Linking high GC content to the repair of double strand breaks in prokaryotic genomes

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

August 15, 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 either high GC content is a selective adaptation to particular habitats, or that certain habitats favor increased rates of gene conversion. 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 DNA damage may be a fundamental driver of GC content, leading in part to the many environmental patterns observed to-date. 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.

Introduction

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]). This bias in content might naturally be assumed to arise from biases in mutation rates, but a puzzle arose when observational studies surprisingly revealed a GC→AT mutational

bias (which implies an expected equilibrium $< 50\%$ GC) in genomes with actual GC content $> 50\%$ [4, 5]; more recently, controlled mutation accumulation experiments showed that even genomes with $< 50\%$ actual GC still 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]. Much debate has resulted over the relative contribution of these two mechanisms to the observed genomic GC content in prokaryotes [8, 9, 5, 4, 7, 10].

This debate between proponents of the selection and BGC hypotheses continues, with many studies focusing on patterns of genetic diversity that by themselves cannot easily differentiate between these two hypotheses because recombination will also locally increase the efficiency of selection [11, 9]; however, the addition of phenotypic information provides the tantalizing clue that GC content correlates with shared environmental factors [2, 3, 12, 13] independent of phylogenetic similarity [3]. Thus, these environmental factors must either lead to an unknown selective advantage for high/low GC content [11] or lead to elevated rates of BGC through an as-yet unknown mechanism.

We noticed that many environments containing high GC microbes, such as soils and aerobic environments [3, 12], induce relatively high rates of DNA damage in the form of double-strand breaks (DSB) that necessitate repair [14, 15]. For instance, in aerobes, this damage typically results from reactive oxygen species produced during metabolism [14] that can lead to DSBs by producing collapsed replication forks [16], as well as via a number of other mechanisms (often in conjunction with other stressors; [17, 18, 19, 20, 21, 22, 23, 24]). In soil-dwelling microbes, DSBs are associated with desiccation and spore formation [25, 26, 27, 28]. Even going back nearly 50 years, it was suggested that the rate of exposure to UV radiation, which can lead to DSBs [29], might be driving observed variation in genomic GC content among microbes [30].

To repair DSBs, microbes may use one of two pathways: homologous recombination (HR), or non-homologous end joining (NHEJ) [15]. HR machinery is ubiquitous across microbes [31], although it requires multiple genome copies to function. To-date, much work on GC content has focused on associating rates of HR locally along a genome (inferred using polymorphism data) with local GC content, which would be taken as evidence for the action of BGC [7]. We might also expect that organisms experiencing many DSBs would have an high overall

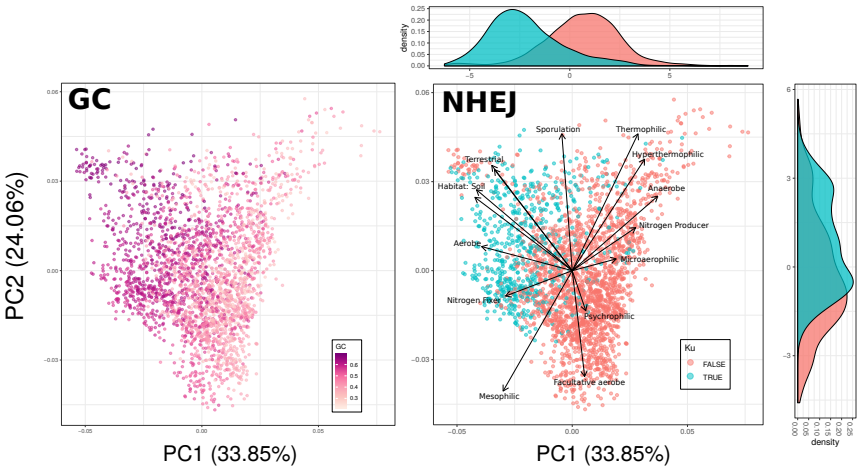
recombination rate in order to repair these breaks. However, average rates of recombination in different genomes do not seem to be correlated with genomic GC content [5], despite the systematic environmental variation in GC content discussed above. It is possible that analyses that correlate global recombination rates and GC content looking across many genomes are too coarse-grained to reveal subtle differences between microbes leading to larger divergence in GC content over evolutionary time. Additionally, it is difficult to get accurate estimates of recombination rates from population-level polymorphism data, and it is unclear how strongly these rates would correlate with DSB formation specifically. Thus, an alternative, complementary indicator of high rates of DSB formation would be useful.

In contrast to HR, 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; [26]). Notably, we expect NHEJ repair to be favored only when HR is not an option, as NHEJ is generally considered a highly error prone pathway [32, 15]. NHEJ repair requires the presence of the highly conserved Ku protein [33, 34], which makes Ku presence/absence a useful indicator of genomes more/less likely to be subjected to high rates of DSBs during especially vulnerable periods (i.e., one genome copy present). We leverage Ku as an indicator of the rate of DSB formation and examine how the incidence of the NHEJ pathway covaries with genomic GC content. We find a strong association between Ku presence and elevated GC content, and go on to discuss several mechanisms that could explain this pattern under a selection or a BGC paradigm.

