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

Nadav Kashtan[1,2]*, Sara E Roggensack[2], Jessie W Berta-Thompson[2], Maor Grinberg[1], Ramunas Stepanauskas[3] & Sallie W Chisholm[2,4]*

**Affiliations:**

[1]Department of Plant Pathology and Microbiology, The Hebrew University of Jerusalem, P.O. Box 12, Rehovot 76100, Israel

[2]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, Massachusetts 02139, USA

[3]Bigelow Laboratory for Ocean Sciences, East Boothbay, Maine 04544, USA

[4] Department of Biology, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, Massachusetts 02139, USA.

*Correspondence to: nadav.kashtan@mail.huji.ac.il or chisholm@mit.edu

**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 BATS Station off Bermuda. We reveal fundamental differences in diversity and the 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 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 plays a significant role in shaping these populations.**

## 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 (Flombaum et al 2013, Partensky et al 1999). *Prochlorococcus* is divided into several major clades, defined by the Intergenic Transcribed Spacer (ITS) region between the 16S and 23S rRNA genes and subsequently mapped to whole genomes and other markers (Biller et al 2015, Malmstrom et al 2013, Martiny et al 2009b, Mühling 2012, Rocap et al 2002, Shibl et al 2016). These clades represent physiologically and ecologically distinct ecotypes that display distinctive seasonal, depth, and geographical patterns (Biller et al 2015, Chandler et al 2016, Johnson et al 2006, Larkin et al 2016, Moore et al 1995, Moore et al 1998, Scanlan et al 2009, West and Scanlan 1999, Zinser et al 2007). An enormous amount of genotypic and phenotypic diversity is found within each of these major clades (Kent et al 2016, Kettler et al 2007, 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 (Farrant et al 2016, Malmstrom et al 2010, Martiny et al 2009b). 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).

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 (Bryant et al 2016, Coleman and Chisholm 2010, Karl and Church 2014, Malmstrom et al 2010). The two most well-studied sites in these ocean ecosystems are the Bermuda Atlantic Time-series Study (BATS) (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 ~150-200m) 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 (Ammerman et al 2003, Friedman et al 1997, Wu et al 2000), but higher fluxes of dust inputs, which bring iron and other metals (Jickells et al 2005). While seasonal changes in the hydrography and the *Prochlorococcus* population at HOT are dampened relative to BATS, relatively deep mixing (~100m 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 200m of the water column, and populations below the mixed layer at different depths (a few tens of meters apart) have the potential to differentiate due to prolonged exposure to different conditions. On average, due to the higher concentrations of inorganic phosphorus (Cavender-Bares et al 2001, Steinberg et al 2001, Wu et

al 2000) in the Pacific site, the N:P ratios are often well below the Redfield ratio of 16:1 suggesting nitrogen limitation (Berube et al 2016, Björkman et al 2000, Karl et al 2001a, Wu et al 2000).

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 (Engel et al 2014, Kalisky et al 2011, Luo 2015, Rodrigue et al 2009, Stepanauskas 2012) 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 over-represented in the population in one ocean compared to the other.

## Materials and Methods

*Water samples*

Atlantic samples: Samples were collected from the Bermuda Atlantic Time Series (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 (Nov 2008), Winter (Feb 2009) and Spring (April 2009). A 60m depth was chosen to ensure all samples were taken within the mixed layer; see Kashtan et al (Kashtan et al 2014) for more information.

Pacific samples: Samples were collected from the Hawaii Ocean Time-series (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 60m depth) and two summer samples at two different depths below the mixed layer (60m and 100m). Seasonality at HOT is dampened relatively to BATS, yet these samples were chosen to maximize seasonal differences at HOT. 60m depth was chosen for a comparison to the Atlantic samples. Originally we tried to analyze three points as a depth profile in the summer stratified sample (at 5m –within mixed layer, 60m and 100m –below mixed layer). However, the cells in the 5m sample were too dim in chlorophyll fluorescence to be flow-sorted.

Samples for single-cell sorting were collected as raw seawater (2x1ml 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), 2.0 μL diluted DNA, 0.25mM each dNTP (NEB), 0.25μM each primer, 1X HF Buffer (Thermo Scientific/NEB), and 0.25X SYBR Green. Reactions were prepared using a Bio-Tek Precision 2000 Liquid Handler. 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* (600-1100bp in comparison to 500-600bp) 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 HL with smaller genomes than the other Low-Light 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 60m and summer 100m 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 Figure1 (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 OTUs with 99% ITS similarity.

*Whole genome sequencing*

Second Round Multiple Displacement Amplification (MDA): Based off of the resulting ITS-sequences, 115 SAGs were selected to undergo a second round of MDA 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 with paired-end reads of length 200bp (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 to 1000bp. CLCbio program "clc_novo_assemble" was used with the command line flags: "-q -p fb ss 150 1000". A minimal contig size of 400bp 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 1.15M bp ± 0.3M bp (Mean ± SD), estimated as ~70% ± 0.18% of the complete genome size. These assembly size statistics are very similar to those of the 96 Atlantic single-cells (1.15M bp ± 0.3M bp (Mean ± SD)). More details on the assembly statistics can be seen in the full QUAST report (Supplementary file S1). Reference-guided assembly: Because 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 to 1000bp (CLCbio program "clc_ref_assemble_long" was used with the command line flags: "-p fb ss 150 1000"). Genome annotation: Annotation of the *denovo*-assembled genomes, as well the cN2-C1 composite genome, were done on the RAST server (Aziz et al 2008). 1971 open reading frames, 3 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 pair-wise 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, ten 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 (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 (i.e. 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 more than 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 HL-adapted cell genomes).
   (ii) Genes that appear in less than 3 cells in total or in more than 75 cells in total.
3. Cluster genes according to their presence/absence in the 115 partial single cell genomes.

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) .

