nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-021-01591-0

Check for updates

Prochlorococcus have low global mutation rate and small effective population size

Zhuoyu Chen[1,6], Xiaojun Wang[2,3,6], Yu Song[1], Qinglu Zeng[4,5], Yao Zhang[1]✉ and Haiwei Luo[2,3,5]✉

Prochlorococcus **are the most abundant free-living photosynthetic carbon-fixing organisms in the ocean.** ***Prochlorococcus*** **show small genome sizes, low genomic G+C content, reduced DNA repair gene pool and fast evolutionary rates, which are typical features of endosymbiotic bacteria. Nevertheless, their evolutionary mechanisms are believed to be different. Evolution of endosymbiotic bacteria is dominated by genetic drift owing to repeated population bottlenecks, whereas** ***Prochlorococcus*** **are postulated to have extremely large effective population sizes (**N_e**) and thus drift has rarely been considered. However, accurately extrapolating** N_e **requires measuring an unbiased global mutation rate through mutation accumulation, which is challenging for** ***Prochlorococcus*****. Here, we managed this experiment over 1,065 days using** ***Prochlorococcus marinus*** **AS9601, sequenced genomes of 141 mutant lines and determined its mutation rate to be** 3.50×10^{-10} **per site per generation. Extrapolating** N_e **additionally requires identifying population boundaries, which we defined using PopCOGenT and over 400 genomes related to AS9601. Accordingly, we calculated its** N_e **to be** 1.68×10^7**, which is only reasonably greater than that of endosymbiotic bacteria but surprisingly smaller than that of many free-living bacteria extrapolated using the same approach. Our results therefore suggest that genetic drift is a key driver of** ***Prochlorococcus*** **evolution.**

Having an annual mean global abundance of 3×10^{27} cells and performing global net primary production of 4 gigatons carbon per year[1], *Prochlorococcus* are the most abundant photosynthetic organisms on Earth and a key driver of global biogeochemical cycles. *Prochlorococcus* have diversified into two major phylogenetic groups with distinct ecology and physiology (for example, pigment ratio, light intensity and water depth), namely the high-light (HL) and low-light (LL) adapted lineages, with the former phylogenetically embedded within the latter[2]. Both lineages have further diversified into multiple clades differentiated by their nutrient and temperature preferences[2]. The HL lineage, for example, is composed of six clades (HLI to HLVI)[2], amongst which the well-characterized HLI and HLII differ in their temperature optima[3,4]. Another major feature is that all genome-sequenced HL and LL clades except LLIV possess highly reduced genomes (1.6–1.8 Mbp). This is due to a major genome reduction event occurring early in the evolutionary history of *Prochlorococcus*, coinciding with the split between LLIV and the remaining clades[5–7].

Given their huge census population sizes (N_c), highly reduced genome sizes, base compositions biased towards A and T and fast sequence evolutionary rates, the population genetic forces driving genome evolution of *Prochlorococcus* have attracted much attention. A key parameter in understanding these ecological, molecular and population genetic features is the effective population size (N_e), defined as the size of an ideal population which harbours the same amount of neutral genetic diversity as is actually observed in the real population[8,9]. For natural bacterial populations, N_e is often substantially lower than N_c by orders of magnitude owing to demographic fluctuations and genomic linkage, amongst other factors[8,9].

The long-term N_e is often approximated by the d_N/d_S ratio[10,11], which is the rate of fixed non-synonymous (amino acid altering) substitutions over the rate of fixed synonymous (silent) substitutions and is calculated by comparing multiple single-copy orthologous genes from different species. Unfortunately, d_N/d_S gives a composite parameter, $N_e s$ (ref. [12]), where s is the selection coefficient of the genes involved in the analysis. There are two main problems with this proxy[7]: (i) the gene function and thus s changes over time and between lineages, and (ii) synonymous changes are assumed to be neutral. For the latter, it has been argued that, in bacteria, truly neutral sites may not exist since weak selection imposed by codon usage and nucleotide composition acts on synonymous sites[13], and changes in the selective pressure on optimizing the codon usages amongst lineages may override the changes of N_e (ref. [7]). In addition, different selective pressures may drive the evolution of nucleotide composition towards opposite directions amongst lineages. For example, nitrogen limitation and carbon limitation act as contrasting selective forces that drive the decrease and increase of genomic G+C content of two co-occurring lineages that dominate heterotrophic bacterial communities in the ocean, the marine SAR11 bacteria[14] and the Roseobacter group members[15], respectively. Together, these between-lineage variations often make d_N/d_S an unreliable proxy. Alternatively, the long-term N_e is often approximated by the neutral intraspecific genetic diversity (π_S)[16,17], which also gives a composite parameter, $N_e\mu$ (ref. [18]), where μ denotes the unbiased global base-substitution mutation rate and needs to be assessed using a mutation accumulation (MA) experiment followed by whole-genome sequencing (WGS) of the mutant lines[19]. In the MA part, a single clonal ancestor is used to initiate multiple replicate lines that each pass through repeated single-cell bottlenecks, usually by repeatedly transferring on solid media. This procedure minimizes the efficiency of natural selection and thus prevents selection from either promoting or eliminating nearly all mutations except

[1]State Key Laboratory of Marine Environmental Science and College of Ocean and Earth Sciences, Xiamen University, Xiamen, China. [2]Simon F. S. Li Marine Science Laboratory, School of Life Sciences and State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong SAR. [3]Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China. [4]Department of Ocean Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong SAR. [5]Hong Kong Branch of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou), Clear Water Bay, Hong Kong SAR. [6]These authors contributed equally: Zhuoyu Chen, Xiaojun Wang. ✉e-mail: yaozhang@xmu.edu.cn; hluo2006@gmail.com

the small subset with unusually large fitness effects. By comparing the genomes of the mutant lines with the genome of the ancestor, a nearly unbiased picture of spontaneous mutations is unveiled. An apparent disadvantage of using π_S as a proxy for N_e is that μ varies over 100-fold amongst prokaryotic species[20], which magnifies the uncertainty of this proxy. Another inevitable problem is that selection on synonymous changes, as outlined above, similarly gives uncertainties on N_e estimates. Since N_e approximated by π_S is based on polymorphism data, it encompasses a shorter timescale covering the past $2N_e$ generations on average[21].

To date, unbiased global mutation rates of 31 prokaryotic species determined by the MA/WGS strategy have become available. This allows disentangling the composite parameter $N_e\mu$ and calculating N_e with the neutral polymorphism according to the equation $\pi_S = 2N_e\mu$ (ref. [18]). On the other hand, the MA procedure is notoriously difficult for many natural bacteria including *Prochlorococcus* that do not readily propagate on solid media, thus the global mutation rate has not been accessible for most natural bacterial species. In a previous calculation of N_e for the *Prochlorococcus* clade HLII[22], for example, the mutation rate (10^{-7}–10^{-8} per gene per generation) was instead determined through fluctuation experiments[23] which use non-synonymous mutations enabling bacterial survival in a selective medium to extrapolate the genome-wide mutations but are known to be biased[7,24,25]. Another challenge to inferring N_e with this approach is that it requires delineation of a genetically isolated and panmictic population (that is, 'species'). This is because π_S needs to be calculated from such a population and can be overestimated if genomes spanning multiple populations are compared[26]. There have been multiple versions of how *Prochlorococcus* lineages are recognized as 'species' based on which N_e was calculated, varying from all *Prochlorococcus* members[16] to the well-characterized ecotype such as eMIT9312 roughly corresponding to the entire HLII clade[22]. A common problem with their species recognition is that the internal population structure was overlooked when genomes were chosen for N_e extrapolations. Specifically, numerous 'backbone subpopulations' were characterized for HLII, each shown to carry conserved core alleles attached with a distinct set of flexible genes[22]. Whereas this genetic feature is consistent with the hypothesis that backbone subpopulations may represent genetically discrete populations, π_S was still estimated from genomes spanning the entire clade HLII instead of a backbone subpopulation, and accordingly the N_e of HLII was derived to be on the order of 10^9 (ref. [22]). The authors further argued that this calculation probably underestimated N_e and that the 'real' N_e should be much closer to the abundance (N_c) of a backbone subpopulation (10^{13}) because they believed that factors such as demography responsible for the mismatch between N_e and N_c are negligible for *Prochlorococcus*[22].

Despite the lack of an unbiased measurement of the global mutation rate and an appropriate delineation of population boundaries in previous estimations of N_e for *Prochlorococcus*, an unusually large N_e has been widely accepted by microbial ecologists and evolutionary biologists. As a consequence, prior discussions on *Prochlorococcus* evolutionary mechanisms were largely built on this untested assumption[7,27–31], whilst only a few studies acknowledged the unresolved nature of this key parameter[6,20,32]. A major goal of the present study is to provide an accurate measure of μ and N_e for a genome-reduced strain *Prochlorococcus marinus* AS9601 (1.67 Mbp) affiliated with the HLII clade, the most abundant lineage in *Prochlorococcus*[33].

Results and discussion

A genome-reduced *Prochlorococcus* member has a very low unbiased global mutation rate. We managed a 1,065-day propagation of this strain on solid media cocultured with a heterotrophic helper bacterium (*Alteromonas* sp. EZ55) that scavenges reactive oxygen species harmful to *Prochlorococcus*[34], and assessed its unbiased global mutation rate using the MA/WGS strategy (Fig. 1a). A total of 150 MA lines were initiated from a single progenitor cell, 141 of which survived after 39 transfers with each line undergoing 1,258 cell divisions (corrected with death rate). Of the surviving lines, 116 accumulated mutations, yielding a total of 170 base-pair substitution mutations (BPSs), 14 deletions and 21 insertions (Supplementary Table 1). Amongst these, 79 BPSs, 9 deletions and 16 insertions fell into a single gene (*RS15215*; Fig. 1b; Supplementary Table 2) which encodes a putative sodium-dependent transporter. All BPSs in this gene were non-synonymous and conferred improved viability, as shown by the survival rate experiment (Fig. 1c), which is evidence that mutations occurring in this gene are under positive selection.

Whilst the hypothesis of neutral accumulation of the remaining 91 BPSs was not rejected (Supplementary Information), there is a possibility that hitchhiking in the background of the beneficial mutations occurring in *RS15215* might have increased the number of visible mutations in this experiment by speeding up their fixation relative to the remaining MA lines without mutations in *RS15215*. It is also possible that the *RS15215* mutants might have epistatic interactions with certain mutations in other genomic regions, which may delay the fixation of the latter and lower the overall mutation rate in the genomic regions other than *RS15215*. To rule out any such potential effects, we focused on the 41 surviving MA lines in which no mutations occurred in *RS15215*. The 30 BPSs accumulated in these 41 lines translate to a global base-substitution mutation rate of $(3.50 \pm 0.76) \times 10^{-10}$ (95% CI 2.36×10^{-10}–4.99×10^{-10}), which was not significantly different from the mutation rate ($(2.91 \pm 0.41) \times 10^{-10}$, 95% CI 2.23×10^{-10}–3.74×10^{-10}) derived from the remaining 100 surviving MA lines mutated at *RS15215* but with these mutations excluded from the calculation. Further, there is evidence that the 30 BPSs accumulated in those 41 lines are truly neutral. First, there was no difference between the ratio of accumulated mutations at protein-coding sites to those at intergenic sites (28 versus 2) and the ratio of the number of protein-coding sites to the number of intergenic sites (1,494,195 versus 161,973) in the AS9601 genome (Fisher's exact test, $P = 0.76$). There was also no difference between the ratio of accumulated non-synonymous to synonymous mutations (23 versus 5) and the risk ratio of the number of all possible non-synonymous changes to the number of all possible synonymous changes (3,297,472 versus 946,657) in the coding part of the AS9601 genome (exact binomial test, $P = 0.66$). We therefore use the truly unbiased global mutation rate ($(3.50 \pm 0.76) \times 10^{-10}$) calculated from the 41 lines in the following analyses.

These careful analyses indicate that the genome-reduced *Prochlorococcus* (1.67 Mbp) has a mutation rate comparable to those of heterotrophic marine bacteria with larger genomes including *Ruegeria* (4.60 Mbp; Alphaproteobacteria)[35], *Vibrio* (4.06, 4.27 and 5.70 Mbp for the three species, respectively; Gammaproteobacteria)[36,37], *Leeuwenhoekiella* (4.05 Mbp; Flavobacteriia)[38] and *Nonlabens* (2.85 Mbp; Flavobacteriia)[38]. The low global mutation rate of *Prochlorococcus* measured here does not support prior studies which postulated that genome-reduced *Prochlorococcus* lineages have high mutation rates owing to the evolutionary losses of multiple DNA repair genes[27,28,39]. To this end, it is useful to mention that *Deinococcus radiodurans*, a deeply branching species in Bacteria in which the mismatch repair system was naturally lost, has a global mutation rate only marginally greater than that of *Escherichia coli*, where most DNA repair genes including mismatch repair are present[40]. These lines of evidence suggest that caution must be taken when using losses of important DNA repair genes as a proxy for mutation rate increase in natural isolates.