Results and Discussion

NHEJ and High GC Content Found in Similar Environments

A number of ecological factors have been associated with GC content in previous works, including aerobicity [12], nitrogen fixation [35], exposure to UV radiation [30], and growth temperature (although this last association has been disputed; [36, 25, 37, 38]). Notably, many of these associations are weak, and in general there is no known universal driver or mechanism that explains the high genomic GC contents seen across many environments. We noted that many of the GC-correlated environmental factors identified in previous analyses are also



(a)

Figure 1: Ku and high GC content share a particular region of trait space. PCA of microbial trait data for select traits with species colored based on either their mean genomic GC content or whether they have known members that encode the Ku protein. Trait loadings signified by arrows. Note the clear separation of Ku and no-Ku organisms in trait space.

associated with high rates of DNA damage, specifically DSBs. Perhaps, then, the unifying driver of GC content is the rate of DSB formation, and the environmental trends observed to-date can be attributed to this underlying driver.

While NHEJ presence is an imperfect indicator of DSB incidence in general, we expect this pathway to be especially common among organisms experiencing many DSBs during periods of slow or no growth [26]. Using a large-scale microbial trait database [39] paired with genomes from RefSeq [40], we identified ecological correlates of NHEJ incidence as well as genomic GC content. A principal component analysis of these traits revealed similar patterns of NHEJ incidence and high genomic GC content in trait-space (Fig 1), consistent with the idea that DNA damage is associated with genomic GC content. In fact, the pairwise correlation between an ecological trait and genomic GC content tracks almost perfectly with correlation between each trait and Ku incidence (S1 Fig).

Nevertheless, the inclusion of Ku along with ecological traits in a linear model to explain genomic GC content resulted in most other environmental

traits still being statistically significant (S1 Table), indicating that either there is some aspect of the environment affecting GC content that is not attributable to DSBs or that NHEJ is an imperfect indicator of the rate of DSB formation (or both). In fact this is trivially true, as Ku presence is a discrete, binary variable whereas the rate of DSB formation is continuous. Despite the fact that Ku is not be the sole predictor of genomic GC content, the shared region of trait-space between NHEJ-capable organisms and high-GC organisms is quite striking (Fig 1).

Organisms with NHEJ Machinery Have High GC Content

Next, we looked directly at the Ku/GC relationship. Using a large set of genomes from RefSeq we found that genomes with Ku have a dramatically shifted GC content relative to genomes without Ku (Fig 2A, S2 Fig), even though Ku presence/absence is sprinkled throughout the prokaryotic phylogeny (Fig 2B). Indeed, this association remains highly significant even after formally correcting for phylogeny using phylogenetic regression with a Brownian motion (BM) model of trait evolution (Table 1). Our analysis is robust to the choice of evolutionary model, as repetition with an Ornstein-Uhlenbeck (OU) model of trait evolution yielded similar results (Table 1). Similarly, restriction of our analysis to a particular phylum (Actinobacteria, Firmicutes, and Proteobacteria, respectively; each with >1000 genomes on our tree) shows that this effect is not attributable to a single branch of the prokaryotic tree but is quite general (Table 1). In sum, the presence of NHEJ on a genome is positively associated with the GC content of that genome.

Importantly, we control for genome length in all our phylogenetic models, which potentially covaries with Ku incidence and is known to be associated with genomic GC content in prokaryotes (Table 1 and S3 Fig). Interestingly, Ku presence and genome length have a significant negative interaction in their effect on GC content (Table 1).

Clearly organisms that encode Ku have a higher genomic GC content than organisms that do not, but can Ku help explain why organisms have a higher GC content than *expected*? Prokaryotes typically have higher GC content than predicted from their mutational biases, which are nearly always skewed towards AT alleles [4, 5, 6]. Does the observed association between NHEJ and genomic GC content contribute to this deviation? In other words, are GC alleles more likely to fix than the neutral expectation in Ku-encoding genomes, and is this

Data	Model	AIC	β_{Ku}	p_{Ku}	$\beta_{\text{GenomeLength}}$	$p_{\text{GenomeLength}}$	$\beta_{\text{Interaction}}$	$p_{\text{Interaction}}$
All	BM	-10754	1.29	1.465×10^{-13}	0.251	2.2×10^{-16}	-0.191	2.522×10^{-13}
All	OU	-10760	1.28	1.938×10^{-13}	0.248	2.2×10^{-16}	-0.190	3.291×10^{-13}
All (GC4)	BM	6719	4.53	3.763×10^{-12}	0.812	2.2×10^{-16}	-0.666	8.247×10^{-12}
All (GC4)	OU	6676	5.58	2.342×10^{-12}	0.827	2.2×10^{-16}	-0.675	5.160×10^{-12}
Actinobacteria	BM	-1997.131	3.26	2.2×10^{-16}	0.610	2.2×10^{-16}	-0.483	2.2×10^{-16}
Actinobacteria	OU	-2011	3.41	2.2×10^{-16}	0.644	2.2×10^{-16}	-0.505	2.2×10^{-16}
Firmicutes	BM	-1541	1.63	0.004541	0.104	0.028670	-0.246	0.004358
Firmicutes	OU	-1536	1.64	0.004242	0.110	0.022632	-0.248	0.004088
Proteobacteria	BM	-3181	0.719	0.01704	0.244	1.054×10^{-11}	-0.105	0.01969
Proteobacteria	OU	-3180	0.716	0.01744	0.238	3.377×10^{-11}	-0.104	0.02011

Table 1: Coefficients and p-values for all phylogenetic regressions performed. All tests significant after Benjamini-Hochberg correction ($\alpha = 0.05$).

