The ISME Journal (2018) 12:1180–1187
https://doi.org/10.1038/s41396-017-0023-7

ARTICLE

Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model

Ferdi L Hellweger[1] · Yongjie Huang[2,3] · Haiwei Luo[2,3]

Received: 13 August 2017 / Revised: 14 November 2017 / Accepted: 21 November 2017 / Published online: 12 January 2018

Abstract

An important unanswered question in evolutionary genomics is the source of considerable variation of genomic base composition (GC content) even among organisms that share one habitat. Evolution toward GC-poor genomes has been considered a major adaptive pathway in the oligotrophic ocean, but GC-rich bacteria are also prevalent and highly successful in this environment. We quantify the contribution of multiple factors to the change of genomic GC content of *Ruegeria pomeroyi* DSS-3, a representative and GC-rich member in the globally abundant *Roseobacter* clade, using an agent-based model. The model simulates 2×10^8 cells, which allows random genetic drift to act in a realistic manner. Each cell has a whole genome subject to base-substitution mutation and recombination, which affect the carbon and nitrogen requirements of DNA and protein pools. Nonsynonymous changes can be functionally deleterious. Together, these factors affect the growth and fitness. Simulations show that experimentally determined mutation bias toward GC is not sufficient to build the GC-rich genome of DSS-3. While nitrogen availability has been repeatedly hypothesized to drive the evolution of GC content in marine bacterioplankton, our model instead predicts that DSS-3 and its ancestors have been evolving in environments primarily limited by carbon.

Introduction

In pelagic marine environments, both the concentration of inorganic nitrogen [1] and the genomic GC content of bacterioplankton cells [2, 3] increase with water depth (Fig. S1). This correlation supports the hypothesis that evolution toward GC-poor genomes is an adaptive strategy for bacteria inhabiting surface oceans because a GC pair uses one more nitrogen (N) than an AT pair [4, 5]. Indeed, several most abundant bacterial lineages [6] including surface water members of the cyanobacterial genus *Prochlorococcus*, the alphaproteobacterial clade SAR11, and the gammaproteobacterial clade SAR86, have %GC of only 0.29–0.32 [7]. Nevertheless, several other successful surface ocean lineages [8, 9], such as the cyanobacterial genus *Synechococcus* and the alphaproteobacterial clade *Roseobacter*, have higher and more variable %GC (0.52–0.66 and 0.37–0.70).

This observation supports an alternate hypothesis that the pattern in GC content is the result from mutation bias toward GC in *Synechococcus* and *Roseobacter*, and toward AT in *Prochlorococcus*, SAR11, and SAR86. This hypothesis has gained support from an experimental study showing mutation bias toward GC in a GC-rich *Roseobacter* strain *Ruegeria pomeroyi* DSS-3 [10] and from a bioinformatics analysis showing that losses of DNA repair genes for correcting GC>AT mutations coincided with genomic GC content reduction during *Prochlorococcus* evolution [11]. Both hypotheses are plausible and the real question is not which one, but what the relative contribution of these (and other) mechanisms is to their evolution.

Several recent studies advanced our understanding on the role of mutation, selection, and recombination in driving

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41396-017-0023-7) contains supplementary material, which is available to authorized users.

✉ Ferdi L Hellweger
ferdi@coe.neu.edu

✉ Haiwei Luo
hluo2006@gmail.com

[1] Civil and Environmental Engineering Department, Northeastern University, Boston, MA 02115, USA

[2] Simon F. S. Li Marine Science Laboratory of School of Life Sciences and Partner State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, China

[3] Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen 518000, China

microbial GC content evolution. For example, numerous mutation accumulation (MA) experiments followed by whole genome sequencing of mutant lines (MA/WGS) provided unequivocal evidence that spontaneous mutations can bias toward either GC or AT depending on the organism under study [12]. While selection in driving GC content evolution was often speculated [12, 13] or directly shown by the experimental results that bacteria have an increased growth rate when they carry GC-richer versions of genes with the same encoded amino acid sequences [14, 15], few studies identified the source and nature of selection. Recent studies also debated on whether recombination acts as a universal force to increase [16] or decrease [17] GC content across microbial lineages. Since these studies (except MA/WGS) use genomic sequences of natural isolates, which are affected by complex evolutionary forces and processes, they generally adopt a reductionist approach by separating the mechanism of interest from other factors and thus can hardly lead to an understanding of the relative contribution of different forces to GC content change.