Analysis of ocean-specific cassette gene functions: All proteins in COGSs 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 *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 to 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 datasets 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 2016), 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 (60m) 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.) (Rocap et al 2002) (Moore et al 1998) across these samples (Kashtan et al 2014, Malmstrom et al 2010). To capture changes due to 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 (60m) and two summer samples at two different depths below the mixed layer (60m and 100m) (Figure 1e). All six samples (three from BATS and three from HOT) were taken within eight months of each other (Nov 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 (60m).

*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 (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 (Figure 1b,c,d,f,g,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 on the diagonal Figure 1f,g,h) as opposed to the Atlantic (Figure 1b,c,d), and the number of 'nearly-identical' ITS clusters is much higher in the Pacific (it has many more small clusters, Figure 1f,g,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 co-existing 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) (Colwell et al 2012, Gotelli and Colwell 2001) corroborates the difference of population richness between the Pacific and Atlantic oceans (Figure 2c). Based on these extrapolations the Pacific populations could be made up of over a thousand co-existing 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 measures. Richness at the Atlantic sample B243 was estimated as 500 (95% confidence interval, [355 to 645]) while richness at the Pacific sample H208 was estimated to be much higher: 2090 (95% CI [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 (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 Figure 1f,g,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, Figure S2) (Chandler et al 2016, Gihring et al 2012).

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

Of the total 1014 ITS-clusters (at 99% ITS similarity, that 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=6.2E-29), indicating that the small number of shared ITS-clusters is unlikely explained by chance only. As might 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 to 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, Figure 1b,c,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 (Figure 1f,g,h) – 14 from the 60m winter sample and 5 from the 60m 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 Figure 1b-d and f-h): In the Atlantic cN2 cells were 23.5% ± 1.6%, 11.5% ± 2.2%, and 24.8% ± 1.9% of the analyzed populations in Autumn, Winter and Spring samples, respectively (Kashtan et al 2014) (Figure 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 60m sample, and none in the summer 100m sample (Figure 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, 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 very 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, even though 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 (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 within 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. (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 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 might 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 playing a role (Bryant et al 2016, Ottesen et al 2014, 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 (Berube et al 2015, Berube et al 2016, Coleman and Chisholm 2010, Martiny et al 2009a). 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. (Berube et al 2015) suggested it is evidence for an early horizontal transfer event prior to divergence of these fine-resolution backbone clades and not independent horizontal gene transfer events. Since 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, i.e., through many independent events of gene loss (García-Fernández et al 2004, Kettler et al 2007, Lee and Marx 2012, Partensky and Garczarek 2010, Sun and Blanchard 2014).

The phylogeny of the partial single-cell genomes indicate 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 (10km diameters) on ecologically relevant time scales (Kashtan et al 2014, Okubo 1971), 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 since the discovery of its co-existing, well-defined coarse-grained ecotypes (Moore et al 1998). Johnson et al (Johnson et al 2006) reported distinct patterns, at coarse scale resolution, along a longitudinal gradient in the Atlantic Ocean, and Martiny et al (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, Follows et al 2007, Kent et al 2016, Larkin et al 2016, Martiny et al 2009b, Martiny et al 2006, Rusch et al 2007). 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. Because 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 to 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.

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. N.K. 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 S.W.C. 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 R.S. from the NSF Biological Oceanography Section.

Supplementary information is available at The ISME Journal's website. Genomic data have been deposited in NCBI GenBank under accession numbers SRX2559063 - SRX2559065, MWOO00000000-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).**

## References

Ammerman JW, Hood RR, Case DA, Cotner JB (2003). Phosphorus deficiency in the Atlantic: An emerging paradigm in oceanography. *Eos, Transactions American Geophysical 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. *The ISME journal* **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. *Limnology and Oceanography* **61:** 482-495.

Biller SJ, Berube PM, Lindell D, Chisholm SW (2015). Prochlorococcus: the structure and function of collective diversity. *Nature Reviews Microbiology* **13:** 13-27.

Björkman K, Thomson-Bulldis AL, Karl DM (2000). Phosphorus dynamics in the North Pacific subtropical gyre. *Aquatic Microbial Ecology* **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. *The ISME journal* **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 Research (Part 1, Oceanographic Research Papers)* **44:** 167-192.

Cavender-Bares KK, Rinaldo A, Chisholm SW (2001). Microbial size spectra from natural and artificial nutrient regimes. *Limnology and oceanography* **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. *Environmental microbiology reports* **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. *Proceedings of the National Academy of Sciences of the United States of America* **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. *Journal of plant ecology* **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. *Proceedings of the National Academy of Sciences* **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. *Proceedings of the National Academy of Sciences* **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. *Proceedings of the National Academy of Sciences* **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, Rousso 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* (2016). CyanoBase: a large-scale update on its 20th anniversary. *Nucleic Acids Research***:** gkw1131.

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. *Microbiology and Molecular Biology Reviews* **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. *Environmental microbiology* **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. *Ecology letters* **4:** 379-391.

Jickells TD, An ZS, Andersen KK, Baker AR, Bergametti G, Brooks N *et al* (2005). Global iron connections netween desert dust, ocean biogeochemistry, and climate. *Science* **308:** 67-71.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Malcolm E, S. SW *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. *Annual Review of Genetics* **45:** 431-445.

Kanehisa M, Sato Y, Morishima K (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of molecular biology* **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 Research Part II: Topical Studies in Oceanography* **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. *Nature Reviews Microbiology* **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 research* **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 research* **40:** D632-D640.

Kent AG, Dupont CL, Yooseph S, Martiny AC (2016). Global biogeography of Prochlorococcus genome diversity in the surface ocean. *The ISME journal* **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 genetics* **3:** e231.

Kumar S, Stecher G, Tamura K (2016). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular biology and evolution***:** msw054.

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. *The ISME Journal* **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. *Frontiers in microbiology* **6:** 174.

Malfatti F, Azam F (2010). Atomic force microscopy reveals microscale networks and possible symbioses among pelagic marine bacteria. *Aquatic Microbial Ecology* **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. *The ISME journal* **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. *The ISME journal* **7:** 184-198.

