environmental
microbiology

Environmental Microbiology (2019) **00**(00), 00–00 doi:10.1111/1462-2920.14552

Marine *Synechococcus* isolates representing globally abundant genomic lineages demonstrate a unique evolutionary path of genome reduction without a decrease in GC content

Michael D. Lee,[1,2] **Nathan A. Ahlgren**,[3]
Joshua D. Kling,[1] **Nathan G. Walworth**,[1]
Gabrielle Rocap,[4] **Mak A. Saito**,[5] **David A. Hutchins**[1]
and Eric A. Webb[1]*

[1]*Department of Biological Sciences, University of Southern California, Los Angeles, CA, USA.*
[2]*Exobiology, Ames Research Center, Moffett Field, CA, USA.*
[3]*Department of Biology, Clark University, Worcester, MA, USA.*
[4]*School of Oceanography, University of Washington, Seattle, WA, USA.*
[5]*Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institute, Woods Hole, MA, USA.*

Summary

***Synechococcus*, a genus of unicellular cyanobacteria found throughout the global surface ocean, is a large driver of Earth's carbon cycle. Developing a better understanding of its diversity and distributions is an ongoing effort in biological oceanography. Here, we introduce 12 new draft genomes of marine *Synechococcus* isolates spanning five clades and utilize ~100 environmental metagenomes largely sourced from the TARA Oceans project to assess the global distributions of the genomic lineages they and other reference genomes represent. We show that five newly provided clade-II isolates are by far the most representative of the recovered *in situ* populations (most 'abundant') and have biogeographic distributions distinct from previously available clade-II references. Additionally, these isolates form a subclade possessing the smallest genomes yet identified of the genus (2.14 ± 0.05Mbps; mean ± 1SD) while concurrently hosting some of the highest GC contents (60.67 ± 0.16%). This is in direct opposition to the pattern in *Synechococcus*'s nearest relative, *Prochlorococcus* – wherein decreasing genome size has coincided with a strong *decrease* in GC content – suggesting this new subclade of *Synechococcus* appears to have convergently undergone genomic reduction relative to the rest of the genus, but along a fundamentally different evolutionary trajectory.**

Introduction

Synechococcus and its closest relative *Prochlorococcus* are the most abundant photoautotrophs on our planet. Broadly speaking, *Prochlorococcus* populations extend to deeper depths and are found in greater abundance in warm, oligotrophic waters, but are more latitudinally and niche-restricted than *Synechococcus* populations – which thrive in a wider range of environments including polar and nutrient-rich habitats (Flombaum *et al*., 2013; Sohm *et al*., 2015; Farrant *et al*., 2016). The successful proliferation of both of these genera across the world's oceans is owed to a plethora of varied genomic lineages that have evolved since the divergence of *Synechococcus* and *Prochlorococcus* – estimated to have occurred between ~0.25 and 1 billion years ago (Dvořák *et al*., 2014). Due to their profusion, ubiquity and carbon-fixing nature – all of which are anticipated to expand under current global change trends (Flombaum *et al*., 2013) – understanding the diversity and ecology of these marine picoplankton is fundamental to our understanding of Earth's carbon cycling. *Synechococcus* abundance in particular has recently been shown to be strongly correlated with carbon export (Guidi *et al*., 2016).

Members of *Synechococcus* have been grouped into various clades based on both physiological characteristics and various phylogenetic markers (Dufresne *et al*., 2008; Ahlgren and Rocap, 2012). Their global distributions have also been studied using marker genes (e.g. the 16S rRNA gene, 16S–23S internal transcribed spacer, *rpoC1*, *ntcA* and *petB*), which has revealed some coherent trends in biogeographic distributions corresponding to clade designations (Sohm *et al*., 2015; Farrant *et al*., 2016; Ahlgren and Rocap, 2012; Toledo and Palenik, 1997;

Received 7 November, 2018; revised 10 January, 2019; accepted 29 January, 2019. *For correspondence. E-mail eawebb@usc.edu; Tel. 213-740-7954; Fax 213-740-8123.

Penno *et al*., 2006; Mazard *et al*., 2012; Zwirglmaier et al., 2008; Kent *et al*., 2018). For instance, clades I and IV are commonly reported at higher latitudes with cooler waters, while clade II is more frequently detected in warmer, lower-latitude waters (Sohm *et al*., 2015; Farrant *et al*., 2016). Individual marker genes are not necessarily indicative of their surrounding genomic content however, and, due to their inherent broad-level resolution, can often mask fine-scale diversity with biogeographical and ecological implications (Chase and Martiny, 2018).

Results and discussion

Nineteen *Synechococcus* isolate genomes sourced from various locations were sequenced, assembled and manually curated resulting in 12 new, distinct (< 98% average nucleotide identity over >90% of the smaller genome), high-quality draft genomes (estimated ~ > 99% complete and ~ < 1% redundancy; Supporting Information Table S1). Phylogenomic analysis with previously available marine *Synechococcus* reference genomes placed these within several clades including I, II, XV, XVI and CRD1 (Fig. 1A, new genomes are bolded and underlined; Supporting Information Fig. S1), with corresponding genome sizes and GC content depicted in Fig. 1B. In order to assess the environmental relevance of these and other currently available marine *Synechococcus* isolate genomic lineages, we created a reference library of 31 non-redundant reference genomes and recruited metagenomic reads from a total of 97 environmental samples (of size fraction 0.2–3 μm) spanning all major oceans and the Mediterranean and Red Seas (Supporting Information Table S2 and Fig. S2). Overall, our *Synechococcus* reference library recruited roughly 1.13% of a total of ~32 billion quality-filtered reads – with ~1.21% mapping from 65 surface samples and ~0.99% mapping from 32 deep-chlorophyll maximum samples (Supporting Information Table S2).