The effect of one or more mechanisms on genome evolution can be quantified using concepts from population genetics [18]. Given sufficient assumptions (e.g., uniform mutation rates, no recombination, constant and independent fitness effect of mutations) simple theories or equations can be developed and applied (e.g., to predict codon usage bias) [19]. For more complex scenarios, including multiple mechanisms and many sites (i.e., base pairs or genes), an alternative approach is to directly simulate a population of individual cells with whole genomes [20–22]. In this individual- or agent-based approach, single cells are simulated explicitly and the population-level properties (e.g., diversity and population size) emerge from the cumulative properties and behaviors of these low-level entities [23]. This is contrast to the more traditional population-level modeling approach, where population-level properties are described directly [19]. For evolution simulations, each individual has a genome that is subject to mutation and recombination, which affects the growth rate and over subsequent divisions the fitness of a strain. This approach is quite flexible and allows for inclusion of various evolutionary mechanisms, but is computationally prohibitive for large population sizes (required for random genetic drift to function in a realistic manner). This can be overcome to some extent by combining the simulation model with concepts from theory (e.g., explicit simulation of two genomes and representation of others using theory [24]). Here we develop a novel method that combines basic concepts from genetics for calculating the number of mutants in a culture [25] with upscaling methods from individual-based ecological modeling [26]. Our model simulates isogenic "super-individuals" with a realistic effective population size. A novel aspect of the model is the explicit consideration of mutation and recombination, the effect on the carbon (C) and N requirements in the DNA and amino acid (aa) pools and the effect of nonsynonymous mutations on the growth rate (Fig. 1, see "Materials and methods" section).

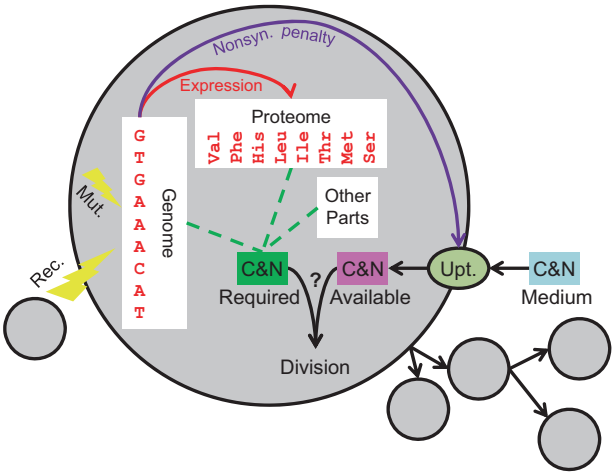

Fig. 1 Model schematic. Each cell has a C and N requirement based on genome, proteome, and other parts (RNA, lipids, etc). Cells take up C and N and divide when the amount of these nutrients meets the requirement. The DNA can change by mutation and recombination. A mutant reducing C or N requirement of the genome or proteome reduces the generation time and provides a competitive advantage to the cell under C- or N-limiting condition. On the other hand, the cell's nutrient uptake rate is penalized if a functionally deleterious nonsynonymous mutation is introduced to the genome. The model includes a large number of cells such that the random effect through genetic drift is incorporated in a realistic manner. See "Materials and methods" section for details

We use the model to understand and quantify the various mechanisms driving GC content evolution on the present-day *R. pomeroyi* DSS-3 genome. Then, we develop a long-term evolutionary trajectory from the most recent common ancestor (MRCA) of the *Roseobacter* clade. Together, our results show that GC content evolution is a complex process affected by the action and interaction of various mechanisms, and that DSS-3 evolved primarily in a C-limiting environment.

Materials and methods

The model simulates a population of individual *R. pomeroyi* DSS-3 cells, a strain that has been the representative heterotrophic bacterium inhabiting surface ocean [27] and numerous genetic, physiological, and ecological data are available to parameterize the model. Cells assimilate C and N nutrients from the extracellular environment at rates depending on the medium concentration (Michaelis–Menten kinetics). When the intracellular levels (quotas) reach the requirements for division the cell divides. Each cell has a complete genome of A/T/C/G-type letters

that is subject to mutation [20] based on an unbiased measure of DSS-3 mutation rate ($\mu = 1.39 \times 10^{-10}$ per site per cell division) and spectrum (the relative frequency of mutations among the four nucleotides) [10]. Homologous recombination is included with parameters (the effect of recombination to mutation or *r/m*) determined from closely related genome sequences of a *Ruegeria* species [28] using ClonalFrameML [29]. The total cellular nutrient requirements are made up of genome, proteome, and other parts. The genome C and N requirements are calculated from the DNA sequence based on the DNA base pair stoichiometry. The proteome requirements are calculated from the DNA based on the amino acid stoichiometry, taking into account the expression level of proteins. Some genes (e.g., housekeeping genes) are expressed at a higher level and thus make up a relatively larger part of the proteome. The level of protein expression is approximated by the relative abundance of mRNA determined using the continuous cultures of DSS-3 when cells are C or N limited [30]. The DNA and protein levels, and the total C and N quotas are based on measurements [31]. In an environment where N or C is limiting, mutations that reduce the total N or C requirement decrease the time needed to acquire the nutrients to make a new cell, which increases the division rate and provides the cell with a competitive advantage. On the other hand, nonsynonymous mutations are more likely to be deleterious than synonymous mutations [32]. Whether a nonsynonymous mutation is neutral or deleterious is a stochastic process with probabilities based on amino acid chemical distances [33]. Functionally deleterious nonsynonymous mutations are assumed to be lethal.