Martínez A, Osburne MS, Sharma AK, DeLong EF, Chisholm SW (2012). Phosphite utilization by the marine picocyanobacterium Prochlorococcus MIT9301. *Environmental microbiology* **14:** 1363-1377.

Martiny AC, Kathuria S, Berube PM (2009a). Widespread metabolic potential for nitrite and nitrate assimilation among Prochlorococcus ecotypes. *Proceedings of the National Academy of Sciences* **106:** 10787-10792.

Martiny AC, Tai AP, Veneziano D, Primeau F, Chisholm SW (2009b). Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus. *Environmental microbiology* **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. *Nature Reviews: Microbiology* **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. *Marine Ecology-Progress Series* **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. *Limnology and Oceanography* **47:** 989-996.

Mühling M (2012). On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. *Environmental microbiology* **14:** 567-579.

Okubo A (1971). Oceanic diffusion diagrams. *Deep Sea Research and Oceanographic Abstracts* **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. *Microbiology and Molecular Biology Reviews* **63:** 106-127.

Partensky F, Garczarek L (2010). Prochlorococcus: Advantages and limits of minimalism. *Annual Review of Marine Science* **2:** 305-331.

Pernthaler J (2005). Predation on prokaryotes in the water column and its ecological implications. *Nature reviews microbiology* **3:** 537-546.

Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002). Resolution of *Prochlorococcus* and *Synechococcus* ecotypes using 16S-23S 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. *Microbiology and Molecular Biology Reviews* **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. *Applied and environmental microbiology* **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. *Frontiers in Marine Science* **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 Research Part II: Topical Studies in Oceanography* **48:** 1405-1447.

Stepanauskas R (2012). Single cell genomics: an individual look at microbes. *Current opinion in microbiology* **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. *Environmental microbiology* **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. *Applied and environmental microbiology* **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. *Limnology and oceanography* **52:** 2205.

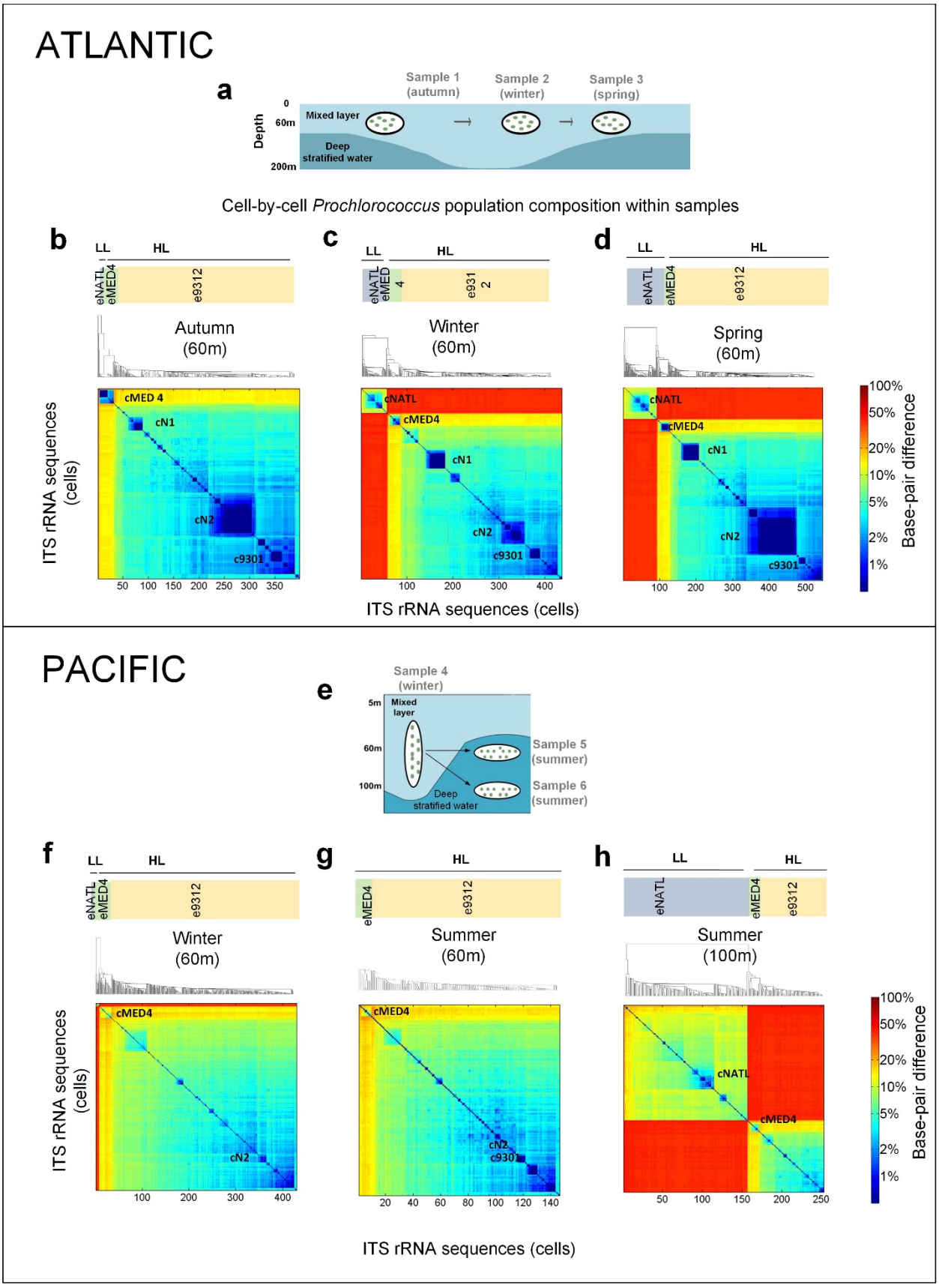
ATLANTIC
a
Sample 1 (autumn)
Sample 2 (winter)
Sample 3 (spring)
Depth
0
60m
200m
Mixed layer
Deep stratified water
Cell-by-cell Prochlorococcus population composition within samples
b
c
d
LL
HL
eNATL
eMED4
e9312
Autumn (60m)
Winter (60m)
Spring (60m)
cMED 4
cN1
cN2
c9301
cNATL
cMED4
ITS rRNA sequences (cells)
Base-pair difference
100%
50%
20%
10%
5%
2%
1%
ITS rRNA sequences (cells)
PACIFIC
e
Sample 4 (winter)
5m
60m
100m
Mixed layer
Deep stratified water
Sample 5 (summer)
Sample 6 (summer)
f
g
h
Winter (60m)
Summer (60m)
Summer (100m)
cMED4
cN2
c9301
cNATL
ITS rRNA sequences (cells)
Base-pair difference
ITS rRNA sequences (cells)

