

ORIGINAL ARTICLE

Fundamental differences in diversity and genomic population structure between Atlantic and Pacific *Prochlorococcus*

Nadav Kashtan^{1,2}, Sara E Roggensack², Jessie W Berta-Thompson², Maor Grinberg¹, Ramunas Stepanauskas³ and Sallie W Chisholm^{2,4}

¹Department of Plant Pathology and Microbiology, The Hebrew University of Jerusalem, Rehovot, Israel;

²Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA,

USA; ³Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA and ⁴Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA

The Atlantic and Pacific Oceans represent different biogeochemical regimes in which the abundant marine cyanobacterium *Prochlorococcus* thrives. We have shown that *Prochlorococcus* populations in the Atlantic are composed of hundreds of genomically, and likely ecologically, distinct coexisting subpopulations with distinct genomic backbones. Here we ask if differences in the ecology and selection pressures between the Atlantic and Pacific are reflected in the diversity and genomic composition of their indigenous *Prochlorococcus* populations. We applied large-scale single-cell genomics and compared the cell-by-cell genomic composition of wild populations of co-occurring cells from samples from Station ALOHA off Hawaii, and from Bermuda Atlantic Time Series Station off Bermuda. We reveal fundamental differences in diversity and genomic structure of populations between the sites. The Pacific populations are more diverse than those in the Atlantic, composed of significantly more coexisting subpopulations and lacking dominant subpopulations. *Prochlorococcus* from the two sites seem to be composed of mostly non-overlapping distinct sets of subpopulations with different genomic backbones—likely reflecting different sets of ocean-specific micro-niches. Furthermore, phylogenetically closely related strains carry ocean-associated nutrient acquisition genes likely reflecting differences in major selection pressures between the oceans. This differential selection, along with geographic separation, clearly has a significant role in shaping these populations.

The ISME Journal advance online publication, 19 May 2017; doi:10.1038/ismej.2017.64

Introduction

The cyanobacterium *Prochlorococcus* is the smallest and most abundant photosynthetic cell in the upper ocean's surface layer, and contributes substantially to marine primary productivity (Partensky *et al.*, 1999; Flombaum *et al.*, 2013). *Prochlorococcus* is divided into several major clades, defined by the intergenic transcribed spacer (ITS) region between the 16S and 23S ribosomal RNA (rRNA) genes and subsequently mapped to whole genomes and other markers (Rocap *et al.*, 2002; Martiny *et al.*, 2009b; Mühling, 2012; Malmstrom *et al.*, 2013; Biller *et al.*, 2015; Shibl *et al.*, 2016). These clades represent physiologically and ecologically distinct ecotypes

that display distinctive seasonal, depth and geographical patterns (Moore *et al.*, 1995, 1998; West and Scanlan, 1999; Johnson *et al.*, 2006; Zinser *et al.*, 2007; Scanlan *et al.*, 2009; Biller *et al.*, 2015; Chandler *et al.*, 2016; Larkin *et al.*, 2016). An enormous amount of genotypic and phenotypic diversity is found within each of these major clades (Kettler *et al.*, 2007; Kent *et al.*, 2016; Larkin *et al.*, 2016). The observed *Prochlorococcus* fine-resolution diversity is not randomly distributed—it reflects abiotic and biotic selection forces and ocean-mixing regimes—both far from being well-understood (Martiny *et al.*, 2009b; Malmstrom *et al.*, 2010; Farrant *et al.*, 2016). A significant association between phylogeny and gene content has been found even at a fine resolution of diversity (Kashtan *et al.*, 2014; Kent *et al.*, 2016; Larkin *et al.*, 2016), and regional differences in both phylogenetic composition and gene content correlating with environmental variables have been observed (Coleman and Chisholm, 2010; Kent *et al.*, 2016; Larkin *et al.*, 2016).

Correspondence: N Kashtan or SW Chisholm, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.

E-mail: nadav.kashtan@mail.huji.ac.il or chisholm@mit.edu

Received 8 November 2016; revised 9 March 2017; accepted 17 March 2017

In a previous study, we reported, based on large-scale single-cell genome sequencing, that *Prochlorococcus* populations in the Atlantic Ocean are composed of at least hundreds of genomically distinct coexisting subpopulations (Kashtan *et al.*, 2014). Each of these subpopulations has a distinct ‘genomic backbone’ consisting of highly conserved (within subpopulation) alleles of the majority of core genes and a small distinct set of flexible genes associated with a particular set of core gene alleles (Kashtan *et al.*, 2014). The functions of the backbone-associated flexible genes, often observed as cassettes within genomic islands, suggest involvement in outer membrane modifications, possibly affecting phage attachment (Avrani *et al.*, 2011), recognition by grazers (Pernthaler, 2005; Strom *et al.*, 2012), cell-to-cell communication or interactions with other bacteria (Malfatti and Azam, 2010). These backbone subpopulations are estimated to have diverged at least a few million years ago (Kashtan *et al.*, 2014), suggesting ancient, stable niche partitioning. That they have different alleles of core genes associated with environmental interactions, carry a distinct set of flexible genes and differ in relative abundance profiles as the environment changes suggests strongly that they are ecologically distinct (Kashtan *et al.*, 2014; Kent *et al.*, 2016; Larkin *et al.*, 2016).

The North Atlantic Subtropical Gyre—the site of our original study (Kashtan *et al.*, 2014)—and the North Pacific Subtropical Gyre represent two different biogeochemical environments where *Prochlorococcus* are abundant (Coleman and Chisholm, 2010; Malmstrom *et al.*, 2010; Karl and Church, 2014; Bryant *et al.*, 2016). The two most well-studied sites in these ocean ecosystems are the Bermuda Atlantic Time Series (BATS) Study (Steinberg *et al.*, 2001) and the Hawaii Ocean Time-Series (HOT) Station ALOHA (Campbell *et al.*, 1997; Karl *et al.*, 2001b; Karl and Church, 2014). Both sites are oligotrophic with similar rates of primary production and carbon export, but they differ in the finer details of their physics and nutrient dynamics, as described below.

The Atlantic site, BATS, experiences stronger seasonality than HOT, displaying substantial seasonal variation in light, temperature and nutrient concentrations as a result of convective deep mixing (to ~150–200 m) during the winter months (Karl *et al.*, 2001b). These mixing events disrupt stratification in the euphotic zone and transport cold, nutrient-rich water to the surface (Steinberg *et al.*, 2001). BATS has lower inorganic phosphorus concentrations than HOT (Friedman *et al.*, 1997; Wu *et al.*, 2000; Ammerman *et al.*, 2003), but higher fluxes of dust inputs, which bring iron and other metals (Jickells *et al.*, 2005). Although seasonal changes in the hydrography and the *Prochlorococcus* population at HOT are dampened relative to BATS, relatively deep mixing (~100 m depth) does occur in winter and waters are highly stratified (the mixed layer is a few tens of meters) throughout the summer. During this summer stratification, light, temperature

and nutrient concentrations display strong gradients over the upper 200 m of the water column, and populations below the mixed layer at different depths (a few tens of meters apart) have the potential to differentiate because of prolonged exposure to different conditions. On average, because of the higher concentrations of inorganic phosphorus (Wu *et al.*, 2000; Cavender-Bares *et al.*, 2001; Steinberg *et al.*, 2001) in the Pacific site, the N:P ratios are often well below the Redfield ratio of 16:1 suggesting nitrogen limitation (Björkman *et al.*, 2000; Wu *et al.*, 2000; Karl *et al.*, 2001a; Berube *et al.*, 2016).

To better understand the differences in ecology of, and selection pressures on, *Prochlorococcus* between these two ocean habitats, we analyzed the cell-by-cell genomic composition of populations sampled from HOT in the Pacific, and compared it with the previously analyzed populations from BATS (Kashtan *et al.*, 2014). Specifically we asked: (i) are the broad genomic structure and the diversity of local populations similar between the two oceans? (ii) Do the two oceans share the same set of backbone subpopulations (Kashtan *et al.*, 2014), or are these ocean specific? (iii) Do closely related clades from different oceans carry genes that are overrepresented in one ocean and not in the other?

To this end, we applied large-scale single-cell genome sequencing (Rodrigue *et al.*, 2009; Kalisky *et al.*, 2011; Stepanauskas, 2012; Engel *et al.*, 2014; Luo, 2015) to *Prochlorococcus* cells collected in six samples: three from the Atlantic Ocean (at BATS, analyzed in our previous study; Kashtan *et al.*, 2014), and three from the Pacific (at HOT). We sorted and sequenced the ITS sequences of 2209 single cells in total (828 new, from HOT, and 1381 from our previous study at BATS; Kashtan *et al.*, 2014), and sequenced 115 nearly complete genomes (19 of them new from HOT and 96 from our previous study from BATS; Kashtan *et al.*, 2014). We then compared the diversity and the genomic composition of local populations from these samples, compared genomes of closely related cells within a single ITS cluster (98% ITS sequence identity) from the different oceans, and looked for genes that were overrepresented in the population in one ocean compared with the other.

Materials and methods

Water samples

Atlantic samples. Samples were collected from the BATS station site (approximate 5 nautical mile radius around 31° 40' N, 64° 10' W). These samples were taken during monthly time series cruises, in addition to the large sample and data collection that is routine at BATS (<http://bats.bios.edu/>), one of the best-characterized regions of the oceans (Steinberg *et al.*, 2001). Three samples were selected for analysis at three different seasons over a period of

5 months: autumn (November 2008), winter (February 2009) and spring (April 2009). A 60 m depth was chosen to ensure all samples were taken within the mixed layer (see Kashtan *et al.*, 2014 for more information).