A large number of cells is simulated so that the effective population size (N_e) and thus the efficiency of natural selection on the model population is comparable to that in nature. This is made computationally feasible by simulating isogenic "super-individuals" that are representative of many "real individuals" [26]. The population is modeled using periodic dilution. During the growth period, each super-individual population grows exponentially. The number of new mutant cells produced in this population over the time interval, by new mutations or division of a new mutant cell, is estimated [25] and used to spawn a new agent. In the following dilution step, the number of real individuals in the two populations is reduced and the next growth period is simulated (see Text S2). The model is tested against theory for a number of simplified scenarios, including growth in a nutrient-limiting chemostat, accumulation of neutral and slightly deleterious mutations, recombination, and linked selection (see Text S3). The genomes computed by the model have similar characteristics as those from real *Roseobacter* populations, including the ratio of nonsynonymous to synonymous substitution rate ($d_N/d_S = 5.0 \times 10^{-2}$) consistent with previous measures of various *Roseobacter* lineages [34], the effect of recombination to mutation ($r/m = 0.5$–3.7), and the effective population size ($N_e = 2 \times 10^8$) calculated from $\pi_S = 2\ N_e\mu$ and close to the previous estimate of a *Ruegeria* lineage [28] related to DSS-3 [10], where π_S (5.6×10^{-2}) is the nucleotide diversity at synonymous site calculated by averaging all pairwise comparisons of d_S for a given single-copy orthologous gene family and then averaging across gene families (see Table S11).

The model is applied to DSS-3 and two other coastal bacteria, *Vibrio fischeri* and *Vibrio cholera*, which have genomic GC content of 0.38 and 0.47, respectively, and mutation spectrum determined using the MA experiments [35]. The GC content of the MRCA of the *Roseobacter* clade is predicted using a maximum-likelihood phylogenetic approach implemented in the *ape* package [36], which is used to infer ancestral states of a continuous character (e.g., GC content) [37, 38].

Results and discussion

Mechanisms driving GC content evolution

We run the model in a time-variable manner, starting with the present-day DSS-3 genome. The model predicts a relatively strong negative slope in GC content (i.e., GC content decreases with time) under N-limiting conditions and a weaker positive slope under C-limiting conditions (Fig. 2a). These slopes are due to a multitude of mechanisms and to understand and quantify their effect, we perform a number of simulations, where factors are added one by one. We compare the various simulations based on the resulting slope of GC, where we break out the genome (All) as well as noncoding and synonymous sites (X&S) and nonsynonymous sites (NS) (Fig. 2b). We also compare changes in the growth rate (Fig. 2c) and the C (Fig. 2d) and N (Fig. 2e) requirements, where we break out total, DNA and aa.

In the first simulation (s), mutations among A/T/C/G are equally likely and neutral. In the GC-rich DSS-3 genome, GC>AT mutations are more frequent than AT>GC mutations and therefore the GC slope is negative (i.e., the GC content decreases over time, Fig. 2b). We refer to this mechanism as site bias (s). The effect is stronger for noncoding and synonymous sites (X&S) compared to nonsynonymous sites (NS), consistent with their GC content ($GC_{X\&S} = 0.74$, $GC_{NS} = 0.59$). When mutation bias is included (s/m), the GC slope increases compared to the previous simulation (s) and is now positive, suggesting that the mutation bias toward GC is strong enough to overcome the site bias toward AT. When nonsynonymous mutations are penalized (s/m/p), the GC slope decreases from the