**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 3 seasons at Bermuda Atlantic Time-series Study station (BATS)**.** Cells were collected within the mixed-layer at 60m depth in November 2008 (autumn), February 2009 (winter) and April 2009 (spring), see Methods. 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 Methods). 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 (e.g. cN2). If a cultured representative falls within a cluster, the cluster name follows its name with a 'c' prefix (e.g. 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 Hawaii Ocean Time-series (HOT)**.** The winter sample was collected within the mixed-layer at 60m depth while the summer samples were collected at two different depths 60m and 100m from the stratified water below the mixed layer, see Methods. (**f**), (**g**), (**h**) Same as **b**, **c**, **d** but for the Pacific samples.

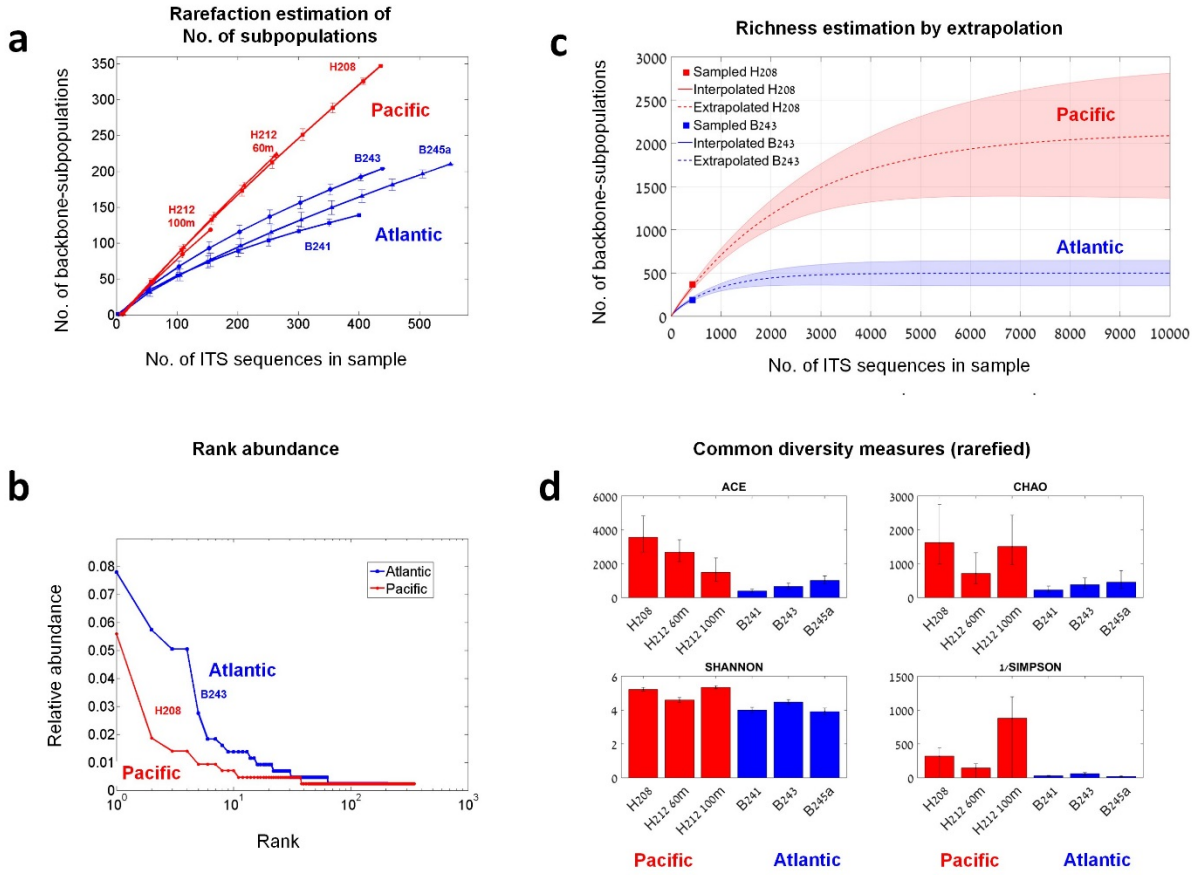


**Figure 2. Comparing the diversity of the Atlantic and Pacific populations.** (**a**) Richness Rarefaction curves estimating the number of co-existing backbone subpopulations within samples. (**b**) Rank abundance profiles of the winter samples from both oceans. Backbone-subpopulations (OTUs) in **a** 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 co-existing 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 6 samples from the two oceans rarefied to 250 ( except from the Pacific summer 60m 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 standard errors (as computed by Mothur (Schloss et al 2009)).

