Potential link between selection for high GC content and repair of double strand breaks in prokaryotic genomes

Jake L. Weissman, William F. Fagan, Philip L.F. Johnson

February 8, 2019

Abstract

Genomic GC content varies widely among microbes for reasons unknown. While mutation bias partially explains this variation, prokaryotes near-universally have a higher GC content than predicted solely by this bias. Debate surrounds the relative importance of the remaining explanations of selection versus biased gene conversion favoring GC alleles. Some environments (e.g. soils) are associated with a high genomic GC content of their inhabitants, which implies that this content may be a selective adaptation to particular habitats. Here, we report a novel association between the presence of the non-homologous end joining DNA double-strand break repair pathway and GC content; this observation suggests that high GC content may be an adaptation to facilitate repair of double strand breaks when homologous recombination is not possible. We discuss potential mechanisms accounting for the observed association, and provide preliminary evidence that sites experiencing higher rates of double-strand breaks are under selection for increased GC content relative to the genomic background.

Prokaryotic genomes vary widely in their GC content, from the small genomes of endosymbionts with low GC content (as low as 16% [1]) to the larger genomes of soil dwelling microbes with high GC content ($> 60\%$ [2, 3]). A puzzle arose when observational studies surprisingly revealed a GC→AT mutational bias in genomes with GC content $> 50\%$ [4, 5]; more recently, controlled mutation accumulation experiments showed that even genomes with $< 50\%$ GC have greater GC content than expected from mutation rates [6]. This discrepancy between mutation rates and GC content implies that GC alleles fix at a higher rate than AT alleles. Two mechanisms could lead to biased fixation: selection directly on GC content [4, 5] or biased gene conversion (BGC), wherein homologous recombination favors GC alleles when resolving heteroduplex DNA mismatches [7].

Patterns of genetic diversity by themselves cannot easily differentiate between these two hypotheses [8, 9]; however, the addition of phenotypic information provides a tantalizing clue that GC content correlates more strongly with

shared environmental factors [2, 3, 10, 11] than with phylogenetic similarity [3]. This environmental correlation suggests there is selection for GC content based on habitat, although the fitness advantage of high/low GC content is unknown [8].

Many environments containing high GC microbes, such as soils and aerobic environments [3, 10], induce relatively high rates of DNA damage in the form of double-strand breaks (DSB) that necessitate repair [12, 13]. For instance, in aerobes, this damage results from the reactive oxygen species produced during metabolism, and in soil-dwelling microbes, this damage is associated with dessication and spore formation [14, 15, 16]. DSB repair primarily proceeds through one of two pathways: homologous recombination (HR), and non-homologous end joining (NHEJ) [13]. HR machinery is ubiquitous across microbes [17], but requires multiple genome copies to function. In contrast, the NHEJ repair pathway is rarer and generally found in organisms experiencing DSBs with only a single copy of the genome present in the cell (e.g. during an extended stationary phase; [15]). NHEJ repair requires the presence of the highly conserved Ku protein [18, 19], which makes Ku presence/absence a useful indicator of genomes with high rates of DSBs during especially vulnerable periods (i.e. one genome copy present; assuming genomes with Ku have a functional NHEJ pathway). In fact, many of the environments and lifestyles with which Ku is associated are also associated with high GC content (soil-dwelling, aerobicity, nitrogen fixation and spore formation; S1 Fig).

We downloaded all available completely sequenced prokaryotic genomes from NCBI's non-redundant RefSeq database and searched for the presence of the gene coding for Ku using a profile hidden Markov model (S1 Text; [20]). We found that the presence of Ku is strongly associated with genomic GC content (Fig 1). This association remains highly significant after correcting for phylogeny (phylogenetic linear regression [21], $m = 1.29$, $p = 1.465 \times 10^{-13}$; S2 Text, S1 Table). This association also holds when restricting to fourfold degenerate sites (S2 Fig). Examining mutation accumulation experiments in detail, Ku shows no effect on GC$\leftrightarrow$AT mutational biases (S3 Fig; data from [6]). Similarly, we saw no positive association between Ku and the rate of homologous recombination, making these differences unlikely to be caused by BGC (S4 Fig; data from [22, 23]). This is consistent with recent work that shows that BGC cannot account for elevated GC content over expectations across bacteria [24]. Additionally, using inferred polymorphism data (S3 Text; [25]), and assuming that recent polymorphisms have not yet been influenced by selection and can thus be used to estimate the GC mutational bias, we saw increased selection for high GC content in genomes with Ku (S5 Fig). Together, these analyses suggest that genomes with Ku are under selection to have higher GC content.

