Bohlin et al. Big Data Analytics (2019) 4:5
https://doi.org/10.1186/s41044-019-0042-7

Big Data Analytics

RESEARCH Open Access

Estimation of AT and GC content distributions of nucleotide substitution rates in bacterial core genomes

Jon Bohlin[1*], Brittany Rose[1,2] and John H.-O. Pettersson[3,4]

* Correspondence: jon.bohlin@fhi.no
[1]Norwegian Institute of Public Health, P.O. Box 4404, Lovisenberggata 8, 0403 Oslo, Norway
Full list of author information is available at the end of the article

Abstract

Background: Genomic GC content varies both within and, substantially, between microbial genomes. While some of this variation can be explained by evolutionary divergence and environmental factors, a notable portion is not understood. To investigate further, we explore a non-linear mathematical model (gcMOD) of single-nucleotide polymorphism (SNP) GC content (sbGC, the GC content of substituted bases) as a function of core genome GC content (cgGC). We estimate the model's parameters using Bayesian inference on empirical genetic data from the microbial core genomes of 35 bacterial species, each of which contains at least 10 representative strains. We utilize 716 bacterial genomes in total. We also explore some possible implications that result from the mathematical properties of gcMOD.

Results: We find that the median GC → AT substitution rates (β) are almost always considerably higher than the corresponding AT → GC substitution rates (α) for all 35 core genomes. The distribution of β is also noticeably more concentrated (i.e. thinner) than the corresponding distribution of α for almost all species, excepting the bacteria with the most GC-rich genomes. We also demonstrate that at the singularity point of gcMOD (where $\alpha = \beta$), the model is reduced to a linear equation. By analyzing the linear model, we show that due to the constraints on gcMOD, the mutation rates can have profound influence on both cgGC as well as sbGC. Moreover, by examining the mathematical properties of gcMOD's inverse function, we find that change in cgGC, and hence in genomic GC content, can potentially occur quite rapidly.

Conclusions: Examining the distributions of the GC → AT and AT → GC substitution rates for 35 bacterial species, we demonstrate that the former (β) are remarkably similar for all species examined. In addition, GC → AT substitution rate distributions were considerably more concentrated for all species, with the mode consistently peaking at higher rates than for AT → GC substitution rates.

Keywords: Bacterial genomics, Core genome analysis, Single nucleotide polymorphisms, Evolutionary biology

Background

Chargaff's parity rules [1] state that the number of adenine (A) nucleotides is similar to the number of thymine nucleotides (T) in double-stranded DNA due to Watson–Crick base pairing. Likewise, the number of guanine nucleotides (G) is approximately the same as the number of cytosine nucleotides (C). It is therefore common to refer to

genomic base composition either as AT or GC content; we shall henceforth do the latter. GC content (%GC) in prokaryotes varies substantially, especially between taxa [2, 3]. The forces behind this variation have not been completely resolved, although both phylogenetic relationships and environmental influences appear to be fundamental aspects [4]. Microbial communities coinhabiting similar environments tend to have similar %GC regardless of taxa [5]. Factors such as nitrogen abundance [6], AT-biased mutations due to loss of DNA repair genes [7], population density [8] and selective pressures [9–11] may explain some of the variance [2, 3, 12, 13] that spans from 13.5% GC in the intracellular symbiont *Candidatus* Zinderia insecticola to 75% GC in the soil bacterium *Anaeromyxobacter dehalogens* [14].

In a previous publication [15], we found an association between %GC (cgGC) of the core genome and %GC of substituted bases (sbGC, predominantly in SNPs, but referred to as sbGC to account for non-removed recombined sites and sequencing errors). The association indicated a strong bias in sbGC for the most AT-rich genomes, while the opposite (AT bias) was observed for the most GC-rich genomes (%GC ≥ 60%). We created a mathematical model (gcMOD) that considered sbGC as a function of cgGC [15]. Two parameters, one for AT → GC mutation rates (α) and the other for GC → AT mutation rates (β), had to be estimated so that gcMOD could make predictions of sbGC. *It should be noted that GC or AT mutation rates should here be taken to mean that an A or a T nucleotide mutates to a C or a G nucleotide (and* vice versa*).* Initially, we estimated α and β using non-linear least squares regression (NLS) to fit gcMOD to empirical data. The data consisted of 716 fully sequenced, closed genomes comprising 35 bacterial species from 6 phyla, each species having more than 10 strains (see [15] for details). The resulting parameter estimates indicated that the best model was obtained when GC → AT mutation rates outnumbered AT → GC mutations approximately 2 to 1.

The purpose of the present study is twofold: first, to expand on the previous study [15] by estimating mutation rate parameters both for each core genome and collectively (e.g. one α and β for all) using Bayesian inference, and second, to delve deeper into the mathematical properties of gcMOD.