Pacific samples. Samples were collected from the HOT site, Station ALOHA (22° 45'N, 158° 00'W, located 100 km north of Oahu, Hawaii): one winter sample from the mixed layer (at 60 m depth) and two summer samples at two different depths below the mixed layer (60 m and 100 m). Seasonality at HOT is dampened relatively to BATS, yet these samples were chosen to maximize seasonal differences at HOT. In all, 60 m depth was chosen for a comparison with the Atlantic samples. Originally, we tried to analyze three points as a depth profile in the summer stratified sample (at 5 m—within mixed layer, 60 m and 100 m—below mixed layer). However, the cells in the 5 m sample were too dim in chlorophyll fluorescence to be flow-sorted.

Samples for single-cell sorting were collected as raw seawater (2 × 1 ml per sample) with glycerol added to a concentration of 10% as a cryoprotectant, flash frozen in liquid nitrogen and stored at −80 °C. See details in Table 1.

Construction of single amplified genome (SAG) libraries Single-cell sorting and whole-genome amplification were performed at the Bigelow Laboratory Single Cell Genomics Center (<http://scgc.bigelow.org>) as previously described in (Kashtan *et al.*, 2014).

ITS rRNA screening and sequencing

ITS screen. The ITS region from Atlantic SAG libraries was amplified as previously described (Kashtan *et al.*, 2014), with the following modifications for the Pacific SAGs: each reaction contained 0.4 Units Phusion DNA Polymerase (Thermo Scientific/NEB, Ipswich, MA, USA), 2.0 µl diluted DNA, 0.25 mM each dNTP (NEB), 0.25 µM each primer, 1X HF Buffer (Thermo Scientific/NEB) and 0.25X SYBR Green. Reactions were prepared using a Bio-Tek Precision 2000 Liquid Handler (BioTek, Winooski, VT, USA). Samples were then sent for Sanger sequencing of the ITS product.

ITS rRNA population composition analysis

In this study, we focused on small-ITS *Prochlorococcus* cells—these are all ribotypes within high-light adapted and LLI/eNATL ecotypes, which all have ITS of 500–600 bp. The ITS sequences of most low-light-adapted cells are much longer than those of the small-genome high-light-adapted *Prochlorococcus* (600–1100 bp in comparison with 500–600 bp) and much more variable in length. This significantly reduced the quality of the multiple alignment and the downstream

analysis. Consequently, the low-light-adapted ITS sequences were discarded from the present study, except for the LLI/eNATL clade, sister to the high-light with smaller genomes than the other low-light-adapted cells, which could be aligned. A total of 2209 ITS sequences (1381 from the Atlantic samples and 828 from the Pacific samples) remained after the removal of partial ITS sequences and all non-NATL low-light-adapted ITS sequences. These 2209 sequences quantitatively represent the population composition of all small-ITS, small-genome *Prochlorococcus* cells in the samples. The number of sequences per sample in the Atlantic was 399, 436 and 546 sequences of the autumn, winter and spring samples, respectively. The number of sequences per sample in the Pacific was 429, 146 and 253 sequences of the winter, summer 60 m and summer 100 m samples, respectively. Sequences were multi-aligned by mafft (Katoh *et al.*, 2002) (<http://mafft.cbrc.jp/alignment/software/>), using the command line flags: 'mafft—auto—ep 0.123'. The ITS trees presented in Figure 1 (main text) were generated by Matlab with 'p-distance' and 'average' linkage. The rarefaction and rank abundance curves as well as the standard richness and diversity measures in Figure 2 were calculated using Mothur (Schloss *et al.*, 2009), based on operational taxonomic units (OTUs) with 99% ITS similarity.

Whole-genome sequencing

Second round multiple displacement amplification. Based off of the resulting ITS sequences, 115 SAGs were selected to undergo a second round of multiple displacement amplification in order to produce enough DNA to construct sequencing libraries (described in detail previously; Kashtan *et al.*, 2014). DNA was purified and prepared for whole-genome sequencing as previously described (Kashtan *et al.*, 2014). The single-cell genomes were sequenced on an Illumina GAIIX (Illumina, San Diego, CA, USA) with paired-end reads of length 200 bp (forward and reverse). Sequencing was done at the BioMicroCenter at MIT (<http://openwetware.org/wiki/BioMicroCenter>).

Genome assembly of single-cell genomes

De novo assembly. *De novo* assembly was done by clc-assembly-cell-3.1.1 (CLCbio, <http://www.clcbio.com/>). Phred quality score of Q=20 was used as a threshold (base call accuracy of 99%) of quality. Reads were considered only if at least 20% of the read was above the Q=20 threshold (CLCbio program 'quality_trim' was used with the command line flags: '-c 20 -l 0.2'). Paired-end reads were assembled assuming insert length is between 150 and 1000 bp. CLCbio program 'clc_novo_assemble' was used with the command line flags: '-q -p fb ss 150 1000'. A minimal contig size of 400 bp was used

for the 19 Pacific SAGs (this threshold was empirically chosen to enhance the quality of the assembly). The resulting assembly size of the 19 new Pacific single-cell partial genomes was 1.15 ± 0.3 Mbp (mean \pm s.d.), estimated as $\sim 70 \pm 0.18\%$ of the complete genome size. These assembly size statistics are very similar to those of the 96 Atlantic single cells (1.15 ± 0.3 Mbp (mean \pm s.d.)). More details on the assembly statistics can be seen in the full QUAST report (Supplementary File S1).

Reference-guided assembly. As we did not have a previously sequenced complete genome of any strain within the cN2 ITS rRNA cluster, a 'composite' genome was constructed to serve as a mediator for referenced-guided assembly. The composite reference genome was created by combining 12 large overlapping contigs, selected by hand, from the *de novo* assemblies of cells within the cN2-C1 cluster (according to their ITS rRNA). These contigs were selected because they had sufficiently large overlaps between each other and they covered the whole genome (determined by alignment to a few high-light-adapted complete genomes). This yielded a composite reference genome of 1 650 354 bp in length which is within the size range of other high-light-adapted genomes (Sequence can be downloaded at: Dryad (doi:10.5061/dryad.9r0p6)). Paired-end reads were assembled assuming insert length is between 150 and 1000 bp (CLCbio program 'clc_ref_assemble_long' was used with the command line flags: '-p fb ss 150 1000').

Genome annotation. Annotation of the *de novo*-assembled genomes, as well the cN2-C1 composite genome, were done on the RAST server (Aziz et al., 2008). In all, 1971 open reading frames, three rRNA genes (1 copy of 5S, 16S and 23S rRNA genes) and 37 tRNAs were identified in the cN2-C1 'composite' genome.

Phylogenetic tree construction. Phylogenetic trees in Figures 3 and 4 were generated by MEGA7 (Kumar et al., 2016). Distances were estimated using 'p-distance'. Positions with pairwise missing data were discarded from the distance calculation. Trees were un-rooted and were generated using 'Neighbor

joining' with bootstrap. For the generation of the tree in Figure 4, 10 sequences with very long branches were omitted, to allow better presentation.

Gene content analysis

Clusters of orthologous genes. Genes were classified into clusters of orthologous genes using the pipeline described in Kelly et al. (2012). Genes from previously sequenced *Prochlorococcus* high-light-adapted cells, as well as all genes from the 115 single-cell partial genomes (the *de novo* assemblies annotated by RAST) were included. Final refinement of the clusters was done manually to improve the clustering (Kashtan et al., 2014).

Detection of ocean-specific gene sets. For the purpose of our analysis, genes that 'differentiate' clades (that is, appear in one or more clades but absent from the other clades) are genes that (a) appear in cells in one or more clades; (b) absent from the other clades; (c) are not shared by the majority of cells regardless of their ITS-phylogeny; and (d) are not rare genes that were found only in one or two cells. Candidate genes were selected by the following steps:

- (1) Choose all genes that pass either (i) or (ii) criteria
 - (i) Genes that appear in at least 50% of the cells of a clade population, in at least one clade population.
 - (ii) Genes that appear in > 7 cells within at least one clade population.
- (2) Omit the following genes from the gene set found in (1):
 - (i) All genes that were found as high-light-adapted core genes (genes that appear in all the culture high-light-adapted cell genomes).
 - (ii) Genes that appear in < 3 cells in total or in > 75 cells in total.
- (3) Cluster genes according to their presence/absence in the 115 partial single-cell genomes.