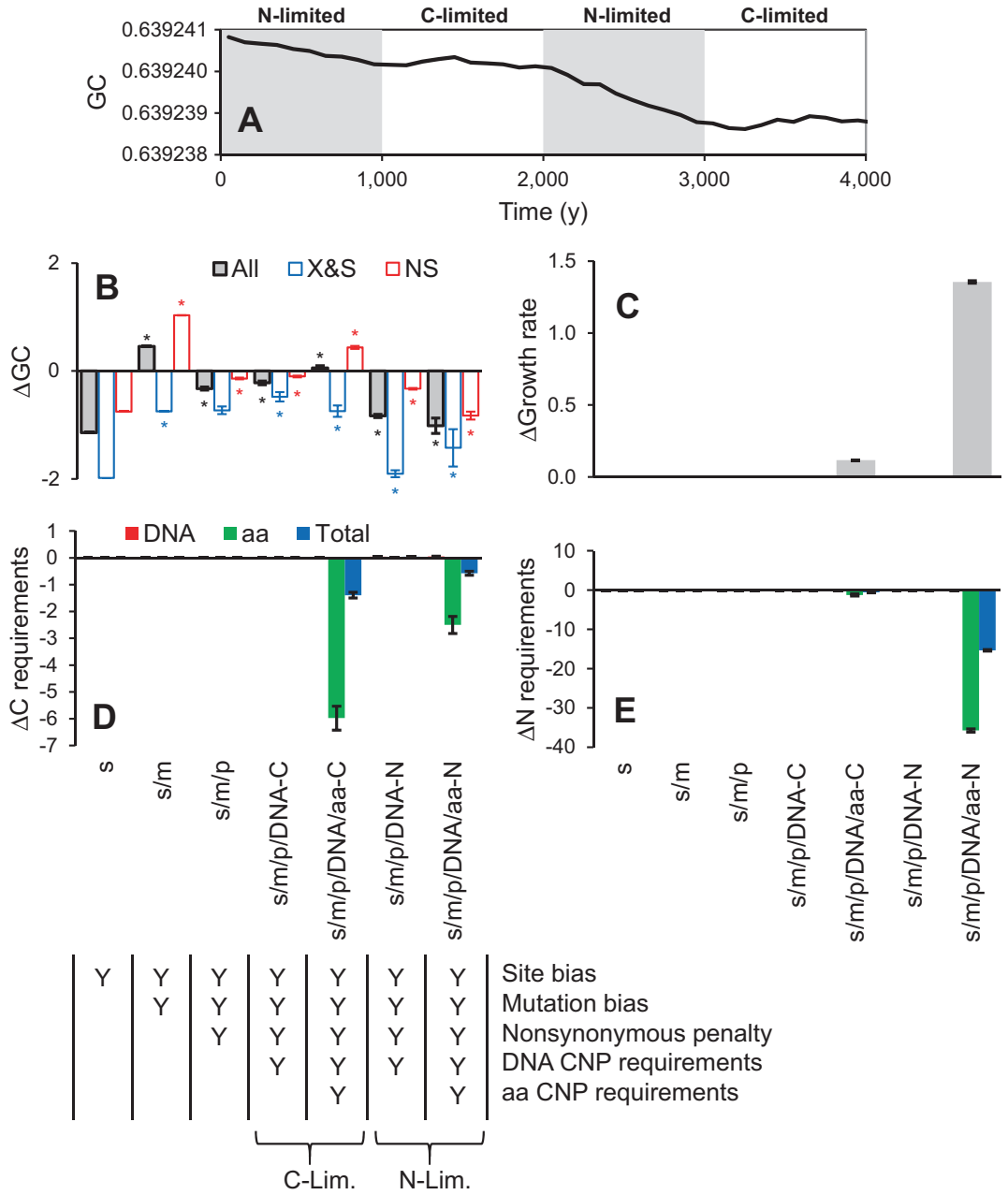

Fig. 2 Evolution of GC content in DSS-3. **a** GC content under alternating N- and C-limiting conditions. **b–e** Effect of various factors on the evolution of GC content. Simulations: s = site bias, m = mutation bias, p = nonsynonymous penalty, DNA = considering change of DNA C and N requirement, aa = considering change of amino acid C and N requirement, C = under C limitation, N = under N limitation. **b** Slope of GC content in fraction over a billion years (1e9y). X&S = noncoding and synonymous sites, NS nonsynonymous sites. **c** Slope of growth rate in d^{-1}/1e9y. **d** Slope of C requirement in relative units. **e** Slope in N requirement in relative units. Error bars are ±2 SEM based on multiple independent simulations. In **b**, results that are statistically different ($p < 0.05$) from the corresponding previous simulation are marked with "*". For example, GC_{NS} of simulation s/m/p/DNA-N is marked with "*" because the GC_{NS} of simulations s/m/p and s/m/p/DNA-N are significantly different

previous simulation (s/m) and is negative again. This is due to the elimination of most nonsynonymous mutants, which gives more weight to the noncoding and synonymous sites.

Next, we perform simulations under C limitation considering DNA C requirement (s/m/p/DNA-C) (Fig. 2b). This gives a slight increase in the GC slope compared to the previous simulation (s/m/p), consistent with selection pressure to reduce the DNA C requirement. A simulation considering both DNA and aa C requirement (s/m/p/DNA/aa-C) leads to a further increase and a positive GC slope. This is consistent with a negative correlation between aa C content and codon GC (see Fig. S4). The largest increase