Environmental *in situ* microbial populations are never identical to the reference genomes we have; microbial diversity is simply too great. However, the value of reference-based metagenomics is that reference genomes grant us access to the genomic lineages they are closely related to from any given sample – these can be considered the genomic lineages they represent. Read recruitment from environmental populations tells us which of our currently available reference genomes are most similar to the *in situ* populations, or most representative of the *in situ* populations (please see Supporting Information Note 1 for a more detailed discussion of this). To mitigate artifactual recruitment due to non-specific mapping, we employed a 'detection' criterion for our reference genomes requiring that at least 50% of the reference base pairs (bps) recruit reads in order for that genome to be considered 'representative of the *in situ* population' (please see Supporting Information Note 2 along with Supporting Information Fig. S3 for an example and further explanation of 'detection'). Boxplots of each reference genome's levels of detection (Supporting Information Table S3) and the number of samples in which they passed this detection threshold are presented in Fig. 1C, which combined with their relative abundances (here defined as the proportion of recruited reads; Fig. 1D) together demonstrate which currently available *Synechococcus* isolates best represent the recovered *in situ* populations of the incorporated samples. For instance, reference genome RS9917's maximum detection in any sample was ~13% (meaning at most only 13% of its bps recruited any reads from any sample; Fig. 1C), and therefore, it was not deemed representative of any of the *in situ Synechococcus* populations recovered from the incorporated 97 metagenomic samples; RS9917 has previously been found to be associated with hypersaline environments rather than the more-typical marine waters analysed here (Dufresne *et al*., 2008). In contrast, isolate genome N32 within clade II recruited reads to more than 50% of its genome in 38 of the 97 incorporated samples (Fig. 1C) and was responsible for >13% of total read recruitment across the entire dataset (Fig. 1D).

The distributions and relative abundances of *Synechococcus* clades varied across the 65 included surface-ocean samples, with locations with highest *Synechococcus* abundance dominated by clade II (Fig. 2); deeper samples showed similar trends but in lower relative abundance; Supporting Information Fig. S4; Supporting Information Table S2. The locations with the greatest recovered abundances of *Synechococcus* were: site 141 just north of the Panama Canal, where ~9% of total reads mapped to our reference genomes; site 33 in the Red Sea with ~7% of all reads recruiting; and site 57 in the southern Indian Ocean with ~6% – all three of which were mostly dominated by clade II (Fig. 2 and Supporting Information Table S2). In fact, the top 13 samples in terms of relative abundance of *Synechococcus* were all dominated by clade II, and consistent with previous literature utilizing marker genes (Sohm *et al*., 2015; Farrant *et al*., 2016), clade II abundance overall was found to be significantly positively correlated with iron (Fe) and temperature (Supporting Information Fig. S5).

Recruitment to clade II was not evenly distributed across its seven representative genomes, however. Each of the five newly contributed genomes within clade II (N32, N5, UW86, N27 and N19; Fig. 1) recruited greater than twice the amount of reads as did the two remaining clade II references, CC9605 and WH8109 (Fig. 1D and Supporting Information Table S4). These five also showed distinct distribution patterns as compared with