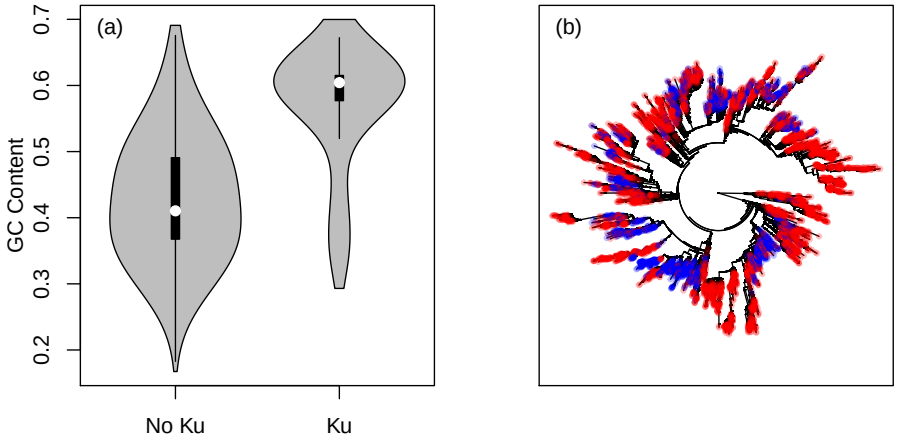
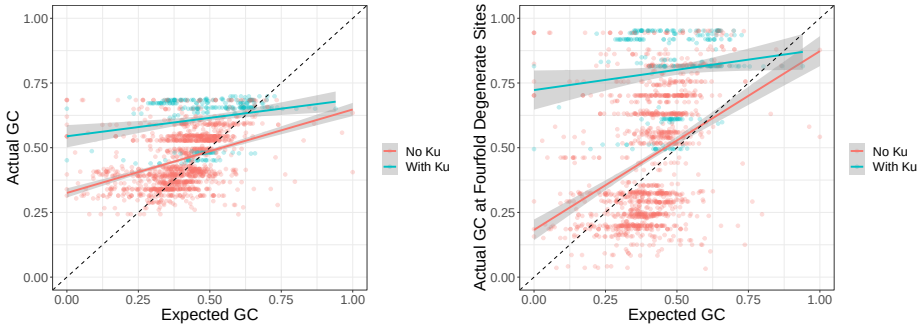


Figure 2: The relationship between genomic GC content and the NHEJ pathway in prokaryotes. (a) Microbes that code for the Ku protein tend to have much higher genomic GC content than those that do not. (b) Ku incidence mapped onto the SILVA Living Tree [41]. While Ku incidence is not randomly distributed 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.

deviance from neutrality larger than in genomes that do not encode Ku? Alternatively, it is possible that the error-prone NHEJ machinery simply leads to an excess of GC mutations. In this case, NHEJ incidence would help explain differences in mutational biases between microbes, but would not help explain the mystery of higher than expected genomic GC content among microbes.

Examining mutation accumulation experiments in detail, Ku shows no effect on GC↔AT mutational biases (S4 Fig; data from [6]). Since mutation accumulation data are limited, we also used the GC bias of inferred polymorphisms as a proxy for mutation. Similar to previous studies of prokaryotic GC content [4], and using the same intuition to that of the McDonald-Kreitman test for selection [42], we assume that polymorphisms within a population have experienced minimal selection and are therefore representative of the mutational biases of a given organism. Thus the GC content of polymorphisms within a population gives an estimate of the GC↔AT mutational bias pre-selection, whereas the genomic GC content of an organism results from a combination of mutational biases and mechanisms that alter the probability of fixation of an allele, such as selection and BGC. We obtained multiple alignments of all orthologous genes



(a) (b)

Figure 3: Genomes with Ku appear to fix GC alleles at a greater rate than expected (either due to BGC to selection). (a,b) Genomes with Ku having, on average, even greater elevation of GC over expectation than genomes without Ku. Expected GC estimated from polymorphism data. This signal is conservative due to observed polymorphisms experiencing some effects of BGC/selection (see Methods for discussion).

for organisms in the ATGC database [43] belonging to clusters that contained at least three genomes (to identify and polarize polymorphisms) and used this dataset to compare the background GC content to the GC content of polymorphisms. We saw greater evidence for BGC/selection in genomes with Ku than without Ku, with the presence of Ku leading to higher observed GC regardless of the expected GC (Fig 3, S6 Fig). Thus, the association between Ku and genomic GC content is not due to differences in mutational bias. This implies that DSBs are either leading to selection for high GC content or influencing the rate and/or biases of homologous recombination to increase the overall action of BGC. In order to minimize the effects of selection on our estimate of expected GC content, we repeated this analysis only using polymorphisms at fourfold degenerate sites and found qualitatively similar results, despite only having only about a third as many informative polymorphisms (S5 Fig, S6 Fig).