Genome-reduced *Prochlorococcus* lineages have unexpectedly small effective population sizes. As mentioned above, calculating N_e requires characterizing the gene flow discontinuities and delineating genetically isolated populations, in addition to determin-

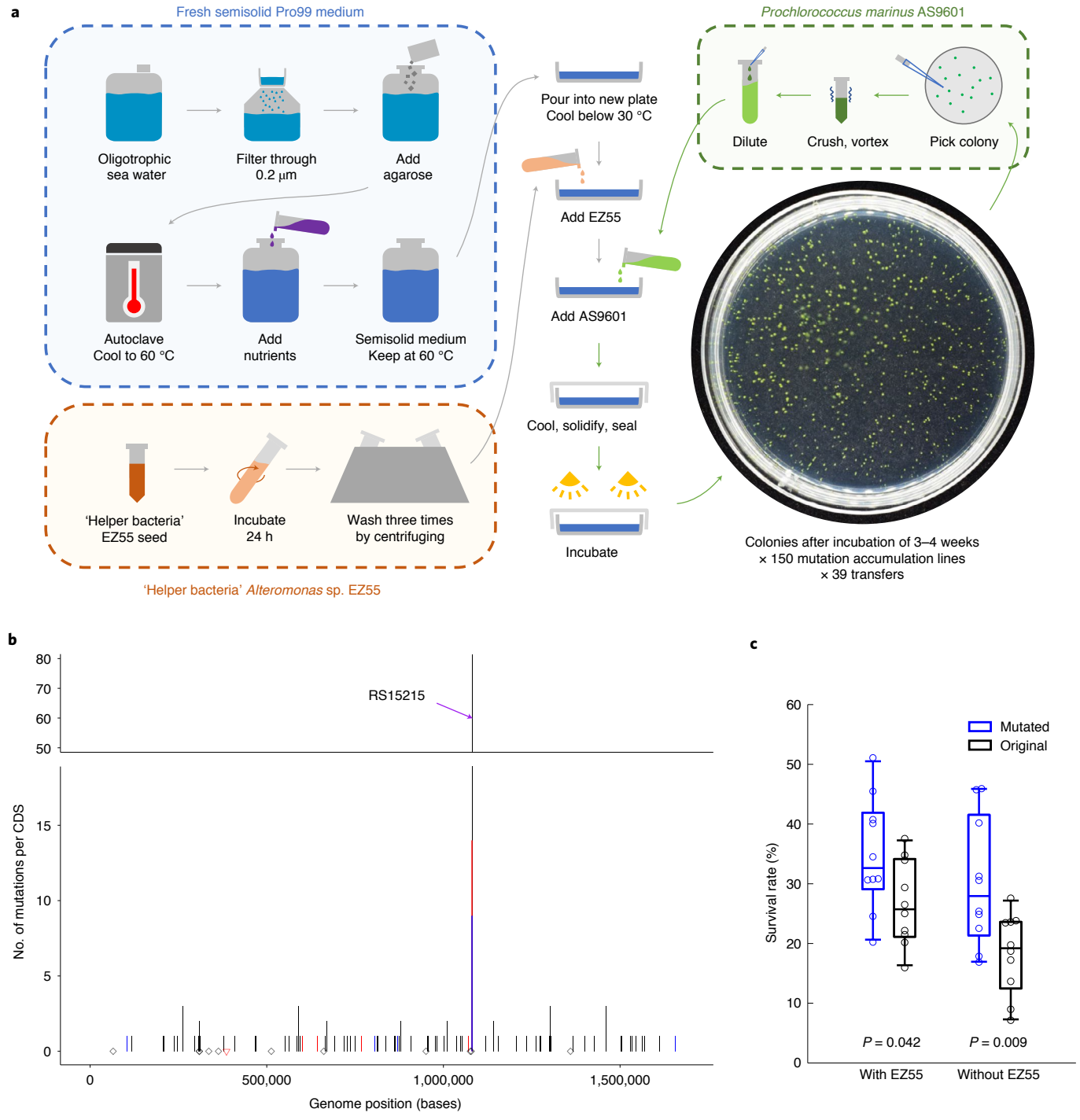


Fig. 1 | Determining the unbiased global mutation rate of *Prochlorococcus marinus* AS9601. a, Experimental procedure of the MA experiment for determining the unbiased global mutation rate of *P. marinus* AS9601. Blue, orange and green dashed boxes illustrate the preparation of fresh semisolid Pro99 medium, helper bacterium *Alteromonas* sp. EZ55 and diluted *P. marinus* AS9601, respectively. EZ55 and AS9601 must be added into the plate immediately and shaken gently to mix after the medium cools down below 30 °C but before it starts to solidify. The MA experiment was initiated by spreading cells from a single founding colony to 150 lines and ended after 39 transfers over 1,065 days. The 141 surviving lines were subjected to WGS. **b**, Base-substitution mutations and insertion/deletion mutations across the whole genome of *P. marinus* AS9601. The height of each bar represents the number of base substitutions (black), insertions (red) and deletions (blue) across all MA lines within each protein-coding gene. Black diamonds and red triangles denote base substitutions and insertions that occurred on the remaining genomic regions (intergenic regions and non-protein-coding genes), respectively; both diamonds and triangles are shown with transparence, thus genomic regions with more mutations show deeper colour than those with less mutations. The genomic position of insertion/deletion mutation refers to the position of the first mutated site. The locus tag of the gene (*RS15215*) with statistical enrichment of mutations is shown. CDS, coding sequence. **c**, The survival rates of 10 lines randomly chosen from the 55 lines, each with their genomic mutations restricted at *RS15215* (blue) and those of 10 lines randomly chosen from the 25 lines each showing no genomic mutations (black). Within each box, the horizontal line marks the median; boxes extend from the 25th to 75th percentile of each group's growth rate; whiskers above and below the box indicate the 10th and 90th percentiles. Lines with mutated *RS15215* have significantly greater survival rates than those without any mutations (two-tailed *t* test), regardless of whether EZ55 was included as a helper.

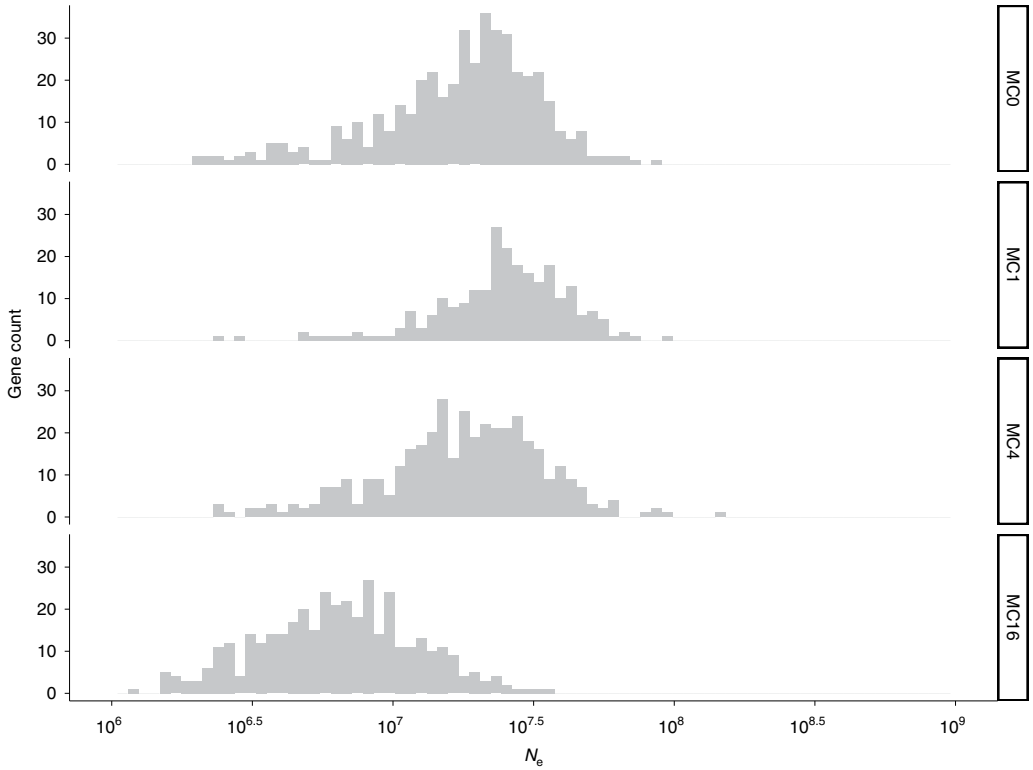


Fig. 2 | Histogram of effective population size (N_e) estimates. Estimates on a gene-by-gene basis for the four *Prochlorococcus* HLII populations (MC0, MC1, MC4 and MC16) related to *Prochlorococcus marinus* AS9601. The population membership for a given gene family depends on whether the population members form a monophyletic group, thus the population membership can vary from gene to gene.

ing μ. Of the 23 isolates' genomes and 395 high-quality single-cell amplified genomes (SAGs; chosen from 557 SAGs) available to HLII, PopCOGenT identified 255 genetically isolated populations, of which 251 each contained few (less than six) non-redundant members (where redundant members are from the same clonal complex and do not contribute to the gene pool of a population). The remaining four populations (MC0, MC1, MC4 and MC16) were composed of 64, 11, 18 and 21 non-redundant members, respectively, and generally matched the *Prochlorococcus* backbone subpopulations (Extended Data Fig. 1). We also used ConSpeciFix to delineate populations (Extended Data Fig. 1), but the results were not supported (Methods and Supplementary Information).

Next, we systematically evaluated the factors that may change the population membership defined by PopCOGenT and thus N_e. Although PopCOGenT can easily accommodate partial genomes such as SAGs[41] which dominate the dataset, this tool is likely sensitive to increased error rates associated with SAG data because it uses enrichment of identical genomic regions as a measure of recent gene transfer[41]. The rationale is that sequence errors may either decrease the identical DNA segments if the errors make the aligned sequences more different and thus may split the original population or increase the identical DNA segments if the errors erase the true differences and thus may merge the originally different populations. To illustrate the effect of using SAG data on PopCOGenT analysis and evaluate the extent to which it may impact the N_e estimates, we simulated SAG assemblies from isolates' genomes from 19 bacterial species (Methods). The simulation results (Extended Data Fig. 2) showed that there were no or negligible changes in ten of these species. For the remaining nine species in which population membership was altered, population splitting occurred much more frequently than population amalgamation (44 events in seven species versus eight events in four species), and the N_e estimates were accordingly changed by 0.60–1.66 times with respect to those based on isolates' genomes (Supplementary Information). These corrections are therefore not negligible but have limited impact on N_e estimates, and our conclusion remains that HLII *Prochlorococcus* have unexpectedly small N_e on the order of 10^7. This simulation analysis also implies that some of the populations defined by PopCOGenT with few members are potential artefacts owing to the extensive use of the SAG data. Given that most of the HLII populations (251 out of 255) each had only fewer than six non-redundant members, there are likely fewer genetically isolated populations than estimated by PopCOGenT.

The impact of population definition on N_e estimates was further tested by progressively adding sister lineages to each of the four main populations (MC0, MC1, MC4 and MC16) defined by PopCOGenT. As expected, N_e increased when including more phylogenetically deeper sister lineages, and this trend was observed for all four populations. However, the N_e estimates increased rapidly for MC4 and MC16, and the changes were rather limited for MC0 and MC1 (Extended Data Fig. 1). This result suggests that population boundary was less clearly defined for MC0 and MC1 compared with MC4 and MC16.

We therefore leveraged the topology of gene trees to test the reliability of the populations defined by PopCOGenT. The rationale is that, since such populations are presumably genetically isolated from each other, orthologous genes in most gene families from a delineated population are expected to form monophyletic groups. On the other hand, if members of a delineated population are engaging in substantial gene flow with other lineages, only a limited number of gene families would support the population defined by PopCOGenT. These arguments assume that recombination is an important driver of the population structure of HLII, which is verified by our analysis showing that the number of single-nucleotide polymorphisms introduced by recombination is approximately 1.72 times that of single-nucleotide polymorphisms caused by mutation

Table 1 | Estimates of intraspecific genetic diversity at fourfold degenerate sites (π_S) and N_e based on populations delineated by two approaches: defined on a cell-by-cell basis (left) and delineated on a gene-by-gene basis (right)

	Population delineated cell by cell		Population delineated gene by gene	
Population	Median π_S	Median N_e	Median π_S	Median N_e
MC0	0.016	2.34×10^7	0.013	1.89×10^7
MC1	0.041	5.87×10^7	0.017	2.47×10^7
MC4	0.013	1.83×10^7	0.012	1.77×10^7
MC16	0.005	7.43×10^6	0.004	5.77×10^6

The key difference is that, in the former, the population membership is directly defined by PopCOGenT and remains constant across gene families, whereas in the latter, the population membership was initially defined by PopCOGenT but may be adjusted for some gene families depending on whether the population members form a monophyletic group, thus the membership can vary from gene to gene. The two approaches yield highly consistent results for all the populations except MC1, which is likely subjected to more frequent recombination with other populations.

(that is, the r/m ratio determined by ClonalFrameML). In general, phylogenetic analysis for 589 single-copy orthologous gene families we analysed (Methods) provided strong support for MC4 and MC16, moderate support for MC0 potentially due to ongoing population subdivision and limited support for MC1, potentially owing to gene flow between MC1 and other populations (Supplementary Information).

If gene flow blurs the population boundary of MC1, population membership may differ from gene to gene. It is therefore useful to estimate N_e at the individual gene level (Fig. 2) by characterizing the population borders for individual gene families based on the topology of gene trees. By redefining the population boundary using this approach (Methods), the N_e estimate of MC1 was reduced by 57% (Table 1). In contrast, the N_e estimates of MC0, MC4 and MC16 remained highly consistent (Table 1) based on the populations defined on a cell-by-cell basis (that is, population membership defined solely by PopCOGenT) and on a gene-by-gene basis (that is, population membership adjusted from gene to gene). Based on the new population boundaries defined on a gene-by-gene basis, the average N_e of the four populations became 1.68×10^7. It is noteworthy that the N_e estimates of these four populations are consistent with the estimates of ten other phylogenetically diverse HLII populations (Supplementary Table 3 and Extended Data Fig. 1) in which only few non-redundant members (varying from three to five) were sampled, suggesting that the average of N_e estimates of the four populations approximates a typical HLII population's N_e.

Our analysis therefore places HLII *Prochlorococcus* at the middle of the wide range of N_e (7.88×10^5–4.18×10^8) assessed for the prokaryotic species, each based on populations defined by PopCOGenT (on a cell-by-cell basis; Methods). The N_e values of the HLII populations estimated here are substantially smaller than the previously reported value (on the order of 10^9 but argued to be closer to 10^{13})[22]. Whilst their N_e estimate was similarly extrapolated from genetic diversity at neutral sites and a mutation rate not dramatically different from the global mutation rate determined here, the genomes used for their calculation of N_e traversed the entire phylogeny of the HLII clade instead of individual backbone subpopulations, some of which corresponded to our delineated populations (Extended Data Fig. 1).

***Prochlorococcus* data help test universal mechanisms of genome evolution across prokaryotes.** The accessibility of two fundamental population genetic parameters, μ and N_e, associated with a representative genome-reduced *Prochlorococcus* member (AS9601) and many other prokaryotic species (Supplementary Table 4) provides an opportunity to test several hypotheses regarding the universal rules governing genome evolution across prokaryotic lineages. A decade ago, Lynch and colleagues proposed the 'drift-barrier' model to explain mutation rate variation amongst species[42]. It posits that selection promotes replication fidelity to a limit set by the power of genetic drift (that is, the inverse of N_e), and further refinements are expected to reduce fitness advantages[19,42]. This model achieves great success in explaining the increases of mutation rate from prokaryotes to unicellular eukaryotes to multicellular eukaryotes[19,42]. In the present study, both the nucleotide-substitution mutation rate (μ) and genome-wide mutation rate (U_P, a proxy for the deleterious mutation load of a genome[19]; Methods) scaled negatively with N_e across phylogenetically and physiologically diverse prokaryotes (Methods) according to both generalized linear model (GLM) and phylogenetic generalized least-squares (PGLS) regression analyses, the latter controlling for the phylogenetic effect on trait evolution (Fig. 3a and b, respectively). Our results therefore provide new evidence for the drift-barrier model, which explains mutation rate variations across prokaryotic lineages.