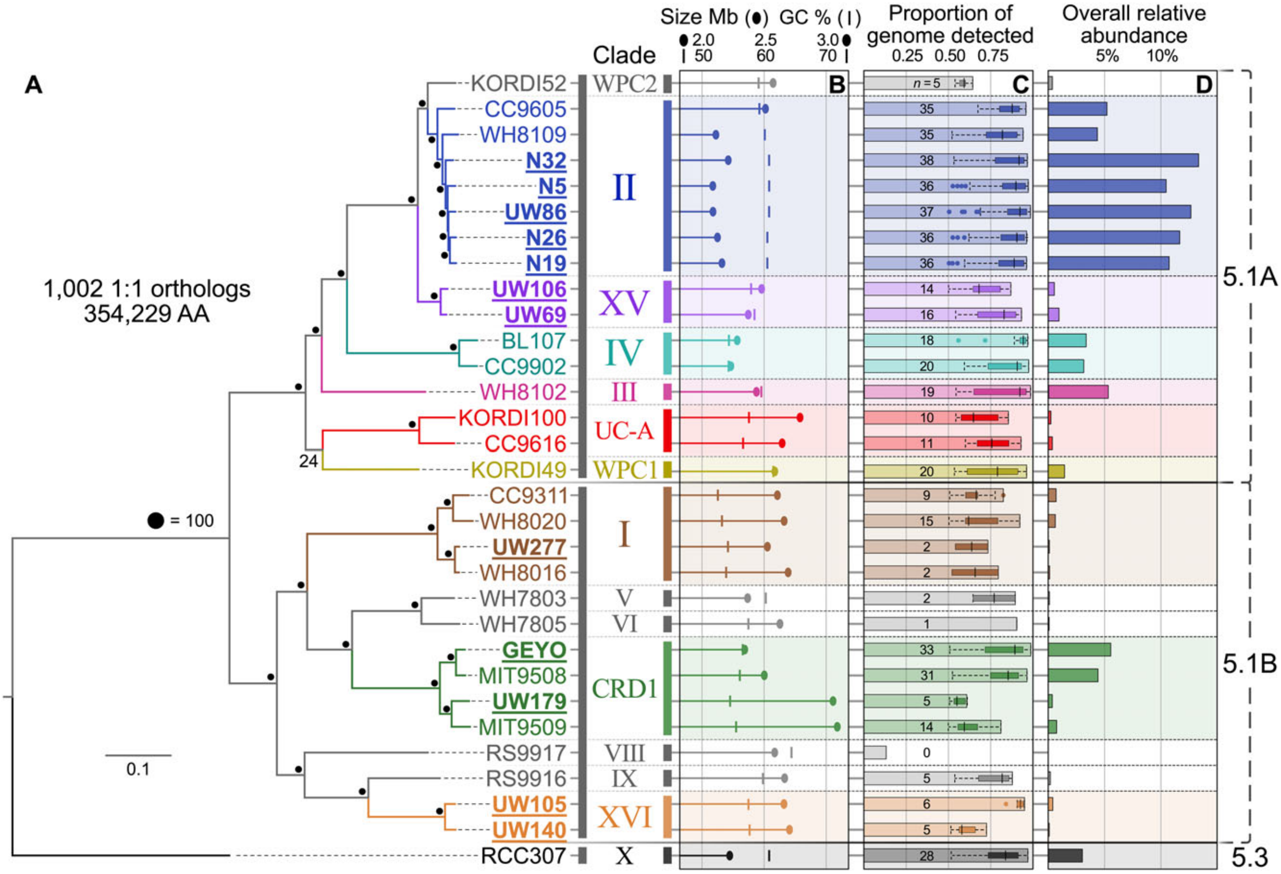

Fig. 1. Phylogenomic tree of the 31 analysed Synechococcus genomes (A) and genome information (B–D).
Phylogenomic tree (A) is maximum-likelihood (100BS) of 1002 orthologs present in single-copy in each of the 31 incorporated Synechococcus reference genomes forming an alignment of 363 980 amino acids. The root depicted is inferred from a similar tree built incorporating Gloeobacter violaceus (Supporting Information Fig. S1). Underlined and bolded genomes are newly provided from the current study, and colours correspond to clades. Size and GC content are plotted in panel B. In panel C, the corresponding genome's maximum detection is plotted (see 'detection' in main text, Supporting Information Note 2 and Supporting Information Fig. S3), and within it are boxplots of the reference genome's detection in all samples for which it was deemed representative of the in situ population (see Supporting Information Note 1), total samples represented by the number shown. Panel D plots each reference's overall relative abundance – defined herein as the proportion of reads recruited to each genome out of the total number of reads recruited to the entire library of the 31 reference genomes (i.e. the 'overall relative abundance' of all reference genomes sums to 100%).

CC9605 and WH8109, which can be seen when viewing the relative abundance of each representative genome per sample (see Fig. 3; and Supporting Information Fig. S6 depicts a map with the relative abundances of just clade II genomes). For example, looking at the sample with the greatest recovered proportion of *Synechococcus* reads (site 141, just north of the Panama Canal in Fig. 2) reveals the majority of read recruitment was to references N19, N26, UW86, N5 and N32 (each accruing a median coverage of greater than 200X), with relatively few reads recruiting to WH8109 or CC9605 (resulting in less than 20X coverage each; in Fig. 3, see row marked by arrow with '141'; also see Supporting Information Fig. S7 and Supporting Information Note 3 for more detail). A similar pattern was seen for site 33 of the Red Sea (Figs 2 and 3 and Fig. S6). Additionally, there were differences in significant correlations among the clade II representatives to environmental parameters – only the five newly contributed genomes had significantly positive correlations with temperature, and only CC9605 had significantly negative correlations with phosphate and nitrogen (Fig. 3). Since one of the primary utilities of clade designations is to delineate distinct, ecologically relevant units, the strongly divergent environmental and geographic distributions of genomes N32, N5, UW86, N26 and N19 from the rest of clade II (Figs 1D and 3; Supporting Information Figs. S6 and S7) suggest that they represent an ecologically distinct subclade; herein we refer to them as 'clade II-S'.

Following their divergence from *Synechococcus*, most known genomic lineages of *Prochlorococcus*, excluding low-light clade IV (LLIV), have proceeded down an evolutionary path of genomic streamlining coinciding with a substantial reduction in GC content (Dufresne *et al*., 2008;

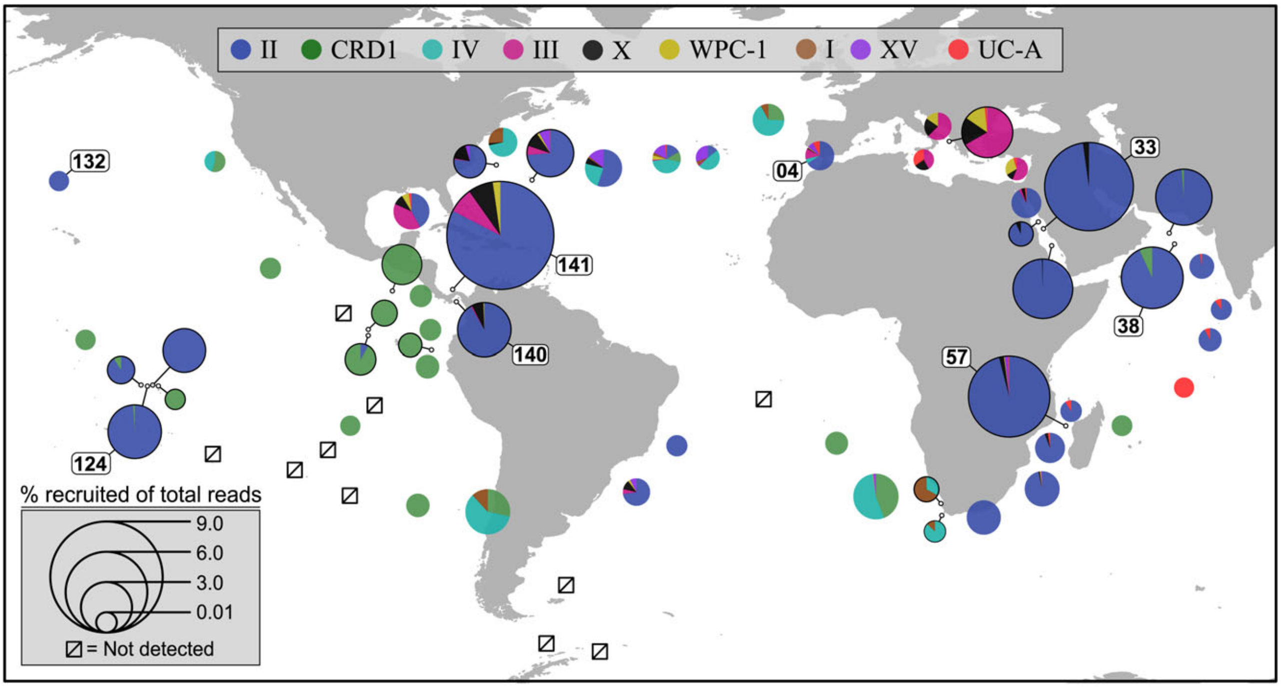

Fig. 2. Distributions of dominant clades from 65 surface ocean samples.
Pie sizes are scaled to represent percent of total sample reads that were recruited to the 31 reference genomes, serving roughly as a metric of how relatively abundant the recovered Synechococcus population was at each sample (this is entirely constrained by the reference genomes). Pies are coloured by proportion of reads recruited to each of the dominant clades in each sample. 'Not detected' refers to sites where no genome's detection surpassed the defined 50% threshold. Select samples have their corresponding numbers to TARA designations. Supporting Information Fig. S2 is the same map with all sample identifiers shown, which are also presented in Supporting Information Table S2 with further sample data.

