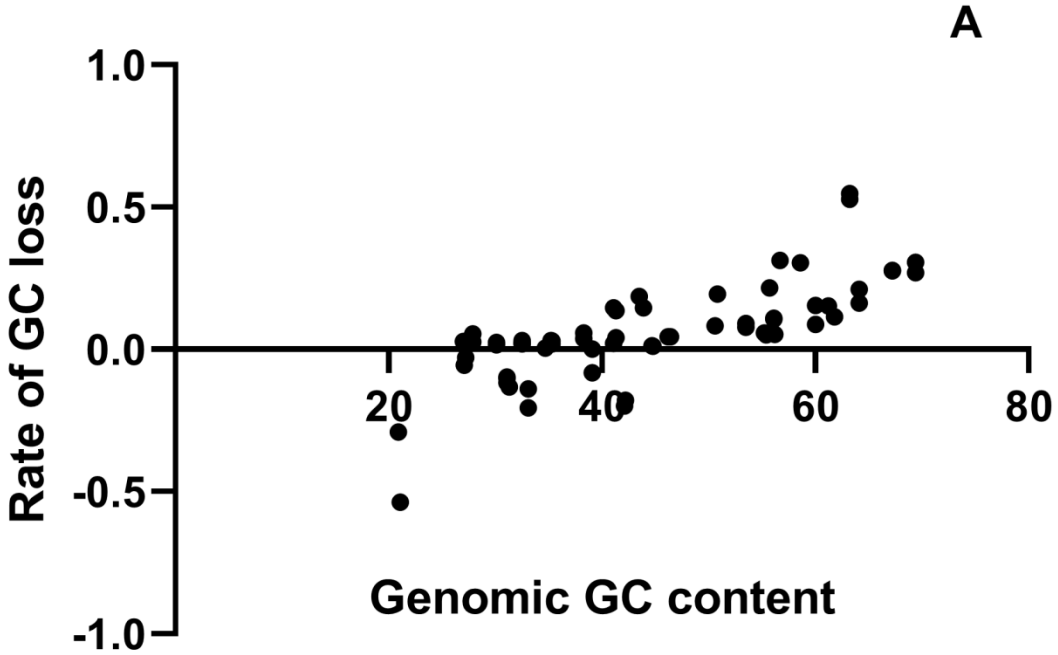
Phylum	Species	Proteobacteria	Species
Actinobacteria	***Corynebacterium diphtheriae***	***Alphaproteobacteria***	***Caulobacter vibrioides***
	Cutibacterium acnes		***Candidatus Sulcia muelleri***
Bacteroidetes	***Flavobacterium psychrophilum***		***Ehrlichia chaffeensis***
	Rhodothermus marinus		***Ochrobactrum anthropi***
Chlamydiae	***Chlamydia pecorum***	***Betaproteobacteria***	***Polynucleobacter asymbioticus***
	Chlamydia trachomatis	***Gammaproteobacteria***	***Haemophilus influenza***
	Chlamydia psittaci		***Pseudomonas putida***
	Chlamydia suis		***Escherichia coli***
Cyanobacteria	***Prochlorococcus marinus***	***Deltaproteobacteria***	***Desulfovibrio vulgaris***
	Synechococcus sp.		
Deinococcus-Thermus	***Thermus thermophilus***		
Firmicutes	***Bacillus cereus***		
	Bacillus subtilis		
Fusobacteria	***Fusobacterium necrophorum***		
	Fusobacterium hwasookii		
Spirochaetes	***Brachyspira pilosicoli***		
	Brachyspira hyodysenteriae		
Thermotogae	***Thermotoga***		
Verrucomicrobia	***Akkermansia muciniphila***		
Archaea	***Methanococcus maripaludis***		
	Thermococcus sp.		