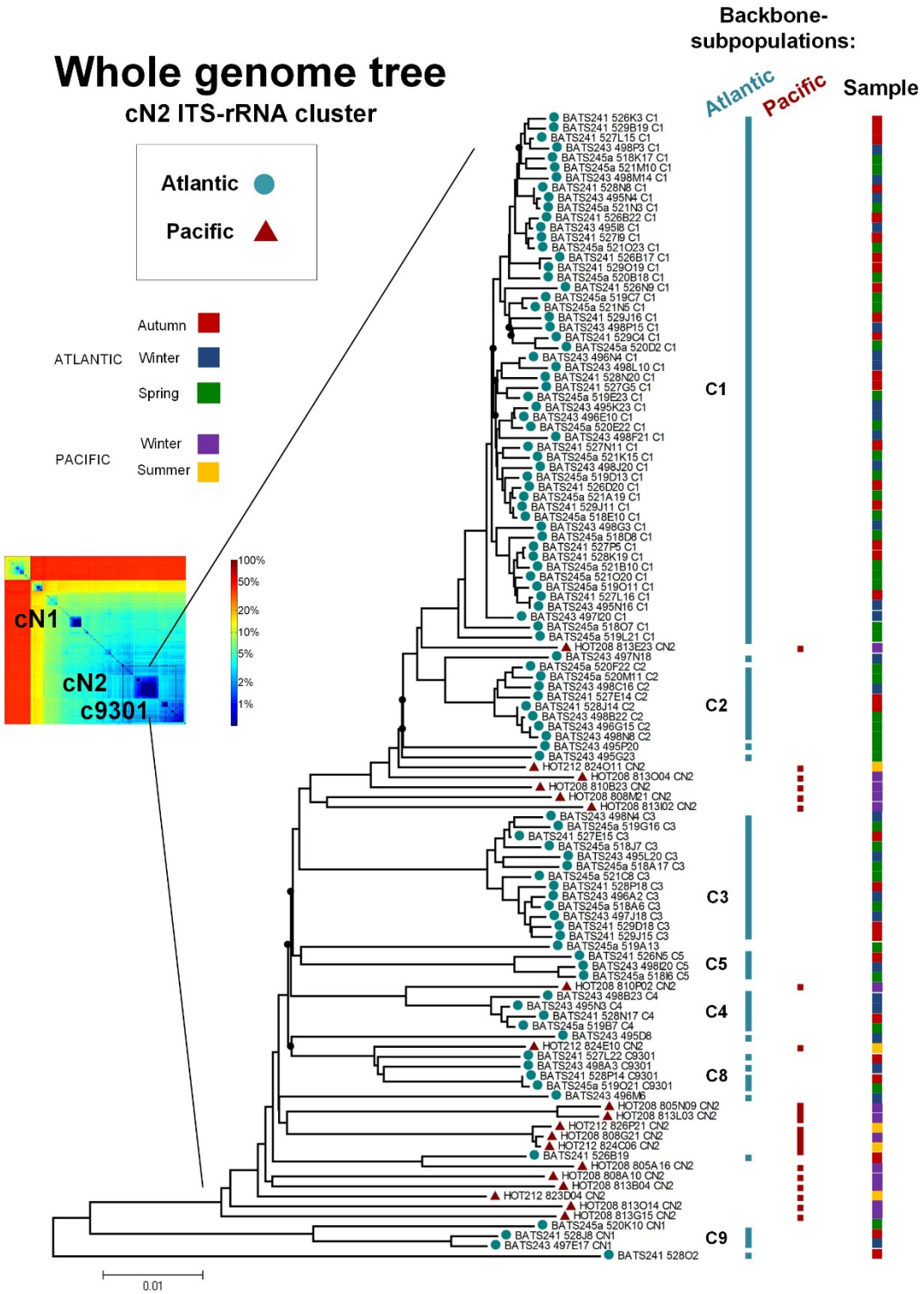


**Figure 3. The genomic backbone clades within the cN2 cluster from the Atlantic and Pacific samples do not overlap.** Neighbor-joining phylogenetic tree of 96 single-cells from the cN2 cluster from the Atlantic samples as well as 19 single-cells from the Pacific based on whole genome sequences. The 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, Methods). Bootstrap values <80 are marked as black dots on the internal nodes. Vertical continous bars on the right represent distinct backbone-subpopulations (blue, Atlantic; red, Pacific). If a subpopulation is represnted 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.