No Apparent Relationship Between Rate of Homologous Recombination and NHEJ

The above analyses suggest that GC alleles fix with a higher probability in organisms experiencing an elevated rate of DSB formation. If BGC were the

primary driver of GC content evolution in prokaryotes, would we expect an association between damage and GC content as we see here? We can think of at least one plausible scenario. The formation of DSBs should stimulate recombination for repair, and assuming that recombination is biased we might expect rates of BGC to increase as the rate of DSB formation increases. We saw no positive association between Ku incidence and inferred rates of homologous recombination looking between genomes, as would be predicted by this hypothesis (S7 Fig; data from [44, 45]). In fact the relationship appeared to be negative regardless of method to measure recombination rate (though not significant). That being said, the effects of BGC are typically only apparent locally, comparing GC content and recombination rates along a genome rather than between genomes [7]. The clearest evidence for BGC leading to high GC content in prokaryotes comes from Lassalle et al [7], who compared the GC content of genes that recombined frequently or rarely within genomes. We checked if any of the 21 taxa they studied typically encode Ku based on the genomes we downloaded from RefSeq above. Four taxa carry Ku at any appreciable frequency (>1%): *Mycobacterium tuberculosis* (100%), *Burkholderia pseudomallei* (100%), *Burkholderia cenocepacia* (87%), and *Bacillus anthracis* (68%). These organisms did not show a consistent association between recombination rate and GC content, unlike most other taxa in the study. *M. tuberculosis* and *B. pseudomallei* are highly clonal and did not present enough diversity for a complete analysis by Lassalle et al [7], though *B. pseudomallei* had a negative association between recombination rate and GC at the third codon position (positive for GC content overall, in both cases not significant). *B. cenocepacia* presented enough data for analysis, but showed no relationship between recombination rate and GC content (again with a negative but non-significant effect at the third codon position). Finally, *B. anthracis* had inconsistent effects, with a significant negative association between recombination rate and GC content overall but a significant positive interaction when restricting to GC content at the third codon position. What to make of this? It seems that the evidence for BGC is not strong among organisms that encode Ku in Lassalle et al.'s dataset, though the total number of species with Ku is extremely limited.

Taking a step back, there are two primary reasons to disfavor BGC as the hypothetical mechanism underlying the observed positive association between NHEJ and genomic GC content. First, the extent to which BGC is a driver of genomic GC content in prokaryotes has been questioned. Studies using alternative methods to quantify recombination across the genome to those of Lassalle et

al. [7] have found either no relationship between recombination rate and local GC content or inconsistent patterns between species (with some species even showing a negative relationship; [46, 47]). More recently, looking specifically at polymorphisms arising via recombination across many prokaryote species, Bobay and Ochman [10] reject the BGC hypothesis outright. Even in some microbial eukaryotes, where the evidence of BGC is thought to be strong, tetrad analysis has been unable to reveal any evidence of BGC leading to elevated GC content, even in genomes where recombination rate and GC content are locally correlated [48]. This suggests that correlative methods may be insufficient to conclusively demonstrate BGC, and it has been suggested that high GC content may actually increase the rate of recombination locally (effectively reversing the logic behind the evidence for BGC; [49]).

Second, while organisms encoding Ku are likely to be experiencing DSBs, they are unlikely to experience high rates of recombination. NHEJ is thought of as an alternative pathway to HR, specifically used when HR cannot proceed because the genome is only present as a single copy [15, 26]. Thus we expect NHEJ to be favored specifically in situations where BGC is unlikely. While it is possible that high rates of damage could still favor both NHEJ and HR, albeit at different points in an organism's life cycle, the extremely strong and specific association between GC and Ku suggests that this relationship may be particular to the specific conditions selecting for Ku (especially considering the absence of an association between HR and GC when looking between genomes [5]; S7 Fig). Nevertheless, we have insufficient information to completely rule out BGC as a mechanism at this time.

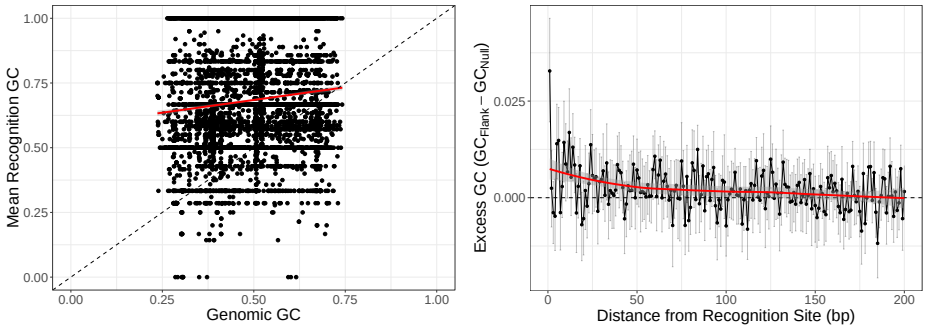
High GC Near Regions with Frequent Breaks

Given our lack of enthusiasm for BGC as a mechanism to explain the association between NHEJ and high GC content, perhaps selection can provide an alternative hypothesis. Could it be that organisms with NHEJ machinery are under stronger selection for high GC content than those without? This leads to us to a puzzle: what fitness advantage might be conferred by GC content? In fact, high GC content may promote DNA repair, both by facilitating canonical NHEJ (i.e., Ku-dependent) and alternative NHEJ (i.e., Ku-independent) pathways.