Fig. 4A–C) – a trend that has been noted in abundant marine organisms beyond *Prochlorococcus* as well, such as SAR11(Biller *et al*., 2014; Giovanonni, 2017). General characteristics of genomic streamlining are typically considered to include reduced genome size, increased coding density, fewer non-essential genes and fewer pseudogenes (Giovannoni *et al*., 2014). The general thinking regarding *Synechococcus* based on the previously available reference genomes is that it has not undergone genomic streamlining relative to its and *Prochlorococcus*'s shared ancestor (Dufresne *et al*., 2005; Scanlan *et al*., 2009). A while not necessarily meeting all current criteria of 'genomic streamlining', the five new clade II-S reference genomes contributed here (N32, N5, UW86, N26 and N19) – that are the most representative of the recovered *in situ* populations (Fig. 1) – do demonstrate some interesting genomic characteristics. They have significantly smaller genomes (~2.14 ± 0.05 Mbps; mean ± 1 SD; Figs 1B and 4B) than the remaining 26 (~2.55 ± 0.23 Mbps; *p* = 5e-9; Welch 2-tail test), hold significantly fewer genes (2469 ± 45 vs 2786 ± 245; *p* = 1e-6) and have a moderately higher coding density (~0.908 ± 0.005 vs ~0.898 ± 0.021; p = 0.04; see Supporting Information Table S1 for full genome summaries). In contrast to 'streamlined'-*Prochlorococcus* as compared with 'non-streamlined' *Prochlorococcus* (MIT9303 and MIT9313, Fig. 4), they demonstrate significantly *higher* GC contents (~60.67 ± 0.16%) as compared with the remaining 26 reference genomes (~57.36 ± 2.93%; p = 5e-6; Figs 1B and 4C) and actually possess a slightly greater percentage of pseudogenes, although not statistically significant (2.2 ± 0.45% as compared with 1.7 ± 0.8%; Supporting Information Table S1). This suggests an environmentally abundant *Synechococcus* genomic lineage has taken an evolutionary path convergent to the majority of known *Prochlorococcus* with regard to reduction in genome size, but divergent with regard to GC content and some other properties typically associated with genomic streamlining (Fig. 4B and C). It is also worth noting that the previously available Clade II reference genome WH8109 is also relatively smaller (Fig. 1B), but its distribution and abundance mirror that of CC9605 rather than being highly prominent like clade II-S (see Figs 1D and 2, and Supporting Information Fig. S6).

Phylogenomic analysis indicates that *Synechococcus* strain RCC307 (coloured black in Fig. 4A–C) diverged before the *Synechococcus* speciation event that led to *Prochlorococcus* (Fig. 4A). Isolate RCC307 possesses a relatively smaller genome compared with the majority of known *Synechococcus* (Fig. 4B) and a relatively higher GC content (Fig. 4C) – characteristics more similar to subclade II-S of clade II (blue in Fig. 4B and C) than to