Another interesting but debated topic is the evolution of prokaryotic genome sizes, which vary over two orders of magnitude across prokaryotic lineages (from 0.1 to more than 10 Mbp). A prevailing hypothesis is that the power of genetic drift scales negatively with genome size across prokaryotes[10,43,44]. In those studies, N_e was often approximated by d_N/d_S, which is problematic as discussed above. It is therefore interesting to re-visit this hypothesis with the absolute N_e values available here. Intriguingly, whilst GLM supported a significant positive scaling between prokaryotic genome sizes and N_e, PGLS did not (Fig. 3c). This difference may arise from the opposing signals between Terrabacteria and Gracilicutes (Extended Data Fig. 3 and Supplementary Information), though factors driving these contrasting patterns remain enigmatic. Importantly, whilst the available data tentatively reject the drift hypothesis of genome size evolution across prokaryotes, a solid conclusion in this regard requires more taxonomically replicated samples owing to the rapid change of genome size and N_e. For example, the three *Vibrio* species, which was the only replication at the genus level, did not form a tight cluster (Fig. 3c).

Compared with genetic drift, mutation rate is an understudied driver of genome reduction. Whilst a correlation between increased mutation rate and increased gene loss rate was recently established in a few free-living and endosymbiotic bacteria[39], whether this is a universal mechanism across prokaryotes remains unknown. It is therefore interesting to find a negative scaling between μ and genome sizes here, regardless of whether phylogeny was controlled or not (Fig. 3d). However, note that, in our dataset, high mutation rates are almost always found in species with small N_e. For example, the highest mutation rates are found in *Mesoplasma florum* and *Thermococcus eurythermalis*, which also have the smallest N_e across the species sampled here. Since high mutation rates are largely a result of low N_e, as discussed above, the possibility that genetic drift is a 'hidden' driver of the negative scaling between μ and genome size observed here cannot be ruled out. More data are needed to test these competing hypotheses.

It is noteworthy that both μ and U_P remained significantly correlated with N_e even without AS9601. Likewise, genome sizes remained significantly correlated with μ but not with N_e, regardless of whether AS9601 was used or not. Although the *Prochlorococcus* data did not drive any of these patterns, it is nevertheless valuable to include these data in the analysis. For example, a prior study considered the drift-barrier model to be inconclusive in explaining mutation rate variations across prokaryotic lineages[20]. This is because *Prochlorococcus* were believed to have very high mutation rates and extremely large N_e, which may violate the negative scaling between μ and N_e (ref. [20]). Hence, the successful fitting of AS9601 to the existing negative scaling strengthens the drift-barrier model.

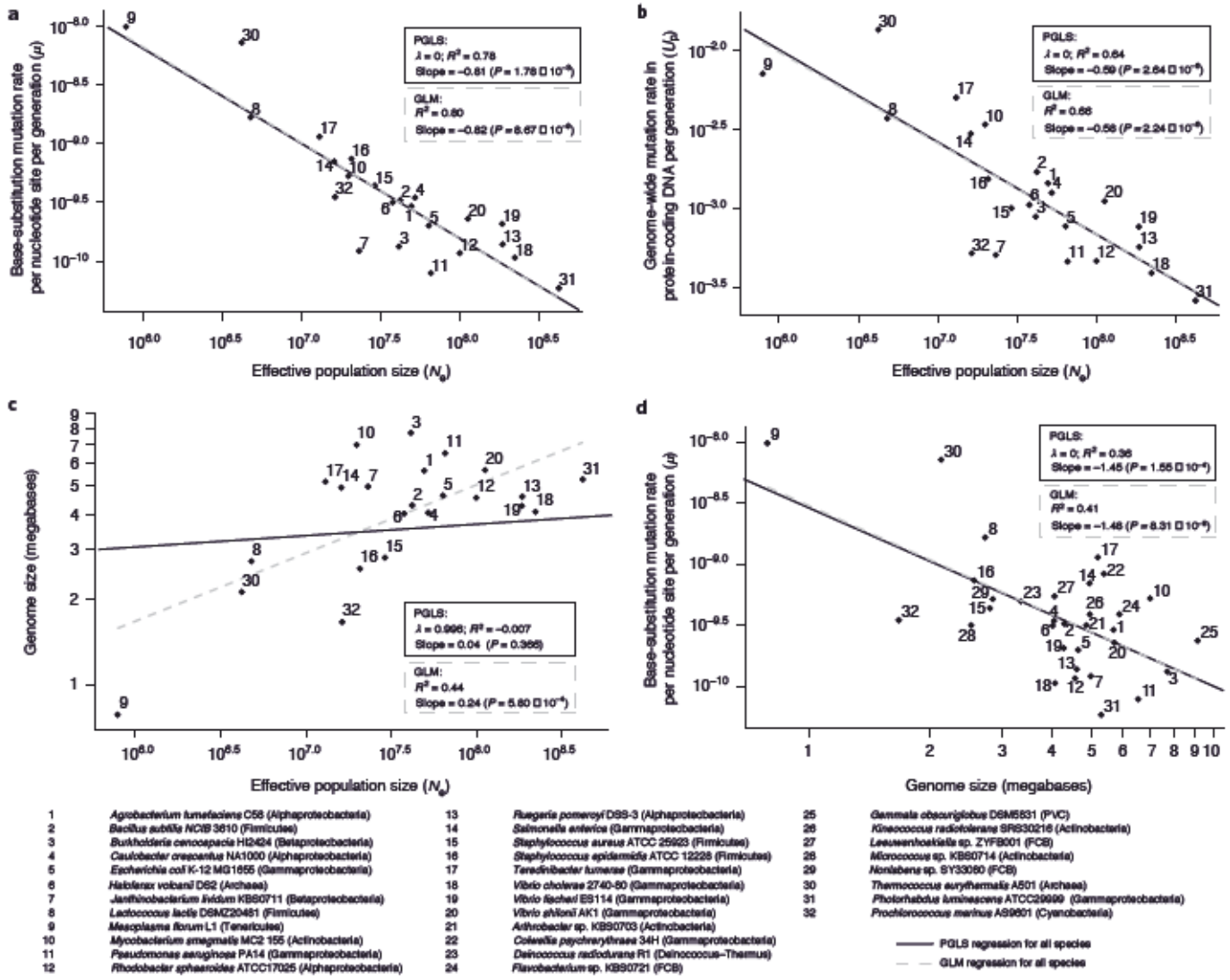


Fig. 3 | Scaling relationships. Scaling relationships involving the base-substitution mutation rate per cell division per nucleotide site (μ), genome-wide mutation rate per cell division per genome (U_p), estimated effective population size (N_e) and genome size across 30 bacterial and two archaeal species, with all traits logarithmically transformed. The mutation rates of species numbered 1–31 are collected from literature, whilst that of species 32 (*Prochlorococcus marinus* AS9601) is from the present study. Mutation rates of all 32 species were determined using the MA/WGS strategy. For the extrapolation of N_e, population boundaries were defined by PopCOGenT. The N_e for *Prochlorococcus marinus* AS9601 shown here is the mean of the four populations (MC0, MC1, MC4 and MC16) related to *P. marinus* AS9601 in the *Prochlorococcus* HLII clade, and the population membership is initially defined by PopCOGenT, followed by correction on a gene-by-gene basis by checking the gene tree topology. Note that these corrections have limited impact and all scaling patterns remain unchanged if the corrections are made. **a**, μ scales negatively with N_e. **b**, U_p scales negatively with N_e. **c**, No significant scaling relationship is found between genome size and N_e. **d**, μ scales negatively with genome size. Numbered data points 21–29 are not shown in **a**–**c** owing to the lack of population-level datasets needed for the estimation of N_e. The dashed grey lines and solid black lines represent the GLM and PGLS regression, respectively. The species name and its affiliation to a larger taxonomic unit (at either class, phylum or domain level) are shown at the bottom.

Periodic selection model helps explain the small effective population size of *Prochlorococcus*. Our finding of small N_e associated with populations within the HLII clade is unusual, given that HLII is the most abundant lineage in *Prochlorococcus*[33]. Metapopulation structure and periodic selection are the known mechanisms that account for low neutral genetic diversity and thus small N_e of natural bacterial populations[8]. In metapopulation structure, the population is divided into multiple patches, each connected with other patches through limited migrations[8]. The best example for metapopulation structure is associated with intracellular and obligately host-dependent populations in which individuals are transmitted between hosts in very small numbers[8]. Metapopulation structure is also characteristic of phytoplankton-associated bacteria such as *Sulfitobacter* spp.[45] and of particle-colonizing bacteria such as *Vibrio* spp.[8]. These marine bacteria explore intensively these nutrient-enriched microenvironments in oligotrophic waters for short bursts, followed by dispersal and colonization of new phytoplankton or particles by very small numbers of cells. Apparently, the above scenarios do not apply to *Prochlorococcus*, which are free-living carbon fixers that do not depend on hosts or colonize particles or other phytoplankton. We therefore turn to periodic selection, in which acquisitions of adaptive genetic variants may lead to fixation (that is, frequency reaching 100%) of the entire genome sequence carrying the adaptive variants owing to the generally low recombination rates in bacteria[46]. In the case of the four HLII populations (MC0, MC1, MC4 and MC16), the relative frequency of recombination to mutation (ρ/θ) is 0.04,

0.05, 0.32 and 0.19, respectively, and the relative effect of recombination to mutation (r/m) is 1.02, 2.96, 1.35 and 1.49, respectively. Whilst recombination has an impact on the population structure, this low rate of recombination is not able to prevent selective sweeps across the entire genome[46,47], suggesting that periodic selection is likely a significant force purging neutral genetic diversity within HLII populations and maintaining low N_e of each.

According to the periodic selection model, the coexistence of numerous genetically isolated populations is probably due to the occurrence of a myriad of niche dimensions, each occupied by a genetically isolated population. The latter is possible and partially indicated by their physiology[2]. For example, many *Prochlorococcus* cells adopt a mixotrophic lifestyle by either using the nitrogen, phosphorus and sulphur moieties of organic compounds (for example, amino acids including methionine and leucine, nucleic acids, dimethylsulphoniopropionate and adenosine triphosphate) or using them as an additional source of carbon and energy (for example, glucose)[48]. In the case of the four HLII populations highlighted here, there was a marginally significant geographical structure ($P=0.047$, Slatkin–Maddison test; Methods). Since most cells composing these four populations were sampled from two nearby sites, BATS and GA03 (Extended Data Fig. 4), dispersal limitation is less likely an important mechanism shaping this spatial population structure. We therefore suggest that different ecological niches harboured at these two sites may be a more important driver, though all the measured ecological factors (temperature, salinity, dissolved oxygen, phosphate, nitrate and nitrite) did not show major differences, except silicate concentrations (Extended Data Fig. 4). No seasonal structure was identified ($P=0.126$). Furthermore, we identified 5, 11, 9 and 4 genes unique to MC0, MC1, MC4 and MC16, respectively (Supplementary Table 5). Whilst most are functionally unknown, four genes involved in urea transport and urease nickel incorporation were unique to MC1. Whilst many population-specific genes were likely missing due to the dominance of partial genomes in the current dataset, the available results tentatively hint at the cryptic niches that supported the genetically isolated populations within the *Prochlorococcus* HLII clade.

Concluding remarks

Nearly all prior discussions on *Prochlorococcus* evolution were built on a key assumption of extremely large effective population sizes (N_e). In the present study, we report the unbiased global mutation rate of a genome-reduced *Prochlorococcus* member belonging to the most abundant high-light-adapted clade II. Based on these data and careful delineation of population borders, we showed that the N_e of clade II populations are only reasonably greater than those of endosymbiotic bacteria and surprisingly smaller than those of many known free-living bacteria. We further inferred that the small N_e is probably due to periodic selection, which is known to lead to fixation of neutral and slightly deleterious variants in linked loci during genome-wide sweeping. These new results challenge the traditional view that natural selection is extremely efficient in modern *Prochlorococcus* populations so that all genomic traits are optimized by selection. Instead, our data imply that genetic drift is a mechanism of paramount importance in today's *Prochlorococcus* populations for determining the fate of new traits that are continuously gained through mutation, homologous recombination and horizontal gene transfer, leading to increased random losses of some beneficial mutants that confer small and moderate competitive advantages and increased chance fixation of some detrimental ones that incur minor and moderate fitness costs. Together with a previous study which presented population genetic evidence that genetic drift was powerful at an early stage of *Prochlorococcus* evolution[49], we suggest that selection may not be as important as previously thought throughout the evolutionary history of *Prochlorococcus*. An improved estimate of the effective population sizes of *Prochlorococcus* lineages is essential to ensure an accurate understanding of the strategies adopted by these picocyanobacteria to become the most abundant photosynthetic organisms, which has important ramifications for global carbon cycles.

Methods

Culture and medium preparation. *Prochlorococcus marinus* AS9601 was isolated from sea water samples from the Arabian Sea[50]. In our research, it was cultured in Pro99 semisolid medium[51], which was made using oligotrophic surface sea water from the South China Sea. Sea water was first filtered through 0.2-μm-pore-size filters, then mixed with 0.375% (w/v) ultra-pure low-melting-point agarose (Invitrogen) before autoclaving. Nutrients were prepared by following a previous study[51] and were added to the warm medium (60 °C); 12 mL of medium was then poured into a new sterile Petri plate (90 mm diameter × 15 mm). After the medium cooled to below 30 °C, 0.5 mL of prepared 'helper bacteria' *Alteromonas* sp. EZ55[34] and a 0.5 mL dilution of AS9601 were added into the plate before the medium started to solidify. The semisolid plates were then mixed by gentle shaking. After 2 h of solidification, plates were sealed and then incubated at 24 °C in continuous light conditions (~30 μmol Q m^{-2} s^{-1}). For the preparation of helper bacteria, strains of EZ55 were cultured in 1/10 ProAC liquid medium[34] at 28 °C for 24 h, then washed with Pro99 liquid medium[51] three times by centrifugation (1,700 *g*, 10 min) before use.