During DSB repair, the NHEJ machinery in prokaryotes takes advantage of homology in any short overhanging regions or nearby microhomology in order to help align the two broken DNA ends [50, 51, 52]. Any factor that stabilizes

this interaction (e.g., high GC via an increased number of hydrogen bonds) may thus increase the efficiency of NHEJ repair. It has also been shown that prokaryotes can employ alternative high-fidelity end-joining pathways that are independent of the NHEJ machinery, and that these pathways are primarily dependent on nearby microhomology to tether the DNA ends together [32, 53]. It stands to reason that high GC content in these regions of microhomology might help stabilize the end-pairings and improve the efficiency of repair. In fact, in eukaryotes, high GC overhangs or microhomologies specifically promote the use of a similar NHEJ-independent, high-accuracy end-joining repair pathway [54, 55]. In these systems high GC content is thought to help tether the DNA together and thus perform a similar role to Ku [54, 55], though this has not yet been confirmed in prokaryotes. Nevertheless, this mechanism suggests that high GC content could help to ameliorate the negative effects of DSBs, especially in environments with high rates of DSBs but only a single genome copy. That being said, high genomic GC content alone cannot protect an entire genome, since most genomes will have at least some AT-rich regions. Thus a combined Ku and high-GC strategy would potentially be favorable for DSB-vulnerable microbes and could explain the strong positive relationship we observed between selection for high genomic GC content and the presence of Ku (Fig 2)

In addition, our hypothesis makes a novel testable prediction: regions of the genome that are especially prone to DSBs should be under selection to have higher GC content. Restriction modification (RM) systems provide an ideal test case as damage due to self-targeting is a known phenomenon (e.g. [56]), and we know the potential locations of self-targeting if we know the restriction enzyme recognition sequence. 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. [57]). We further predicted that, for restriction enzymes with low GC recognition sequences, the bases flanking restriction sites on the genome would have elevated GC. Both of these predictions were borne out. We analyzed the complete set of genomes and their listed restriction enzymes in the REBASE database [58] and found that restriction enzymes indeed tend to have higher GC recognition sequences than their genomic background, and that the bases immediately flanking AT-rich recognition sites have elevated GC content (Fig 4).



(a) (b)

Figure 4: Restriction sites are associated with elevated GC content. (a) Restriction enzymes tend to target sequences with GC content higher than the genomic average. (b) Bases immediately flanking AT-rich restriction sites ($\geq$ 75% AT, $n = 214$ genomes) have an elevated mean GC content. This signature mostly decays within 50bp of the recognition site. This pattern is particularly striking when looking at the first flanking base. Error bars represent bootstrapped 95% confidence intervals of the mean.

We emphasize that the idea that DSB formation selects for high GC content, while consistent with our data, is at this point largely speculative. By outlining this scenario we hope to enable experimentalists to design specific, mechanistic studies. Much of the debate over genomic GC content (including our present contribution), and especially its ecological role, has relied on population-scale correlative studies and has largely avoided mechanisms. Our "GC-tethering" hypothesis has the advantage of being amenable to laboratory-based investigations.

Conclusions

We found a strong positive association between the presence of the NHEJ pathway on a genome and genomic GC content across prokaryotes. This association holds controlling for phylogeny and genome length and cannot be explained by mutational biases. The NHEJ repair pathway is broadly but sparsely distributed across the prokaryotic tree (Fig 2), showing up in only about a quarter of genomes [33], and is expected to be favored among organisms experiencing high rates of DSB formation during periods of no or slow growth (i.e., only a

single genome copy present so that HR is impossible; [26, 15]). This suggests that high GC content may be an adaptation to deal with DSBs when HR is not feasible, especially when rates of damage are high. In fact, we find that in regions of the genome where DSBs are likely to occur, GC content is locally elevated (Fig 4). Alternatively, the presence of NHEJ may be an indicator of high rates of DSBs in general, so that these organisms are also experiencing higher rates of HR for repair, and subsequently increased BGC. We discussed the relative merits of these two hypothetical mechanisms linking DSBs to genomic GC content, though at this point it is not possible to state conclusively which mechanism is the primary driver of the pattern we observe. It is also possible that some combination of BGC and selection is acting to increase genomic GC content in organisms experiencing DNA damage.

Regardless of the underlying mechanism, high risk of DSB formation is a common factor in many of the habitats that high GC microbes have been shown to inhabit. While the presence of NHEJ cannot single-handedly explain high GC content in all organisms (there are many organisms incapable of NHEJ that still have high GC content, Fig 2), it is possible that DSB formation can (or at least come close). For example, *Deinococcus radiodurans* is resilient to extremely high rates of DSB formation [59] and has high genomic GC content, but lacks Ku. It is difficult to directly assess the rate of DSB formation, but not impossible [60]. We hope to see future work that assays rates of damage in the environment. DNA damage is an important challenge for microbes to overcome, and a systematic understanding of how damage varies between environments is of broad ecological interest.