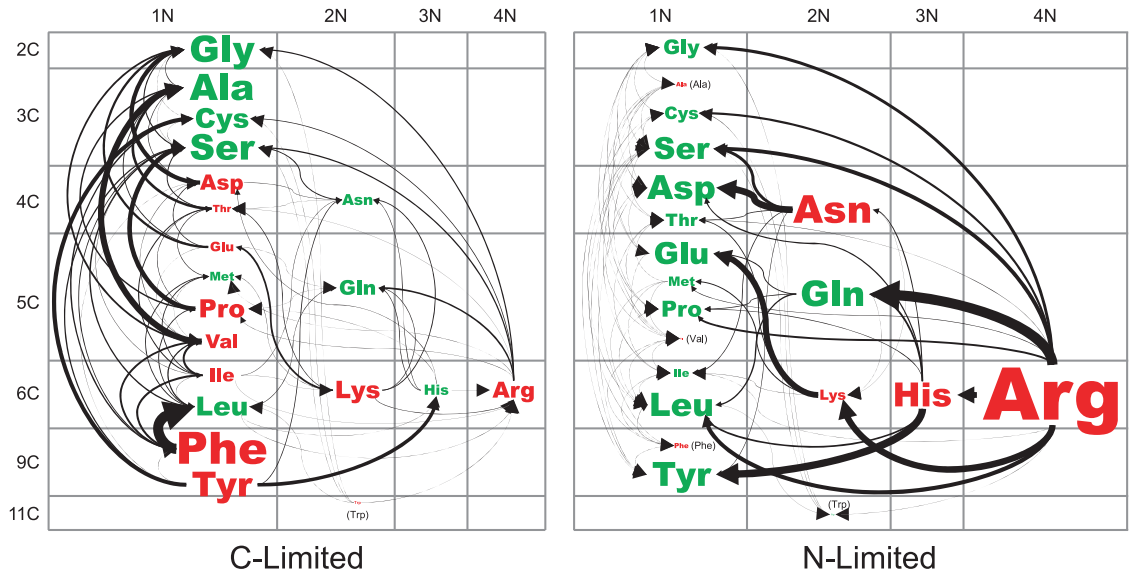

Fig. 3 Evolution of amino acid composition in DSS-3 under C- and N-limiting conditions. Amino acid labels are colored green for increase and red for decrease and scaled by magnitude of change. Arrows indicate net change and thickness corresponds to magnitude

and decrease is in the amino acid Ala (C = 3, GC = 0.96) and Phe (C = 9, GC = 0.23), respectively (Fig. 3). The overall GC slope of this simulation reflects the net effect of several drivers, including site bias (decreases GC), mutation bias (increases GC), nonsynonymous penalty (decreases GC), DNA C requirement (increases GC), and aa C requirement (increases GC). The overall slope is relatively weak, which suggests these drivers are presently in approximate equilibrium. The weak slope should not be interpreted as a weak influence of C limitation. Rather, the effect of C limitation is reflected in the difference of the simulations with it (s/m/p/DNA/aa-C) and without it (s/m/p).

Likewise, simulations are also performed under N limitation (Fig. 2b). By considering DNA N requirement (s/m/p/DNA-N), the GC slope decreases from the previous simulation (s/m/p) due to selection pressure to reduce DNA N requirements. The effect is larger than the comparable simulation under C-limiting conditions (s/m/p/DNA-C), because DNA N is a larger fraction of the total N requirement (6.4%) compared to C (2.6%). A simulation where both DNA and aa N requirement is considered (s/m/p/DNA/aa-N) further decreases the GC slope compared to s/m/p/DNA-N. This is due to a positive correlation between aa N content and codon GC (see Fig. S4). The largest increase is in Gln ($N = 2$, GC = 0.60) and the largest decrease is in Arg ($N = 4$, GC = 0.93) (Fig. 3).

For both C- and N-limiting conditions, a different GC slope is predicted for simulations with and without considering the aa nutrient requirement (Fig. 2b), suggesting that processes and mechanisms changing aa composition affect GC evolution. Multiple factors impact the evolution of the aa pool. An important consideration is the topology of the mutation network (Fig. 3). For example, under N limitation most low-N aa ($N = 1$) increase, but Phe ($N = 1$, GC = 0.23) actually decreases. Despite its low-N content, there is no fitness advantage of cells that have mutations toward Phe, because all possible one-step mutations to it are neutral with respect to N (Fig. 3). However, Phe has a low codon GC, which means it is eliminated by mutation bias. Another important factor is the interaction between selection acting to lower N requirement by reducing aa N content, selection for conservation of aa physicochemical property and mutation bias. The observation that the largest increase under N limitation is in Gln and not one of the 14 low-N aa ($N = 1$) (Fig. 3) is because mutations between Arg and Gln are more likely to be neutral than those between Arg and any of the low-N aa (see Fig. S2). Lys has an even smaller chemical distance to Arg, but it gains less mutations from Arg due to codon usage bias and loses more to Glu due to mutation bias (Fig. S20) so it actually decreases (Fig. 3). These results are consistent with the observation of increased low-N aa usage in oligotrophic oceans [39], but our analysis provides additional insights into the multitude and complexity of the mechanisms driving the evolution of aa pool and highlights the need for a model that accounts for all of these factors.

When both DNA and aa C and N requirements are considered, the growth rate increases under C- (s/m/p/DNA/aa-C) and N (s/m/p/DNA/aa-N)-limiting conditions (Fig. 2c). It suggests that both C and N limitations are strong selective forces that can overcome the intrinsic mutation biases and drive Darwinian evolution of marine bacteria by adjusting aa composition. The N limitation leads