In fact, high GC content may serve a similar, and somewhat complementary role to the Ku protein itself. When DSBs form with high GC overhangs or microhomologies, organisms frequently employ alternative, high-accuracy end-joining repair pathways that do not suffer from the same error rates as NHEJ with Ku [26, 27, 28]. It is thought that the high GC content can help to tether the DNA together in place of Ku which typically performs that function [27].

Thus high GC content could help to ameliorate the negative effects of DSBs, especially in cases where there are many DSBs but only a single genome copy. This mechanism might also explain why longer genomes tend to have higher GC content than shorter ones [29]. The selective advantage of high GC content would be stronger in longer genomes, where the number of breaks per generation is likely to be higher. All breaks must be repaired for replication to proceed, so the negative fitness effects of DSBs are felt on a per-genome rather than per-base level. We observed that Ku presence and genome length both had independent positive effects on genomic GC content (S2 Text, S1 Table, S6 Fig, and within phyla S7 Fig).

Several predictions arise from this hypothesized mechanism. We would expect that regions of the genome that are especially prone to DSBs would be under selection to have higher GC content. Restriction modification (RM) systems offer a convenient test case because damage due to self-targeting is a known phenomenon (e.g. [30]) and the potential locations of self-targeting will be known if the recognition sequence of the restriction enzyme is known. We hypothesized that restriction enzymes on a genome would be selected to target sites higher in GC content than expected from the genomic background to help ameliorate the effects of autoimmunity (otherwise they should match host background GC content because phage typically track their host nucleotide composition, often as a byproduct of optimizing codon usage bias for transcription in their host, e.g. [31]). We also predicted that the bases flanking restriction sites on the genome (corresponding to enzymes encoded on that genome) would have elevated GC, especially for AT-rich recognition sites. Both of these predictions were borne out. We analyzed the complete set of genomes and their listed restriction enzymes in the REBASE database [32] and found that restriction enzymes tend to have higher GC recognition sequences than their genomic background, and that the bases immediately flanking recognition sites have elevated GC content (Fig 2, S8 Fig, S9 Fig).

Given the association between Ku and GC content, DNA damage could well be the mechanism driving the association between various environmental factors and GC content. In environments in which microbes experience high rates of DSBs but often have only one copy of their genome, we hypothesize that high GC is selected for to help maintain genomic integrity and aid repair. An attractive element of this hypothesis is its ability to potentially explain a large number of previously observed GC versus trait correlations, because many of these traits are associated with high rates of DSBs and/or extended stationary phases.

Acknowledgements JLW was supported by a GAANN Fellowship from the U.S. Department of Education and the University of Maryland as well as a COMBINE Fellowship from the University of Maryland and funded by NSF DGE-1632976. WFF was partially supported the U.S. Army Research Laboratory and the U.S. Army Research Office under Grant W911NF-14-1-0490. PLFJ was supported in part by NIH R00 GM104158.

Conflict of Interest The authors declare no conflict of interest.

References

[1] Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, et al. The 160-kilobase genome of the bacterial endosymbiont Carsonella. Science. 2006;314(5797):267–267.

[2] Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. EMBO reports. 2005 Dec;6(12):1208–1213.

[3] Reichenberger ER, Rosen G, Hershberg U, Hershberg R. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. Genome Biology and Evolution. 2015 Apr;7(5):1380–1389.

[4] Hershberg R, Petrov DA. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLOS Genetics. 2010 Sep;6(9):e1001115.

[5] Hildebrand F, Meyer A, Eyre-Walker A. Evidence of Selection upon Genomic GC-Content in Bacteria. PLOS Genetics. 2010 Sep;6(9):e1001107.

[6] Long H, Sung W, Kucukyildirim S, Williams E, Miller SF, Guo W, et al. Evolutionary determinants of genome-wide nucleotide composition. Nature Ecology & Evolution. 2018 Feb;2(2):237–240.