Table 1 Sample details

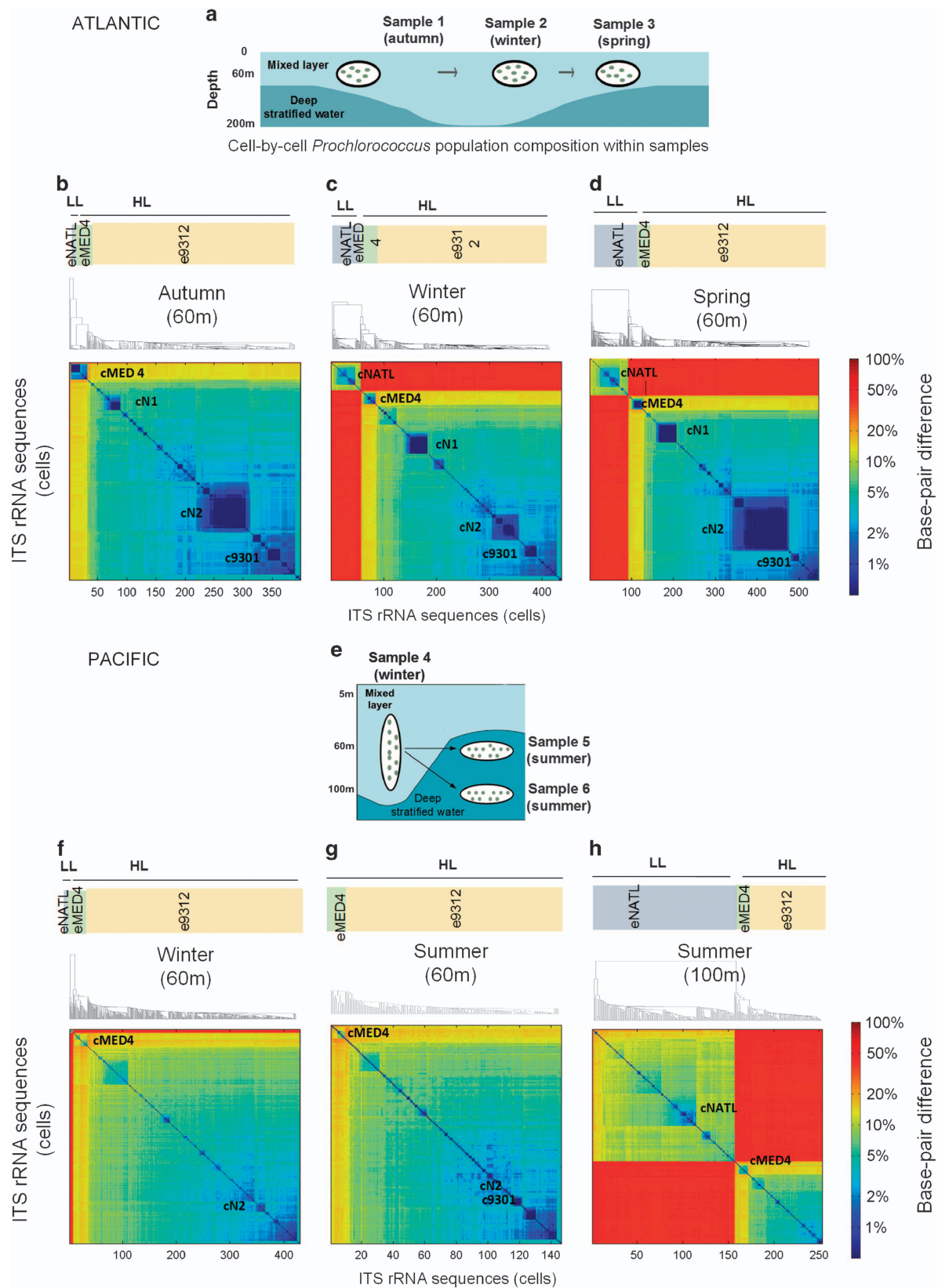
Sample	Ocean	Date	Season	Cruise	Depth	Mixed layer	No. of SAGs	<i>Prochlorococcus</i> abundance cells mL ⁻¹ (mean \pm s.e.)
1	Atlantic	8 November 2008	Autumn	BATS 241	60 m	Within	399	41 350 \pm 750
2	Atlantic	8 February 2009	Winter	BATS 243	60 m	Within	436	33 100 \pm 800
3	Atlantic	1 April 2009	Spring	BATS 245a	60 m	Within	546	33 000 \pm 1350
4	Pacific	20 January 2009	Winter	HOT 208	60 m	Within	429	200 000
5	Pacific	3 July 2009	Summer	HOT 212	60 m	Below	146	230 000
6	Pacific	3 July 2009	Summer	HOT 212	100 m	Below	253	172 500

Abbreviations: BATS, Bermuda Atlantic time series; HOT, Hawaii Ocean time series; SAG, single amplified genome.

The abundance of *Prochlorococcus* cells was determined by flow cytometry. BATS abundance is from (Kashtan et al., 2014). HOT abundance is obtained from HOT site (http://hahana.soest.hawaii.edu/hot/hot_jgofs.html).

Steps 1 and 2 yielded a set of 549 genes. The genes were then clustered using standard hierarchical clustering, using 'hamming distance' and 'complete'

linkage in Matlab (Figure 5, Table 2). Predictions of the location on the genome of the differential genes was done as described previously (Kashtan *et al.*, 2014).



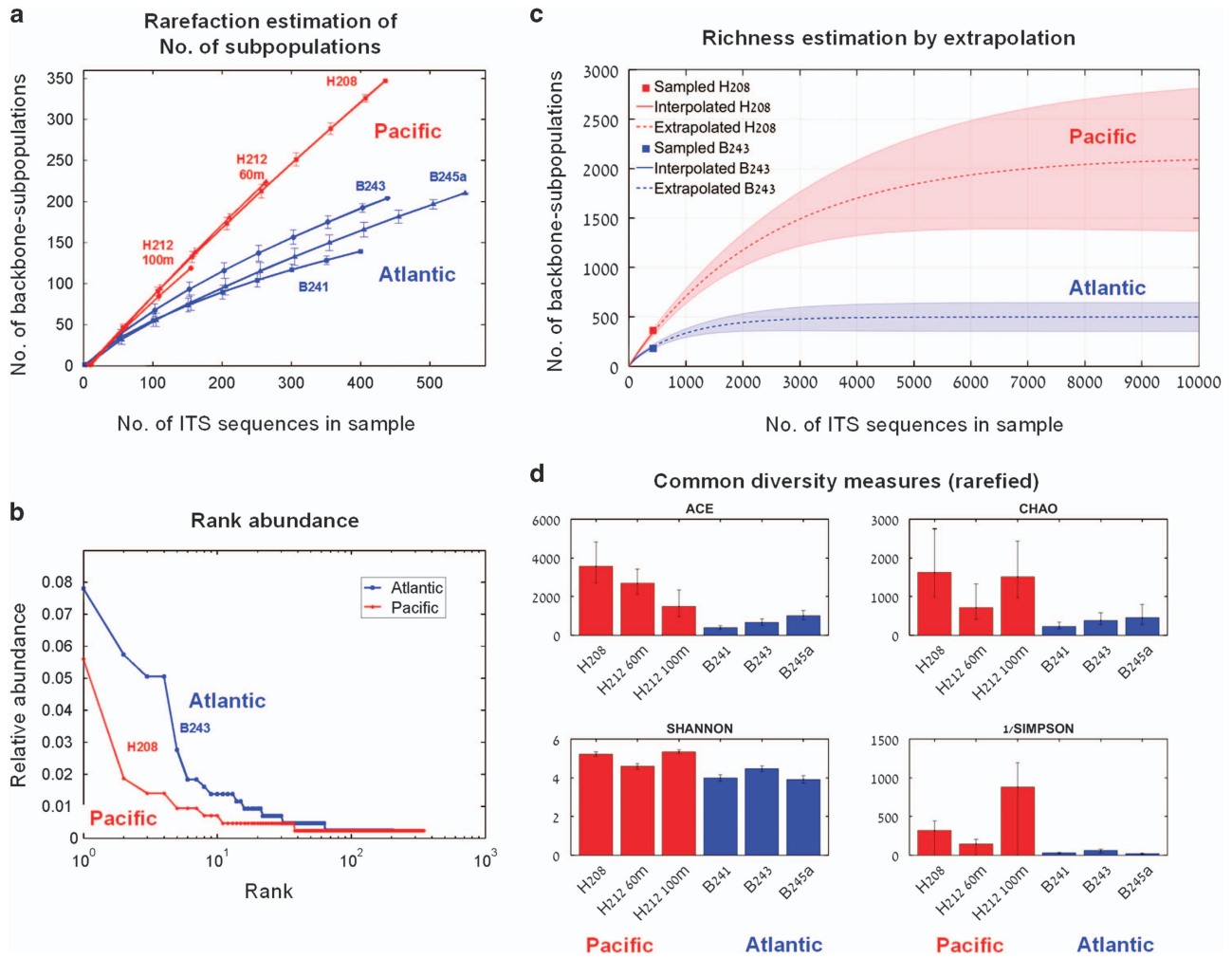


Figure 2 Comparing the diversity of the Atlantic and Pacific populations. **(a)** Richness rarefaction curves estimating the number of coexisting backbone subpopulations within samples. **(b)** Rank abundance profiles of the winter samples from both oceans. Backbone subpopulations (OTUs) in **a** and **b** were identified as those cells with 99% ITS similarity in the full set of 2209 ITS sequences. **(c)** Richness estimation of the number of coexisting backbone subpopulations (OTUs) within samples. Interpolated values are computed by rarefaction (as in **a**). Extrapolated values are the estimated number of backbone subpopulations in an augmented set of sample sizes, based on observed rank abundance profiles, as computed by EstimateS (Statistical estimation of species richness and shared species from samples. Version 9. <http://purl.oclc.org/estimates>). Shaded areas indicate the range of extrapolated population richness, which corresponds to 95% confidence interval of mean sampled richness. **a–c** indicate that the Pacific samples are composed of significantly more subpopulations, and lacks dominant ones. **(d)** Richness (ACE,Chao1) and diversity (Shannon, 1/Simpson) estimators of the *Prochlorococcus* populations within the six samples from the two oceans rarefied to 250 (except from the Pacific summer 60 m sample that was smaller and thus rarefied to 150) when defining units as sets of cells that are 99% similar in their ITS sequences. Error bars indicate s.e. (as computed by Mothur; Schloss *et al.*, 2009).