Results

As previously shown [15], gcMOD describes sbGC as a function of cgGC (more details in the Methods section). More precisely, core genome SNP GC content is assumed to be a function of total core genome GC content. Both sbGC and cgGC are found to comply with Chargaff's parity laws [15]. Change in sbGC with respect to cgGC is thus modeled as a fraction of AT → GC mutation rates (α) and GC → AT mutation rates (β). In a recent publication [15], this was carried out using NLS against empirical data, as described above. Here we re-estimate α and β with Bayesian inference using the same empirical data, and we expand on the previous analyses by estimating empirical distributions for α and β for each of the 35 core genomes. Thus, we estimate mutation rate parameters for each core genome by considering sbGC for each strain with respect to its corresponding cgGC. We also estimate α and β collectively for mean sbGC and cgGC for all core genome/species in bulk. This was also done in the previous study (see additional file 5 in [15]), and we compare those results to our present results below. We choose this approach in order to explore the more asymptotic (i.e. long-

term) properties of SNP GC content in microbial core genomes. We assign normal prior distributions to α and β, and we assign normally distributed hyperparameters to the prior distributions' respective means. We also assume that gcMOD's model errors follow a normal distribution, but with a fixed mean $\mu = 0$ and precision $1/\sigma$ modeled as a gamma distribution. More details regarding the model setup can be found in the Methods section.

Parameter estimates based on all species/core genome sbGC (i.e. parameter estimates based on bulk sbGC from all strains constituting each core genome/species) do not indicate substantial deviations from previous NLS-based estimates. Figure 1 shows slight deviations from the posterior distribution implicitly assumed by the NLS-based method (a Student's *t*-distribution; see [15], additional file 5). Furthermore, we find (see Fig. 2) that $\alpha = -1.443$ 95%CredI(− 2.680, − 0.638) and $\beta = 2.645$ 95%CredI (2.093, 3.660). Previous NLS estimates were $\alpha = -1.35$ *95%CI(− 2.16, − 0.54)* and $\beta = 2.59$ *95%CI (1.99, 3.19)* for the bulk dataset. This suggests both higher (specifically, 2.645/1.443 = 1.83 times higher) and more concentrated GC → AT mutation rates than AT → GC mutation rates (95%CredI (2.093, 3.660) for β versus 95%CredI(− 2.680, − 0.638) for α; see Methods section for details).

Using Bayesian statistics, we estimate AT → GC and GC → AT mutation rates for each species/core genome (35 α and β parameter pairs in total). In other words, we fit gcMOD for each core genome. We assign hyperpriors to the precisions (gamma) of both α and β, but we fix rather than assign hyperpriors to the means (more details in the Methods section). Figure 3 shows that the general trend of thinner distributions for β is largely the trend for all species. Actual estimates can be found in Additional file 1. Closer examination of these results, however, reveals that the most GC-rich microbes (e.g. *Brucella* spp., *Pseudomonas* spp. *and Mycobacterium tuberculosis)* have estimated posterior distributions for α and β that are more similar in terms of precision. For most other species, the estimated distributions for α vary considerably—more than for the

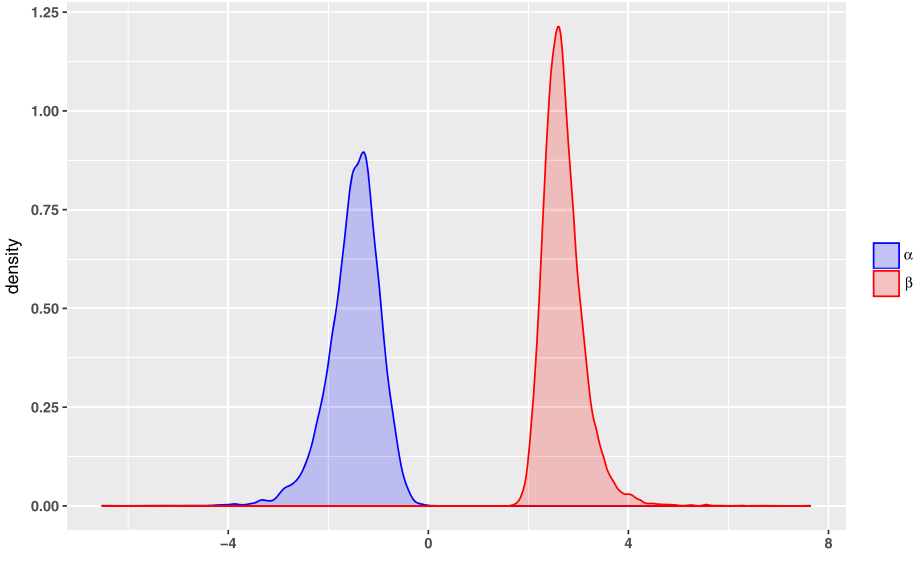

Fig. 1 Density plot of the posterior distributions of bulk estimated α (blue) and β (red) for all 35 species/core genomes. The red distribution represents GC→ AT mutation rates while the blue distribution AT→ GC mutation rates