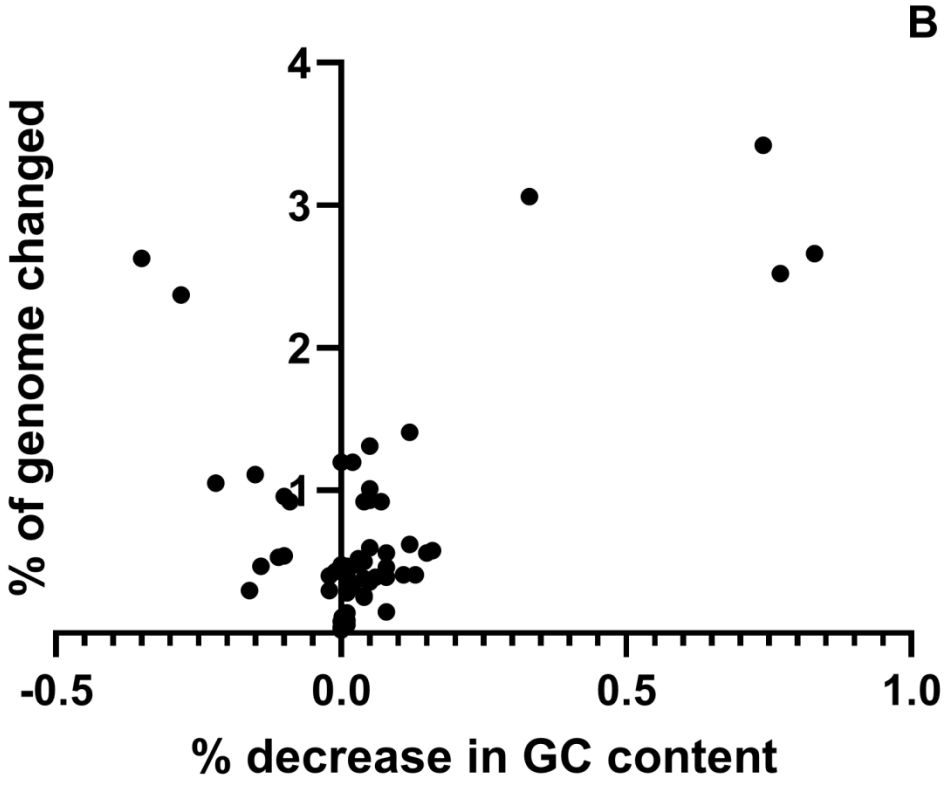
Table 2. Loss of GC content in *Alphaproteobacteria.*

Phyla and Bacteria	Net Loss of GC Content	Total Mutations	Genome Size (Mb)	Genomic %GC Content	Percent Change in GC	SNP % of Genome	Loss of GC/SNPs
Caulobacter vibrioides							
C. vibrioides NA1000	6420	23250	4.0	67.7	0.16	0.58	0.276
C. vibrioides CB2	6310	22793	4.1	67.2	0.15	0.56	0.277
C. vibrioides CB13			4.1	67.1			
Candidatus Sulcia muelleri							
C. Sulcia muelleri CARI	-439	817	0.28	21.1	-0.16	0.30	-0.537
C. Sulcia muelleri PSPU	-392	1346	0.29	20.9	-0.14	0.47	-0.291
C. Sulcia muelleri DMIN			0.24	22.5			
Ehrlichia chaffeensis							
E. chaffeensis Liberty	13	517	1.18	30.1	0.00	0.04	0.025
E. chaffeensis Jax	7	449	1.18	30.1	0.00	0.04	0.016
E. chaffeensis Heartland			1.17	30.1			

Figure

Fig. 1 A) Net change in genomic GC content divided by the total number of SNPs analyzed versus the genomic GC content. B) Change in genomic GC content (X axis) versus the number of SNPs analyzed relative to the size of the genome.


A
1.0
0.5
0.0
-0.5
-1.0
Rate of GC loss
20
40
60
80
Genomic GC content


B
4
3
2
1
% of genome changed
-0.5
0.0
0.5
1.0
% decrease in GC content