Figure 1 Cell-by-cell *Prochlorococcus* population composition in the samples selected for study. **(a)** Schematic of seasonal dynamics at the Atlantic sampling site, and sampling design, showing a typical mixed layer depth and seasonal context of our three samples. Samples were collected during three seasons at BATS Study station. Cells were collected within the mixed layer at 60 m depth in November 2008 (autumn), February 2009 (winter) and April 2009 (spring), see Materials and methods section. Winter deep mixing brings cold nutrient-rich water to the surface. **(b)** A Phylogenetic tree from pairwise genetic distances of individual cells ITS-rRNA sequences from the autumn sample (neighbor-joining trees, see Materials and methods section). The relevant sub-tree range of the ‘traditional’ ecotypes (Moore *et al.*, 1998) are marked above the tree if cells belonging to that ecotype were found, as is the division into low-light-adapted (LL) and high-light-adapted (HL) (Moore *et al.*, 1998) clades. The heatmap below the tree describes the pairwise distance matrix between ITS-rRNA sequences of individual cells from that sample. Rows and columns are arranged according to the order of leaves of the tree. The color map represents genetic distances as percentage of base substitutions per site (log-scale), such that the blue blocks identify very closely related ITS-ribotypes. Names of the largest clusters are marked in bold (for example, cN2). If a cultured representative falls within a cluster, the cluster name follows its name with a ‘c’ prefix (for example, cMED4). **(c)** Same as **b** for the winter sample. **(d)** Same as **b** for the spring sample. **(e)** Schematic of seasonal dynamics at the Pacific sampling site, and sampling design. A typical mixed layer depth profile and seasonal and depth context of our three samples. The Pacific samples were collected in January 2009 (winter) and July 2009 (summer) at Station Aloha with the HOT. The winter sample was collected within the mixed layer at 60 m depth while the summer samples were collected at two different depths 60 m and 100 m from the stratified water below the mixed layer, see Materials and methods section. **(f, g, h)** Same as **b, c, d** but for the Pacific samples.

Analysis of ocean-specific cassette gene functions. All proteins in COGs assigned to ocean-specific categories (Table 2) were given KEGG Orthology

assignments through the BlastKOALA web service of the KEGG database (<http://www.kegg.jp/blastkoala/>) (Kanehisa *et al.*, 2016). Taxid was set to 1218 (for

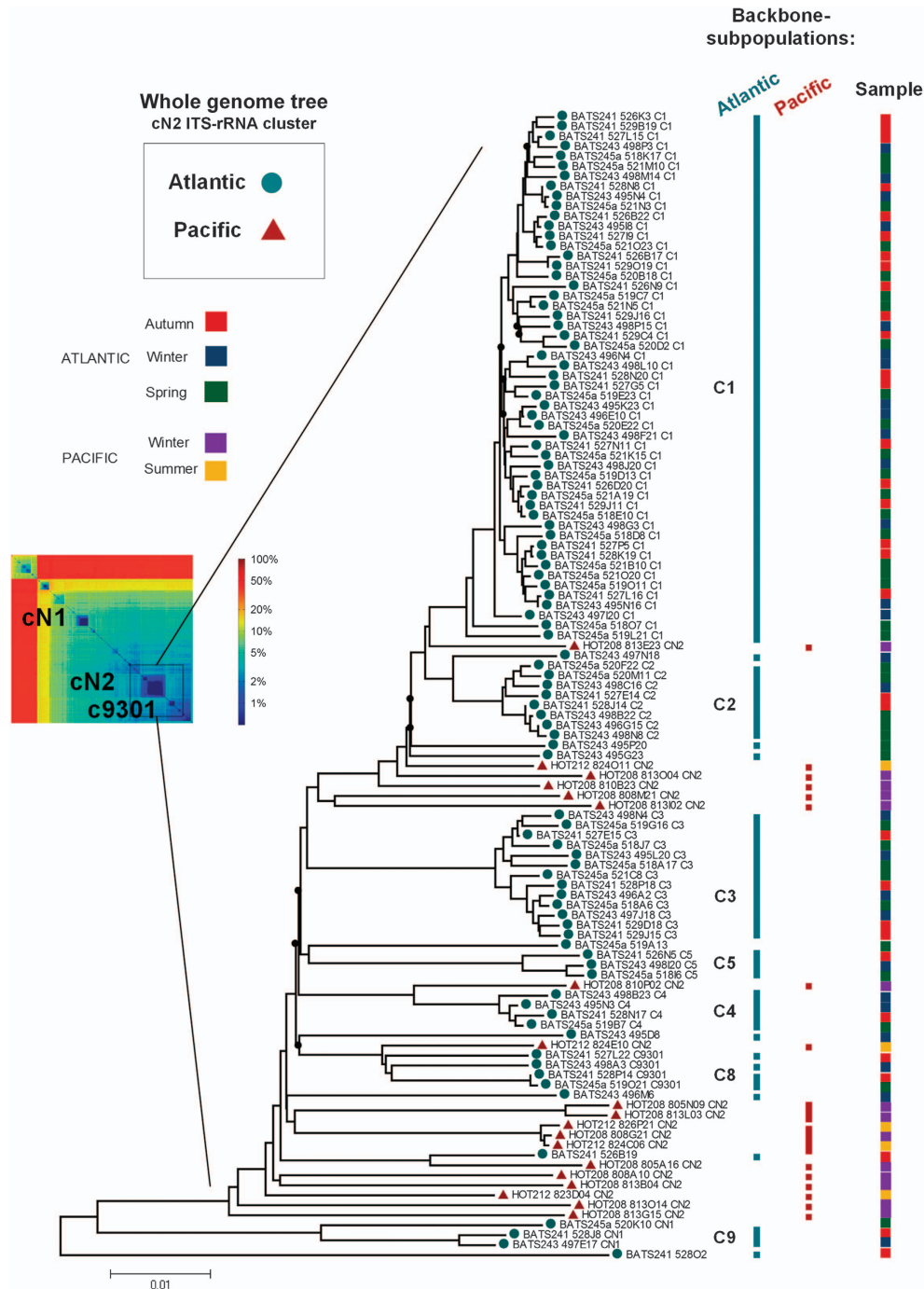


Figure 3 The genomic backbone clades within the cN2 cluster from the Atlantic and Pacific samples do not overlap. Neighbor-joining phylogenetic tree of 96 single cells from the cN2 cluster from the Atlantic samples as well as 19 single cells from the Pacific based on whole-genome sequences. The heatmap to the left describes the pairwise distance matrix between ITS-rRNA sequences of all six samples combined indicating the cN2 cluster. The colored symbols to the left of the leaf labels represent the ocean of origin of each sample (blue circles, Atlantic; red triangles, Pacific). Distance units are percent base substitutions per site (see scale bar, Materials and methods section). Bootstrap values < 80 are marked as black dots on the internal nodes. Vertical continuous bars on the right represent distinct backbone subpopulations (blue, Atlantic; red, Pacific). If a subpopulation is represented by only one cell, it is marked as a square. The cN2 clades C1 to C5, identified in Kashtan *et al.* (2014) are marked to the left of the Atlantic bars. Note the lack of subpopulations with cells from both oceans, as indicated by the gaps in the Atlantic column where the Pacific cells are present in the tree ($P < 0.001$, permutation test). Colors on the right-most column represent the sample origin.

Prochlorococcus), and two searches were performed, one using the family_eukaryotes+genus_prokaryote database and one using the species_prokaryote database (redundancy removed from the KEGG Genes database at the given taxonomic level). Summary results in most cases represent unanimous assignments across the COGs. To refine their annotations, all proteins in ocean-specific COGs were compared with a number of databases through blastp searches. An e-value cutoff of 0.0001 was used to filter the results, an inclusive limit as the main goal was information gathering across many very different databases. In most cases, only the top hit was examined. Searches were conducted either with command line BLAST 2.3.0+ or the Biopython NCBI API tool Bio.Blast.NCBIWWW qblast. The following data sets were searched: Swissprot (curated protein database), the Transporter Classification Database (organized curated transporters), Protein Data Bank (proteins with structure data), Conserved Domains Database (curated protein database), the Synechocystis PCC6803 genome with updated annotations from CyanoBase (cyanobacterium with best functional and genetic characterization) (Fujisawa *et al.*, 2017), nr, refseq and env_nr (not useful for annotation, demonstrated additional instances of these proteins in nature). All programs used the newest versions accessible in January 2017. Targeted bidirectional searches were also performed comparing Atlantic-specific COGs to phosphonate/phosphite uptake and utilization genes in MIT9301, ptxABC/phnCDE, ptxD, phnY and phnZ (P9301_RS15030, P9301_RS15035, P9301_RS15040, P9301_RS15045, P9301_RS15055, P9301_RS15060); only phnZ matched (e-values 1e-39-1e-41).

Results

Features of the sampling sites and rationale for sample selection

Our sampling scheme was designed to capture some of the characteristic environmental gradients in the two oceans. To capture the strong seasonal changes in the North Atlantic, we analyzed samples collected from the same depth within the mixed layer (60 m) at three different times in the year: autumn (before winter mixing), winter (during mixing) and spring (shortly after winter mixing) (Figure 1a). These populations represent cells that have experienced significant changes in their environment over tens of generations, causing shifts in relative abundance of traditional coarse-grained ecotypes (ITS-defined, HLI, LLIV etc.) (Moore *et al.*, 1998; Rocap *et al.*, 2002) across these samples (Malmstrom *et al.*, 2010; Kashtan *et al.*, 2014). To capture changes because of water stratification, as well as other seasonal changes to some extent, we collected three water samples at HOT: one winter sample from the mixed layer (60 m) and two summer samples at two different depths below the mixed layer (60 and 100 m) (Figure 1e). All six samples (three from BATS and three from HOT)

were taken within 8 months of each other (November 2008 to July 2009; see Table 1). Note that the two winter samples were taken at almost the same time of the year in winter 2009 (Table 1), within the mixed layer, and from the same depth (60 m).