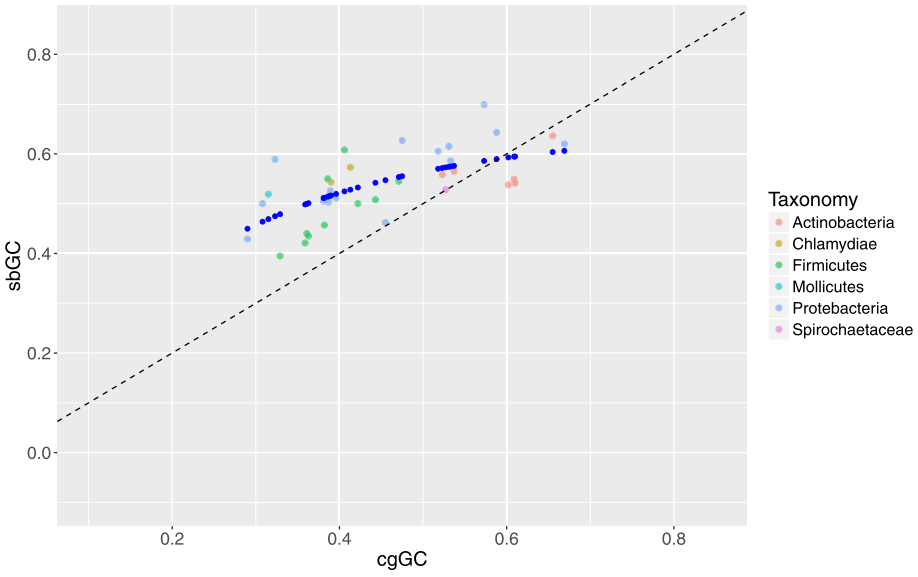

Fig. 2 Bulk SNP GC content (sbGC) for each of the 35 core genomes (vertical axis) plotted against corresponding core genome GC content (cgGC, horizontal axis). Dots are colored according to taxonomic group (see legend). The dark blue points represent gcMOD with median α (AT→GC mutation rates) and β (GC→AT mutation rates) estimated using Bayesian inference. The dashed line represents an alternative hypothesis, namely identical SNP GC- and core genome GC content (sbGC = cgGC)

corresponding estimated distributions for β. A notable exception is the case of the pathogen *Francisella tularensis,* which is known for its ability to acquire DNA horizontally [16]. Additionally, only 12 closed strains of *F. tularensis* were available, so genomic heterogeneity between its strains might bias estimates.

It is evident from the formulation of gcMOD that if $\alpha = \beta$ (i.e. if the AT → GC and GC → AT mutation rates are identical), the equation approaches a singularity point. A mathematical inquiry into this singularity, however, reveals that the equation degenerates into a linear equation with a slope dependent on the mutation rates (see Methods section).

Since sbGC is modeled as a function of cgGC, linearizing gcMOD by setting $\alpha = \beta$ indicates that the mutation rates impose constraints on cgGC (x in gcMOD), as $\beta\ x < 1 \Rightarrow x < 1/\beta;\ \beta \neq 0$. Indeed, as can be seen from Fig. 4 and the Methods section, the derivative of gcMOD is decreasing and approaches 0 only as $x \to \infty$, which implies (by the inverse function theorem) that gcMOD has an inverse function. Figure 5 demonstrates that the inverse of gcMOD increases in an exponential-like manner. Therefore, change in sbGC can induce rapid change in cgGC, and thus in genomic GC content in general. This can be mediated, for example, by loss of mismatch and repair genes (MMR), something that was observed by Lind and Anderson in a knockout experiment on *Salmonella typhimurium* [17].

Discussion

Figure 1 shows that the absolute values of GC → AT mutation rates are considerably larger than those of AT → GC mutation rates. Closer examination of all mutation rates reveals that the more frequent GC → AT mutation rates appear to be consistent across species, while AT → GC mutation rates exhibit considerable variation between species.