Methods

Data

We downloaded all available completely sequenced prokaryotic genomes from NCBI's non-redundant RefSeq database FTP site on December 23, 2017 [61] and searched for the presence of the gene coding for the Ku protein, which is central to the NHEJ pathway, using hmmsearch (E-value cutoff $10^{-2}/\#$ Genomes, Pfam family PF02735; [62]). This identified 21389 genomes containing Ku out of a total 104297 genomes analyzed.

For phylogenetic analyses, we downloaded the SILVA Living Tree 16s rRNA tree [41]. We obtained mutational bias estimates from Long et al. (m; [6]), of

which 22 could be matched with a genome in our dataset. We obtained estimates of the rate of homologous recombination from [44, 45]. For our analyses on the fate of polymorphism we downloaded alignments from the Alignable Tight Genomic Cluster (ATGC) database [43].

Trait data were obtained from the ProTraits microbial trait database (2679 species; [39]). ProTraits scores are expressed as separate confidences that a particular species does or does not have a trait based on the results of an automated text-mining algorithm, giving two scores per binary trait. We combined these scores to obtain what is essentially a probability that a microbe has a given trait using the equation in Weissman et al. [63], yielding a single score between zero and one for each trait for each species. We selected a suite of traits known to be associated with either the incidence of Ku or genomic GC content (Soil-dwelling, aerobicity, growth temperature, nitrogen fixation, spore formation) to include in our analysis.

We downloaded the complete list of genomes from the REBASE restriction enzyme database [58], which includes all RM enzymes found on a given genome, along with their target sequence if known. Using the listed accession numbers, we then downloaded each corresponding genome from RefSeq in order to assess GC content near restriction sites on the genome (potential sites where DSBs would occur).

Phylogenetic Linear Models

Using the 6648 organisms on the SILVA tree for which we had a genome and could assess Ku presence/absence (2051 with, 4597 without) we built a series of phylogenetically corrected linear models of genomic GC content using the phylolm package in R [64]. First we logit transformed our %GC values

$$GC_l = \log\left(\frac{\%\text{GC}}{1 - \%\text{GC}}\right) \tag{1}$$

so that values were in the range $(-\infty, \infty)$. We then used Ku incidence as a binary predictor along with $\log_{10}$ genome length as a continuous predictor to predict GC_l using phylogenetic regression:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{2}$$

where

$$\varepsilon \sim \mathcal{N}(0, \sigma_p^2 \mathbf{V} + \sigma_e^2 \mathbf{I}) \tag{3}$$

so that $\mathbf{V}$ is the phylogenetic covariance matrix and σ_e^2 is the variance of the measurement error [64]. Brownian motion (BM) models of trait evolution are inappropriate when trait values are bounded. While GC content is theoretically bounded at zero and one, there are no species that approach these bounds, and our logit transforming of the GC values should ameliorate this issue. Sometimes Ornstein-Uhlenbeck (OU) models of trait evolution are also used in these cases. We found that using an OU model had no qualitative effect on our result (1), though this model had a lower AIC than the Brownian motion model (-10760 versus -10754).

We also applied the above analyses independently to the three best-represented phyla in the dataset, each with >1000 genomes: Actinobacteria (614 with Ku, 446 without), Firmicutes (320 with Ku, 699 without), and Proteobacteria (539 with Ku, 1384 without).

Fate of Polymorphism

The ATGC database groups closely-related genomes into "clusters" and provides alignments of their core genes [43]. We downloaded multiple alignments corresponding to clusters in the ATGC database that had at least three genomes. In this way we could, at a minimum, identify polymorphisms between two genomes while using a third, more distantly related genome to polarize these polymorphisms. We restricted our analysis to orthologous genes (COGs) that were present in all members of a cluster. For each genome in a cluster we obtained a set of polymorphisms for that genome by comparing to the most similar genome in that cluster, using the most diverged genome in the cluster to polarize these polymorphisms (assuming that the diverged genome represented the ancestral state, and ignoring cases where neither of the other two genomes matched this "ancestral" allele). Similarity was calculated as the percent identity over the entire aligned core genome provided by ATGC. In order to ensure that polymorphisms were recent and had not yet undergone selection, we discarded genomes that were not within 1% pairwise divergence of any other genome in their respective cluster (calculated over the set of core genes provided in the ATGC alignments). We also discarded pairs of genomes that had fewer than 5 informative sites (either GC→AT or AT→GC) in order to avoid extreme expected GC estimates. Thus we obtained a set of 1868739 polarized polymorphisms for

1643 pairs of genomes for which GC content could be assessed and compared to the background genomic GC content (Fig 3). Expected GC was calculated as in Long et al [6].

To obtain expected GC content at fourfold degenerate sites we repeated the above analysis only looking at polymorphisms at fourfold degenerate sites (S5 Fig). There are about a third as many polymorphisms in this dataset (574944) but a similar number of genome pairs are retained (1351).