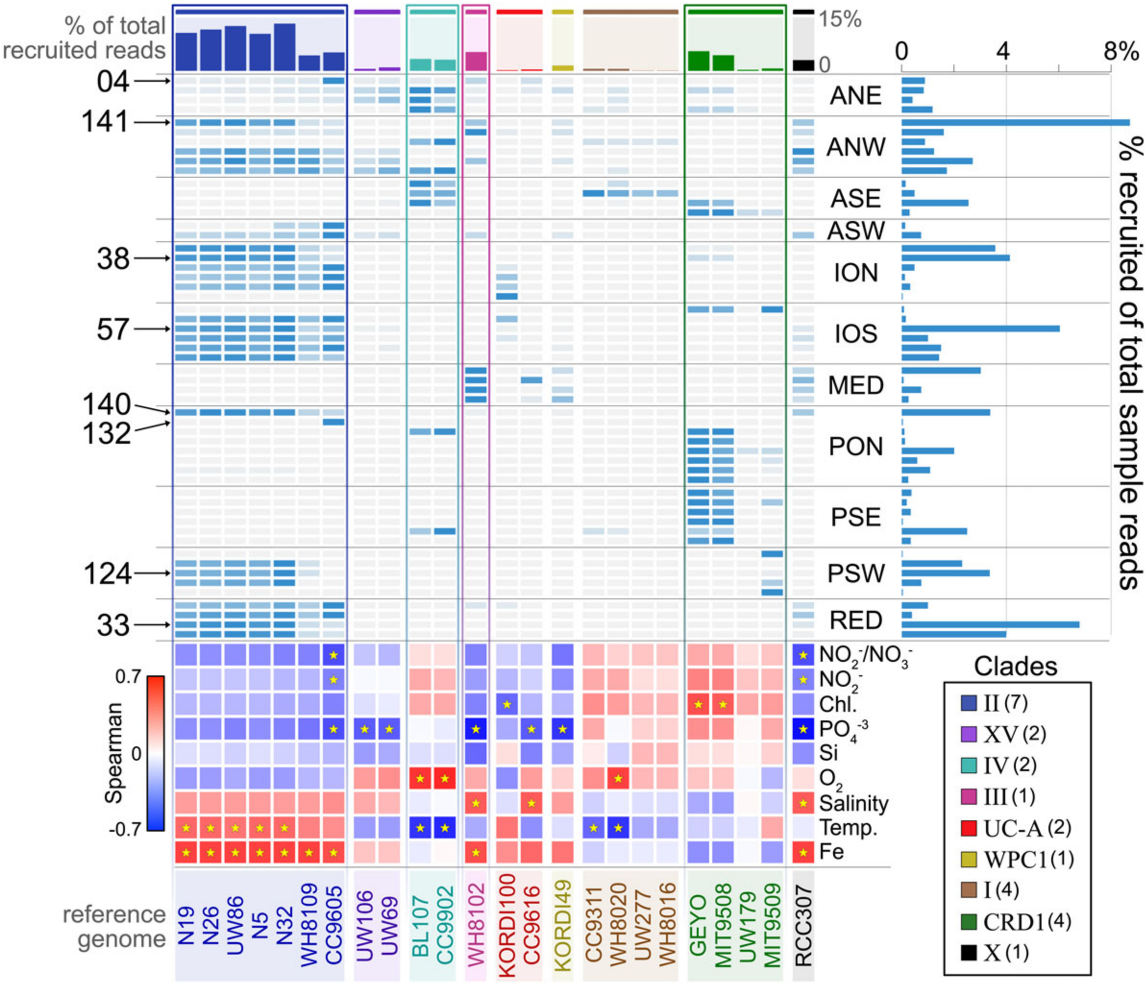

Fig. 3. Individual distributions and correlations of most* representative genomes across select surface samples.
Starting from the bottom, each individual genome is labelled followed by: a Spearman correlation heatmap of genome relative abundance to environmental data (with stars denoting significance at an adj. *P* value of ≤ 0.05); the relative distributions of each genome across select surface samples (normalized by row – darker blue indicates that genome recruited a greater proportion of reads from that sample); at the top is the overall relative abundance of that reference genome in the entire 97 metagenomes (same as Fig. 1D, but valuable here for interpreting the relative distributions across samples); and lastly, the horizontal barplot at right depicts the percent of reads recruited of total reads for the corresponding row (sample) – as a metric of relative abundance of the Synechococcus population recovered by the 31 reference genomes for each specified sample. ANE = Atlantic northeast; ANW = Atlantic northwest, ASE = Atlantic southeast, ION = Indian Ocean north, IOS = Indian Ocean south, MED = Mediterranean Sea, PON = Pacific Ocean north, PSE = Pacific southeast, PSW = Pacific southwest and RED = Red Sea. *Genomes were included if any member of their clade was responsible for at least 1% of total reads recruited across the entire dataset.

the rest of the *Synechococcus* isolates (red in Fig. 4B and C). This begs the question of whether: (i) the ancestral *Synechococcus* population that diverged into RCC307 and the current clades of *Synechococcus* (and eventually *Prochlorococcus*) possessed a larger genome, and RCC307, this subclade of clade II, and the majority of known *Prochlorococcus* isolates subsequently independently reduced in genome size while the majority of *Synechococcus* and clade LLIV of *Prochlorococcus* did not (Fig. 4A and B); or (ii) the ancestral population possessed a smaller genome, like that of the current RCC307 and the new subclade of clade II, while the majority of other *Synechococcus* along with *Prochlorococcus* clade LLIV expanded their genome sizes and the remaining known *Prochlorococcus* subsequently further reduced their genomes (Fig. 4A and B). Regardless of their origins, these newly accessed *Synechococcus* genomic lineages that are represented by isolates N32, N5, UW86, N26 and N19 are highly abundant in the surface ocean that the TARA dataset covers, and they demonstrate an evolutionary history of genomic reduction *without* a concurrent decrease in GC content (Fig. 4A and C).

One of the possible contributing differences between *Prochlorococcus* and *Synechococcus* regarding GC content may lie in genes involved in mutation repair. It has been shown that mutations in Bacteria are biased towards