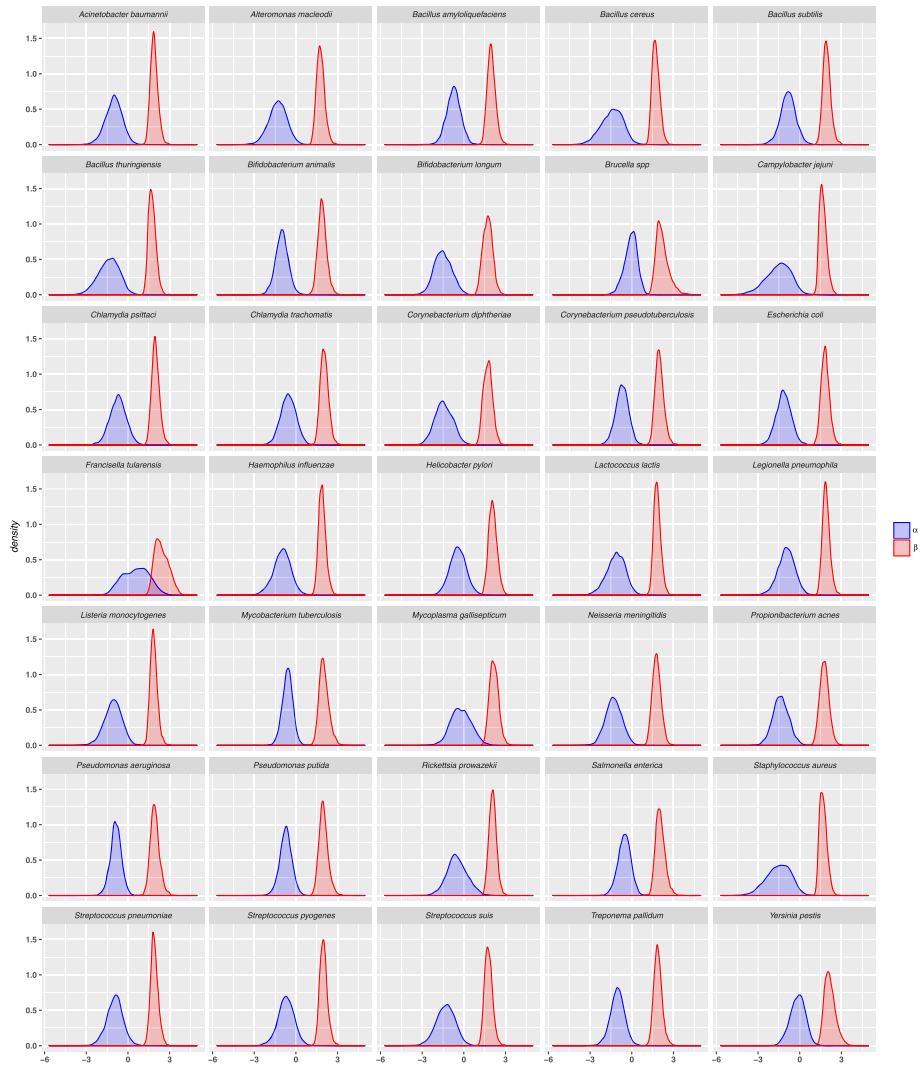

Fig. 3 Density plots of the posterior distributions of estimated α (blue) and β (red) for each of 35 core genomes. The red distribution represents GC→ AT mutation rates while the blue distribution AT→ GC mutation rates. Each core genome consists of at least 10 strains of a distinct species, with the exception of *Brucella* spp., which contains multiple species due to their close genomic similarity

The remarkably similar and consistently high GC → AT rates observed for the majority of the 35 microbial species are assumed to be a consequence of the evidence-based hypothesis that mutations are universally AT-biased [18]. Due to the fact that β is negative, one interpretation of the higher variance in AT → GC mutation rates between core genomes/species is that selection and/or a selective neutral force [19, 20] acts to retain G/C nucleotides [9]. Hence, an interpretation of gcMOD (from the lower absolute median/mode β estimates in Figs. 1 and 3) is that G/C nucleotides leave microbial genomes less frequently than A/T nucleotides enter. This suggests that a mutation in a clonal isolate of a bacterial species (e.g. a pathogenic strain in a disease outbreak) would most likely be an A/T mutation with a high probability of not being fixed, as G/C nucleotides are typically retained over longer periods of time. The species with the most similar distributions for α and β are all GC-rich, i.e. *M. tuberculosis, P. aeruginosa, P. putida* and *Brucella spp.* However, even for these species, we find that absolute

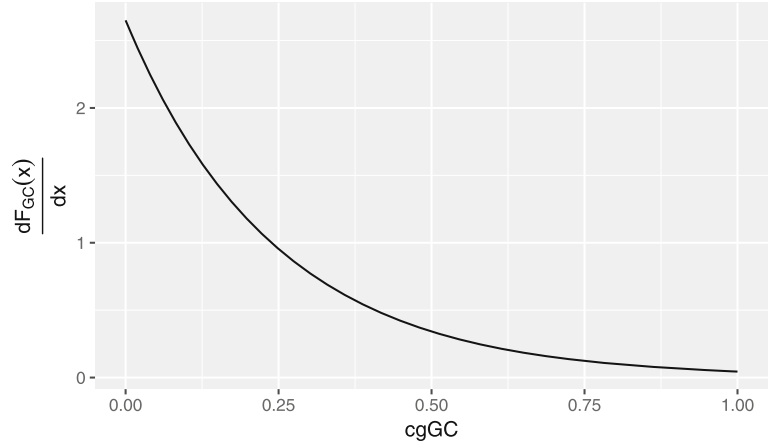

Fig. 4 The derivative of the SNP GC content model (gcMOD) (vertical axis) with respect to core genome GC content plotted against core genome GC content (cgGC, vertical axis) for bulk estimated median α (AT→ GC mutation rates) and β (GC→ AT mutation rates) parameters for all 35 core genomes