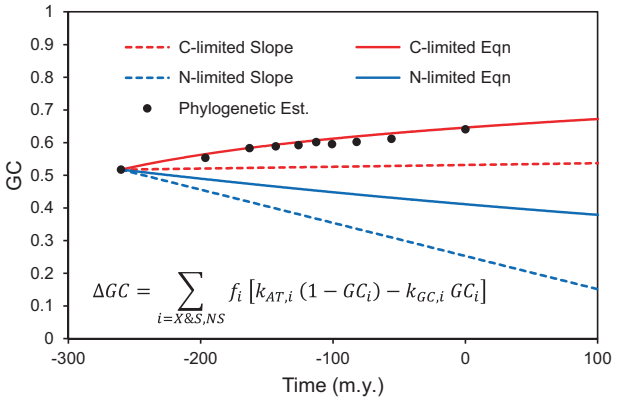

Fig. 4 Evolution of DSS-3 under C- and N-limiting conditions. Phylogenetic-predicted values (symbols) are based on a maximum-likelihood method for inferring ancestral states of a continuous character (see Text S6). Model estimates are based on slope (see Fig. 2b) or equation, which integrates the contribution of noncoding and synonymous (X&S) and nonsynonymous (NS) sites, with parameters estimated from the model using a perturbation analysis (see Text S6)

to a much stronger increase in growth rate than the C limitation (Fig. 2c). That is because proteome N represents a larger fraction of the total N requirement (45%) compared to C (24%). N limitation also leads to a decrease in aa C requirements (Fig. 2d), consistent with a positive correlation between aa C and N content (Fig. S4). These simulations were also performed with the GC-poor *Vibrio fischeri* and the resulting patterns, including a positive GC slope in C-limiting conditions and a stronger negative slope in N-limiting conditions, are consistent (Fig. S23).

Long-term evolutionary trajectory

An important question is whether these processes can explain the divergence of DSS-3 from the MRCA of the *Roseobacter* clade with a predicted GC content of 0.52 (Fig. 4). To make predictions over this time frame, the model could theoretically be run from this MRCA, but this would require specifying the genome sequence of the MRCA, including fast-evolving sites such as synonymous and noncoding sites that are difficult to reconstruct due to repeated and back mutations. Even if the MRCA sequence were known, it would be computationally prohibitive to run the model for such a long time. Alternatively, the GC slope from the model can be extrapolated (dashed lines in Fig. 4), but this ignores the effect of changing GC on the slope via site bias. The GC content of the MRCA is predicted to be close to 0.5, so site bias would have been low at that time. To account for the change of GC content during the evolution of DSS-3, a simple equation that considers the change of various fractions of the genome proportional to the GC content is developed and parameterized based on the results of the model (see Text S6). For C-limiting conditions, that equation predicts a present GC content of 0.645 (95% confidence interval: 0.627–0.664), which is the same as DSS-3 (GC = 0.641), a remarkable result for a straight prediction (no calibration). Moreover, the projected GC content at different evolutionary stages from the model matches well with those predicted using a phylogenetic approach (Fig. 4) [37, 38]. Therefore, the mechanisms included in our model are sufficient to explain the evolution of high GC content in DSS-3. The present slope is only slightly positive (see also Fig. 2a), which suggests the system is in approximate equilibrium. For the MRCA, there was a relatively strong net driving force to increase GC content (i.e., due to mutation bias and selection to reduce DNA and aa C requirements). As the GC content increased over time, the effect of site bias driving to decrease GC content became stronger until today, when the forces approximately cancel out.

In contrast to the C-limiting conditions, the model predicts a decreasing GC content for N-limiting conditions, which is inconsistent with the high GC content of DSS-3 (Fig. 4). This suggests that the evolution of DSS-3 occurred in a predominantly C-limiting (N-rich) environment. This is consistent with the working hypothesis that N became more available after the appearance of the major groups of N-fixing cyanobacteria during the Neoproterozoic or early Cambrian (542–485 Mya), including *Crocosphaera*, *Cyanothece*, *Trichodesmium*, and UCYN-A [40], though N may have become stressful occasionally during the Permian–Triassic transition [41] and during oceanic anoxia events (OAEs) in the past 200 million years (m.y.) [42] due to increased loss of fixed N through enhanced microbial activities of denitrification and anammox (anaerobic ammonium oxidation by nitrite) [41, 42]. If the C-limiting conditions and other aspects (e.g., mutation pattern) are maintained in the future, the model predicts a gradually decreasing slope as the GC content approaches the equilibrium level (Fig. 4). However, a more careful projection of DSS-3 GC content needs to consider the effect of ocean acidification in a future ocean. While ocean acidification may not change the concentration and composition of dissolved organic carbon [43] available to DSS-3, it may change the mutation pattern of DSS-3. Recent MA experiments on a coral pathogen *Vibrio shilonii* AK1 showed that reduced pH from 8.14 to 6.67 drives spontaneous mutation toward a decreased rate and an altered spectrum in the direction of generating more G/C nucleotides [44]. As *V. shilonii* AK1 has a GC-poor (0.44) genome, it remains unknown how ocean acidification changes the mutation rate and spectrum on the GC-rich (0.64) DSS-3.