Inferring population composition from ITS sequences of hundreds of single cells within a sample

From our Atlantic study (Kashtan *et al.*, 2014), as well as more recent work by Kent *et al.* (2016), ITS sequences serve as a good proxy for whole-genome content and phylogeny in *Prochlorococcus* even at a fine level of resolution. ITS-ribotype clusters at the 99% similarity level coincide, in most cases, with distinct genomic backbones (Kashtan *et al.*, 2014). The composition of a local population of co-occurring cells could thus be inferred from the ITS sequences of *Prochlorococcus* cells within a given sample. Through fluorescence-activated cell sorting, DNA amplification and sequencing of hundreds of single cells within each sample, we revealed the presence of finely resolved clusters within the broadly defined ecotypes (Figures 1b–d and f–h). In each ocean, the populations were composed of a large number of ‘nearly identical’ ITS clusters (>99% similar) that likely correspond to distinct genomic backbones. These can be seen in the heat-maps in Figure 1 as the blue blocks on the diagonal. As can be easily seen by eye, we observed clear differences in the population structure between the two oceans: the Pacific samples lack very abundant ‘nearly identical’ ITS clusters (no large blue blocks

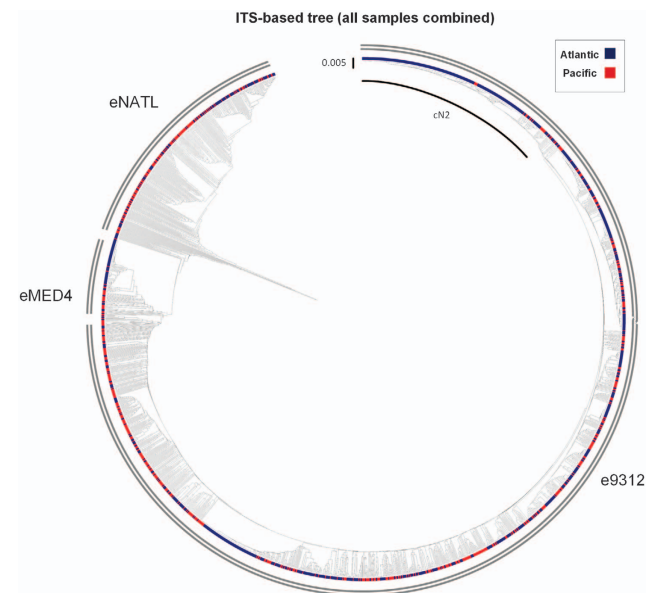


Figure 4 The global phylogenetic pattern observed in *Prochlorococcus* ITS-rRNA sequences from all single cells sampled at both sites. Colors represent sample origin: HOT (red leaf coloring), BATS (blue leaf coloring). Neighbor-joining phylogenetic tree is based on multiple alignment of 2209 ITS sequences. The tree was generated by MEGA7 with p-distance (Kumar *et al.*, 2016). Scale bar represent 0.005 substitutions per bp.

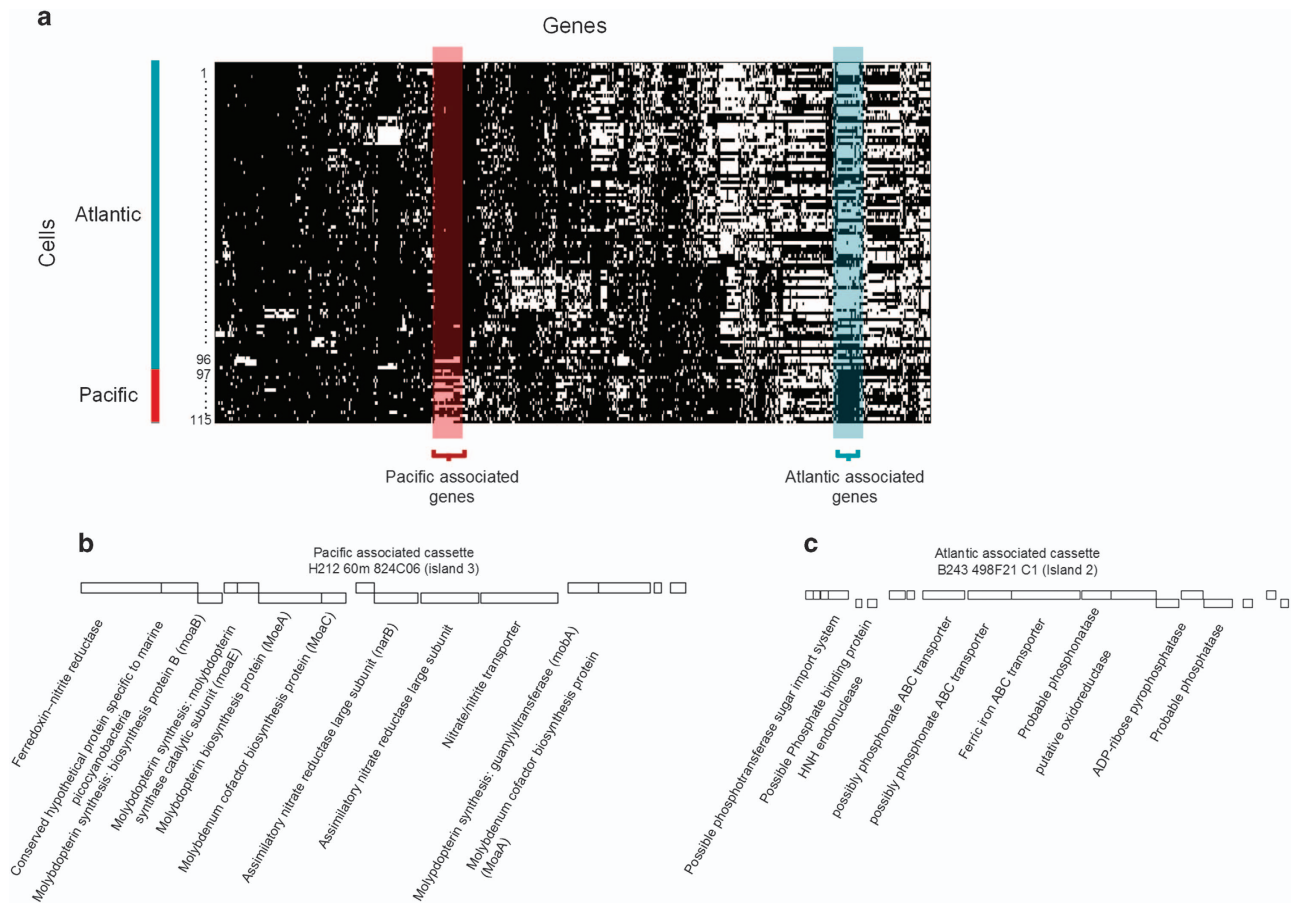


Figure 5 Differential gene sets distinguishing the Atlantic and Pacific sites. **(a)** Matrix representation: each column is a gene and each row represents a single-cell. White/black dots represent the presence/absence of a gene in the partial genome of a single-cell, respectively. Single cells are grouped according to the sample location. The order of the single cells within each of the Atlantic and Pacific groups is according to the leaf order of the whole-genome phylogenetic tree. Note that as the genomes are partial genomes, the absence of a gene may be due to the partiality rather than true absence. Genes were clustered using standard hierarchical clustering. The orders of genes (that is, of columns in the matrix) do not reflect location on the genome; the order is determined by the clustering (that is, the similarity between the existence/absence pattern of genes). Bracketed color-shaded sets of genes show genes differentially abundant in a pattern associated with a particular ocean. **(b)** The Pacific-specific genes fall into three cassettes. The majority of these genes (13 out of 22) appear as a single cassette—found in most of the HOT cells that we sequenced (whole-genome). This cassette is related to nitrate/nitrite assimilation. It appears in different cells on two different islands (island 2.2 and island 3). **(c)** The Atlantic-specific genes form a single gene cassette, which contains possible phosphonate transport and utilization genes. This cassette appears in most cN2 Atlantic cells in our sample. It is located, however, within one of three different islands (island 1, 2 or 5) in different cells.

on the diagonal Figures 1f–h) as opposed to the Atlantic (Figures 1b–d), and the number of ‘nearly identical’ ITS clusters is much higher in the Pacific (it has many more small clusters, Figures 1f–h)—as described below.

Diversity within Pacific populations is much higher than in the Atlantic

We estimated the number of coexisting backbone subpopulations based on ITS clusters at 99% identity (equivalent to OTU at 99% ITS identity) in our samples through rarefaction analysis. In the following text, we refer to ‘populations’ as the whole set of coexisting backbone subpopulations in a given site (locality). The Pacific populations had a significantly higher diversity: they were composed of a much larger number of coexisting backbone

subpopulations (Figure 2a). At the level we sampled—hundreds of cells per sample—the rarefaction curve of each of the three Pacific samples was far from reaching an asymptote. However, an extrapolation of the sample rarefaction curves based on the sampled rank abundance distributions (Figure 2b) (Gotelli and Colwell, 2001; Colwell *et al.*, 2012) corroborates the difference of population richness between the Pacific and Atlantic sites (Figure 2c). Based on these extrapolations, the Pacific populations could be made up of over a thousand coexisting subpopulations, whereas the number of Atlantic subpopulations was only in the hundreds. The extrapolated rarefaction curves for H208 and B243 samples (the two winter samples, the most ‘fair’ comparison) approach significantly different richness values, implying that the total populations have significantly different richness

Table 2 Gene cassettes that are found to be ocean-associated within the cN2 cells