median/mode β estimates are substantially higher than α estimates. Indeed, *Brucella spp.* has the lowest AT → GC mutation rates, together with *Mycoplasma gallisepticum* and *Yersinia pestis.*

gcMOD is the nonlinear solution of a linear differential equation, thus requiring nonlinear statistical methods for the estimation of its parameters. If a likelihood function can be produced, Bayesian inference is an easy and efficient way to do this. While we previously used nonlinear regression to estimate α and β for gcMOD, it was sometimes challenging to find appropriate starting values for these parameters such that the NLS method converged. Once suitable starting values were identified, the NLS method converged and provided estimates for both α and β. Random effects can be added to NLS models with R packages like nlme [21]. Unfortunately, it can be difficult to get models with a hierarchical structure of random effects (including gcMOD) to converge with NLS-based methods.

By utilizing Markov chain Monte Carlo (MCMC) based Bayesian inference, on the other hand, one can make several adjustments with regards to how the model is specified and choose appropriate prior distributions for the parameters being estimated. Although gcMOD has a singularity point where $\alpha = \beta$, posterior estimates of these parameters are straightforward to obtain with Bayesian inference. The parameters estimated by nonlinear regression are assumed to follow an asymptotic Student's *t*-distribution. As discussed in the Results section, the 95% CI intervals are somewhat smaller than the 95% CredI intervals obtained from Bayesian analysis. This is likely due to the fact that no assumptions are made about the types of posterior distributions that are empirically estimated using MCMC simulations. Testing the models with uninformative uniform priors did not change the conclusion of the analysis or the distributions of the estimated parameters to any substantial degree.

A closer inspection of gcMOD's mathematical properties reveals that setting mutation rates for AT→ GC and GC → AT equal to one another reduces gcMOD to a linear equation. Despite this, the change in sbGC with respect to cgGC is still determined by $\alpha = \beta$; high rates increase the slope, while low rates decrease it (see Fig. 2, dashed

line). The different α and β estimates observed for all bacterial species suggest that AT → GC and GC → AT mutation rates are not necessarily equal, thereby indicating that the alternative model, or null hypothesis, of a linear relationship between sbGC and cgGC (described in Fig. 2) does not adequately explain the observed data.

Further inspection of gcMOD's properties reveals that its derivative approaches 0 only as cgGC approaches infinity. This implies that gcMOD has a mathematical inverse (as outlined in the Methods section). A bit of algebra (also in the Methods section) reveals that gcMOD predicts that cgGC may change abruptly with respect to sbGC. Thus, core genome GC content, and therefore also genomic GC content, may change quickly as α and β vary, provided there is strong enough selection, or lack thereof, on sbGC. From our data, we see that the distributions of GC → AT mutation rates are similar for all species, suggesting that cgGC can change rapidly according to how often C/G nucleotides are retained. Indeed, as mentioned above, Lind and Anderson [17] demonstrated a fast decrease in *S. typhimurium* GC content by knocking out MMR genes.

Conclusions

Using Bayesian inference, we find that across all 35 microbial core genomes examined, GC → AT mutation rates (β) are remarkably similar to one another, while AT → GC mutation rates (α) are considerably more heterogeneous. Only for the most GC-rich species are the GC → AT and AT → GC mutation rate distributions similar in shape. Median GC → AT mutation rates, however, are substantially higher than AT → GC mutation rates at the species level, by a factor of roughly two on average. Based on estimates of α and β, we speculate that G/C mutations are more often retained within bacterial genomes, possibly due to selection, and that A/T nucleotides enter the same genomes more frequently but are less seldom fixed.

Inspection of gcMOD's mathematical properties reveals that it is possible to obtain its inverse function, and examination of this inverse indicates that cgGC, and thus total genomic GC content, can potentially change rapidly. We also find that cgGC appears to be strongly constrained by mutation rates. The above results taken together seem to suggest that the presence of higher GC → AT than AT → GC mutation rates has a profound effect on genomic GC content in bacteria.

Methods

Data preparation

All genetic material considered here was obtained from Genbank/NCBI [22] and is described in a recent study [15]. That study also describes in detail how SNPs (e.g. sbGC) are extracted, and all data used in the present study is available there (additional files 3 and 4 in [15]). Due to the stringency required for estimating SNPs, we only considered closed genomes and required that each core genome was based on more than 10 strains. This resulted in a dataset consisting of a total of 716 different bacterial genomes divided amongst 35 core genomes. With one exception, all the core genomes consist of different species. The core genome of *Brucella* spp. also consists of different genera that have strong genomic similarity to one another (see [15]).