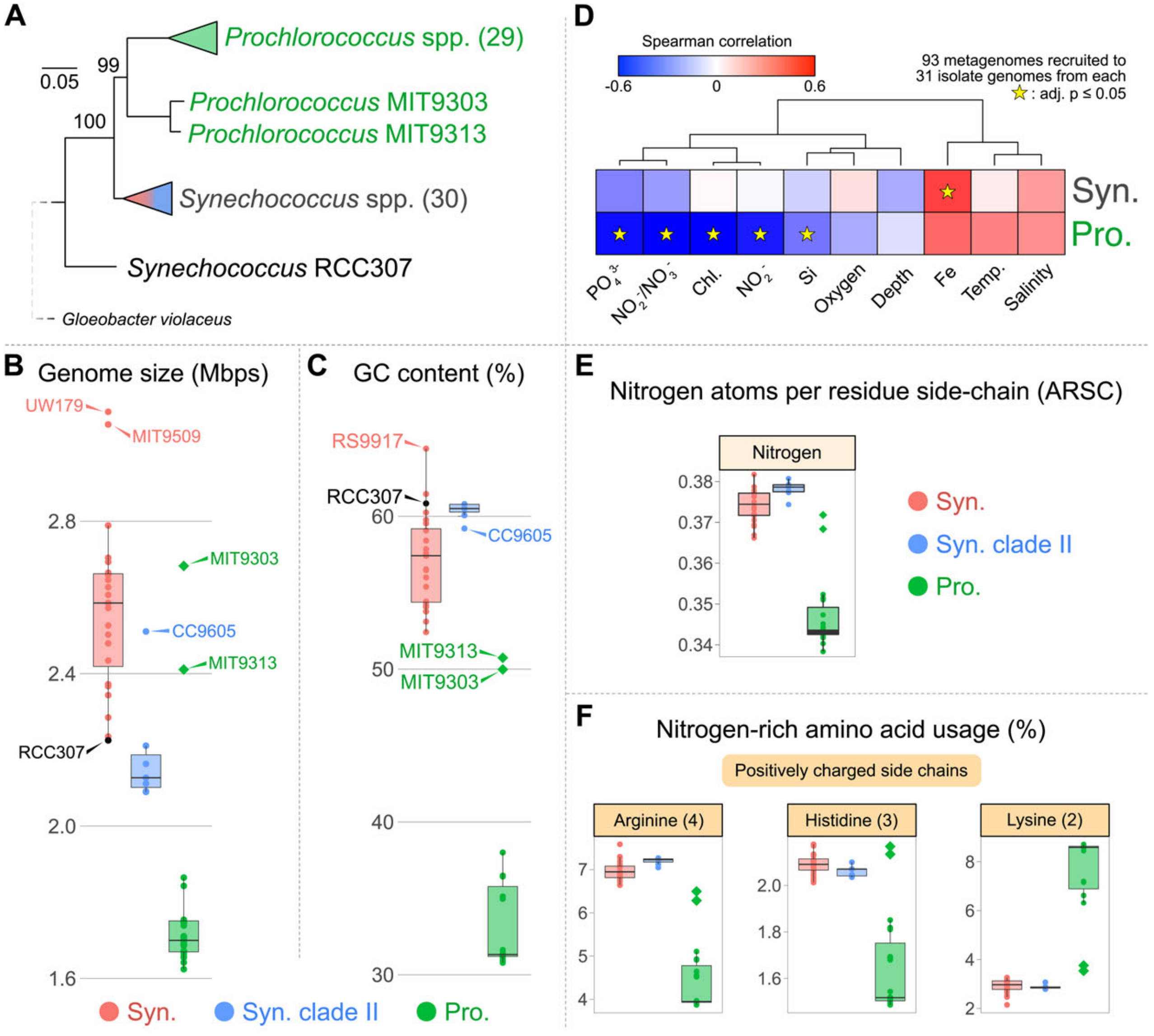

Fig. 4. Contrasting marine Synechococcus and Prochlorococcus.

A. Maximum likelihood phylogenomic tree of 57 shared single-copy ribosomal proteins comprising an alignment of 5866 amino acid positions (100BS).

B and C. Panes (B) and (C) represent genome sizes and GC contents with points coloured as described in the bottom legend, both genera include 31 reference isolate genomes.

D. Spearman correlations of normalized read recruitment to the 31 reference genomes of each genus to environmental parameters from the 93 incorporated TARA samples (Prochlorococcus recruitment data from O. Delmont and Eren 2018).

E. Nitrogen atoms per residue side-chain of groups (all residue atoms presented in Supporting Information Fig. S8), and (F) percentage of positively charged nitrogen-rich amino acid usage (frequencies of all amino acids presented in Supporting Information Fig. S9).

increasing AT content (Hershberg and Petrov, 2010). In the case of *Prochlorococcus*, all streamlined members are missing one or both of *ada* and *mutY* – genes involved in the repair of G:C to A:T transversions (Biller *et al*., 2014; Rocap *et al*., 2003), whereas all currently available *Synechococcus* reference genomes, including the members of subclade II-S, possess copies of both of these genes (Supporting Information Table S5). An additional, interrelated factor may be nitrogen (N). A reduction in GC content has been directly shown to be associated/intertwined with N-cost minimization (Lightfield *et al*., 2011; Grzymski and Dussaq, 2012), and it has been put forward that N availability may have been one of the key selective factors that helped drive the speciation of *Prochlorococcus* from *Synechococcus* (Dufresne *et al*., 2005; Grzymski and Dussaq, 2012). In considering the same, 93 TARA Oceans metagenomic samples recruited to 31 *Prochlorococcus* isolate genomes (data from Delmont and Eren 2018), *Prochlorococcus* abundance is significantly negatively correlated with N, while *Synechococcus* abundance is not (Fig. 4D). In terms of atoms of N, an A–T base pair requires 7, while a G–C base pair requires 8. Therefore, all other things being equal, an organism with a lower GC content in general would be better suited for a lower-N-quota lifestyle – selection for this lower-N-quota lifestyle by the majority of *Prochlorococcus* may have been aided by the lack of G:C to A:T transversion repair genes. Moreover, codons rich in GC tend to code for amino acids with greater N-usage (Lightfield *et al*., 2011).

In considering N-content in terms of atoms per residue side-chain [ARSC; following (Grzymski and Dussaq, 2012; Baudouin-cornu *et al*., 2001)], clade II-S genomes do not show a decrease in N ARSC relative to the rest of the *Synechococcus* reference genomes, but rather are on the higher end (Fig. 4E; all residue atoms presented in Supporting Information Fig. S8). Contrasting amino acid usage between *Synechococcus* and *Prochlorococcus* revealed that of the six amino acids that contain additional N atoms in their side-chains, *Synechococcus* members – including the five subclade II-S genomes – encode for proportionately more per genome than *Prochlorococcus* in all except for lysine and asparagine (positively charged side-chains presented in Fig. 4F; frequencies of all amino acids presented in Supporting Information Fig. S9 and Supporting Information Table S6). When considering the characteristics of these amino acids, it is also fitting that *Prochlorococcus* would use proportionately more lysine than *Synechococcus* (Fig. 4F). The three amino acids with positively charged side-chains are arginine (with 4 N atoms), histidine (3 N atoms) and lysine (2 N atoms). As *Prochlorococcus* would still need amino acids with positively charged side-chains, it might be anticipated it would favour lysine at a cost of fewer N atoms. Fig. 4E and F show that the trend of streamlined *Prochlorococcus*, wherein the corresponding reduction in GC content is accompanied with N-cost minimization, is mirrored in the genome-reduced *Synechococcus* – wherein maintained higher GC content is accompanied with maintained N-usage (Fig. 4E and F). This suggests the evolutionary pressures that led to genome reduction in *Synechococcus* are independent of, or at least not as largely driven by, N availability.