Mutation accumulation experiment and genome sequencing. As unicellular cyanobacteria, *Prochlorococcus* cells have been shown to form single colonies in pour plates containing low-melting-point agarose, and one colony represents clonal replicates from a single cell[52]. A total of 150 MA lines were initiated, all of which started from a single founding colony of *Prochlorococcus marinus* AS9601. After an incubation period of ca. 3–4 weeks, a single colony from each MA line was randomly picked using a pipette and placed into a tube containing liquid Pro99 medium. These individual colonies were crushed with a pipette tip against the wall of the tube, vortexed adequately, diluted (~3,000-fold) then transferred into a new plate. The cells after vortexing were visualized under a fluorescence microscope to confirm that individual cells were well separated from each other and cell aggregates were not seen. The number of inoculated cells was limited to make sure that no more than 300 colonies formed in the new plate. Hence, each colony in the plate was in theory developed from a single cell (that is, the bottleneck size N_b is 1), and our MA procedure followed the 'single-cell bottleneck' rule during each transfer. If no colony formed in the new plate, an additional colony was transferred from the last plate of the corresponding line. This occurred 28 times throughout the MA experiment. Of the 150 MA lines, 141 survived throughout.

The effective population size (N_e) of an MA line was calculated based on the harmonic mean according to the equation

$$N_e = \frac{n+1}{\sum_{i=0}^{n} \frac{1}{2^i}},$$

where n is the number of cell divisions from the initial population (with a population size of 1) to the final population[53–56] (with a size of ~2 (ref. [25])), measured by flow cytometry for a single colony of *Prochlorococcus*). Here, n is 25 according to the logarithm base 2 of the final population size. Accordingly, N_e was calculated to be 13, a number at which the action of natural selection is negligible.

At the end of the MA experiment, each survived MA line was subjected to WGS. For each MA line, samples of the semisolid medium containing colonies were mixed with buffer to solubilize agarose gel and release the cells. The suspended cells were then collected on 0.2-μm-pore-size filter membranes and stored at −80 °C. DNA was extracted using the QIAamp DNA Mini Kit (Qiagen) and stored at −80 °C until subsequent sequencing with the Illumina NovaSeq platform with 150 bp paired-end.

Estimation of division frequency, death rate, total cell divisions and survival rate. After incubation for a period of time (t) of ca. 3–4 weeks, a colony from each of the 50 randomly selected MA lines was picked using a pipette and placed into a separate tube containing Pro99 liquid medium to measure the division frequency (f_d) of *P. marinus* AS9601. These samples were crushed, vortexed, fixed with 0.5% glutaraldehyde and stored at −80 °C. The total number of *Prochlorococcus* cells (N) in each colony was measured by flow cytometry (BD Accuri C6). The average division frequency was calculated by the equation $f_d = \frac{\log_2 N}{t}$.

The death rate (r_d) is approximated by the proportion of dead cells in a single colony, which was measured using a cell digestion assay[57]. Briefly, DNAse I and trypsin were added successively to each sample to digest any dead cells; the remaining live *Prochlorococcus* cells were counted by flow cytometry (N_l). The death rate was calculated by the equation $r_d = \frac{N-N_l}{N}$.

The number of cell divisions (g) was estimated on the basis of total cultivation time (T), division frequency and death rate[58]. During the MA experiment, division frequency was measured five times, with an interval of at least 3 months between each measurement. The average division frequency was 0.947 ± 0.041 per day. Given that the mean death rate and total cultivation time is 19.8% and 1,065 days,

respectively, the mean number of cell divisions was estimated as 1,258 across the 141 survived MA lines according to the equation $g = T \times \frac{f_d}{1-r_d}$.

To detect the effect of mutations at the gene locus *RS15215* on cell survival, the survival rate was measured in 20 MA lines, of which 10 lines accumulated mutations only at *RS15215* whereas the remaining 10 lines did not accumulate any mutations across the whole genomes. The survival rate was defined as the proportion of cells that successfully formed a colony after transfer. A single colony was picked using a pipette and placed into a separate tube containing liquid Pro99 medium. These colonies were then crushed, vortexed, and filtered through a sterile 100-µm-pore-size sieve to remove any traces of agarose. A 2-µL sample of each filtrate was diluted and inoculated into two semisolid medium plates (one including helper bacteria EZ55, the other without). The rest of the filtrate was fixed with 0.5% glutaraldehyde and stored at −80 °C prior to cell counting by flow cytometry (BD Accuri C6). Based on the number of colonies (n_{cln}) formed in plates after a 4-week incubation and the number of cells (n_c) inoculated into the plate according to the cell abundance measured by flow cytometry, the survival rate (r_s) was calculated by the equation $r_s = \frac{n_{cln}}{n_c}$.

Mutation calling and mutation rate determination. Raw reads were first processed by Trimmomatic 0.32 (ref. [59]) to remove adaptors and trim low-quality bases. Then the paired-end reads of the 141 MA lines were individually mapped to the *P. marinus* AS9601 reference genome using BWA-mem v.0.7.17 (ref. [60]). The resulting pileup files were converted to SAM format with SAMTOOLS v.1.4.1 (ref. [61]). Next, these SAM files were processed using Picard MarkDuplicates (http://broadinstitute.github.io/picard/) to remove duplicate reads which may arise during the sequencing process, like polymerase chain reaction (PCR) duplication artefacts. To adjust the base quality score affected by systematic technical errors, BaseRecalibrator in GATK-4.0 (ref. [62]) was used for base quality recalibration. Then base substitutions and small indels were called using HaplotypeCaller implemented in GATK-4.0 (ref. [62]). Variants were further filtered with standard parameters described by GATK best practice recommendations, except that the Phred-scaled quality score QUAL > 100 and root-mean-square mapping quality MQ > 59 were set, which followed previous studies[62–65]. An in-house script was implemented to identify genes showing significant enrichment of mutations (bootstrap test, $P < 0.05$ for each gene).

PCR primers were then designed with Primer Premier 5.0 (ref. [66]), and PCR was performed to confirm the mutations identified above. As the gene locus *RS15215* accumulated the most mutations during the MA experiment, 20 lines showing mutations in *RS15215* were chosen for validation by Sanger sequencing of the PCR products. Sixteen base substitutions and six indels in *RS15215* were sampled from these lines and validated (Supplementary Table 2). The average number of analysable sites (genomic regions with mapped reads) and the average coverage per site in the *P. marinus* AS9601 MA lines were 1,664,230 ± 203 and 440 ± 151, respectively.

The base-substitution mutation rate per nucleotide site per cell division (μ) for each line was calculated according to the equation

$$\mu = \frac{m}{ng},$$

where m is the number of observed base substitutions, n is the number of nucleotide sites analysed and g is the mean number of cell divisions estimated during the MA process. Following a previous study[35], the standard error of base-substitution mutation rate across all MA lines was calculated as

$$\text{s.e.}_{\text{pooled}} = \frac{s}{\sqrt{N}},$$

where s is the standard deviation of the mutation rate across all lines and N is the number of lines analysed.

An important observation from this MA/WGS work is that, of the 170 BPSs accumulated in the 141 surviving lines, 79 occurred in a single gene (*RS15215*) and all 79 mutations are non-synonymous. This is evidence that the mutations in *RS15215* are under positive selection. To minimize or even eradicate the impact of selection on our estimate of the global base-substitution mutation rate, we calculated the mutation rate based on two subsets of the mutation data. One subset used all 141 surviving lines but does not count BPSs occurring at *RS15215*. The other subset left out the 100 lines in which mutations occurred at *RS15215* and instead used the remaining 41 lines without any mutations at *RS15215*. To evaluate whether these two datasets produce a significant difference, the mutation rate for each was estimated based on maximum likelihood using poisson.test() in R (ref. [67]), and the 95% CI of mutation rate estimates were calculated using the Poisson cumulative distribution function.

To confirm that the effect of selection is minimized on the remaining genomic regions during the MA experiments, we calculated the ratio of accumulated mutations at protein-coding sites to those at intergenic sites and compared it with the ratio of the number of protein-coding sites to the number of intergenic sites using the χ^2 test. Within protein-coding genes, we calculated the ratio of accumulated non-synonymous to synonymous mutations and compared it with the 'risk' ratio of the number of all possible non-synonymous changes to the number of all possible synonymous changes using the exact binomial test. Here, we use risk ratio following previous studies[40,63,68]. The rationale is that, since each site within protein-coding genes has different probabilities of changing the amino acid sequence due to the redundancy of the genetic code, there is an unequal risk of synonymous versus non-synonymous changes at each site. Besides, these sites changed over time and may accumulate more than one mutation. These factors are controlled when the risk ratio is used. In practice, the number of sites at risk for synonymous and non-synonymous changes were estimated using 'gdtools count -b' syntax in breseq v.0.35.7 (ref. [69]). We implemented this procedure for the two subsets of the mutation data mentioned above separately.

Population delineation and effective population size estimation. Estimation of effective population size (N_e) for a prokaryotic species followed the equation $\pi_S = 2N_e\mu$, where π_S is the intraspecific nucleotide diversity at fourfold degenerate sites and μ is the unbiased global base-substitution mutation rate. It is challenging to define a microbial species boundary as genetically structured populations are common in microbial species, which has a major influence on π_S and thus N_e estimation. Members recombining freely therefore compose the ideal population to estimate N_e for a prokaryotic species[26]. Two programs were recently developed to delineate microbial populations: ConSpeciFix based on homoplastic single-nucleotide polymorphisms[70,71] and PopCOGenT based on enrichment of identical genomic regions[41]. PopCOGenT[41] is able to differentiate recent gene flow from historical events, whereas ConSpeciFix[70] integrates too long evolutionary timeframes, thus historical gene transfer events may blur the boundaries between closely related but genetically isolated populations[72]. Both were implemented here to define panmictic populations for the HLII clade.

Briefly, 580 public genomes affiliated with the HLII clade were downloaded from the National Center for Biotechnology Information (NCBI) GenBank database on 12 December 2020, and the quality of genome assembly was assessed by CheckM v.1.1.3 (ref. [73]). Whilst most data were derived from SAGs, which are often very incomplete, PopCOGenT can handle the partial genome sequences because the inference of recent gene flow is based on pairwise genome comparisons[41]. However, it is unknown whether and how SAGs may affect the performance of ConSpeciFix. Assemblies showing less than 50% completeness and over 5% contamination were filtered, and the remaining 418 high-quality genomes (23 cultured isolates and 395 SAGs) were used to delineate populations by PopCOGenT[41] in 'single-cell' mode and by ConSpeciFix[70,71] with default parameters, respectively. The former revealed fine population structures amongst the 418 HLII genomes, whereas the latter defined all these genomes as members of an entire species. Our analyses provided evidence for the reliability of the populations delineated by PopCOGenT (Supplementary Information) and against the delineation by ConSpeciFix (Supplementary Information), thus the subsequent N_e estimation was based on the populations defined by PopCOGenT.

Next, we estimated π_S values for each of the four populations with more than six non-redundant members (MC0, MC1, MC4 and MC16) defined by PopCOGenT. In practice, orthologous gene families were clustered using OrthoFinder-2.2.1 (ref. [74]) across genomes of each population. Owing to the fact that most members are represented by SAGs with high incompleteness, few single-copy orthologous gene families were present in all members within each population (0, 119, 2 and 7 for MC0, MC1, MC4 and MC16, respectively). As π_S derived from a small number of gene families may be biased, the selected gene families were not required to be shared by all population members. Briefly, the single-copy orthologous gene families were chosen if they were found in a certain proportion of the population members. This proportion was set to be equivalent to the average completeness of the genomes in each population, so that the selected genes are more likely shared by sequenced genomic regions. For instance, since the average completeness of MC0 members was 77%, each selected single-copy orthologous gene family was required to cover at least 77% of the MC0 members. In total, 403, 519, 783 and 666 single-copy orthologous gene families were identified in MC0, MC1, MC4 and MC16, respectively.

For each population, amino acid sequences of each single-copy orthologous gene family were aligned using MAFFT v.7.464 (ref. [75]) then imposed on nucleotide sequences. Next, fourfold degenerate sites were identified for each gene family using 'get4foldSites' available at https://github.com/brunonevado/get4foldSites. π_S was calculated using an in-house script based on the following formulas:

$$h_i = \frac{n}{n-1} \times \left(1 - \sum p^2\right), \quad (1)$$

$$\pi_S = \frac{\sum_{i=1}^{S} h_i}{N}, \quad (2)$$

where n is the number of strains, p is the allele frequency of each nucleotide at a segregating fourfold degenerate site, h_i is the heterozygosity at the ith segregating fourfold degenerate site, S is the number of segregating sites within all fourfold degenerate sites and N is the number of all fourfold degenerate sites. Finally, the median π_S across all single-copy gene families were used to calculate the N_e for each population.

Evaluating the effect of using SAGs on population delineation and N_e estimation through simulations. The population delineation and subsequent N_e

estimation of *Prochlorococcus* clade HLII relied heavily on the SAGs. However, the single-cell amplification procedure is often subjected to amplification bias, chimeric reads and read pairs that complicate the following assembly[76], so mismatches between the assembly of SAGs and that of reference genomes are not avoidable. These SAG-associated errors are expected to have an impact on population delineation by PopCOGenT, which defines populations based on enrichment of identical DNA segments. What is not clear is the extent to which these errors affect the population delineation and N_e estimation of the HLII clade. This calls for a computer simulation analysis to simulate SAGs from isolates' genomes (of other prokaryotic species), perform population delineation and N_e estimation based on the simulated SAGs and compare them with those derived from isolates' genomes.

Specifically, the SAG assemblies were simulated by importing a portion of variants, breaking, and removing a portion of sequences from original assemblies based on isolates' genomes. The magnitude of imported variants was derived from a benchmark dataset[77], where the highest mismatch error rate between SAG assemblies and the reference genome is about 15 mismatches per 100 kb, which was used in our simulation. Next, the completeness and the degree of fragmentation in the simulated SAGs were derived from a reference distribution modelled from the 557 publicly available SAGs of *Prochlorococcus* clade HLII, regardless of their quality. With these parameters, SAG assemblies were simulated from isolates' genomes, and population membership was subsequently determined by PopCOGenT and compared with the results based on isolates' genomes.