Sequences were prepared via the process described in [12, 15]. Briefly, core genomes (containing coding and non-coding regions) were extracted using Harvesttools [23], and corresponding core genome SNPs were retained with Gubbins v. 1.4.5 [24]. Seaview v. 4.5.4 was used to manually call sbGC and cgGC for all species and strains [25]. In the present study, we consider mean sbGC from all strains in each core genome, finding that sbGC represents bulk SNP GC content of all strains in each species (see additional information 4–6 in [15]). We also estimate mutation rates with respect to each strain in each core genome (see Fig. 3 and figures. 1–2 in [15]). We created all figures with the free statistical software package R [26] and its library ggplot2 [27].

Derivation of gcMOD and estimation of parameters

We derived the mathematical model used in this study, gcMOD, according to the guidelines set forth in [15]. In broad terms, gcMOD describes the linear difference between, on one hand, the change in sbGC ($F_{GC}(x)$) with respect to cgGC (x), and on the other, the AT → GC and GC → AT mutation rates (α and β, respectively). α and β are taken to be to be scalars multiplied by sbGC. Written as an equation,

$$\frac{dF_{GC}(x)}{dx} = \alpha F_{GC}(x) + \beta(1 - F_{GC}(x)) \quad (1)$$

We estimated α and β using Bayesian inference, with the genomic data described above, which allowed us to model the mutation rate distributions of each core genome with considerably more detail than the previously used NLS method. Bayesian inference is performed by assuming prior distributions on the parameters (i.e. "expert" knowledge) of a particular model to be estimated. Next, to produce a posterior distribution, the priors are combined with empirical data and the model's given likelihood function. The resulting posterior distributions represent data and prior driven model estimates (i.e. α and β) for each core genome. More specifically, we fit two models using R 3.4.4 [26] and JAGS V.4.3.0 [28, 29]. The first was fitted for all species using the bulk sbGC/cgGC for each core genome (i.e. one bulk/mean α and β parameter estimate for all species; see Fig. 1 and Methods section). The second model was fitted strain-wise for each of the 35 species (i.e. 35 different α and β parameter estimates, one for each core genome; see Fig. 3 and Additional file 1).

Priors for α and β were, for the bulk case, set to be normal with a normal hyperprior for the mean (with − 1 and 1 set as the α and β hyperprior means, respectively). Precision was set to 1E-2 for both. Errors were assumed to be normally distributed with gamma-distributed precision 1E-3 for both scale and shape parameters. One chain was run for 1 million iterations, half of which were discarded as burn-in. Thinning was set to $n = 50$, and so a total of 10,000 iterations were saved. Effective sample size (ESS) was $n = 4486$ for α and $n = 4407$ for β. The hyperpriors for the means of both α and β were set differently due to the singularity point at $\alpha = \beta$. Starting values were also selected differently for each parameter, namely $\alpha = -1$ and $\beta = 1$. The estimated posterior distributions for α and β can be seen in Fig. 1. Median estimates were used in the model fit seen in Fig. 2.

We set up the Bayesian model estimating α and β strain-wise with 5 chains, each run for 5 million iterations (half discarded as burn-in) and thinning set to 100, resulting in 25,000 saved iterations. α and β were estimated for each of the 35 species, and ESS was