Summary and outlook

In summary, our study provides evidence that evolution of genomic GC content of a representative GC-rich

heterotrophic marine bacterium is the result of complex evolutionary processes in the past ~260 m.y. [45]. A number of mechanisms, including mutation bias, selection against functionally deleterious nonsynonymous mutations, and selection to lower nutrient requirements in the DNA and protein pool, act and interact to drive the evolution of GC content. A surprising finding is that the low availability of C, rather than low N, has been driving the long-term evolution of genomic base composition in this bacterium. This does not conflict with our previous finding that N stress is a key factor in maintaining low genomic GC content in another abundant marine alphaproteobacterial lineage SAR11. In that study, we quantified a stronger bias of substitution toward AT in marine SAR11 populations compared to their freshwater counterparts, and provided evidence for stronger selection for reduced N requirement in marine populations [46] because N is more limiting in seawater [47, 48] compared to freshwater [49]. As all extant marine and freshwater SAR11 members have nearly identical GC content (29–30%), that study addressed the question what evolutionary mechanisms maintain the GC content in its current state, but it does not inform mechanisms driving the massive GC content reduction which presumably occurred in the distant past [46]. A relevant study showed that the *ung* gene responsible for repairing cytosine deamination that is considered as the strongest cause to AT-biased mutational spectrum is present in SAR11 but absent in >80% of phylogenetically diverse bacteria with GC-poor genomes, which argues against selection driving GC content reduction in the long-term evolution of SAR11 [50].

Our individual-based, genome-scale modeling approach is immune to the overly simplistic and limiting assumptions of existing population-level models, and is able to integrate various mechanisms underlying GC content evolution in a quantitative manner. It can be readily applied to explore genome evolution in other bacteria and environments, and it can be extended to include additional mechanisms, including insertion–deletion mutations, hyper-variable regions, and lysogenic viruses. This can be readily integrated into global ocean circulation and biogeochemistry models and used to explore questions in biogeography and microdiversity [20, 51, 52].

Acknowledgements We thank Sahar Shirani and Jijun Tang for help with the code development, Rong Zhao for reconstructing the ancestral states of the *Roseobacter* GC content, Way Sung for providing the *Vibrio* MA data, and Leong Keat Chan for explaining his microarray data. FLH is supported by the National Science Foundation (1240894 and 1404163); HL is supported by the Hong Kong RGC Early Career Scheme (24101015), the National Natural Science Foundation of China (41576141), the Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16), and the Direct Grant of the Chinese University of Hong Kong (4930062).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Libes S. Introduction to marine biogeochemistry, 2nd ed. Academic Press: Burlington, MA; 2009.
2. Mizuno CM, Ghai R, Saghaï A, López-García P, Rodriguez-Valera F. Genomes of abundant and widespread viruses from the deep ocean. mBio. 2016;7:e00805–16.
3. Romero H, Pereira E, Naya H, Musto H. Oxygen and guanine–cytosine profiles in marine environments. J Mol Evol. 2009;69:203–6.
4. Giovannoni SJ, Thrash JC, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8:1553–65.
5. Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. Proc Natl Acad Sci USA. 2013;110:11463–68.
6. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348:1261359.
7. Luo H, Huang Y, Stepanauskas R, Tang J. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. Nat Microbiol. 2017;2:17091.
8. Luo H, Moran MA. Evolutionary ecology of the marine *Roseobacter* clade. Microbiol Mol Biol Rev. 2014;78:573–87.
9. Coutinho F, Tschoeke DA, Thompson F, Thompson C. Comparative genomics of *Synechococcus* and proposal of the new genus *Parasynechococcus*. PeerJ. 2016;4:e1522.
10. Sun Y, Powell KE, Sung W, Lynch M, Moran MA, Luo H. Spontaneous mutations of a model heterotrophic marine bacterium. ISME J. 2017;11:1713–18.
11. Partensky F, Garczarek L. *Prochlorococcus*: advantages and limits of minimalism. Ann Rev Mar Sci. 2010;2:305–31.
12. Long H, Sung W, Kucukyildirim S, Williams E, Miller S, Guo W et al. Evolutionary determinants of genome-wide nucleotide composition. Nature Ecol Evol. 2017;in press.
13. Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010;6:e1001107.
14. Kelkar YD, Phillips DS, Ochman H. Effects of genic base composition on growth rate in G+C-rich genomes. G3: Genes|Genomes|Genet. 2015;5:1247–52.
15. Raghavan R, Kelkar YD, Ochman H. A selective force favoring increased G + C content in bacterial genes. Proc Natl Acad Sci USA. 2012;109:14504–507.
16. 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 Genet. 2015;11:e1004941.
17. Bobay L-M, Ochman H. Impact of recombination on the base composition of bacteria and archaea. Mol Biol Evol. 2017;34:2627–36.
18. Loewe L, Hill WG. The population genetics of mutations: good, bad and indifferent. Philos Trans R Soc Lond B Biol Sci. 2010;365:1153–67.
19. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129:897–907.
20. Hellweger FL, van Sebille E, Fredrick ND. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. Science. 2014;345:1346–49.