Amongst the 31 species collected from prior MA studies, 22 each had genome sequences of multiple intraspecific isolates for N_e estimation. The isolates' genomes used here are identical to the ones used in scaling analyses. We therefore implemented the above simulation procedure to these 22 species, each with ten replicates (this replication number balances the need for the subsequent statistical assessment with computational efficiency). For each species, if population splits occurred in the simulation, only the populations which contained the largest number of strains were used to compare with the original population based on isolates' genomes when assessing the impact of using SAGs on N_e estimates. Before the simulation, isolates with redundant genomes from clonal complexes were identified by PopCOGenT and were excluded. The rationale is that a clonal complex consists of nearly identical genomes and these clonal replications do not contribute to the gene pool of a population and will underestimate π_S and thus N_e. During SAG simulation, the imported variants may diversify the clonal replicates, thus the resulting simulated SAGs may no longer be grouped into a clonal complex but instead counted as non-redundant members of a population by PopCOGenT. This assignment may also lead to underestimation of π_S and thus N_e because most genomic regions remain identical between members originally from clonal complex. To control for these effects, only one genome from each clonal complex of the original isolates' genomes was kept for further SAG simulation and population delineation. This procedure was implemented for 12 (out of 22) species, where the largest defined population harbours at least one clonal complex.

Amongst the 22 species with simulated SAG assemblies, three (*Agrobacterium tumefaciens*, *Pseudomonas aeruginosa* and *Haloferax volcanii*) were not further pursued for population delineation with the simulated SAG assemblies. In *Agrobacterium tumefaciens*, originally available strains showed a large genetic distance (the least average nucleotide identity was about 78%), and some simulated SAGs turned out to be very different from the remaining members, which led to failure of pairwise genome alignment and thus population delineation by PopCOGenT. The *Pseudomonas aeruginosa* dataset contained over 300 genomes, each with approximately 6.5 Mbp, and PopCOGenT exceeded our computing capacity (available maximum memory of 512 GB RAM). In the case of *Haloferax volcanii*, only two genomes were available for simulation after removing redundant strains from clonal complexes. Hence, the remaining 19 species were used for simulation of SAG assemblies and population delineation.

Next, π_S and N_e were estimated for the delineated population from each simulated replicate dataset. In detail, the simulated SAG assemblies in the delineated population were first annotated by Prokka-1.14.6 (ref. [78]). Then, orthologous gene families were identified based on annotated protein-coding genes using OrthoFinder-2.2.1 (ref. [74]). Considering the incompleteness of simulated genomes, the criterion for retrieving single-copy orthologs was the same as that described in the last section. Finally, the same approach for calculating π_S mentioned above was employed for the selected single-copy orthologs, and N_e was further estimated for each newly defined population based on simulated SAG assemblies.

Adjusting population membership defined by PopCOGenT and thus N_e estimates according to gene tree topology. Since population membership may differ from gene to gene owing to recombination, we improved the population delineation by checking the gene tree topology and subsequently estimated N_e at the individual gene family level. The rationale is that population members should in theory form a monophyletic group for each gene family, though this may not be observed for some gene families due to the lack of phylogenetic signals associated with short and sometimes incomplete sequences.

We first selected 589 orthologous gene families for gene tree constructions. These genes were chosen because (i) they each had a single-copy ortholog in all 23 HLII isolates' genomes, which ensures that the chosen genes from SAGs had high diversity (since the isolates span the whole phylogeny of the HLII clade) and that the chosen gene families were most likely single-copy across the HLII cells; (ii) they each covered at least 50% members in at least one of the four main populations (MC0, MC1, MC4 and MC16) defined by PopCOGenT, which ensures that selected gene families each had a sufficient number of gene members that allow us to reconstruct gene flow patterns for at least one of the four populations. Of the 589 families, 169 each consisted of single-copy genes, whereas the remaining 420 families included 'multi-copy' genes that can be found in an average of 3.5 SAGs. We argue that these 'multi-copy' genes may not be true gene duplicates for two reasons: (i) genomes of the HLII clade are highly reduced, thus these SAGs are less likely to harbour multi-copy genes given that their orthologs are single-copy in all isolates; (ii) the lengths of the 'multi-copy' genes were much shorter than their single-copy orthologs, suggesting that fragmented SAG assemblies may lead to multiple sequences of a single gene dispersing in different contigs. Despite these compelling reasons, we were not able to rule out the possibility of true gene duplicates and thus simply filtered out the 'multi-copy' genes for downstream analyses. Next, the protein-coding genes were aligned at the amino acid sequence level using MAFFT v.7.464 (ref. [75]) and the alignment imposed on the nucleotide sequences for each gene family. The gene trees were constructed using IQ-TREE 2.0 (ref. [79]) with ModelFinder[80] assigning the best substitution model and with 1,000 ultrafast bootstrap replicates. The gene tree topology was checked with a strict criterion: all members in each of the four populations present in the gene tree form a monophyletic group.

For the gene families in which members from a population defined by PopCOGenT formed monophyletic groups, population membership remained unchanged. However, for those families that did not, we used the following criteria to adjust population membership for a given gene family based on gene tree topology: Assume that M strains included in a gene tree are from MC*i* (MC*i* = MC0, MC1, MC4 or MC16) defined by PopCOGenT, that N strains compose a monophyletic group which is enriched in members from MC*i* and that m out of N strains in this monophyletic group are from MCi. Such a monophyletic group is defined as a population for this gene family if it meets the following criteria:

(i) Members from MC*i* account for at least 70% of the members in this monophyletic group (that is, $m/N \geq 70\%$);
(ii) At least 70% of the members from MC*i* (defined by PopCOGenT) are included in this monophyletic group (that is, $m/M \geq 70\%$);
(iii) No more than three strains from each of the remaining three populations are included in this monophyletic group;
(iv) The last common ancestor of this monophyletic group should be the same as that of the m members from MC*i*.

For the population MC0, PopCOGenT further separated it into three subclusters (MC0.0, MC0.1 and MC0.2, including 52, 11 and 2 members, respectively), and the separation of MC0.1 from the other two was supported by 82 (out of 589) gene trees. For these genes, each split into two subclades, we used the N_e calculated from the subclade showing greater neutral genetic diversity to represent the N_e of this particular gene population.

π_S for each gene was estimated based on the gene-specific population, and N_e was subsequently calculated for each gene using the global mutation rate. Next, the median value of the calculated N_e across the gene families was designated as the N_e of the population.

Regression analysis between mutation rate, effective population size and genome size. Another 31 phylogenetically diverse species were included to investigate the relationship between mutation rate, effective population size (N_e) and genome size (G) across prokaryotic lineages. The base-substitution mutation rate per site per generation (μ) was collected from prior MA/WGS studies (Supplementary Table 4), and the genome size was collected from the NCBI GenBank database[81] if not mentioned in the corresponding study. Amongst the 31 species, 22 each had multiple isolates' genomes available from the NCBI RefSeq database[82], and thus were used for estimating N_e. Eight out of the 22 species had numerous isolates' genomes available (varying from 79 to 4,221 genomes), and feeding so many genomes into PopCOGenT is computationally intractable. We therefore started our analysis with a subset of these genomes which were previously characterized for population boundaries by ConSpeciFix[71] (Supplementary Table 4). For the remaining 14 species, all isolates' genomes available from the NCBI RefSeq database (last accessed June 2020) were used. Within each species, population boundaries were characterized by PopCOGenT, and in the case of multiple genetically isolated populations identified, the one consisting of the largest number of non-redundant members was used for further analyses. Gene-by-gene correction of the population membership was not performed. Next, single-copy orthologous gene families shared by all members within the population were identified using OrthoFinder-2.2.1 (ref. [74]) and were further used for the calculation of π_S and N_e with the same approach as mentioned above.

It is worth mentioning that the prokaryotic species included here showed high phylogenetic diversity. They included both Bacteria and Archaea. Within Bacteria, multiple major branches were included in the two deeply branching clades (Extended Data Fig. 5)[83], namely Terrabacteria (for example, Cyanobacteria,

Firmicutes, Actinobacteria and Tenericutes) and Gracilicutes (for example, Proteobacteria, Fibrobacteres–Chlorobi–Bacteroidetes (FCB) group and Planctomycetes–Verrucomicrobia–Chlamydiae (PVC) group). The included species also showed high physiological diversity; they encompassed both aerobes and obligate anaerobes, both heterotrophic members and photosynthetic carbon-fixing cyanobacteria, species adopting both free-living and obligate host-dependent lifestyles and species with temperature and salinity optima varying over a wide range. The use of this comprehensive dataset strengthened the conclusion derived from the regression analysis.

Amongst the pairwise scaling analyses between mutation rate, N_e and genome size, the negative scaling between mutation rate and N_e was used to support the 'drift-barrier' model. Whilst we initially used base-substitution mutation rate per site per generation (μ) to represent mutation rate in this analysis, more direct support for this model should instead use the genome-wide deleterious mutation rate, U_D, which selection operates on. Since a precise estimate of U_D is not available, the genome-wide mutation rate in all protein-coding genes per generation, U_p, was used as a proxy for U_D, which is the product of μ and the number of nucleotides in all protein-coding genes of the strain subjected to MA/WGS analysis[19]. The length of protein-coding genes of each species was collected from the NCBI GenBank database[81] if not available in the corresponding study.

The pairwise linear relationship between μ, U_p, N_e and G across 32 prokaryotic species was assessed with both GLM and PGLS methods, implemented in the 'stats'[67] and 'caper'[84] packages in R v.4.0.2 (ref. [67]), respectively. For μ versus G, all 32 species were used. In the cases of N_e versus μ, N_e versus U_p and N_e versus G, only the 23 species each with multiple strains' genomes were used because N_e can only be estimated when multiple genomes are available. In terms of GLM regression, the outlier data point (Bonferroni $P < 0.05$) was identified using the 'outlierTest' function in 'car' package v.3.0-11 (ref. [85]). For the PGLS regression, a species tree must be used as an input. Since our dataset included 30 bacterial species and 2 archaeal species and since a recent study clarified that using archaea to root the bacterial tree of life can easily distort the topological structure within the latter[83], we derived the topological structure of the 32 or 23 species from the Genome Taxonomy Database Toolkit. Specifically, the tree topology for the 30 bacterial species and the one for the 2 archaeal species were each pruned from Genome Taxonomy Database release95 (ref. [86]) then combined manually. Next, the branch length of this combined tree was estimated based on this fixed tree topology and the protein sequences of 27 conserved marker genes shared by both Bacteria and Archaea[83] using IQ-TREE[79] under the best-fitting model (LG+C60+R8+F), which followed a recent study[83]. Since the autocorrelation between species of close evolutionary relationship may lead to artificial scaling relationship of traits, the potential association of trait evolution with the phylogeny (that is, phylogenetic signal, represented by Pagel's λ (ref. [87])) was evaluated using the 'pgls' function of the 'caper' package v.1.0.1 (ref. [84]), which took the phylogeny of the 32 or the 23 species as an input. The value of λ ranges from 0 to 1, with 0 indicating no phylogenetic signal and 1 indicating a strong phylogenetic signal due to Brownian motion. The P values for the lower and upper bounds represent whether λ is significantly different from 0 and 1, respectively. There was a marginally significant phylogenetic effect on the relationship of N_e versus G ($\lambda = 0.996$, lower bound $P = 0.060$, upper bound $P = 0.600$) and limited effect on the relationship for N_e versus μ ($\lambda = 0$, lower bound $P = 1$, upper bound $P = 0.151$), N_e versus U_p ($\lambda = 0$, lower bound $P = 1$, upper bound $P = 0.002$) and μ versus G ($\lambda = 0$, lower bound $P = 1$, upper bound $P = 8.505 \times 10^{-11}$).

Functional annotation and population structure analysis of HLII strains. Protein-coding genes of the 418 HLII genomes were predicted using Prokka 1.14.6 (ref. [78]), and functional annotation of these genes was performed using the RAST server[88,89]. The clade-specific gene families were identified based on their presence and absence in the four main populations (MC0, MC1, MC4 and MC16). For a gene family claimed to be specific to one population, gene members should be present in at least 60% of the members from that population but be present in no more than five members from any of the remaining three populations.

Permutation test for correlation between population distribution and sampling location or season. To test whether geographical and seasonal factors contributed to the population structure, the Slatkin–Maddison test was implemented for the four main populations (MC0, MC1, MC4 and MC16) defined by PopCOGenT. This test allows statistical detection of enrichment of certain labels (for example, sampling location or season) through permutation of the labels across the phylogenetic tree. The sampling depth information was missing in about one-third of strains, so the test along depth could not be performed. Members of the four populations came from five sampling locations, amongst which two at BATS were very close so they were counted as the same site. Thus, four different labels were set for these sampling locations. Since all members of the four populations were sampled from the Northern Hemisphere, sampling time was divided into spring, summer, autumn and winter accordingly. The phylogenetic tree of the four populations was pruned from the phylogenomic tree of the whole HLII clade (Extended Data Fig. 1).

Assessing the rate and effect of recombination relative to mutation. To measure the recombination rates within each of the four main populations (MC0, MC1, MC4 and MC16) defined by PopCOGenT, the whole genome alignment for each population was produced using progressiveMauve-2.4.0 (ref. [90]) and the core genome alignments longer than 500 bp were extracted using the stripSubsetLCB module provided by Mauve[91] and concatenated. The phylogeny of each population was pruned from the phylogenomic tree of the whole HLII clade (Extended Data Fig. 1). With these inputs, ClonalFrameML[92] was implemented to estimate the relative frequency of recombination to mutation (ρ/θ) and the relative effect of recombination to mutation (r/m) for each population.

The effect of recombination was also measured for the 418 genomes (23 isolates' genomes and 395 high-quality SAGs) across the HLII clade. Owing to the extensive use of SAG partial genomes, there was not a core genome alignment and thus progressiveMauve could not be used. We therefore turned to the core genome alignment of the 23 HLII isolates, which spread over the genome-based phylogeny of HLII (Extended Data Fig. 1) and thus can be a proxy for the entire HLII clade. In general, the r/m ratio of these four main populations was 1.02, 2.96, 1.35 and 1.49, which is consistent with that (1.72) of the entire HLII clade represented by the 23 HLII isolates.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Source data are provided with this paper. All the datasets generated, analysed and presented in the current study are available in the Supplementary Information. Raw reads of the 141 surviving lines are available at the NCBI SRA under accession no. PRJNA733321. The 589 gene trees are available at https://doi.org/10.6084/m9.figshare.c.5638369.

Code availability

All the scripts are deposited at https://doi.org/10.6084/m9.figshare.c.5638369.