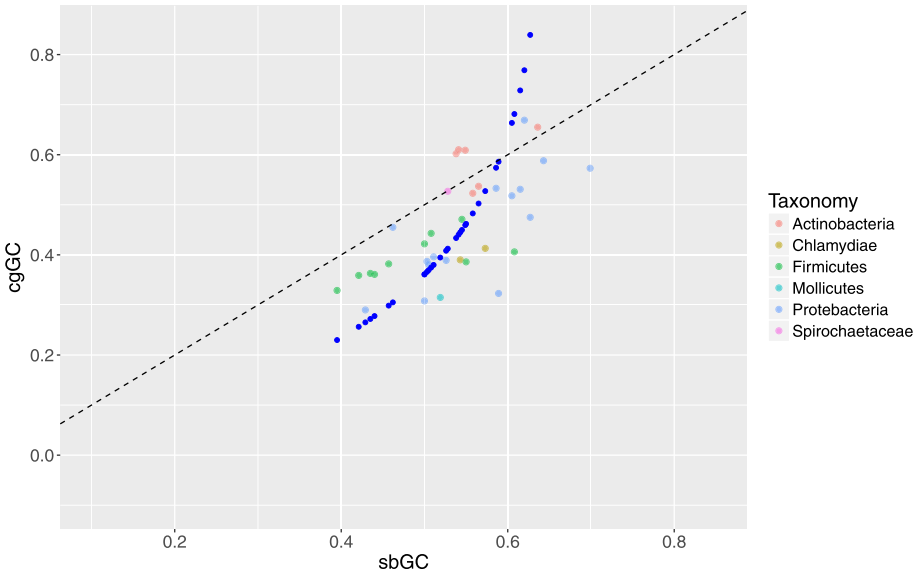

Fig. 5 gcMOD's inverse Core genome GC content (cgGC, vertical axis) plotted against SNP GC content (sbGC, horizontal axis), together with the SNP GC content model's (gcMOD) inverse function (dark blue points) for estimated median α (AT→GC mutation rates) and β (GC→AT mutation rates). Except for gcMOD's inverse, the dots are colored according to phylogenetic group membership stated in the legend to the right

at least$n = 7164$ for all parameters. Priors and starting values ($\alpha = -1$, $\beta = 1$) were the same as in the bulk model, but the means of the priors for α and β were not set using hyperpriors; rather, they were assumed to be 0, but with different starting values. Precision was presumed to have gamma-distributed hyperpriors with both shape and scale parameters set to 1E-3. Normal errors were again anticipated with gamma-distributed precision 1E-3 for both scale and shape parameters.

Bayesian parameter estimates are reported with median values and 95% credible intervals (CredI), while results based on standard frequentist methods (e.g. NLS) are reported with means and 95% confidence intervals (CI).

Some mathematical properties of gcMOD

It was shown in [15] that gcMOD (1) can be written as

$$F_{GC}(x) = \frac{\beta}{\alpha-\beta}\left(e^{(\alpha-\beta)x}-1\right) \tag{2}$$

which is subject to the constraints

$$0 < F_{GC}(x) < 1 \text{ and } 0 < x < 1.$$

Using the chain rule, we see that (2) has the derivative

$$F'_{GC}(x) = \beta e^{(\alpha-\beta)x} \tag{3}$$

Since $|\beta| > |\alpha|$, the derivative will always be positive and approach zero as $x \to \infty$. This implies, by the inverse function theorem, that (2) has an inverse, which after a bit of algebra can be expressed as

$$x = H_{GC}(y) = \frac{1}{\alpha-\beta}\ \log\left(\frac{\alpha y-\beta(y-1)}{\beta}\right) \tag{4}$$

Furthermore, closer inspection of (2) at the singularity $\alpha = \beta$ *reveals that.*

$$\begin{aligned} F_{GC}(x) &= \frac{\beta}{\alpha-\beta}\left(e^{(\alpha-\beta)x}-1\right) \\ &= \frac{\beta}{\alpha-\beta}\left(1+(\alpha-\beta)x+\frac{(\alpha-\beta)^2x^2}{2!}+\cdots+\frac{(\alpha-\beta)^nx^n}{n!}+\cdots-1\right) \\ &= \beta\left(\frac{1}{\alpha-\beta}+x+\frac{(\alpha-\beta)^1x^2}{2!}+\cdots+\frac{(\alpha-\beta)^{n-1}x^n}{n!}+\cdots-\frac{1}{\alpha-\beta}\right) \\ &= \beta\left(x+\frac{(\alpha-\beta)^1x^2}{2!}+\cdots+\frac{(\alpha-\beta)^{n-1}x^n}{n!}+\cdots\right) \end{aligned} \tag{5}$$

It should now be clear that $\lim_{\alpha\to\beta} F_{GC}(x) = \beta x$. For equal mutation rates $\alpha = \beta$, *(2)* is restricted by $0 < \beta x < 1$, which implies that cgGC is subject to the condition $x = 1/\beta$, $0 < x < 1$, $\beta \neq 0$.

Additional file

Additional file 1: Output from JAGS run of α and β estimates for each of the 35 species/core genomes. (TXT 6 kb)

Abbreviations

AT/GC content: A + T or G + C nucleotides divided by DNA sequence length; cgGC: Core genome GC content; CI: Confidence interval; CredI: Credible interval; gcMOD: The mathematical model for sbGC; mbp: Mega base pairs; millions of base pairs; MCMC: Markov-chain Monte Carlo; NLS: Non-linear least squares regression method; sbGC: The GC content of core genome substituted bases; SNP: Single nucleotide polymorphism; α: Parameter that designates AT → GCsubstitution rates in gcMOD; β: Parameter that designates GC → ATsubstitution rates in gcMOD

Acknowledgements

Not applicable.

Authors' contributions

Initiated the project: JB. Wrote the paper: JB and BR. Evolutionary and genomic analyses: JB, JP. Statistical and mathematical analyses: JB. All co-authors contributed to the writing of the manuscript. All authors have read and approved the final manuscript.