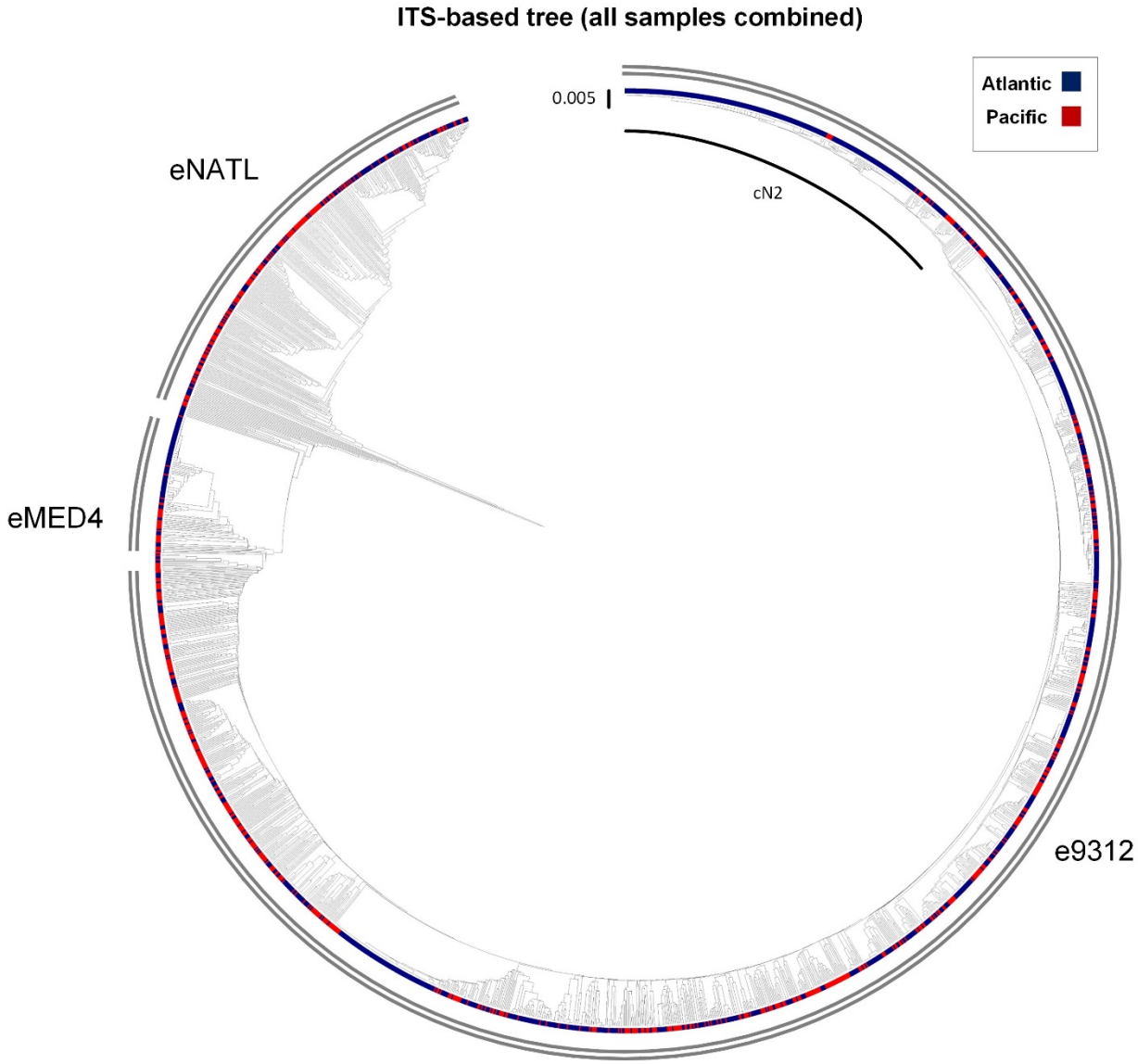


**Figure 4. The global phylogenetic pattern observed in *Prochlorococcus* ITS-rRNA sequences from all single cells sampled at both sites.** Colors represent sample origin: HOT (red leaf coloring), BATS (blue leaf coloring). Neighbor-joining phylogenetic tree is based on multiple alignment of 2209 ITS-sequences.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

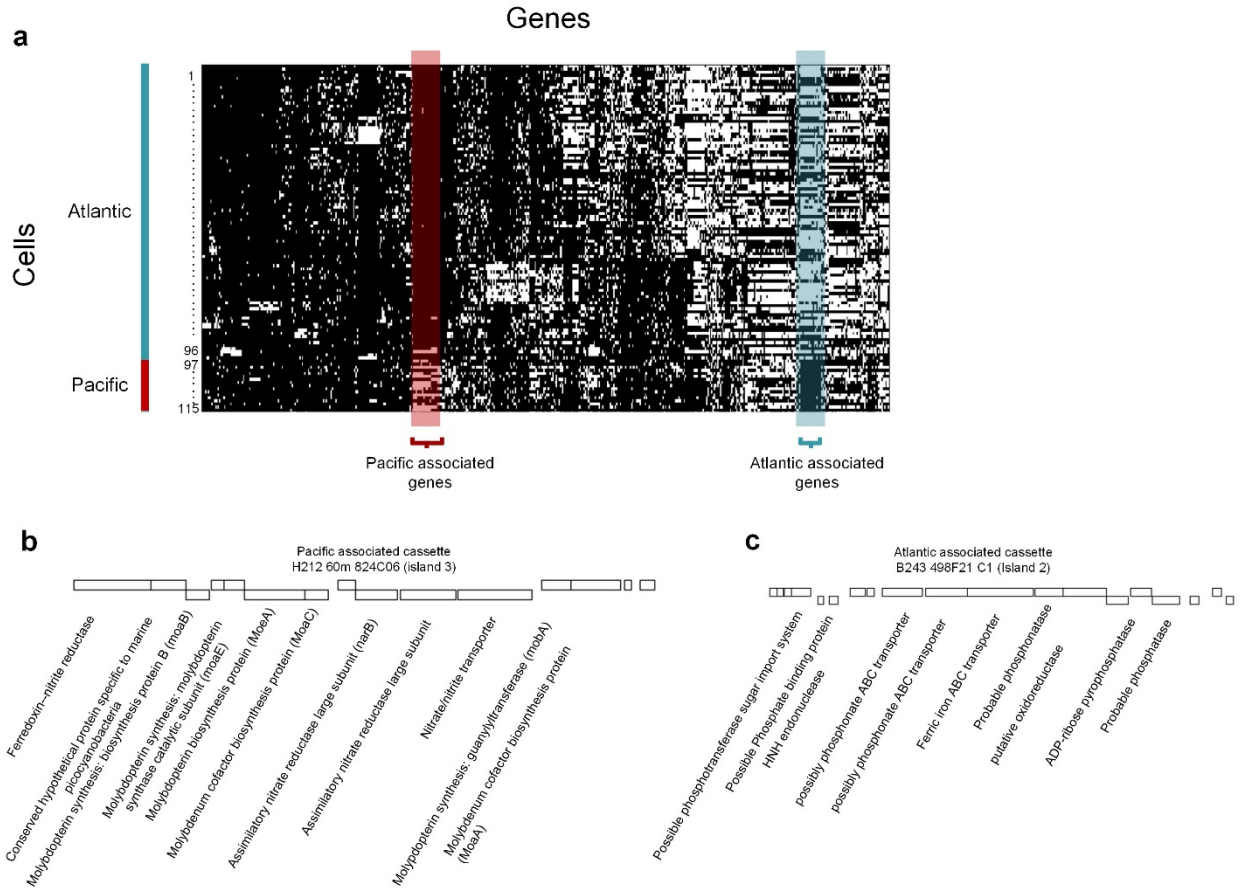


**Figure 5. Differential gene sets distinguishing the Atlantic and Pacific sites** (**a**) Matrix representation: Each column is a gene and each row represents a single-cell. White/black dots represent the presence/absence of a gene in the partial genome of a single-cell, respectively. Single cells are 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 since 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 (i.e. of columns in the matrix) do not reflect location on the genome; the order is determined by the clustering (i.e. 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.