The rationale behind this test is that recent polymorphisms will not have had a chance to undergo selection (or BGC). This is similar to the intuition underlying the McDonald-Kreitman test for selection [42], and similar analyses have been performed in past work on GC content [4, 5]. Therefore we can obtain an estimate of the expected GC content based on mutational bias, and infer that selection (or BGC) is acting if the realized genomic GC content differs from the expectation. In practice, because we are looking at alignable coding sequence, selection is likely to be strong, and may bias our estimates (S8 Fig). This is further compounded by the fact that genomes in a cluster can still be quite diverged, although we control for this by restricting to genomes within 1% sequence divergence from each other. In any case, the direction of bias will be towards the equilibrium GC content, as estimated via the genomic background. Thus, this test for selection suffers somewhat in that it has an increased probability of false-negatives, but this bias should not cause a false signal of selection to occur.

While our estimates do not perfectly align with those found in mutation accumulation experiments (S8 Fig), we note that even within a genus there can be extremely high heterogeneous values for GC$\leftrightarrow$AT mutation bias. For example, Long et al. [6] estimate the ratio of the rate of GC$\rightarrow$AT mutation to the rate of AT$\rightarrow$GC mutation to be 4.5 in Vibrio fischeri but 2.3 in Vibrio cholerae. Similarly, they find values of 6.6 in Staphylococcus epidermidis but 4.6 in Staphylococcus aureus, despite having very similar values for genomic GC (0.33 vs. 0.32) and GC at fourfold degenerate sites (0.20 vs 0.19). This implies that closely related organisms may have very different mutational biases, making comparisons between datasets challenging.

Restriction Sites

We identified all genomes encoding restriction enzymes with known restriction sequences in our dataset using the REBASE database [58]. We then restricted

our analyses to genomes encoding enzymes that had low-GC restriction sequences ($\geq$ 75% AT, $n = 214$, no genomes with multiple AT-rich enzymes). For each remaining genome we mapped the corresponding restriction sequence to the genome itself to find all potential sites of self-targeting. We then calculated the mean GC content of the sites directly flanking these self-targets across the genome, obtaining a value for average GC content for each distance (1-200bp) from the target for each genome.

In order to generate an adequate null for comparison, for each genome-restriction sequence pair in our dataset we generated a novel restriction sequence. To do this, we took each restriction recognition sequence and randomly permuted it to obtain a new sequence with identical base composition (if the permuted sequence was identical to the original, we continued drawing until a different sequence was obtained). We then repeated the above flank-analysis with this set of "fake" restriction recognition sequences.

Finally, for each flank-distance (1-200bp) in each genome we calculated the difference in mean GC of the bases flanking true restriction sites from bases flanking the null sites. We bootstrapped the mean of this distribution for each flank distance across genomes to obtain 95% confidence intervals (Fig 4).

Acknowledgments 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.

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. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001115.

[5] Hildebrand F, Meyer A, Eyre-Walker A. Evidence of Selection upon Genomic GC-Content in Bacteria. PLOS Genetics. 2010 Sep;6(9):e1001107. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001107.

[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. Available from: https://www.nature.com/articles/s41559-017-0425-y.

[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. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1004941.

[8] Raghavan R, Kelkar YD, Ochman H. A selective force favoring increased G+ C content in bacterial genes. Proceedings of the National Academy of Sciences. 2012;109(36):14504–14507.

[9] Rocha EPC. Neutral Theory, Microbial Practice: Challenges in Bacterial Population Genetics. Molecular Biology and Evolution. 2018 Jun;35(6):1338–1347. Available from: https://academic.oup.com/mbe/article/35/6/1338/4976545.

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

[11] Rocha EPC, Feil EJ. Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria? PLOS Ge-

netics. 2010 Sep;6(9):e1001104. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001104.

[12] 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.

[13] 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. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2722718/.

[14] 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. Available from: http://www.sciencedirect.com/science/article/pii/S096098220200684X.

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

[16] Charbon G, Bjørn L, Mendoza-Chamizo B, Frimodt-Møller J, Løbner-Olesen A. Oxidative DNA damage is instrumental in hyperreplication stress-induced inviability of Escherichia coli. Nucleic acids research. 2014;42(21):13228–13241.

[17] Dianov GL, Timehenko TV, Sinitsina OI, Kuzminov AV, Medvedev OA, Salganik RI. Repair of uracil residues closely spaced on the opposite strands of plasmid DNA results in double-strand break and deletion formation. Molecular and General Genetics MGG. 1991;225(3):448–452.

[18] Kozmin SG, Sedletska Y, Reynaud-Angelin A, Gasparutto D, Sage E. The formation of double-strand breaks at multiply damaged sites is driven by the kinetics of excision/incision at base damage in eukaryotic cells. Nucleic acids research. 2009;37(6):1767–1777.

[19] Hong Y, Li L, Luan G, Drlica K, Zhao X. Contribution of reactive oxygen species to thymineless death in Escherichia coli. Nature microbiology. 2017;2(12):1667.

[20] Henrikus SS, Henry C, McDonald JP, Hellmich Y, Wood EA, Woodgate R, et al. DNA double-strand breaks induced by reactive oxygen species promote DNA polymerase IV activity in Escherichia coli. bioRxiv. 2019;p. 533422.

[21] Bonura T, Town CD, Smith KC, Kaplan HS. The influence of oxygen on the yield of DNA double-strand breaks in X-irradiated Escherichia coli K-12. Radiation research. 1975;63(3):567–577.

[22] Tilby MJ, Loverock PS. Measurements of DNA double-strand break yields in E. coli after rapid irradiation and cell inactivation: the effects of inactivation technique and anoxic radiosensitizers. Radiation research. 1983;96(2):309–321.