<i>Ocean</i>	<i>COG ID</i>	<i>Description</i>	<i>Position</i>
Atlantic cassette 1	8440	Hypothetical protein	Island 1 or Island 2 or Island 5 ^a
	8544	Hypothetical protein	
	4300	Possible phosphotransferase sugar import system; Lactose/Cellobiose specific	
	5925	Hypothetical protein	
	201713	Hypothetical protein	
	3041	HAD-superfamily hydrolase; subfamily IA; variant 3, possible sugar phosphatase	
	2097	Hypothetical protein	
	200104	HNH endonuclease	
	101610	Possible phosphate-binding protein	
	8967	Hypothetical protein	
	2055	ABC transporter permease, possibly phosphonate. AfuB Fe(III) transport family	
	1270	Possible phosphonate	
	3042	ABC transporter periplasmic substrate-binding protein, possibly phosphonate, AfuA Fe(III) transport family	
	104240	Hypothetical protein	
	8113	Putative oxidoreductase associated with phosphonate genes	
	101354	ADP-ribose pyrophosphatase (EC 3.6.1.13, Nudix)	
Pacific cassette 1	13152	Nitrate/nitrite transporter	Island 2.2 or Island 3
	20126	Molybdopterin synthesis: guanylyltransferase (moaA)	
	51985	Hypothetical protein	
	17161	Conserved hypothetical nitrate reductase-associated protein	
	29798	Molybdopterin synthesis: cyclic pyranopterin monophosphate synthase (moaC)	
	50277	Molybdopterin synthesis: molybdopterin synthase sulfur carrier subunit (moaD)	
	30998	Molybdopterin synthesis: biosynthesis protein B (moaB)	
	15904	Hypothetical protein	
	25024	Molybdopterin synthesis: GTP 3',8-cyclase (moaA)	
	34565	Assimilatory nitrate reductase large subunit (EC:1.7.99.4) (narB)	
	43381	Ferredoxin-nitrite reductase (EC 1.7.7.1)	
	9750	Possible homeobox domain	
		Conserved hypothetical protein specific to marine picocyanobacteria	
	40686	Molybdopterin synthesis: molybdotransferase (moaA)	
	26607	Molybdopterin synthesis: molybdopterin synthase catalytic subunit (moaE)	
Pacific cassette 2	21021	Diadenosine tetraphosphatase and related serine/threonine protein phosphatases	Island 2.2 or Island 3 or Island 5
	28325	Predicted ATPase specific for cyanobacteria/signal recognition particle GTPase	
	42039	Diadenosine tetraphosphatase and related serine/threonine protein phosphatases	
Pacific cassette 3 ^b	19523	Hydrogenase/urease nickel incorporation protein (hypA)	Island 2.2 or Island 3 or Island 5
	30352	ABC transporter; periplasmic substrate-binding protein possibly urea carboxylase-related, or nitrate, sulfonate, bicarbonate, taurine import	
	60426	Putative rieske (2Fe-2 S) family protein, possible ring hydroxylating dioxygenase	
	50117	Hydrogenase/urease nickel incorporation protein (hypB)	
	59708	ABC transporter, permease protein	

^aIslands names (2.2, 3, 5 etc. are defined as in Kashtan *et al.*, 2014). ^bPacific cassette 3 genes appear also in some cN1 and c9301 cells in the Atlantic (but not within cN2 Atlantic cells).

measures. Richness at the Atlantic sample B243 was estimated as 500 OTUs (95% confidence interval, (355–645)), whereas richness at the Pacific sample H208 was estimated to be much higher: 2090 OTUs (95% confidence interval, (1370–2810)) at the extrapolated sample size of 10 000 (Figure 2c). A different approach for estimating the richness in these two samples yielded very similar values (Supplementary Figure S1). Another observed difference between the two oceans samples is that the Pacific populations lack dominant subpopulations, as can be seen in the rank abundance curves (Figure 2b) and in the absence of large dark blue squares on the diagonals in Figures 1f–h. These significant differences in diversity were also observed based on standard

richness and diversity measures when all samples were rarified to identical size (Figure 2d, Supplementary Figure S2) (Gihring *et al.*, 2012; Chandler *et al.*, 2016).

The Atlantic and Pacific samples showed almost completely distinct sets of genomic backbone subpopulations

Of the total 1014 ITS clusters (at 99% ITS similarity, which corresponds to backbone subpopulations or OTUs) in our six samples only 75 (~7% of all clusters) were detected in both oceans. This is significantly lower than the number expected as determined by a permutation test, where cells were

randomly shuffled between samples (154 shared clusters after shuffling, ~15% of all clusters, $P=6.2\text{E-}29$), indicating that the small number of shared ITS clusters is unlikely explained by chance only. As may be expected, the fraction of shared ITS clusters between samples of *different* oceans is significantly lower than the fraction shared between samples of the *same* ocean (7% between oceans in comparison with 14–18% within the Atlantic). We note, however, that even the fraction of ITS clusters shared *within* each ocean is still lower than expected by chance ($P < 0.0001$, permutation test). Thus, because of the immense diversity observed within these samples, our sampling of even hundreds of single cells per sample was not deep enough; deeper sampling is required to estimate the number (or the relative number) of shared genomic backbone clades between oceans.

To further understand the genomic differences between closely related cells from the two oceans, we sequenced the partial genomes of a total of 115 single cells. For our Atlantic study (Kashtan *et al.*, 2014), we selected 90 cells of the largest ‘nearly identical’ ITS cluster (cN2; 98% ITS similarity, Figures 1b–d) for sequencing, as well as 6 cells from two other clusters (cN1 and c9301). To compare these cells with closely related cells from the Pacific, we sequenced the partial genomes of 19 single cells from the cN2 ITS cluster (Figures 1f–h)—14 from the 60 m winter sample and 5 from the 60 m summer sample. Cells within the cN2 cluster were far less abundant in the Pacific relative to the Atlantic, where it was the most abundant cluster (compare Figures 1b–d and f–h): In the Atlantic, cN2 cells were $23.5 \pm 1.6\%$, $11.5 \pm 2.2\%$ and $24.8 \pm 1.9\%$ of the analyzed populations in autumn, winter and spring samples, respectively (Kashtan *et al.*, 2014) (Figures 1b–d), whereas in the Pacific there were 14 cN2 cells (out of 429 cells, ~3.2% of the population) in the winter sample, 5 cells (out of 146, ~3.4% of the population) in the summer 60 m sample and none in the summer 100 m sample (Figures 1f–h). It is noteworthy that the cN2 cells are members of the HLII ecotype of *Prochlorococcus*, which is the most abundant ecotype at both of these sites.

As in our previous study on the Atlantic samples (Kashtan *et al.*, 2014), we used a reference-guided assembly by mapping the reads to a composite genome of the cN2 cells (Kashtan *et al.*, 2014). We analyzed between-cell variation in the recovered partial genomes, by estimating the average genome-wide bp similarity between any two genomes, consider only the positions recovered for both along the whole genome. Based on these genome-wide pair-similarities spanning the entire genome (1 650 354 bp), we generated a phylogenetic tree and examined the genome-wide tree structure of these closely related single cells. Remarkably, we found that the genomes of these 115 closely related cells fall into ocean-specific clades (Figure 3, Supplementary Figure S3). When clustered by 99%

whole-genome similarity, these Atlantic and Pacific cells do not share a single backbone-clade, as opposed to 4.3 ± 1.2 shared clades expected by random shuffling of these single cells between samples (permutation test, $P < 0.0001$). By contrast, samples from the same ocean all share backbone subpopulations (permutation test, $P > 0.1$). Thus, cells belonging to the cN2 ITS-rRNA cluster from each ocean seem to be composed of clades consisting mostly of different backbone subpopulations (Figure 3). These ocean-specific clades are spread along the phylogenetic tree of cN2 cells, rather than splitting the tree into two deep ocean-related branches (Figure 3). The same picture can be observed in the global phylogenetic tree that is based on all the ITS sequences (not restricted to cN2) from all six samples (Figure 4). It can be readily seen by eye that there are no major single-colored sub-clades from the same ocean, but rather the differentiation between oceans is close to the leaves of the tree.

Ocean-specific genes within the cN2 clade

We next asked whether closely related strains from different oceans carry different sets of flexible genes. For that we used *de novo* assemblies to capture regions not present in the reference assemblies. We previously reported that distinct coexisting subpopulations carry small sets of distinct genes—typically in the form of cassettes within genomic islands (Kashtan *et al.*, 2014). We used a clustering analysis to analyze the gene content of the 115 single-cell partial genomes and found that there are several groups of genes that are associated with one of the oceans and not the other (Figure 5, Table 2). Interestingly, these genes are not backbone subpopulation specific but rather found in different closely related backbone subpopulations. The only genes that appear to be overrepresented in the Atlantic and under-represented in the Pacific single cells appear on a gene cassette within a genomic island that contains possible phosphonate transporters and utilization genes, distinct from a previously characterized *Prochlorococcus* phosphonate operon, (Martínez *et al.*, 2012) (Figure 5, Table 2 Atlantic cassette 1). This cassette is found on almost all cN2 single cells from the Atlantic samples, although it belongs to at least five distinct backbone subpopulations. On the other hand, most cells from the Pacific carry a gene cassette with nitrogen acquisition genes (Moore *et al.*, 2002), including a nitrate assimilation gene cassette (similar to the cassette described in detail in Berube *et al.*, 2015) (Figure 5, Table 2 Pacific cassette 1). This cassette was observed only in one single cell in the Atlantic that belonged to the cN1 clade and none of the 90 cN2 single cells (Berube *et al.*, 2015). The same gene order and genomic location of the nitrate assimilation gene cassette was shared in all the Pacific cells and the nitrate assimilation gene cassette observed in the HLII clade genomes examined by Berube *et al.*

(2015). An ABC transporter and a pair of genes for the insertion of nickel into metalloenzymes (urease or hydrogenase) were found in some of the Pacific cN2 cells and none of the Atlantic cN2 cells (Table 2 Pacific cassette 3).