**Table 1. Samples details.** 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).

| Sample | Ocean | Date | Season | Cruise | Depth | Mixed layer | No. of SAGs | *Prochloro-coccus abundance* cells/ml (mean±SE) |
|---|---|---|---|---|---|---|---|---|
| 1 | Atlantic | November 8th 2008 | Autumn | BATS 241 | 60m | Within | 399 | 41,350±750 |
| 2 | Atlantic | February 8th 2009 | Winter | BATS 243 | 60m | Within | 436 | 33,100±800 |
| 3 | Atlantic | April 1st 2009 | Spring | BATS 245a | 60m | Within | 546 | 33,000±1350 |
| 4 | Pacific | Jan 20th 2009 | Winter | HOT 208 | 60m | Within | 429 | 200,000 |
| 5 | Pacific | July 3rd 2009 | Summer | HOT 212 | 60m | Below | 146 | 230,000 |
| 6 | Pacific | July 3rd 2009 | Summer | HOT 212 | 100m | Below | 253 | 172,500 |

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

| **Ocean** | **COG ID** | **Description** | **Position** |
|---|---|---|---|
| **Atlantic Cassette 1** | 8440 | hypothetical protein | Island 1 or Island 2 or Island 5 ** |
| | 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 phosphonatase | |
| | 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 | Molypdopterin synthesis: guanylyltransferase (mobA) | |
| | 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<br><br>Conserved hypothetical protein specific to marine picocyanobacteria | |
| | 40686 | Molybdopterin synthesis: molybdotransferase (moeA) | |
| | 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 *** | 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-2S) family protein, possible ring hydroxylating dioxygenase | |
| | 50117 | hydrogenase/urease nickel incorporation protein (hypB) | |
| | 59708 | ABC transporter, permease protein | |

*Pacific cassette 3 genes appear also in some cN1 and c9301 cells in the Atlantic (but not within cN2 Atlantic cells).

**Islands names (2.2, 3, 5 etc. are defined as in (Kashtan et al 2014))

Figure S1

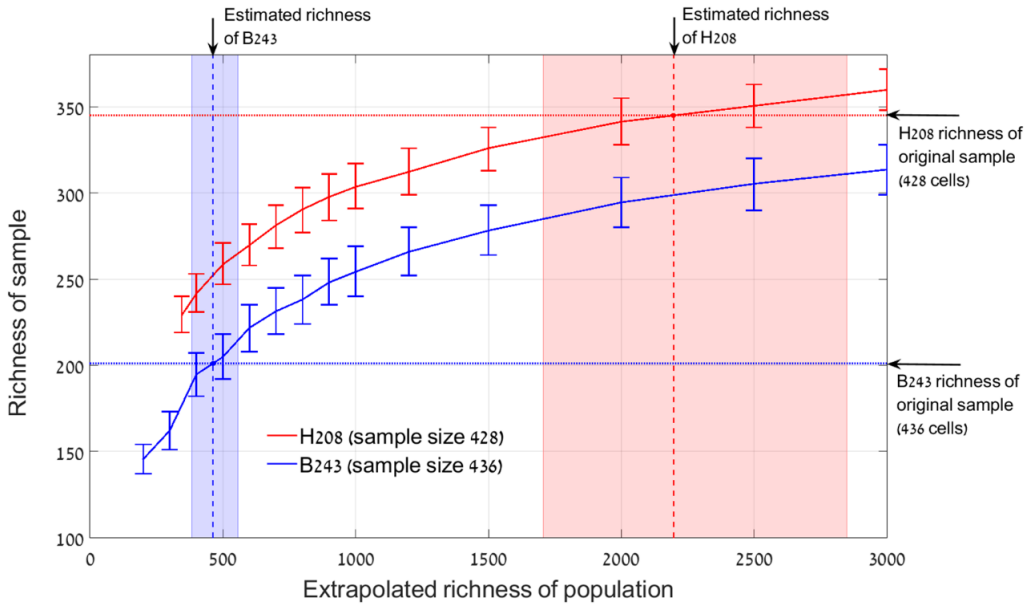


**Figure S1. Estimation of entire population richness of the B243 and H208 samples.** At each data point, a synthetic population is generated by extrapolating the observed rank abundance curve while maintaining the observed evenness of the original sample. The synthetic population is sampled at the original sample size. The mean richness of samples of the original size is plotted against the richness of the synthetic population. The extrapolated richness value at which the mean richness of the samples is equal to the value of observed richness of the original sample (shown as a horizontal line), is the estimated total richness of the population (shown as a vertical line). The resulting values are 460 (95% CI [385 – 555]) for the B243 site and 2200 (95% CI

[1700 – 2850]) for the H208 site. Error bars indicate the standard deviation of 1000 samples at the original sample size. Shaded areas indicate the range of synthetic population richness which corresponds to 95% confidence interval of mean sampled richness. OTUs are defined as sets of cells that are 99% similar in their rRNA-ITS sequences.