Received: 22 May 2021; Accepted: 14 October 2021;
Published online: 23 December 2021

References

1. Flombaum, P. et al. Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl Acad. Sci. USA* **110**, 9824–9829 (2013).
2. Biller, S. J., Berube, P. M., Lindell, D. & Chisholm, S. W. *Prochlorococcus*: the structure and function of collective diversity. *Nat. Rev. Microbiol.* **13**, 13–27 (2015).
3. Johnson, Z. I. et al. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740 (2006).
4. Zinser, E. R. et al. Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol. Oceanogr.* **52**, 2205–2220 (2007).
5. Kettler, G. C. et al. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet.* **3**, 2515–2528 (2007).
6. Luo, H., Friedman, R., Tang, J. & Hughes, A. L. Genome reduction by deletion of paralogs in the marine cyanobacterium *Prochlorococcus*. *Mol. Biol. Evol.* **28**, 2751–2760 (2011).
7. Batut, B., Knibbe, C., Marais, G. & Daubin, V. Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat. Rev. Microbiol.* **12**, 841–850 (2014).
8. Fraser, C., Alm, E. J., Polz, M. F., Spratt, B. G. & Hanage, W. P. The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**, 741–746 (2009).
9. Bobay, L.-M. & Ochman, H. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* **18**, 153 (2018).
10. Kuo, C.-H., Moran, N. A. & Ochman, H. The consequences of genetic drift for bacterial genome complexity. *Genome Res.* **19**, 1450–1454 (2009).
11. Luo, H., Swan, B. K., Stepanauskas, R., Hughes, A. L. & Moran, M. A. Comparing effective population sizes of dominant marine Alphaproteobacteria lineages. *Environ. Microbiol. Rep.* **6**, 167–172 (2014).
12. Kryazhimskiy, S. & Plotkin, J. B. The population genetics of d_N/d_S. *PLoS Genet.* **4**, e1000304 (2008).
13. Rocha, E. P. C. & Feil, E. J. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of bacteria? *PLoS Genet.* **6**, e1001104 (2010).
14. Luo, H., Thompson, L. R., Stingl, U. & Hughes, A. L. Selection maintains low genomic GC content in marine SAR11 lineages. *Mol. Biol. Evol.* **32**, 2738–2748 (2015).
15. Hellweger, F. L., Huang, Y. & Luo, H. Carbon limitation drives GC content evolution of a marine bacterium in an individual-based genome-scale model. *ISME J.* **12**, 1180–1187 (2018).

16. Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
17. Andreani, N. A., Hesse, E. & Vos, M. Prokaryote genome fluidity is dependent on effective population size. *ISME J.* **11**, 1719–1721 (2017).
18. Kimura, M. Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. *Genet. Res.* **11**, 247–270 (1968).
19. Lynch, M. et al. Genetic drift, selection and the evolution of the mutation rate. *Nat. Rev. Genet.* **17**, 704–714 (2016).
20. Gu, J. et al. Unexpectedly high mutation rate of a deep-sea hyperthermophilic anaerobic archaeon. *ISME J.* **15**, 1862–1869 (2021).
21. Charlesworth, B. Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009).
22. Kashtan, N. et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**, 416–420 (2014).
23. Osburne, M. S., Holmbeck, B. M., Coe, A. & Chisholm, S. W. The spontaneous mutation frequencies of *Prochlorococcus* strains are commensurate with those of other bacteria: mutation frequencies in *Prochlorococcus*. *Environ. Microbiol. Rep.* **3**, 744–749 (2011).
24. Lee, H., Popodi, E., Tang, H. & Foster, P. L. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. *Proc. Natl Acad. Sci. USA* **109**, E2774–E2783 (2012).
25. Williams, A. B. Spontaneous mutation rates come into focus in *Escherichia coli*. *DNA Repair* **24**, 73–79 (2014).
26. Rocha, E. P. C. Neutral theory, microbial practice: challenges in bacterial population genetics. *Mol. Biol. Evol.* **35**, 1338–1347 (2018).
27. Marais, G. A. B., Calteau, A. & Tenaillon, O. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica* **134**, 205–210 (2008).
28. Marais, G. A. B., Batut, B. & Daubin, V. Genome evolution: mutation is the main driver of genome size in prokaryotes. *Curr. Biol.* **30**, R1083–R1085 (2020).
29. Morris, J. J., Lenski, R. E. & Zinser, E. R. The black queen hypothesis: evolution of dependencies through adaptive gene loss. *mBio* **3**, e00036–12 (2012).
30. Giovannoni, S. J., Cameron Thrash, J. & Temperton, B. Implications of streamlining theory for microbial ecology. *ISME J.* **8**, 1553–1565 (2014).
31. Wernegreen, J. J. Endosymbiont evolution: predictions from theory and surprises from genomes: endosymbiont genome evolution. *Ann. N. Y. Acad. Sci.* **1360**, 16–35 (2015).
32. Bobay, L.-M. & Ochman, H. The evolution of bacterial genome architecture. *Front. Genet.* **8**, 72 (2017).
33. Malmstrom, R. R. et al. Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J.* **4**, 1252–1264 (2010).
34. Morris, J. J., Kirkegaard, R., Szul, M. J., Johnson, Z. I. & Zinser, E. R. Facilitation of robust growth of *Prochlorococcus* colonies and dilute liquid cultures by 'helper' heterotrophic bacteria. *Appl. Environ. Microbiol.* **74**, 4530–4534 (2008).
35. Sun, Y. et al. Spontaneous mutations of a model heterotrophic marine bacterium. *ISME J.* **11**, 1713–1718 (2017).
36. Dillon, M. M., Sung, W., Sebra, R., Lynch, M. & Cooper, V. S. Genome-wide biases in the rate and molecular spectrum of spontaneous mutations in *Vibrio cholerae* and *Vibrio fischeri*. *Mol. Biol. Evol.* **34**, 93–109 (2017).
37. Strauss, C., Long, H., Patterson, C. E., Te, R. & Lynch, M. Genome-wide mutation rate response to pH change in the coral reef pathogen *Vibrio shilonii* AK1. *mBio* **8**, e01021–17 (2017).
38. Xue, C.-X. et al. Ancestral niche separation and evolutionary rate differentiation between sister marine flavobacteria lineages. *Environ. Microbiol.* **22**, 3234–3247 (2020).
39. Bourguignon, T. et al. Increased mutation rate is linked to genome reduction in prokaryotes. *Curr. Biol.* **30**, 3848–3855.e4 (2020).
40. Long, H. et al. Background mutational features of the radiation-resistant bacterium *Deinococcus radiodurans*. *Mol. Biol. Evol.* **32**, 2383–2392 (2015).
41. Arevalo, P., VanInsberghe, D., Elsherbini, J., Gore, J. & Polz, M. F. A reverse ecology approach based on a biological definition of microbial populations. *Cell* **178**, 820–834 (2019).
42. Sung, W., Ackerman, M. S., Miller, S. F., Doak, T. G. & Lynch, M. Drift-barrier hypothesis and mutation-rate evolution. *Proc. Natl Acad. Sci. USA* **109**, 18488–18492 (2012).
43. Daubin, V. & Moran, N. A. Comment on 'The origins of genome complexity'. *Science* **306**, 978–978 (2004).
44. Kirchberger, P. C., Schmidt, M. L. & Ochman, H. The ingenuity of bacterial genomes. *Annu. Rev. Microbiol.* **74**, 815–834 (2020).
45. Qu, L. et al. Metapopulation structure of diatom-associated marine bacteria. Preprint at *bioRxiv* https://doi.org/10.1101/2021.03.10.434754 (2021).
46. Wiedenbeck, J. & Cohan, F. M. Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. *FEMS Microbiol. Rev.* **35**, 957–976 (2011).
47. Gillespie, J. H. Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* **155**, 909–919 (2000).
48. Muñoz-Marín, M. C. et al. Mixotrophy in marine picocyanobacteria: use of organic compounds by *Prochlorococcus* and *Synechococcus*. *ISME J.* **14**, 1065–1073 (2020).
49. Luo, H., Huang, Y., Stepanauskas, R. & Tang, J. Excess of non-conservative amino acid changes in marine bacterioplankton lineages with reduced genomes. *Nat. Microbiol.* **2**, 1–9 (2017).
50. Shalapyonok, A., Olson, R. J. & Shalapyonok, L. S. Ultradian growth in *Prochlorococcus* spp. *Appl. Environ. Microbiol.* **64**, 1066–1069 (1998).
51. Moore, L. R. et al. Culturing the marine cyanobacterium *Prochlorococcus*. *Limnol. Oceanogr.* **5**, 353–362 (2007).
52. Lindell, D. in *The Prokaryotes: Other Major Lineages of Bacteria and The Archaea* (eds Rosenberg, E. et al.) 829–845 (Springer, 2014).
53. Long, H. et al. Antibiotic treatment enhances the genome-wide mutation rate of target cells. *Proc. Natl Acad. Sci. USA* **113**, E2498 (2016).
54. Dillon, M. M., Sung, W., Lynch, M. & Cooper, V. S. The rate and molecular spectrum of spontaneous mutations in the GC-rich multichromosome genome of *Burkholderia cenocepacia*. *Genetics* **200**, 935–946 (2015).
55. Wahl, L. M. & Gerrish, P. J. The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution* **55**, 2606–2610 (2001).
56. Hall, D. W., Mahmoudizad, R., Hurd, A. W. & Joseph, S. B. Spontaneous mutations in diploid *Saccharomyces cerevisiae*: another thousand cell generations. *Genet. Res.* **90**, 229–241 (2008).
57. Agustí, S. Viability and niche segregation of *Prochlorococcus* and *Synechococcus* cells across the central Atlantic Ocean. *Aquat. Microb. Ecol.* **36**, 53–59 (2004).
58. Frenoy, A. & Bonhoeffer, S. Death and population dynamics affect mutation rate estimates and evolvability under stress in bacteria. *PLoS Biol.* **16**, e2005056 (2018).
59. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
61. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
62. McKenna, A. et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
63. Long, H., Behringer, M. G., Williams, E., Te, R. & Lynch, M. Similar mutation rates but highly diverse mutation spectra in ascomycete and basidiomycete yeasts. *Genome Biol. Evol.* **8**, 3815–3821 (2016).
64. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
65. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinform.* **43**, 11.10.1–33 (2013).
66. Singh, V. K., Mangalam, A. K., Dwivedi, S. & Naik, S. Primer premier: program for design of degenerate primers from a protein sequence. *BioTechniques* **24**, 318–319 (1998).
67. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2021).
68. Shewaramani, S. et al. Anaerobically grown *Escherichia coli* has an enhanced mutation rate and distinct mutational spectra. *PLoS Genet.* **13**, e1006570 (2017).
69. Deatherage, D. E. & Barrick, J. E. in *Engineering and Analyzing Multicellular Systems: Methods and Protocols* (eds Sun, L. & Shou, W.) 165–188 (Springer, 2014).
70. Bobay, L.-M., Ellis, B. S.-H. & Ochman, H. ConSpeciFix: classifying prokaryotic species based on gene flow. *Bioinformatics* **34**, 3738–3740 (2018).
71. Bobay, L.-M. & Ochman, H. Biological species are universal across life's domains. *Genome Biol. Evol.* **9**, 491–501 (2017).
72. VanInsberghe, D., Arevalo, P., Chien, D. & Polz, M. F. How can microbial population genomics inform community ecology? *Philos. Trans. R. Soc. B* **375**, 20190253 (2020).
73. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
74. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157 (2015).
75. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
76. Lasken, R. S. & McLean, J. S. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat. Rev. Genet.* **15**, 577–584 (2014).
77. Clingenpeel, S., Clum, A., Schwientek, P., Rinke, C. & Woyke, T. Reconstructing each cell's genome within complex microbial communities—dream or reality? *Front. Microbiol.* **5**, 771 (2015).

78. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
79. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
80. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).
81. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. GenBank. *Nucleic Acids Res.* **35**, D21–D25 (2007).
82. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
83. Coleman, G. A. et al. A rooted phylogeny resolves early bacterial evolution. *Science* **372**, eabe0511 (2021).
84. Orme, D. et al. The caper package: comparative analysis of phylogenetics and evolution in R (2013).
85. Fox, J. & Weisberg, S. *An R Companion to Applied Regression* (SAGE, 2019).
86. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).
87. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).
88. Aziz, R. K. et al. The RAST server: rapid annotations using subsystems technology. *BMC Genom.* **9**, 75 (2008).
89. Overbeek, R. et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **42**, D206–D214 (2014).
90. Darling, A. E., Mau, B. & Perna, N. T. Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **5**, e11147 (2010).
91. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).
92. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).

Acknowledgements

The authors thank the three reviewers for providing constructive suggestions that substantially improved the manuscript, H. Long and J. Pan for sharing their script to calculate π_S based on fourfold degenerate sites and X. Feng for contributing the script to simulate the SAG assemblies based on the isolates' genomes. Y.Z. was supported by the National Science Fund for Distinguished Young Scholars (42125603) and NSFC project 92051114. H.L. was supported by the Shenzhen Science and Technology Committee (JCYJ20180508161811899), the Hong Kong Branch of Southern Marine Science and Engineering Guangdong Laboratory (Guangzhou) (SMSEGL20SC02) and the Hong Kong Research Grants Council Area of Excellence Scheme (AoE/M-403/16). Z.C. was supported by the PhD Fellowship of the State Key Laboratory of Marine Environmental Science at Xiamen University.

Author contributions

H.L. conceptualized the work and strategy, directed the bioinformatics analyses, interpreted the data and wrote the main manuscript. Y.Z. directed the experimental analyses and related writing, co-interpreted the data and provided comments on the manuscript. Z.C. performed all the experiments with contributions from Y.S., co-interpreted the data, drafted the experimental methods and prepared Fig. 1. X.W. performed all the bioinformatics analyses, co-interpreted the data, drafted the bioinformatics methods and prepared Figs. 2 and 3, and all the supplementary figures. Q.Z. co-directed the experimental work.

Competing interests

The authors declare no competing interests.