21. Kaiser VB, Charlesworth B. The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet. 2009;25:9–12.
22. Shirani S, Hellweger FL. Neutral evolution and dispersal limitation produce biogeographic patterns in *Microcystis aeruginosa* populations of lake systems. Microb Ecol. 2017;74:416–26.
23. Hellweger FL, Clegg RJ, Clark JR, Plugge CM, Kreft J-U. Advancing microbial sciences by individual-based modelling. Nat Rev Micro. 2016;14:461–71.
24. Dixit PD, Pang TY, Studier FW, Maslov S. Recombinant transfer in the basic genome of Escherichia coli. Proc Natl Acad Sci USA. 2015;112:9070–75.
25. Luria SE, Delbrück M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics. 1943;28:491–11.
26. Scheffer M, Baveco JM, DeAngelis DL, Rose KA, van Nes EH. Super-individuals a simple solution for modelling large populations on an individual basis. Ecol Model. 1995;80:161–70.
27. Moran MA, Buchan A, Gonzalez JM, Heidelberg JF, Whitman WB, Kiene RP, et al. Genome sequence of *Silicibacter pomeroyi* reveals adaptations to the marine environment. Nature. 2004;432:910–13.
28. Sonnenschein EC, Nielsen KF, D'Alvise P, Porsby CH, Melchiorsen J, Heilmann J, et al. Global occurrence and heterogeneity of the *Roseobacter*-clade species *Ruegeria mobilis*. ISME J. 2017;11:569–83.
29. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015;11:e1004041.
30. Chan L-K, Newton RJ, Sharma S, Smith CB, Rayapati P, Limardo AJ, et al. Transcriptional changes underlying elemental stoichiometry shifts in a marine heterotrophic bacterium. Front Microbiol. 2012;3:159.
31. Zimmerman AE, Allison SD, Martiny AC. Phylogenetic constraints on elemental stoichiometry and resource allocation in heterotrophic marine bacteria. Environ Microbiol. 2014;16:1398–10.
32. Nei M, Suzuki Y, Nozawa M. The neutral theory of molecular evolution in the genomic era. Annu Rev Genom Hum Genet. 2010;11:265–89.
33. Xia X, Xie Z. Protein structure, neighbor effect, and a new index of amino acid dissimilarities. Mol Biol Evol. 2002;19:58–67.
34. Luo H, Swan BK, Stepanauskas R, Hughes AL, Moran MA. Comparing effective population sizes of dominant marine alphaproteobacteria lineages. Environ Microbiol Rep. 2014;6:167–72.
35. Dillon MM, Sung W, Sebra R, Lynch M, Cooper VS. Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in *Vibrio cholerae* and *Vibrio fischeri*. Mol Biol Evol. 2017;34:93–109.
36. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
37. Schluter D, Price T, Mooers AØ, Ludwig D. Likelihood of ancestor states in adaptive radiation. Evolution. 1997;51:1699–11.
38. Felsenstein J. Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet. 1973;25:471–92.
39. Grzymski JJ, Dussaq AM. The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. 2012;6:71–80.
40. Sánchez-Baracaldo P, Ridgwell A, Raven John A. A neoproterozoic transition in the marine nitrogen cycle. Curr Biol. 2014;24:652–57.
41. Luo G, Wang Y, Algeo TJ, Kump LR, Bai X, Yang H, *et al* Enhanced nitrogen fixation in the immediate aftermath of the latest Permian marine mass extinction. Geology. 2011;39:647–50.
42. Jenkyns HC. Geochemistry of oceanic anoxic events. Geochem Geophys Geosystems. 2010;11:Q03004.
43. Zark M, Riebesell U, Dittmar T. Effects of ocean acidification on marine dissolved organic matter are not detectable over the succession of phytoplankton blooms. Sci Adv. 2015;1:e1500531.
44. Strauss C, Long H, Patterson CE, Te R, Lynch M. Genome-wide mutation rate response to pH change in the coral reef pathogen *Vibrio shilonii* AK1. mBio. 2017;8:e01021–01017.
45. Luo H, Csűros M, Hughes AL, Moran MA. Evolution of divergent life history strategies in marine Alphaproteobacteria. mBio. 2013;4:e00373–00313.
46. Luo H, Thompson LR, Stingl U, Hughes AL. Selection maintains low genomic GC content in marine SAR11 lineages. Mol Biol Evol. 2015;32:2738–48.
47. Moore CM, Mills MM, Arrigo KR, Berman-Frank I, Bopp L, Boyd PW, et al. Processes and patterns of oceanic nutrient limitation. Nat Geosci. 2013;6:701–10.
48. Sterner RW. On the phosphorus limitation paradigm for lakes. Int Rev Hydrobiol. 2008;93:433–45.
49. McMahon KD, Read EK. Microbial contributions to phosphorus cycling in eutrophic lakes and wastewater. Annu Rev Microbiol. 2013;67:199–19.
50. Viklund J, Ettema TJG, Andersson SGE. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. Mol Biol Evol. 2012;29:599–15.
51. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science. 2014;344:416–20.
52. Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. Emergent biogeography of microbial communities in a model ocean. Science. 2007;315:1843–46.