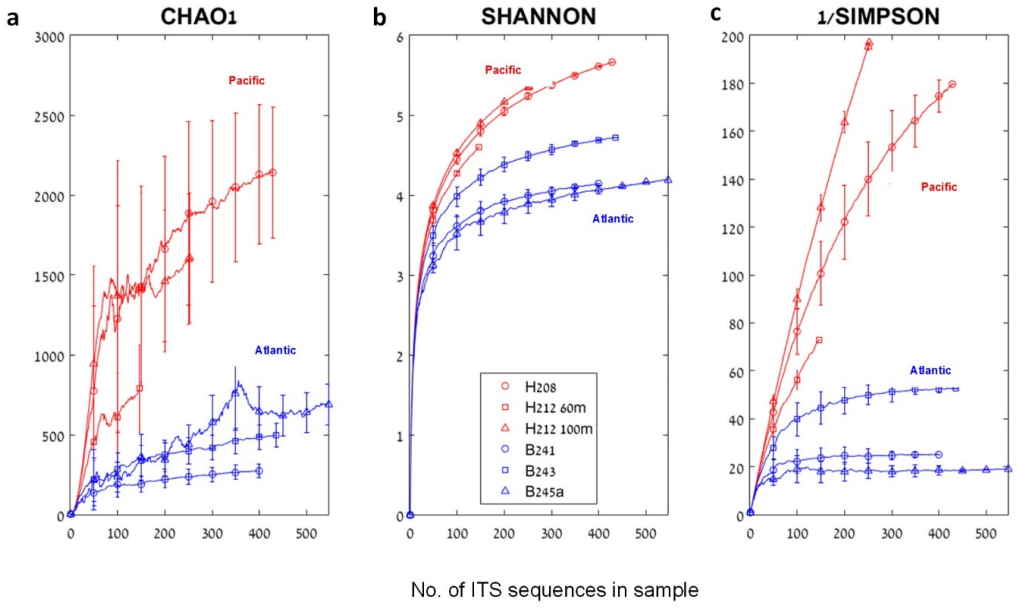


**Figure S2.** Richness (Chao1) and diversity (Shannon, 1/Simpson) estimators of *Prochlorococcus* as a function of sampled population size within the 6 samples from the two oceans. Defining units are sets of cells that are 99% similar in their rRNA-ITS sequences. Error bars of diversity estimators indicate the standard error of 100 computations for each estimator, and error bars of richness indicate 95% confidence intervals, as computed by EstimateS version 9 (http://purl.oclc.org/estimates) (Colwell, R. K. 2013. EstimateS: Statistical estimation of species richness and shared species from samples.)

Figure S3

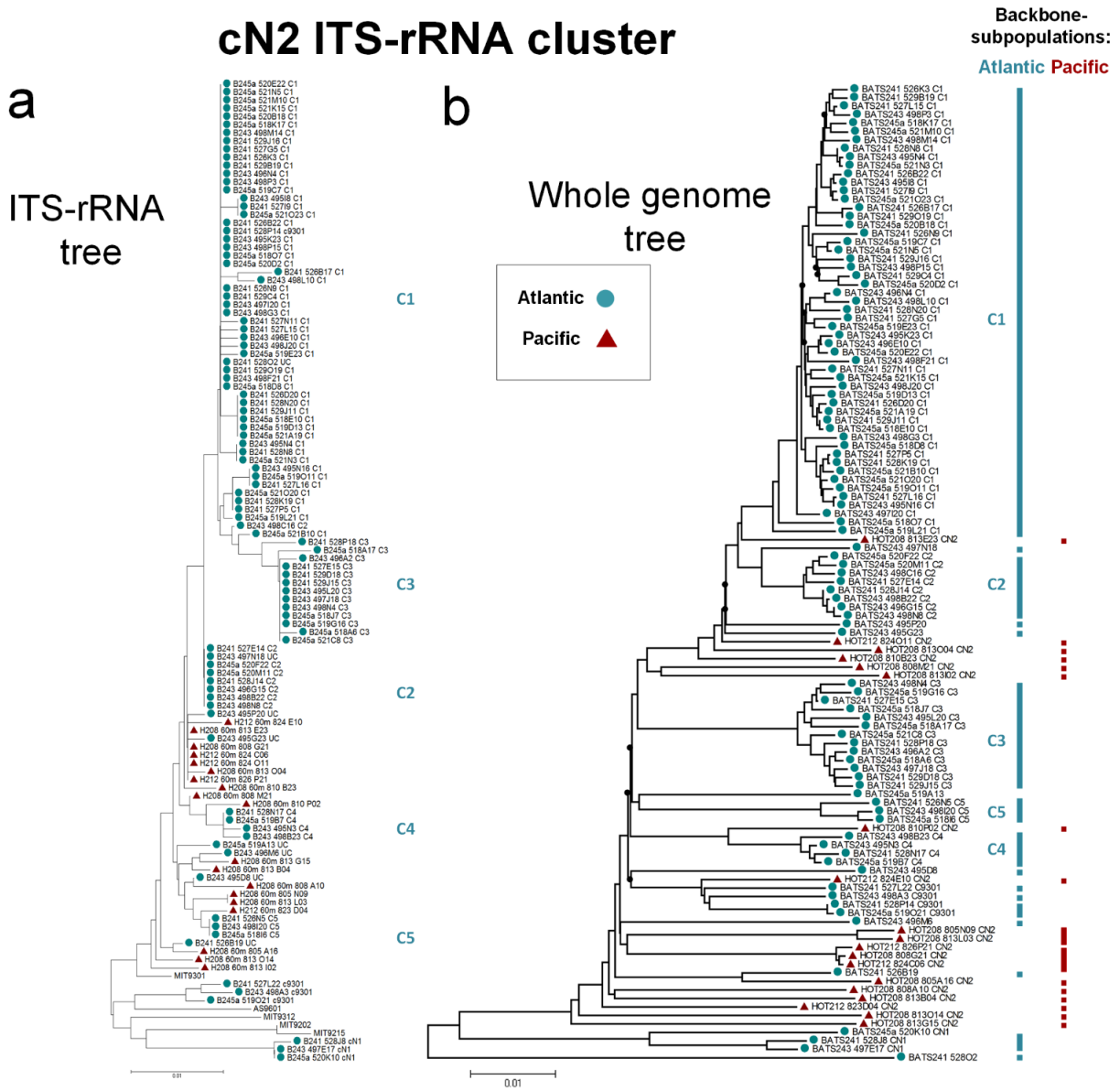


**Figure S3. ITS-rRNA sequence and whole-genome neighbor-joining phylogenetic trees at a fine resolution of diversity.** (**a**) Phylogenetic tree based on ITS-rRNA sequences of 115 single cells (96 from BATS and 19 from HOT) as well as additional five High-Light adapted cultured strains. (**b**) Phylogenetic tree of the 115 single cells based on whole genome sequences. Distance units are base substitutions per site (see scale bar). Bootstrap values <80 are marked as black dots on the internal nodes in **b**. Neighbor-joining trees in **a** and **b** were constructed using p-distance.