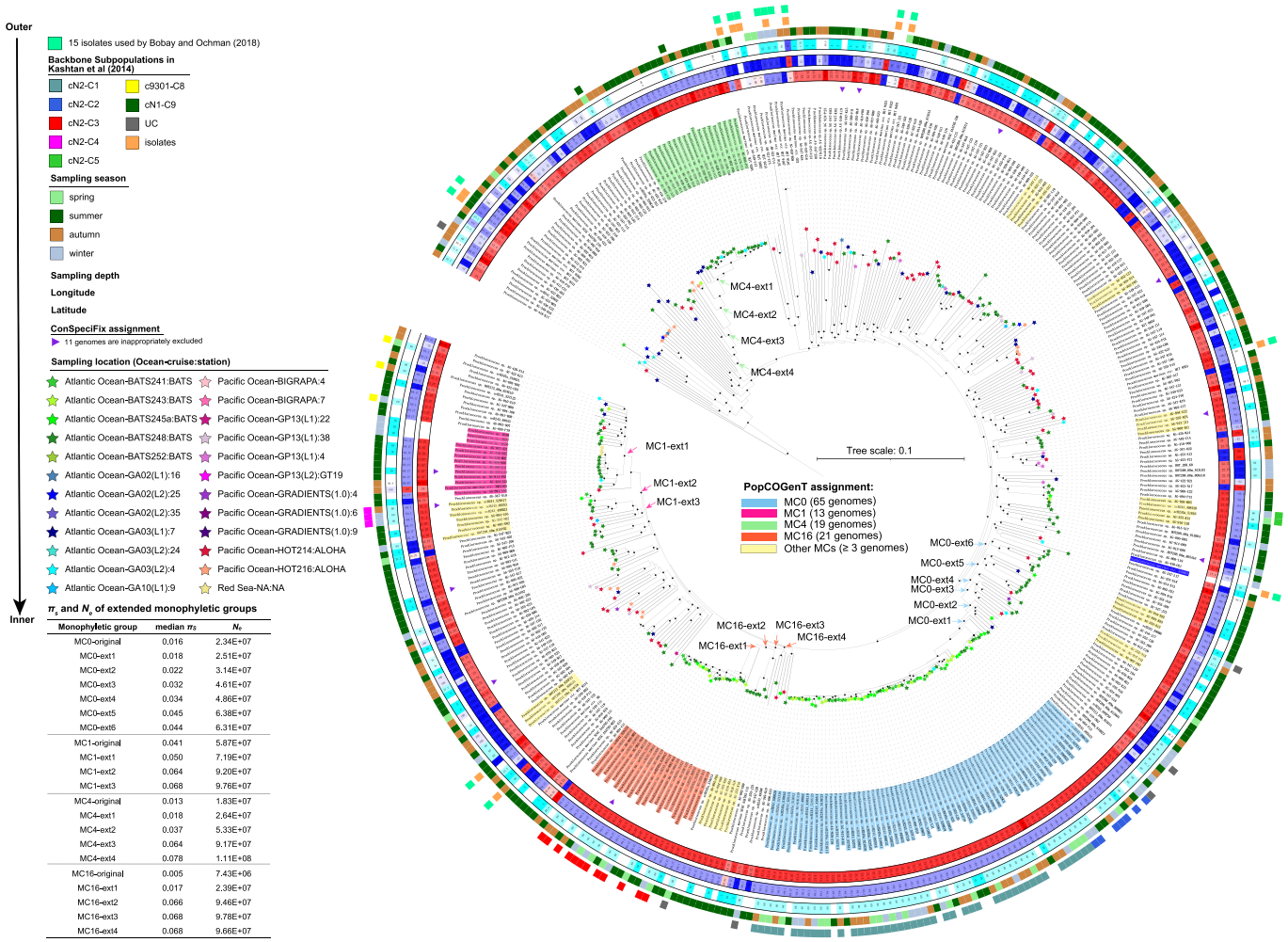
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-021-01591-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41559-021-01591-0.

Correspondence and requests for materials should be addressed to Yao Zhang or Haiwei Luo.

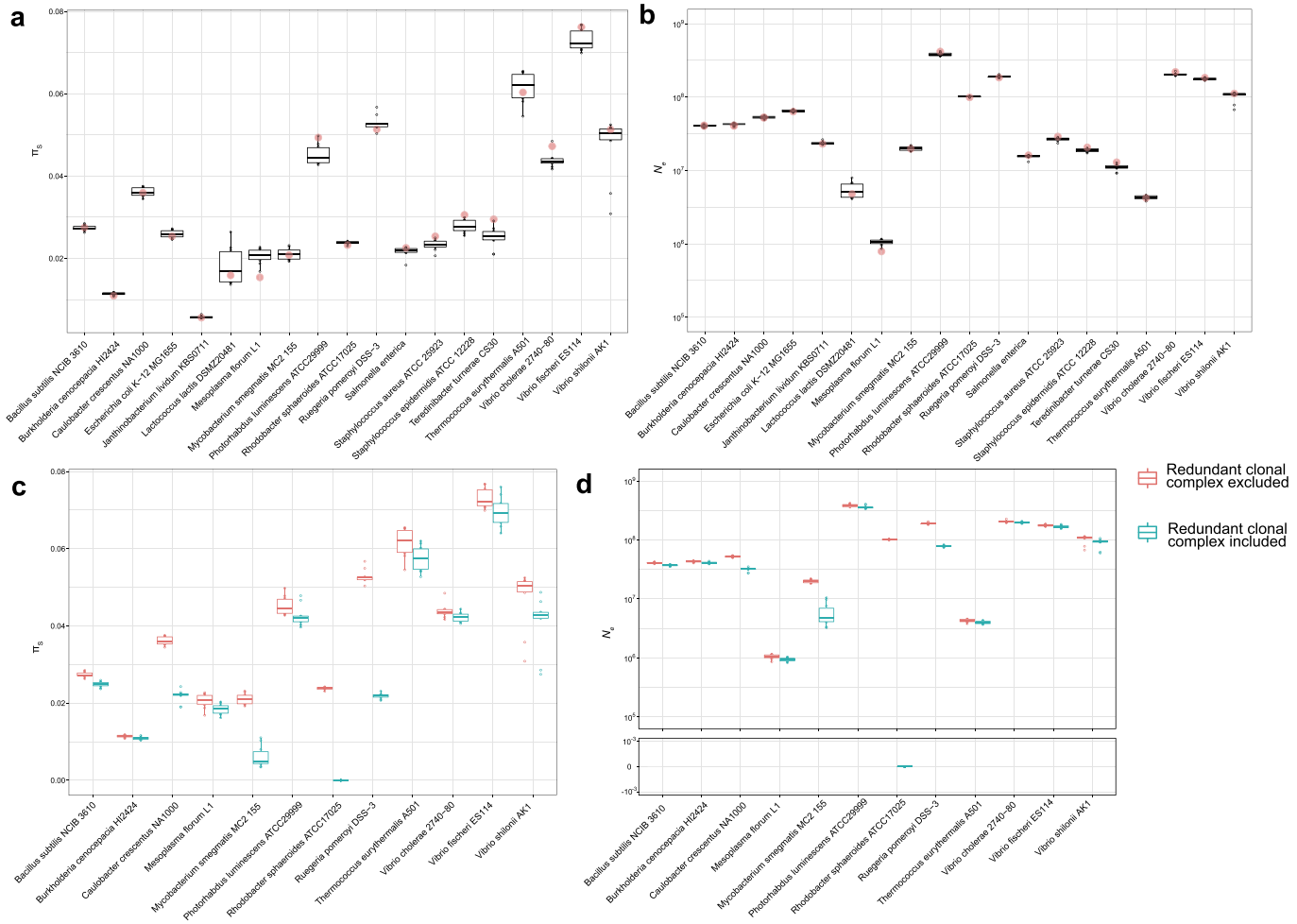
Peer review information *Nature Ecology & Evolution* thanks Louis-Marie Bobay, Sébastien Wielgoss and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



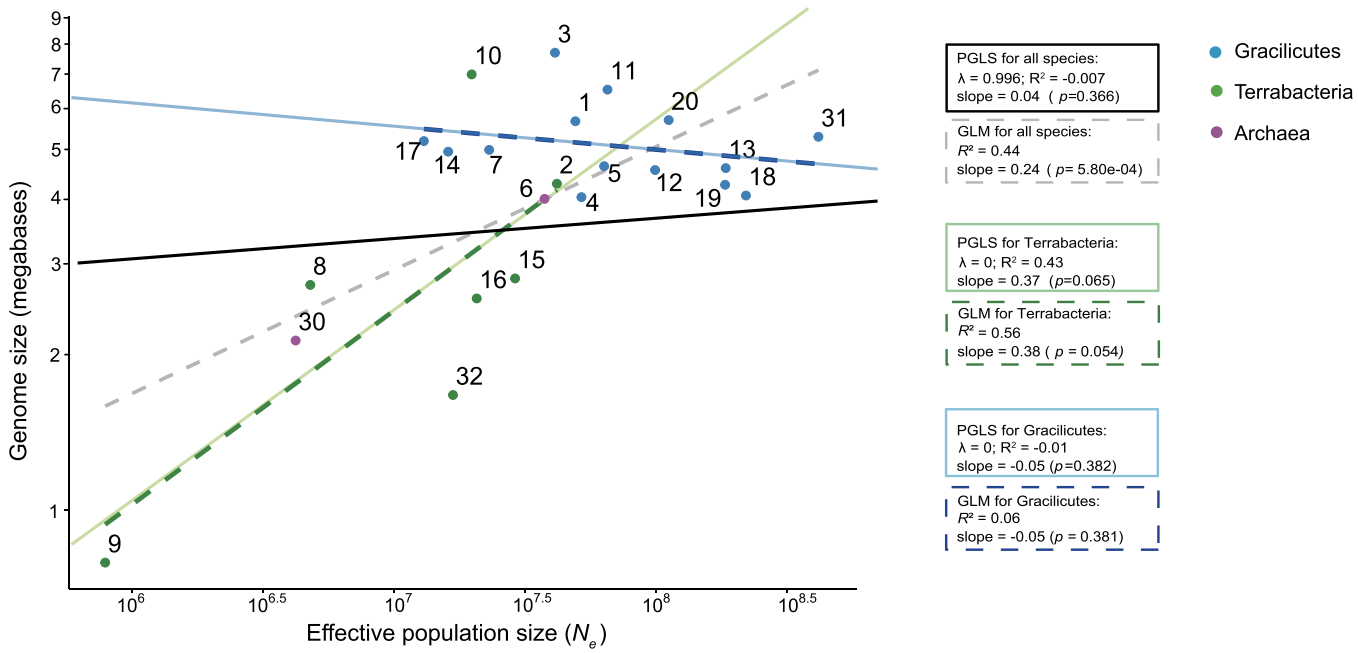
π_s and N_e of extended monophyletic groups

Monophyletic group	median π_S	N_e
MC0-original	0.016	2.34E+07
MC0-ext1	0.018	2.51E+07
MC0-ext2	0.022	3.14E+07
MC0-ext3	0.032	4.61E+07
MC0-ext4	0.034	4.86E+07
MC0-ext5	0.045	6.38E+07
MC0-ext6	0.044	6.31E+07
MC1-original	0.041	5.87E+07
MC1-ext1	0.050	7.19E+07
MC1-ext2	0.064	9.20E+07
MC1-ext3	0.068	9.76E+07
MC4-original	0.013	1.83E+07
MC4-ext1	0.018	2.64E+07
MC4-ext2	0.037	5.33E+07
MC4-ext3	0.064	9.17E+07
MC4-ext4	0.078	1.11E+08
MC16-original	0.005	7.43E+06
MC16-ext1	0.017	2.39E+07
MC16-ext2	0.066	9.46E+07
MC16-ext3	0.068	9.78E+07
MC16-ext4	0.068	9.66E+07

Extended Data Fig. 1 | The maximum-likelihood phylogenomic tree of 523 high-light (HL) adapted *Prochlorococcus* genomes (26 isolates' genomes and 497 high-quality SAGs from HLII, HLI, HLIII/IV), constructed with IQ-TREE based on concatenated single-copy orthologous genes at the amino acid level and rooted with low-light (LL) adapted clade I (LLI) genomes. To help show the fine phylogenetic structure within the HLII clade, all distant relatives (LLI, HLIII/IV, HLI and HLVI) were pruned. Solid circles at the nodes indicate the frequency of the group defined by that node greater than 90 out of the 100 bootstrapped replicates. The phylogeny is visualized and annotated with iTOL. From the outer to inner rings: (1) denotes the 15 HLII isolates used in Bobay and Ochman (2018) for defining the species '*Prochlorococcus marinus*' by ConSpeciFix. (2) denotes the 93 genomes (13 isolates and 80 high quality SAGs) used for N_e estimation by Kashtan et al (2014), among which 80 HLII SAGs were classified into seven backbone subpopulations and are marked with corresponding color strips; (3)-(6) represents the season, water depth, longitude and latitude of the samples they collected, respectively; (7) shows the identity of the SAGs or isolates. Colored stars at the tips of the phylogeny differentiate the strains of distinct sources (that is, ocean, cruise, and station). Strains without the above information are not marked at the tips. The strain *P. marinus* AS9601 subjected to unbiased global mutation rate determination is highlighted in deep blue. Cells with the genome id highlighted with blue, pink, green, and orange compose the population MC0, MC1, MC4, and MC16, respectively, delineated by PopCOGenT. Other populations delineated by PopCOGenT were highlighted with light yellow, with each consisting of at least three non-redundant genomes (Supplementary Table 3). (8) illustrates the progressive extensions of the four main populations (MC0, MC1, MC4, and MC16) defined by PopCOGenT, with arrows marking the most recent common ancestor of each extension. The purpose of this analysis is to estimate the impact of population delineation on N_e estimates. Both π_S and N_e were estimated for the extended groups (left bottom). The ConSpeciFix grouped all 418 HLII members (23 isolates and 395 high-quality SAGs) into one species except 11 strains (marked with purple triangles), which ConSpeciFix inappropriately reported as distantly related relatives to the defined species.