Discussion

There is a fundamental difference in the diversity of, and the genomic structure within, populations of *Prochlorococcus* at HOT and BATS sites. One possible explanation for the observed difference in genomic backbone subpopulation-level *Prochlorococcus* diversity between HOT and BATS is the stronger seasonality at the Atlantic site (Giovannoni and Vergin, 2012). Pronounced seasonal changes could lead to shorter exclusion times and thus to a smaller number of coexisting subpopulations (Barton *et al.*, 2010). Moreover, seasonal changes can lead to a transient rapid increase in the abundance of specific subpopulations, resulting in more pronounced abundance dynamics and, possibly, temporal appearance of dominant subpopulations as we observed (Kashtan *et al.*, 2014) in the Atlantic samples (Barton *et al.*, 2010). Another factor that may explain the differences in diversity is that *Prochlorococcus* cell abundance per milliliter in the Pacific samples is about five times higher than in the Atlantic (Table 1). This could contribute to the higher diversity in the Pacific, if larger carrying capacity may support a greater number of ecologically differentiated subpopulations. It also means that while the relative abundance of the most abundant clades in the Pacific is smaller than the Atlantic, the absolute number of cells ml^{-1} of the most abundant clades in the Pacific is higher than the most abundant clades in the Atlantic (Figure 2b). Correlation between latitude and *Prochlorococcus* diversity has recently been reported (Larkin *et al.*, 2016) as predicted by modeling (Barton *et al.*, 2010). It may be that the differences in latitude—as HOT (22°) is located at a lower latitude than BATS (31°)—account for part of the increase in diversity, but we suggest that the large difference is mostly due to differences in the ecology of these two ocean habitats. In addition, other factors including temporal and spatial heterogeneity and water mixing could also be having a role (Ottesen *et al.*, 2014; Bryant *et al.*, 2016; Soccodato *et al.*, 2016).

The observation of sets of genes that are ocean specific, with almost no relation to phylogeny, shows clearly that there are selection pressures to keep the genes for nitrate acquisition in the Pacific, and genes related to acquisition of phosphorus in the Atlantic, as shown in earlier work (Martiny *et al.*, 2009a; Coleman and Chisholm, 2010; Berube *et al.*, 2015, 2016). It is puzzling that these ocean-specific traits seem to be almost independent of fine-scale phylogeny. What evolutionary scenario could

explain the diversity pattern where Pacific lineages carry the nitrate acquisition gene cassette with a striking degree of similarity in the order, phylogeny and location of these genes at a very specific location (end of Island 3) on the genome? Berube *et al.* (2015) suggested it is evidence for an early horizontal transfer event before divergence of these fine-resolution backbone clades and not independent horizontal gene transfer events. As all the closely related cN2 cells from BATS lack these genes, one explanation is that the last common ancestor of the cN2-Pacific clade acquired the nitrate assimilation gene cassette. Yet, the existence of a single cN1-Atlantic clade genome containing the nitrate assimilation gene cassette with the same gene order and genomic location suggests an even earlier acquisition followed by the loss of these genes as the clades and backbone subpopulations diverged, that is, through many independent events of gene loss (García-Fernández *et al.*, 2004; Kettler *et al.*, 2007; Partensky and Garczarek, 2010; Lee and Marx, 2012; Sun and Blanchard, 2014).

The phylogeny of the partial single-cell genomes indicates that the Pacific and Atlantic lineages went through different evolutionary trajectories that led to the many extant backbone subpopulations. One possible scenario could have been that the cN2-Pacific clade and cN2 Atlantic clade diverged long ago, forming two distinct ocean-specific clades. Instead it seems that there are many independent branches that diverged into ocean-specific backbone clades (with distinct genomic backbones). This provides evidence for the mixed evolutionary histories of the two ocean populations, and that this mixing is only observed at millions of years of evolution, as indicated by the independent branches of ocean-related backbone clades on the tree (Figures 3 and 4).

The biogeography of *Prochlorococcus* emerges from selection forces and the ocean-mixing regimes that govern cell dispersal (Denman and Gargett, 1983; Kashtan *et al.*, 2014). Populations can be thought of as well-mixed over large oceanic water parcels (10 km diameters) on ecologically relevant time scales (Okubo, 1971; Kashtan *et al.*, 2014), and as dispersing over large ocean provinces, through turbulence and ocean currents, within weeks to months (Doblin and van Sebille, 2016). Mixing of water masses between the Pacific and Atlantic Oceans is largely dependent on global ocean circulation, and is much slower (Martiny *et al.*, 2009b).

Evidence for the pronounced biogeography among *Prochlorococcus* populations has been growing steadily as the discovery of its coexisting, well-defined coarse-grained ecotypes (Moore *et al.*, 1998). Johnson *et al.* (2006) reported distinct patterns, at coarse scale resolution, along a longitudinal gradient in the Atlantic Ocean, and Martiny *et al.* (2009b) revealed that biogeographical resolution as a function of different nutrient regimes was dependent on the cut-offs used for diversity measures. As genomic

resolution became possible, different patterns emerged at finer scales of resolution (Coleman *et al.*, 2006; Martiny *et al.*, 2006, 2009b; Follows *et al.*, 2007; Rusch *et al.*, 2007; Kent *et al.*, 2016; Larkin *et al.*, 2016). Here we show that this biogeography is even more pronounced at extremely fine-scale diversity. It is quite remarkable that the two oceans seem to contain largely distinct sets of genomic backbone clades. As these clades are estimated to have diverged at least a few million years ago (Kashtan *et al.*, 2014), our results suggest that the populations that occupy the two oceans today have been going through largely separate evolutionary paths for a few million years.

Conclusions

The Atlantic and Pacific Oceans represent not only geographically separated habitats of the globally abundant cyanobacterium *Prochlorococcus*, but also environments with different ecologies and selection pressures. *Prochlorococcus* populations in the Pacific have much higher diversity, and lack dominant subpopulations, compared with those in the Atlantic. We suggest this is due at least in part to the weaker seasonality in the Pacific. There seems to be little overlap in the subpopulation composition between the two oceans, as if each ocean is a home for ocean-specific strains. The fine-scale phylogeny is disrupted by ocean-specific genes that confer selection advantages with respect to nitrate acquisition in the Pacific and phosphorous acquisition in the Atlantic. The exact history of gene gain and loss throughout *Prochlorococcus* evolution within closely related cells is still not clear. Future studies of extensive single-cell genome sequencing on smaller geographical scales will provide more insights into the evolution, population mixing and selection pressures that shape global *Prochlorococcus* diversity.

Genomic data have been deposited in NCBI GenBank under accession numbers SRX2559063 - SRX2559065, MW000000000-MWPA00000000, MWPC00000000-MWPG00000000. Additional data files have been deposited to figshare (https://figshare.com/projects/Supplementary_data_-_Kashtan_et_al_Fundamental_differences_in_diversity_and_genomic_population_structure_between_Atlantic_and_Pacific_Prochlorococcus/19477).

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Paul M Berube for helpful comments on the manuscript. We thank the Bermuda Atlantic time-series

Study and the Hawaii Ocean Time-Series (HOT) for sample collection and the BioMicroCenter facility at MIT for their contributions to the generation of genomic data. We thank Huiming Ding for his contribution to the generation of Clusters of Orthologous Genes database. NK acknowledges the Rothschild Foundation (Yad Hanadiv) and the NOAA 'Climate and Global Change' Postdoctoral Research Fellowships. This work was supported in part by grants to SWC from the National Science Foundation (NSF) Evolutionary Biology Section and Biological Oceanography Section, the NSF Center for Microbial Oceanography Research and Education (C-MORE), the Gordon and Betty Moore Foundation Marine Microbiology Initiative, the Simons Foundation (SCOPE award ID 329108, and LIFE Award ID: 337262) and to RS from the NSF Biological Oceanography Section.

References

- Ammerman JW, Hood RR, Case DA, Cotner JB. (2003). Phosphorus deficiency in the Atlantic: an emerging paradigm in oceanography. *Eos, Transac Am Geophys Union* **84**: 165–170.
- Avrani S, Wurtzel O, Sharon I, Sorek R, Lindell D. (2011). Genomic island variability facilitates *Prochlorococcus*-virus coexistence. *Nature* **474**: 604–608.
- Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. *BMC genomics* **9**: 75.
- Barton AD, Dutkiewicz S, Flierl G, Bragg J, Follows MJ. (2010). Patterns of diversity in marine phytoplankton. *Science* **327**: 1509–1511.
- Berube PM, Biller SJ, Kent AG, Berta-Thompson JW, Roggensack SE, Roache-Johnson KH *et al.* (2015). Physiology and evolution of nitrate acquisition in *Prochlorococcus*. *ISME J* **9**: 1195–1207.
- Berube PM, Coe A, Roggensack SE, Chisholm SW. (2016). Temporal dynamics of *Prochlorococcus* cells with the potential for nitrate assimilation in the subtropical Atlantic and Pacific oceans. *Limnol Oceanogr* **61**: 482–495.
- Biller SJ, Berube PM, Lindell D, Chisholm SW. (2015). *Prochlorococcus*: the structure and function of collective diversity. *Nat Rev Microbiol* **13**: 13–27.
- Björkman K, Thomson-Bulldis AL, Karl DM. (2000). Phosphorus dynamics in the North Pacific subtropical gyre. *Aquat Microb Ecol* **22**: 185–198.
- Bryant JA, Aylward FO, Eppley JM, Karl DM, Church MJ, DeLong EF. (2016). Wind and sunlight shape microbial diversity in surface waters of the North Pacific Subtropical Gyre. *ISME J* **10**: 1308–1322.
- Campbell L, Liu H, Nolla HA, Vaulot D. (1997). Annual variability of phytoplankton and bacteria in the subtropical North Pacific Ocean at Station ALOHA during the 1991–1994 ENSO event. *Deep-Sea Res (Pt 1 Oceanogr Res Papers)* **44**: 167–192.
- Cavender-Bares KK, Rinaldo A, Chisholm SW. (2001). Microbial size spectra from natural and artificial nutrient regimes. *Limnol Oceanogr* **46**: 778–789.
- Chandler JW, Lin Y, Gainer PJ, Post AF, Johnson ZI, Zinser ER. (2016). Variable but persistent coexistence of *Prochlorococcus* ecotypes along temperature gradients in the ocean's surface mixed layer. *Environ Microbiol Rep* **8**: 272–284.

- Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, DeLong EF *et al.* (2006). Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* **311**: 1768–1770.
- Coleman ML, Chisholm SW. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. *Proc Natl Acad Sci USA* **107**: 18634–18639.
- Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao CX, Chazdon RL *et al.* (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J Plant Ecol* **5**: 3–21.
- Denman K, Gargett A. (1983). Time and space scales of vertical mixing and advection of phytoplankton in the upper ocean. *Oceanography* **28**: 801–815.
- Doblin MA, van Sebille E. (2016). Drift in ocean currents impacts intergenerational microbial exposure to temperature. *Proc Natl Acad Sci USA* **113**: 5700–5705.
- Engel P, Stepanauskas R, Moran NA. (2014). Hidden diversity in honey bee gut symbionts detected by single-cell genomics. *PLoS Genet* **10**: e1004596.
- Farrant GK, Doré H, Cornejo-Castillo FM, 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 JL, Gordillo RA, Rincón J, Zabala LL, 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.
- Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. (2007). Emergent biogeography of microbial communities in a model ocean. *Science* **315**: 1843–1846.
- Friedman N, Rouso I, Sheves M, Fu X, Bressler S, Druckmann S *et al.* (1997). Time-resolved titrations of ASP-85 in bacteriorhodopsin: the multicomponent kinetic mechanism. *Biochemistry* **36**: 11369–11380.
- Fujisawa T, Narikawa R, Maeda S-i, Watanabe S, Kanesaki Y, Kobayashi K *et al.* (2017). CyanoBase: a large-scale update on its 20th anniversary. *Nucleic Acids Res* **45**: D551–D554.
- García-Fernández JM, de Marsac NT, Diez J. (2004). Streamlined regulation and gene loss as adaptive mechanisms in *Prochlorococcus* for optimized nitrogen utilization in oligotrophic environments. *Microbiol Mol Biol Rev* **68**: 630–638.
- Gihring TM, Green SJ, Schadt CW. (2012). Massively parallel rRNA gene sequencing exacerbates the potential for biased community diversity comparisons due to variable library sizes. *Environ Microbiol* **14**: 285–290.
- Giovannoni SJ, Vergin KL. (2012). Seasonality in ocean microbial communities. *Science* **335**: 671–676.
- Gotelli NJ, Colwell RK. (2001). Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* **4**: 379–391.
- Jickells TD, An ZS, Andersen KK, Baker AR, Bergametti G, Brooks N *et al.* (2005). Global iron connections between desert dust, ocean biogeochemistry, and climate. *Science* **308**: 67–71.
- Johnson ZI, Zinser ER, Coe A, McNulty NP, Malcolm E, SW S *et al.* (2006). Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**: 1737–1740.
- Kalisky T, Blainey P, Quake SR. (2011). Genomic analysis at the single-cell level. *Annu Rev Genet* **45**: 431–445.
- Kanehisa M, Sato Y, Morishima K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* **428**: 726–731.
- Karl DM, Björkman KM, Dore JE, Fujieki L, Hebel DV, Houlihan T *et al.* (2001a). Ecological nitrogen-to-phosphorus stoichiometry at station ALOHA. *Deep Sea Res Pt II Topical Stud Oceanogr* **48**: 1529–1566.
- Karl DM, Dore JE, Lukas R, Michaels AF, Bates NR, Knap A. (2001b). Building the long-term picture. *Oceanography* **14**: 6.
- Karl DM, Church MJ. (2014). Microbial oceanography and the Hawaii Ocean time-series programme. *Nat Rev Microbiol* **12**: 699–713.
- Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A *et al.* (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* **344**: 416–420.
- Katoh K, Misawa K, Kuma Ki, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066.
- Kelly L, Huang KH, Ding H, Chisholm SW. (2012). ProPortal: a resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Res* **40**: D632–D640.
- Kent AG, Dupont CL, Yooseph S, Martiny AC. (2016). Global biogeography of *Prochlorococcus* genome diversity in the surface ocean. *ISME J* **10**: 1856–1865.
- Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S *et al.* (2007). Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**: e231.
- Kumar S, Stecher G, Tamura K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**: 1870–1874.
- Larkin AA, Blinebry SK, Howes C, Lin Y, Loftus SE, Schmaus CA *et al.* (2016). Niche partitioning and biogeography of high light adapted *Prochlorococcus* across taxonomic ranks in the North Pacific. *ISME J* **10**: 1555–1567.
- Lee M-C, Marx CJ. (2012). Repeated, selection-driven genome reduction of accessory genes in experimental populations. *PLoS Genet* **8**: e1002651.
- Luo H. (2015). The use of evolutionary approaches to understand single cell genomes. *Front Microbiol* **6**: 174.
- Malfatti F, Azam F. (2010). Atomic force microscopy reveals microscale networks and possible symbioses among pelagic marine bacteria. *Aquat Microb Ecol* **58**: 1.
- Malmstrom RR, Coe A, Kettler GC, Martiny AC, Frias-Lopez J, Zinser ER *et al.* (2010). Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J* **4**: 1252–1264.
- Malmstrom RR, Rodrigue S, Huang KH, Kelly L, Kern SE, Thompson A *et al.* (2013). Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. *ISME J* **7**: 184–198.
- Martínez A, Osburne MS, Sharma AK, DeLong EF, Chisholm SW. (2012). Phosphite utilization by the marine picocyanobacterium *Prochlorococcus* MIT9301. *Environ Microbiol* **14**: 1363–1377.

- Martiny AC, Kathuria S, Berube PM. (2009a). Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci* **106**: 10787–10792.
- Martiny AC, Tai AP, Veneziano D, Primeau F, Chisholm SW. (2009b). Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. *Environ Microbiol* **11**: 823–832.
- Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL et al. (2006). Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102–112.
- Moore LR, Goericke R, Chisholm SW. (1995). Comparative physiology of *Synechococcus* and *Prochlorococcus* - influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar Ecol Progr Ser* **116**: 259–275.
- Moore LR, Rocap G, Chisholm SW. (1998). Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**: 464–467.
- Moore LR, Post AF, Rocap G, Chisholm SW. (2002). Utilization of different nitrogen sources by the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Limnol Oceanogr* **47**: 989–996.
- Mühling M. (2012). On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. *Environ Microbiol* **14**: 567–579.
- Okubo A. (1971). Oceanic diffusion diagrams. *Deep Sea Res Oceanogr Abst* **18**: 789–802.
- Ottesen EA, Young CR, Gifford SM, Eppley JM, Marin R, Schuster SC et al. (2014). Multispecies diel transcriptional oscillations in open ocean heterotrophic bacterial assemblages. *Science* **345**: 207–212.
- Partensky F, Hess WR, Vaulot D. (1999). *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* **63**: 106–127.
- Partensky F, Garczarek L. (2010). *Prochlorococcus*: advantages and limits of minimalism. *Annu Rev Mar Sci* **2**: 305–331.
- Pernthaler J. (2005). Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* **3**: 537–546.
- Rocap G, Distel DL, Waterbury JB, Chisholm SW. (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes using 16 S-23 S rDNA Internal Transcribed Spacer (ITS) sequences. *Appl Env Microbiol* **68**: 1180–1191.
- Rodrigue S, Malmstrom RR, Berlin AM, Birren BW, Henn MR, Chisholm SW. (2009). Whole genome amplification and de novo assembly of single bacterial cells. *PLoS ONE* **4**: e6864.
- Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S et al. (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**: e77.
- Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR et al. (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249–299.
- Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB et al. (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.
- Shibl AA, Haroon MF, Ngugi DK, Thompson LR, Stingl U. (2016). Distribution of *Prochlorococcus* ecotypes in the Red Sea basin based on analyses of rpoC1 sequences. *Front Marine Sci* **3**: 104.
- Soccodato A, d'Ovidio F, Lévy M, Jahn O, Follows MJ, De Monte S. (2016). Estimating planktonic diversity through spatial dominance patterns in a model ocean. *Marine Genomics* **29**: 9–17.
- Steinberg DK, Carlson CA, Bates NR, Johnson RJ, Michaels AF, Knap AH. (2001). Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. *Deep Sea Res Pt II: Topical Stud Oceanogr* **48**: 1405–1447.
- Stepanaukas R. (2012). Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* **15**: 613–620.
- Strom SL, Brahamsha B, Fredrickson KA, Apple JK, Rodríguez AG. (2012). A giant cell surface protein in *Synechococcus* WH8102 inhibits feeding by a dinoflagellate predator. *Environ Microbiol* **14**: 807–816.
- Sun Z, Blanchard JL. (2014). Strong genome-wide selection early in the evolution of *Prochlorococcus* resulted in a reduced genome through the loss of a large number of small effect genes. *PLoS ONE* **9**: e88837.
- West NJ, Scanlan DJ. (1999). Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. *Appl Environ Microbiol* **65**: 2585–2591.
- Wu J, Sunda W, Boyle EA, Karl DM. (2000). Phosphate depletion in the western North Atlantic Ocean. *Science* **289**: 759–762.
- Zinser ER, Johnson ZI, Coe A, Karaca E, Veneziano D, Chisholm SW. (2007). Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol Oceanogr* **52**: 2205.

Supplementary Information accompanies this paper on The ISME Journal website (<http://www.nature.com/ismej>)