Funding

The work was funded by the Norwegian Institute of Public Health.

Availability of data and materials

All genomes used in the present study are publicly available at NCBI: https://www.ncbi.nlm.nih.gov/genome/microbes/ The present study is based on the data generated from a previous study [15] (Additional files 3 and 4) of which all data is available in open-access format.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors declare that they have no competing interests.

Author details

[1]Norwegian Institute of Public Health, P.O. Box 4404, Lovisenberggata 8, 0403 Oslo, Norway. [2]Department of Biostatistics, University of Oslo, 0317 Oslo, Norway. [3]Department of Medical Biochemistry and Microbiology, Zoonosis Science Center, Uppsala University, 751 05 Uppsala, Sweden. [4]Marie Bashir Institute for Infectious Diseases and

Biosecurity, Charles Perkins Centre, School of Life and Environmental Sciences, and Sydney Medical School, The University of Sydney, Sydney, New South Wales 2006, Australia.

Received: 11 December 2018 Accepted: 31 July 2019
Published online: 14 August 2019

References

1. Chargaff E. Structure and function of nucleic acids as cell constituents. Fed Proc. 1951;10(3):654–9.
2. Bentley SD, Parkhill J. Comparative genomic structure of prokaryotes. Annu Rev Genet. 2004;38:771–92.
3. Agashe D, Shankar N. The evolution of bacterial DNA base composition. J Exp Zool B Mol Dev Evol. 2014;322(7):517–28.
4. Reichenberger ER, Rosen G, Hershberg U, Hershberg R. Prokaryotic nucleotide composition is shaped by both phylogeny and the environment. Genome Biol Evol. 2015;7(5):1380–9.
5. Foerstner KU, von Mering C, Hooper SD, Bork P. Environments shape the nucleotide composition of genomes. EMBO Rep. 2005;6(12):1208–13.
6. 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):173–8.
7. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 2012;10(1):13–26.
8. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. 2016;17(11):704–14.
9. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010; 6(9):e1001107.
10. Raghavan R, Kelkar YD, Ochman H. A selective force favoring increased G+C content in bacterial genes. Proc Natl Acad Sci U S A. 2012;109(36):14504–7.
11. Bobay LM, Ochman H. Impact of recombination on the base composition of Bacteria and archaea. Mol Biol Evol. 2017; 34(10):2627–36.
12. Bohlin J, Eldholm V, Pettersson JH, Brynildsrud O, Snipen L. The nucleotide composition of microbial genomes indicates differential patterns of selection on core and accessory genomes. BMC Genomics. 2017;18(1):151.
13. Bohlin J, Snipen L, Hardy SP, Kristoffersen AB, Lagesen K, Donsvik T, Skjerve E, Ussery DW. Analysis of intra-genomic GC content homogeneity within prokaryotes. BMC Genomics. 2010;11(1):464.
14. Nishida H. Evolution of genome base composition and genome size in bacteria. Front Microbiol. 2012;3:420.
15. Bohlin J, Eldholm V, Brynildsrud O, Petterson JH, Alfsnes K. Modeling of the GC content of the substituted bases in bacterial core genomes. BMC Genomics. 2018;19(1):589.
16. Bohlin J, Sekse C, Skjerve E, Brynildsrud O. Positive correlations between genomic %AT and genome size within strains of bacterial species. Environ Microbiol Rep. 2014;6(3):278–86.
17. Lind PA, Andersson DI. Whole-genome mutational biases in bacteria. Proc Natl Acad Sci U S A. 2008;105(46):17878–83.
18. Hershberg R, Petrov DA. Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet. 2010;6(9): e1001115.
19. Lassalle F, Perian S, Bataillon T, Nesme X, Duret L, Daubin V. GC-content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet. 2015;11(2):e1004941.
20. Lawrence JG, Ochman H. Amelioration of bacterial genomes: rates of change and exchange. J Mol Evol. 1997;44(4):383–97.
21. Pinheiro J, Bates D, DebRoy S, Sarkar D: R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1–117. *See http://CRAN R-project org/package= nlme* 2014.
22. Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2014;42(1):D32–7.
23. Treangen TJ, Ondov BD, Koren S, Phillippy AM. The harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. Genome Biol. 2014;15(11):524.
24. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43(3):e15.
25. Gouy M, Guindon S, Gascuel O. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27(2):221–4.
26. Team RDC: R: a language and environment for statistical computing. R Foundation for Statistical Computing. 2011;2 14.
27. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2016.
28. Depaoli S, Clifton JP, Cobb PR. Just another Gibbs sampler (JAGS) flexible software for MCMC implementation. J Educ Behav Stat. 2016;41(6):628–49.
29. Su Y-S, Yajima M: R2jags: A Package for Running jags from R. *R package version 003-08, URL http://CRAN R-project org/ package= R2jags* 2012.

Publisher's Note