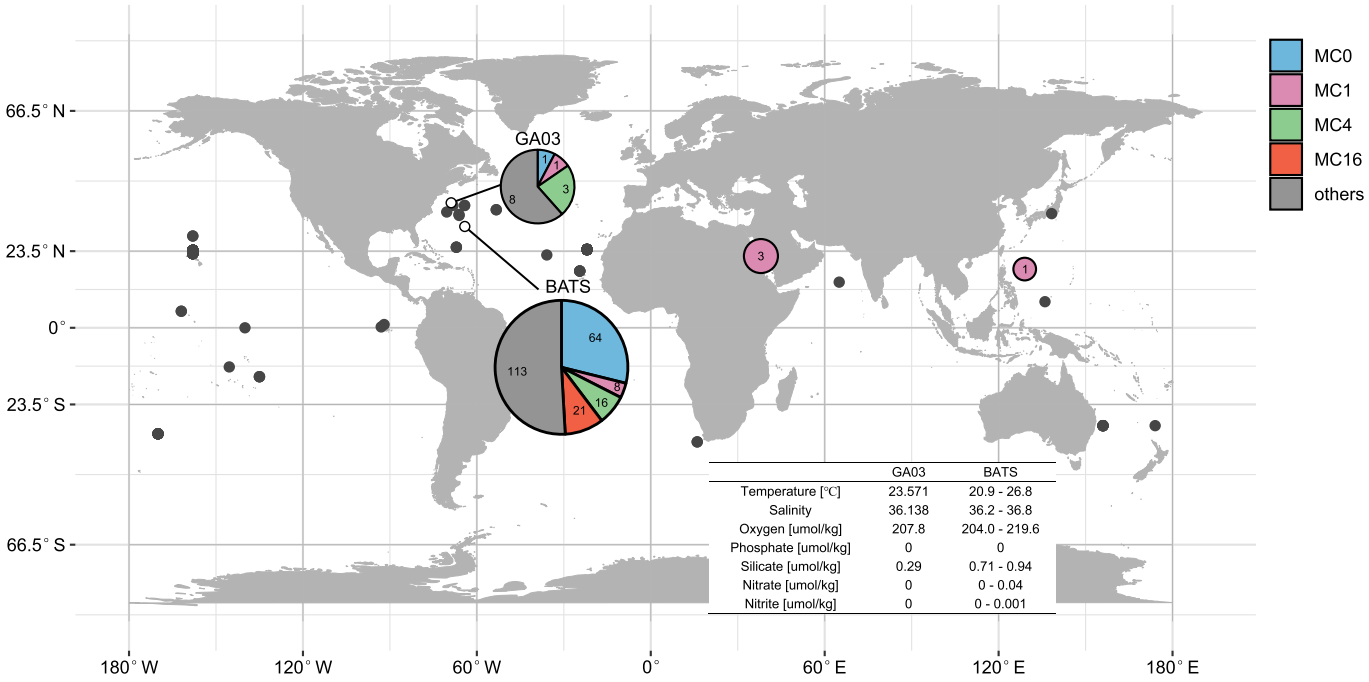


Extended Data Fig. 2 | The SAG simulation analysis evaluates the impact of using error-prone SAG data on population delineation by PopCOGenT and subsequently on the estimates of π_S and N_e. Of the 31 prokaryotic species with their unbiased global mutation rate data publicly available, 19 are used in the simulation analysis because these species each have multiple members and thus are amenable for population delineation. SAGs are simulated from the isolates' genomes in each of these 19 species by incorporating the realistic error rates (collected from literature) associated with SAG sequences and the genomic statistics of all available *Prochlorococcus* clade HLII SAGs (without quality filter). For each species, the whole procedure was replicated for 10 times. **(a-b)** Summary of π_S and N_e estimates for each of the 19 species based on populations delineated by PopCOGenT using simulated SAGs (box and whisker plot) and the original isolates' genomes (red solid circles). Within each box of the SAG data, the horizontal line marks the median; boxes extend from the 25th to the 75th percentile of each group's π_S or N_e; whiskers above and below the box indicate the 10th and 90th percentiles. **(c-d)** Including redundant members from a clonal complex underestimates π_S and N_e. Before each simulation, the clonal complex identified in original genomes by PopCOGenT was preprocessed by excluding redundant strains. This is because these strains can be erroneously identified as non-redundant population members by PopCOGenT when simulated SAG data are used, which underestimates π_S and N_e owing to the extremely close relationship between members from a clonal complex. In total, 12 species have the problem of clonal complex. As expected, the inclusion of redundant strains in a clonal complex before SAG simulations results in a decrease of π_S and N_e estimates compared with the exclusion of the redundant strains from a clonal complex.

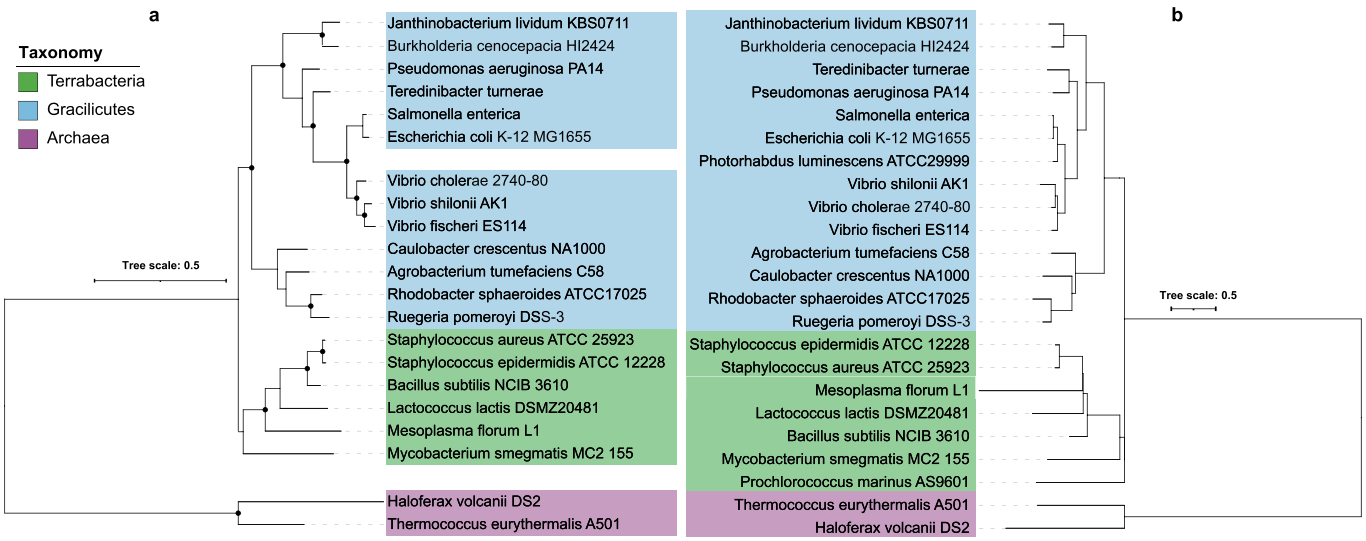


1	*Agrobacterium tumefaciens* C58 (Alphaproteobacteria)	17	*Teredinibacter turnerae* (Gammaproteobacteria)
2	*Bacillus subtilis NCIB* 3610 (Firmicutes)	18	*Vibrio cholerae* 2740-80 (Gammaproteobacteria)
3	*Burkholderia cenocepacia* HI2424 (Betaproteobacteria)	19	*Vibrio fischeri* ES114 (Gammaproteobacteria)
4	*Caulobacter crescentus* NA1000 (Alphaproteobacteria)	20	*Vibrio shilonii* AK1 (Gammaproteobacteria)
5	*Escherichia coli* K-12 MG1655 (Gammaproteobacteria)	21	*Arthrobacter* sp. KBS0703 (Actinobacteria)
6	*Haloferax volcanii* DS2 (Archaea)	22	*Colwellia psychrerythraea* 34H (Gammaproteobacteria)
7	*Janthinobacterium lividum* KBS0711 (Betaproteobacteria)	23	*Deinococcus radiodurans* R1 (Deinococcus-Thermus)
8	*Lactococcus lactis* DSMZ20481 (Firmicutes)	24	*Flavobacterium* sp. KBS0721 (FCB)
9	*Mesoplasma florum* L1 (Tenericutes)	25	*Gemmata obscuriglobus* DSM5831 (PVC)
10	*Mycobacterium smegmatis* MC2 155 (Actinobacteria)	26	*Kineococcus radiotolerans* SRS30216 (Actinobacteria)
11	*Pseudomonas aeruginosa* PA14 (Gammaproteobacteria)	27	*Leeuwenhoekiella* sp. ZYFB001 (FCB)
12	*Rhodobacter sphaeroides* ATCC17025 (Alphaproteobacteria)	28	*Micrococcus* sp. KBS0714 (Actinobacteria)
13	*Ruegeria pomeroyi* DSS-3 (Alphaproteobacteria)	29	*Nonlabens* sp. SY33080 (FCB)
14	*Salmonella enterica* (Gammaproteobacteria)	30	*Thermococcus eurythermalis* A501 (Archaea)
15	*Staphylococcus aureus* ATCC 25923 (Firmicutes)	31	*Photorhabdus luminescens* ATCC29999 (Gammaproteobacteria)
16	*Staphylococcus epidermidis* ATCC 12228 (Firmicutes)	32	*Prochlorococcus marinus* AS9601 (Cyanobacteria)

Extended Data Fig. 3 | No scaling relationship was found between the logarithmically transformed estimated effective population size (N_e) and the logarithmically transformed genome size across 21 bacterial and two archaeal species. The generalized linear model (GLM) regression and the phylogenetic generalized least square (PGLS) regression of the 23 species are identical from what is presented in Fig. 3c. The Pagel's λ among 23 species is near to 1, suggesting strong phylogenetic signal (that is, traits evolve in close association with the phylogeny). Thus, the scaling was further investigated at a lower taxonomic rank. Specifically, the 21 bacterial species were divided into two deep lineages, the Terrabacteria clade (seven species marked by green dots) and the Gracilicutes clade (14 species marked by blue dots), and both GLM and PGLS regression analyses were applied to each. Again, no scaling relationship between N_e and genome size was found for either Terrabacteria or Gracilicutes.



Extended Data Fig. 4 | The global distribution of 418 HLII members (23 isolates and 395 high-quality SAGs). Members of the four main populations (MC0, MC1, MC4, and MC16) defined by PopCOGenT are represented with blue, pink, green, and orange, respectively. These members were largely sampled from two sites (BATS and GA03) in North Atlantic Ocean marked by white dots, where pie charts are used to illustrate the proportion of each population in the sampled cells. Another two pink circles with numbers denote the sites where four members of MC1 were from. The sampling sites of members from other minor populations defined by PopCOGenT are marked as black dots, and at least one cell was sampled in these sites. The table on the right bottom lists the available environmental factors of the two sites. Genomes from BATS were derived from four independent samples, and thus the range of each environmental factor is provided.



Extended Data Fig. 5 | The phylogenetic trees used for phylogenetic generalized least squares (PGLS) regression analyses. Species affiliated with Terrabacteria, Gracilicutes, and Archaea were shadowed with green, blue, and pink, respectively. **a**, The maximum likelihood phylogeny built from the 16S rRNA gene sequences of 21 prokaryotic species, which was generated in a recently published study. Solid circles at the nodes indicate the frequency of the group defined by that node greater than 90 out of the 100 bootstrapped replicates. **b**, In this phylogeny, the tree topology was pruned from GTDB release95, followed by branch length estimation with the fixed tree topology under maximum likelihood framework.

nature research

Corresponding author(s): Haiwei Luo, Yao Zhang

Last updated by author(s): Sep 30, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We generated most of the data ourselves. External, published information is referenced with accession numbers. The availability of all these data is specified in 'Data availability'.
Data analysis	For almost all genomic analyses, we used custom bash, R, perl, and python codes, all of which are provided following the link https://doi.org/10.6084/m9.figshare.c.5638369

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All genomic data generated for this study are archived in the sequence read archive under bioproject ID PRJNA733321 at the National Centre of Biotechnology Information (www.ncbi.nlm.nih.gov/sra)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We experimentally determined the unbiased global mutation rate of a Prochlorococcus strain affiliated with the most abundant highlight-adapted clade II (HLII). We also characterized the gene flow discontinuities and delineated population boundaries within the Prochlorococcus HLII. Based on these parameters, we estimated the effective population sizes (Ne) of the Prochlorococcus HLII lineages. The estimated Ne is surprisingly low, potentially resulting from periodic selection in the ocean waters. The low Ne and the

	unexpected population structure indicate that random genetic drift must be considered in future studies of the evolutionary mechanisms shaping the genomic features of Prochlorococcus.
Research sample	The samples for growth rate, death rate, and survival rate measurements were derived from the colonies formed in semisolid plates.
Sampling strategy	For the determination of mutation rate, DNA of all 141 survived mutant lines was extracted and sequenced. For the estimation of cell division frequency and death rate, 50 colonies from 50 different mutant lines were sampled. For the estimation of survival rate, 10 colonies from 10 mutant lines with mutations restricted to the gene RS15215 and another 10 colonies from 10 mutant lines showing no mutations were sampled.
Data collection	The DNA was sequenced using Illumina Novaseq platform with 150 bp pair-end. The cell number from flow cytometry and the colony number on plates were recorded by Zhuoyu Chen in Excel and with photos.
Timing and spatial scale	The mutation accumulation experiments were performed from Feb 2018 to Jan 2021. The DNA samples was extracted from 03 Jan 2021 to 17 Feb 2021. These experiments were performed at Xiamen University.
Data exclusions	All data are presented, including the mutations concentrated at the gene RS15215.
Reproducibility	The experimental procedure, as described in the manuscript, is reproducible.
Randomization	For genome sequencing, all 141 survived mutant lines were sampled. For the mutation accumulation experiment, a random single colony was picked and transfered. To estimate cell division rate and death rate, 50 mutant lines (out of the 141 survived lines) were randomly selected and a random colony from each was sampled. To estimate the survival rate, 10 lines were randomly selected from the 25 lines without accumulating any mutations and another 10 lines were randomly selected from the 42 lines where mutations were restricted to the gene RS15215. For each of the selected lines, a random colony was sampled.
Blinding	Blinding was not relevant to this study as all experimental data was included for analysis and procedures were standardized so there was no subjective manipulation possible by the experimenter.

Did the study involve field work? ☐ Yes ☒ No

Field work, collection and transport

Field conditions	*Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).*
Location	*State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).*
Access & import/export	*Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).*
Disturbance	*Describe any disturbance caused by the study and how it was minimized.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	No animals were involved. The cyanobacterium Prochlorococcus marinus AS9601 was originally isolated from Arabian Sea.
Wild animals	The study did not involve wild animals.
Field-collected samples	The study did not involve samples collected from the field.
Ethics oversight	No ethical approval or guidance was required.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	*Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."*
Recruitment	*Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.*
Ethics oversight	*Identify the organization(s) that approved the study protocol.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*

Outcomes | *Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	The colonies of Prochlorococcus marinus AS9601 were picked into separate tubes containing Pro99 liquid medium, crushed, and vortexed. If they were not counted immediately, they were fixed with 0.5% glutaraldehyde and stored at -80 °C.
Instrument	BD Accuri C6 Flow Cytometry
Software	BD Accuri C6 software (version 1.0.264.21)
Cell population abundance	The pure culture of Prochlorococcus marinus AS9601 was counted. The colony samples were obtained from the laboratory. No fluorochrome was used.
Gating strategy	The clustering events with red fluorescence between 2,000 and 16,000 and SSC between 100 and 12,000 were counted as Prochlorococcus.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*

Volume censoring

Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.

Statistical modeling & inference

Model type and settings

Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).

Effect(s) tested

Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.

Correction

Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.