It is possible the ecological niche-space this reduced-genome *Synechococcus* lineage holds may be less N-limited than the vast areas where the majority of *Prochlorococcus* dominates, but that the large effective population size of *Synechococcus* still provides an evolutionary benefit to overall metabolic efficiency and reduced genome size. This would still be distinct from the most common known other driving force thought to be behind the genome reduction that many symbionts, parasites and commensals can be subject to – genetic drift due to small effective population sizes. However, the characteristic of retained higher GC content alongside genome reduction has been reported in an alphaproteobacterial symbiont (McCutcheon *et al*., 2009). While there is a mutational bias towards increasing AT content in Bacteria (Hershberg and Petrov, 2010), it has also demonstrated that there of course must exist as-yet-not-understood selective pressures in favour of maintaining various levels of GC content for different organisms in different environments (Hildebrand *et al*., 2010). Clearly there is much more to be investigated in genome reduction in general, as well as in this new subclade of *Synechococcus.*

Conclusion

Descendants of the ancestral genomic lineage that led to today's *Synechococcus* and *Prochlorococcus* populations have successfully explored many avenues of evolution while establishing these genera as the most abundant photoautotrophs on the planet. This exploration has spanned varying light-harvesting systems, cell and genome sizes, elemental compositions, and more, both within each genus and between them. To the best of our knowledge, the environmentally abundant genomic lineages represented by the new *Synechococcus* subclade II-S reference genomes are the first evidence of another successfully exploited evolutionary path – one of genome reduction without a concurrent reduction in GC content. This adds further context to our developing understanding of the ways in which genomic reduction can occur, and the availability of these isolates for experimentation opens up a new area of investigation into microbial evolution in the marine environment.

Experimental procedures

Isolate source and genome sequencing

Details of the newly sequenced isolates from the current study, including isolation source, maintenance, accessions of assembled genomes and input reads, are presented in Supporting Information Table S1. Paired-end, 2 × 150 bps sequencing was performed by the Joint Genome Institute (JGI) on the Illumina HiSeq platform. Isolates are currently maintained and openly available to any members of the scientific community for experimentation.

Genomic sequence processing, assembly and curation

All programs were run with default settings unless otherwise noted. Starting from the quality-filtered and decontaminated reads provided by JGI, reads for each genome were kmer-depth normalized with bbnorm (B. Bushnell; https://sourceforge.net/projects/bbmap/files/) and assembled with SPAdes (v3.11.1)(Bankevich *et al*., 2012) with the – meta flag specified (as all cultures were enrichments), and – error-correction and – careful mode turned on. For manual curation and elimination of contaminating contigs, assemblies and read coverages generated via bowtie2 (Langmead and Salzberg, 2012) v2.3.4 were input into anvi'o (Eren *et al*., 2015) v3. Coding sequences were identified by Prodigal (Hyatt *et al*., 2010) v.2.6.2, and the program centrifuge (Kim *et al*., 2016) v1.0.1 was used to taxonomically classify them. Percent completeness and redundancy were estimated based on gene copies identified by hidden Markov models of conserved single-copy genes (Rinke *et al*., 2013; Campbell *et al*., 2013). The interactive framework provided by anvi'o (Eren *et al*., 2015), including clustering of contigs based on tetranucleotide frequency and coverage, and manual identification and assessment of target-cultivar assembled contigs was very clean and straightforward due to the relatively much higher coverage of the target cultivars compared with contaminants (target-cultivar coverage was consistently orders of magnitude greater).

Incorporated reference genomes

All currently available marine *Synechococcus* genomes (accessible in October of 2017) were downloaded from NCBI (NCBI Resource Coordinators 2018). To minimize cross-mapping between references to some extent, we collapsed redundant genomes at the arbitrary threshold of 98% average nucleotide identity (ANI) or greater across at least 90% of the shorter genome and retained solely the larger genome. Therefore, all incorporated genomes share less than 98% average ANI. ANI was calculated with pyani (Pritchard *et al*., 2016). All included genomes and relevant information are presented in Supporting Information Table S1.

Functional annotation, pangenomics and phylogenomics of all incorporated genomes

All incorporated genomes were processed through anvi'o (Eren *et al*., 2015) as described above. All open reading frames were functionally annotated with NCBI's Cluster of Orthologous Groups (COGs; Galperin *et al*., 2015) and Kyoto Encyclopedia of Genes and Genomes (KEGG; Kanehisa *et al*., 2016) Orthologs (KOs). All identified genes were clustered using the Markov CLuster (MCL) algorithm (Van Dongen and Abreu-Goodger, 2012) within anvi'o to generate gene clusters with the –mcl-inflation parameter set to 4. These gene clusters were used to identify the 1002 1:1 orthologs utilized in the phylogenomic tree of Fig. 1. Amino acid sequences of each individual gene of these 1:1 orthologs were aligned with muscle (Edgar, 2004), then the alignments were concatenated together and the maximum-likelihood tree was built with RAxML (Stamatakis, 2014) v8.2.0 set to 100 bootstraps.

Construction of reference library and recruitment of environmental metagenomic data

Nucleotide fasta files for the 31 incorporated reference genomes (Supporting Information Table S1) were used to generate a reference library with bowtie2(Langmead and Salzberg, 2012) v2.3.4. Metagenomic short reads from 93 samples from the TARA Oceans project and four samples from the Costa Rica Upwelling Dome, sequenced by JGI, were downloaded, quality filtered with the iu-filter-quality-minoche program within the illumina-utils package (Eren *et al*., 2013), and recruited to the reference library with bowtie2 (Langmead and Salzberg, 2012) v2.3.4 default settings. Supporting Information Table S2 contains environmental and accession information for all 97 of the incorporated samples. There was no substantive cross-recruitment from *Prochlorococcus* (see Supporting Information Note 3).

Data availability and reproducibility

Draft genomes of the 12 newly sequenced *Synechococcus* isolates have been deposited in the European Nucleotide Archive (ENA) database under study accession: PRJEB26976. A publicly available repository containing annotated bash and R scripts for performing the presented analyses, required inputs for those scripts, additional data files such as amino acid fasta files for all called genes, and all supplemental files can be found at https://figshare.com/projects/Synechococcus_Lee_et_al_2018/34388.

Acknowledgements

We would like to thank Tom O. Delmont, Chris Dupont, A. Murat Eren and Jed A. Fuhrman for their insight and expertise on many aspects of this work. We also thank the anvi'o team for providing such a rich platform to the world, and the TARA Oceans project for providing such a rich dataset to the world. Funding sources: partial from National Science Foundation (NSF) grants OCE1657757 and OCE1260490NSF; and Joint Genome Institute CSP #1118223. Part of this work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant number ACI-1548562, via JetStream through allocation TG-BCS180020 (Towns *et al*., 2014).