[7] Lassalle F, Périan S, Bataillon T, Nesme X, Duret L, Daubin V. GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. PLOS Genetics. 2015 Feb;11(2):e1004941.

[8] Rocha EPC, Feil EJ. Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria? PLOS Genetics. 2010 Sep;6(9):e1001104.

[9] Rocha EPC. Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. Molecular Biology and Evolution. 2018 Jun;35(6):1338–1347.

[10] Naya H, Romero H, Zavala A, Alvarez B, Musto H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. Journal of Molecular Evolution. 2002 Sep;55(3):260–264.

[11] Romero H, Pereira E, Naya H, Musto H. Oxygen and Guanine–Cytosine Profiles in Marine Environments. Journal of Molecular Evolution. 2009 Aug;69(2):203–206.

[12] Karanjawala ZE, Murphy N, Hinton DR, Hsieh CL, Lieber MR. Oxygen Metabolism Causes Chromosome Breaks and Is Associated with the Neuronal Apoptosis Observed in DNA Double-Strand Break Repair Mutants. Current Biology. 2002 Mar;12(5):397–402.

[13] Pitcher RS, Brissett NC, Doherty AJ. Nonhomologous end-joining in bacteria: a microbial perspective. Annual Review of Microbiology. 2007;61:259–282.

[14] Wang HC, Susko E, Roger AJ. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: Data quality and confounding factors. Biochemical and Biophysical Research Communications. 2006 Apr;342(3):681–684.

[15] Pitcher RS, Green AJ, Brzostek A, Korycka-Machala M, Dziadek J, Doherty AJ. NHEJ protects mycobacteria in stationary phase against the harmful effects of desiccation. DNA repair. 2007;6(9):1271–1276.

[16] Vriezen JA, De Bruijn FJ, Nüsslein K. Responses of rhizobia to desiccation in relation to osmotic stress, oxygen, and temperature. Applied and Environmental Microbiology. 2007;73(11):3451–3459.

[17] Rocha EP, Cornet E, Michel B. Comparative and evolutionary analysis of the bacterial homologous recombination systems. PLoS genetics. 2005;1(2):e15.

[18] Aravind L, Koonin EV. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. Genome Research. 2001 Aug;11(8):1365–1374.

[19] Doherty Aidan J , Jackson Stephen P , Weller Geoffrey R . Identification of bacterial homologues of the Ku DNA repair proteins. FEBS Letters. 2001 Jul;500(3):186–188.

[20] Eddy SR. Profile hidden Markov models. Bioinformatics (Oxford, England). 1998;14(9):755–763.

[21] Ho LsT, Ané C. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Systematic Biology. 2014 May;63(3):397–408.

[22] Vos M, Didelot X. A comparison of homologous recombination rates in bacteria and archaea. The ISME journal. 2009;3(2):199.

[23] Rendueles O, de Sousa JAM, Bernheim A, Touchon M, Rocha EP. Genetic exchanges are more frequent in bacteria encoding capsules. PLoS genetics. 2018;14(12):e1007862.

[24] Bobay LM, Ochman H. Impact of recombination on the base composition of bacteria and archaea. Molecular biology and evolution. 2017;34(10):2627–2636.

[25] Kristensen DM, Wolf YI, Koonin EV. ATGC database and ATGC-COGs: an updated resource for micro-and macro-evolutionary studies of prokaryotic genomes and protein family annotation. Nucleic acids research. 2016;p. gkw934.

[26] Sandoval A, Labhart P. High G/C content of cohesive overhangs renders DNA end joining Ku-independent. DNA repair. 2004;3(1):13–21.

[27] Daley JM, Wilson TE. Rejoining of DNA double-strand breaks as a function of overhang length. Molecular and cellular biology. 2005;25(3):896–906.

[28] Aniukwu J, Glickman MS, Shuman S. The pathways and outcomes of mycobacterial NHEJ depend on the structure of the broken DNA ends. Genes & development. 2008;22(4):512–527.

[29] Bohlin J, Sekse C, Skjerve E, Brynildsrud O. Positive correlations between genomic% AT and genome size within strains of bacterial species. Environmental microbiology reports. 2014;6(3):278–286.