[23] Van der Schans G, Blok J. The influence of oxygen and sulphhydryl compounds on the production of breaks in bacteriophage DNA by gamma-rays. International Journal of Radiation Biology and Related Studies in Physics, Chemistry and Medicine. 1970;17(1):25–38.

[24] Mahaseth T, Kuzminov A. Prompt repair of hydrogen peroxide-induced DNA lesions prevents catastrophic chromosomal fragmentation. DNA repair. 2016;41:42–53.

[25] 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. Available from: `http://www.sciencedirect.com/science/article/pii/S0006291X06003214`.

[26] 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.

[27] 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.

[28] Dupuy P, Gourion B, Sauviac L, Bruand C. DNA double-strand break repair is involved in desiccation resistance of Sinorhizobium meliloti, but is not essential for its symbiotic interaction with Medicago truncatula. Microbiology. 2017;163(3):333–342.

[29] Slieman TA, Nicholson WL. Artificial and solar UV radiation induces strand breaks and cyclobutane pyrimidine dimers in Bacillus subtilis spore DNA. Appl Environ Microbiol. 2000;66(1):199–205.

[30] Singer CE, Ames BN. Sunlight ultraviolet and bacterial DNA base ratios. Science. 1970;170(3960):822–826.

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

[32] Gong C, Bongiorno P, Martins A, Stephanou NC, Zhu H, Shuman S, et al. Mechanism of nonhomologous end-joining in mycobacteria: a low-fidelity repair system driven by Ku, ligase D and ligase C. Nature structural & molecular biology. 2005;12(4):304.

[33] 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.

[34] 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. Available from: `https://febs.onlinelibrary.wiley.com/doi/full/10.1016/S0014-5793%2801%2902589-3`.

[35] Mcewan CE, Gatherer D, Mcewan NR. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas. 1998;128(2):173–178.

[36] Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. Genomic GC level, optimal growth temperature, and genome size in prokaryotes. Biochemical and biophysical research communications. 2006;347(1):1–3.

[37] Galtier N, Lobry J. Relationships between genomic G+ C content, RNA secondary structures, and optimal growth temperature in prokaryotes. Journal of molecular evolution. 1997;44(6):632–636.

[38] Hurst LD, Merchant AR. High guanine–cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes.

Proceedings of the Royal Society of London Series B: Biological Sciences. 2001;268(1466):493–497.

[39] Brbić M, Piškorec M, Vidulin V, Kriško A, Šmuc T, Supek F. The landscape of microbial phenotypic traits and associated genes. Nucleic Acids Research. 2016 Dec;44(21):10074–10090. Available from: https://academic.oup.com/nar/article/44/21/10074/2290929.

[40] Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic acids research. 2013;42(D1):D553–D559.

[41] 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. Available from: http://www.sciencedirect.com/science/article/pii/S072320200800060X.

[42] McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351(6328):652.

[43] 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.

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

[45] 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.

[46] Yahara K, Didelot X, Jolley KA, Kobayashi I, Maiden MC, Sheppard SK, et al. The landscape of realized homologous recombination in pathogenic bacteria. Molecular biology and evolution. 2015;33(2):456–471.

[47] González-Torres P, Rodríguez-Mateos F, Antón J, Gabaldón T. Impact of homologous recombination on the evolution of prokaryotic core genomes. mBio. 2019;10(1):e02494–18.

[48] Liu H, Huang J, Sun X, Li J, Hu Y, Yu L, et al. Tetrad analysis in plants and fungi finds large differences in gene conversion rates but no GC bias. Nature ecology & evolution. 2018;2(1):164.

[49] Marsolier-Kergoat MC, Yeramian E. GC content and recombination: reassessing the causal effects for the Saccharomyces cerevisiae genome. Genetics. 2009;183(1):31–38.

[50] Brissett NC, Pitcher RS, Juarez R, Picher AJ, Green AJ, Dafforn TR, et al. Structure of a NHEJ polymerase-mediated DNA synaptic complex. Science. 2007;318(5849):456–459.

[51] Brissett NC, Doherty AJ. Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway. Biochemical Society transactions. 2009 Jun;37:539–545.

[52] Della M, Palmbos PL, Tseng HM, Tonkin LM, Daley JM, Topper LM, et al. Mycobacterial Ku and ligase proteins constitute a two-component NHEJ repair machine. Science. 2004;306(5696):683–685.

[53] 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.

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

[55] 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.

[56] 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.

[57] 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.

[58] 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. Available from: `https://academic.oup.com/nar/article/38/suppl_1/D234/3112229`.

[59] Cox MM, Battista JR. Deinococcus radiodurans-the consummate survivor. Nature Reviews Microbiology. 2005;3(11):882.

[60] Rohwer F, Azam F. Detection of DNA damage in prokaryotes by terminal deoxyribonucleotide transferase-mediated dUTP nick end labeling. Appl Environ Microbiol. 2000;66(3):1001–1006.

[61] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research. 2016 Jan;44(Database issue):D733–D745. Available from: `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702849/`.

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

[63] Weissman JL, Laljani RM, Fagan WF, Johnson PL. Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. The ISME journal. 2019;.

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