References

Ahlgren, N.A., and Rocap, G. (2012) Diversity and distribution of marine *Synechococcus*: multiple gene phylogenies for consensus classification and development of qPCR assays for sensitive measurement of clades in the ocean. *Front Microbiol* **3**: 1–24.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., *et al*. (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**: 455–477.

Baudouin-cornu, A.P., *et al*. (2001) Molecular evolution of protein atomic composition. *Science* **293**: 297–300.

Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., Awad, L., *et al*. (2014) Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. *Sci Data* **1**: 140034.

Campbell, J., *et al*. (2013) UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota. *Proc Natl Acad Sci USA* **110**: 5540–5545.

Chase, A., and Martiny, J.B.H. (2018) The importance of resolving biogeographic patterns of microbial microdiversity. *Microbiol Aust* **39**: 5–8. https://doi.org/10.1071/MA18003.

Coordinators, N.R. (2018) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **46**: D8–D13.

Delmont, T.O., and Eren, A.M. (2018) Linking pangenomes and metagenomes: the *Prochlorococcus* metapangenome. *PeerJ* **6**: e4320.

Dufresne, A., Garczarek, L., and Partensky, F. (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* **6**: R14.

Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P., *et al*. (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* **9**: R90.

Dvořák, P., Casamatta, D.A., Poulíčková, A., Hašler, P., Ondřej, V., and Sanges, R. (2014) *Synechococcus*: 3 billion years of global dominance. *Mol Ecol* **23**: 5538–5551.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H. G., Sogin, M.L., and Delmont, T.O. (2015) Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**: e1319.

Eren, A.M., Vineis, J.H., Morrison, H.G., and Sogin, M.L. (2013) A filtering method to generate high quality short reads using Illumina paired-end technology. *PLoS One* **8**: 1–6.

Farrant, G.K., Doré, H., Cornejo-Castillo, F.M., Partensky, F., Ratin, M., Ostrowski, M., *et al*. (2016) Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proc Natl Acad Sci USA* **113**: E3365–E3374.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincon, J., Zabala, L.L., Jiao, N., *et al*. (2013) Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci USA* **110**: 9824–9829.

Galperin, M.Y., Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* **43**: D261–D269.

Giovannoni, S.J. (2017) SAR11 bacteria: the Most abundant plankton in the oceans. *Ann Rev Mar Sci* **9**: 231–255.

Giovannoni, S.J., Cameron Thrash, J., and Temperton, B. (2014) Implications of streamlining theory for microbial ecology. *ISME J* **8**: 1–13.

Grzymski, J.J., and Dussaq, A.M. (2012) The significance of nitrogen cost minimization in proteomes of marine microorganisms. *ISME J* **6**: 71–80.

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., *et al*. (2016) Plankton networks driving carbon export in the oligotrophic ocean. *Nature* **532**: 465–470.

Hershberg, R., and Petrov, D.A. (2010) Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* **6**: e1001115.

Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010) Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet* **6**: e1001107.

Hyatt, D., Locascio, P.F., Hauser, L.J., and Uberbacher, E. C. (2010) Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics* **28**: 2223–2230.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**: D457–D462.

Kent, A.G., Baer, S.E., Mouginot, C., Huang, J.S., Larkin, A. A., Lomas, M.W., and Martiny, A.C. (2018) Parallel phylogeography of Prochlorococcus and Synechococcus. *ISME J* **13**: 430–441.

Kim, D., Song, L., Breitwieser, F., and Salzberg, S.L. (2016) Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Res* **26**: 1721–1729.

Langmead, B., and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lightfield, J., Fram, N.R., and Ely, B. (2011) Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One* **6**: e17677.

Mazard, S., Ostrowski, M., Partensky, F., and Scanlan, D.J. (2012) Multi-locus sequence analysis, taxonomic resolution and biogeography of marine *Synechococcus*. *Environ Microbiol* **14**: 372–386.

McCutcheon, J.P., McDonald, B.R., and Moran, N.A. (2009) Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. *PLoS Genet* **5**: e1000565.

Penno, S., Lindell, D., and Post, A.F. (2006) Diversity of *Synechococcus* and *Prochlorococcus* populations determined from DNA sequences of the N-regulatory gene ntcA. *Environ Microbiol* **8**: 1200–1211.

Pritchard, L., Glover, R., Humphris, S., Elphinstone, J., and Toth, I. (2016) Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* **8**: 12–24.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., *et al*. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., *et al*. (2003) Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. *Nature* **424**: 1042–1047.

Scanlan, D.J., Ostrowski, M., Mazard, S., Dufresne, A., Garczarek, L., Hess, W.R., *et al*. (2009) Ecological genomics of marine Picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249–299.

Sohm, J.A., Ahlgren, N.A., Thomson, Z.J., Williams, C., Moffett, J.W., Saito, M.A., *et al*. (2015) Co-occurring *Synechococcus* ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *ISME J* **10**: 1–13.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–1313.

Toledo, G., and Palenik, B. (1997) *Synechococcus* diversity in the California current as seen by RNA polymerase (rpoC1) gene sequences of isolated strains. *Appl Environ Microbiol* **63**: 4298–4303.

Towns, T., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., et al. (2014) XSEDE: Accelerating Scientific Discovery. *Comput Sci Eng* **16**: 62–74. https://doi.org/10.1109/MCSE.2014.80.

Van Dongen, S., and Abreu-Goodger, C. (2012) Using MCL to extract clusters from networks. *Methods Mol Biol* **804**: 281–295.

Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S., Garczarek, L., Vaulot, D., *et al*. (2008) Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ Microbiol* **10**: 147–161.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1 Genome info
Table S2 Sample information
Table S3 Genome detections*
Table S4 Genome relative abundances*, **
Table S5 Genes table
Appendix S1: Supplementary Information
Appendix S2: Figures