[30] Pleška M, Qian L, Okura R, Bergmiller T, Wakamoto Y, Kussell E, et al. Bacterial autoimmunity due to a restriction-modification system. Current Biology. 2016;26(3):404–409.

[31] Limor-Waisberg K, Carmi A, Scherz A, Pilpel Y, Furman I. Specialization versus adaptation: two strategies employed by cyanophages to enhance their translation efficiencies. Nucleic acids research. 2011;39(14):6016–6028.

[32] Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Research. 2010 Jan;38(suppl_1):D234–D236.

[33] Yarza P, Richter M, Peplies J, Euzeby J, Amann R, Schleifer KH, et al. The All-Species Living Tree project: A 16S rRNA-based phylogenetic tree of all sequenced type strains. Systematic and Applied Microbiology. 2008 Sep;31(4):241–250.

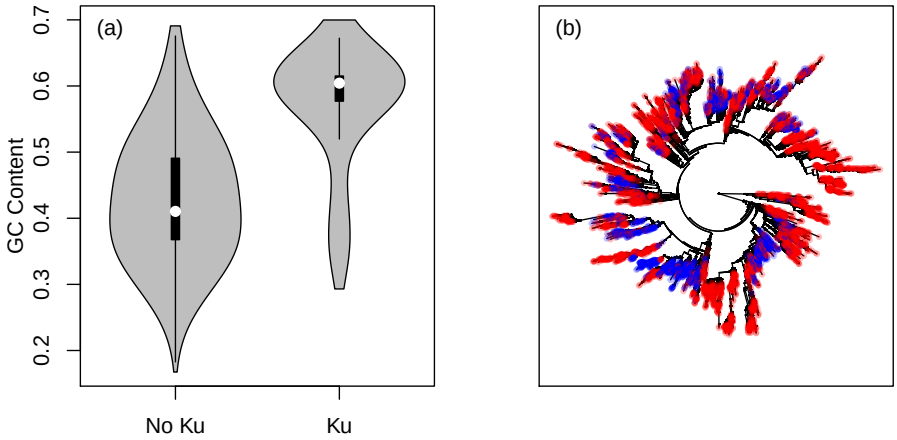
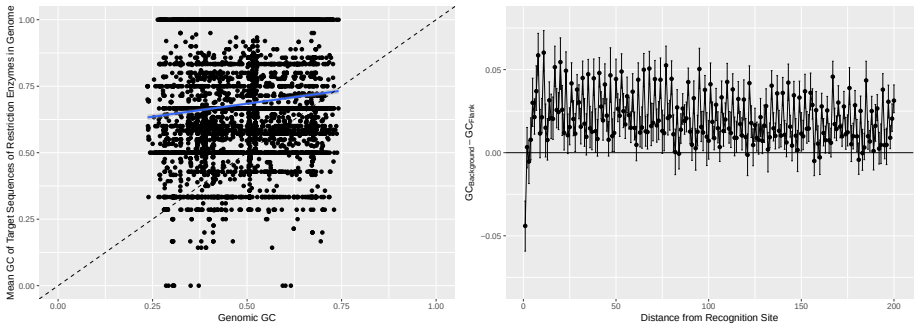


Figure 1: The relationship between genomic GC content and the NHEJ pathway in prokaryotes. (a) Microbes that code for the Ku protein tend to have much higher genomic GC content than those that do not. (b) Ku incidence mapped onto the SILVA Living Tree [33]. While Ku incidence is not randomly distributed across the prokaryotic tree, neither is it isolated to a particular clade. Organisms coding for Ku shown in blue, those not coding for Ku shown in red.



(a) (b)

Figure 2: Restriction sites are associated with elevated GC content. (a) Restriction enzymes tend to target higher GC sites than the average GC content of the genomes their are found on. (b) Bases immediately flanking restriction sites for restriction enzymes found on a given have an elevated GC content. This is seen especially strongly when looking at the most extreme values at each distance from the restriction site (S8 Fig), possibly because only some RM systems frequently target the host genome [30]. Shown here (b) are only restriction enzymes with 100% AT recognition sequences ($n = 102$), for which this effect is strongest, but it is also seen more broadly across all recognition sequences ($n = 16001$ recognition sequence-genome pairs; S9 Fig). Error bars represent bootstrapped 95% confidence intervals of the mean.