# *Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus*

**Citation:** Kashtan, N., S. E. Roggensack, S. Rodrigue, J. W. Thompson, S. J. Biller, A. Coe, H. Ding, et al. "Single-Cell Genomics Reveals Hundreds of Coexisting Subpopulations in Wild Prochlorococcus." Science 344, no. 6182 (April 24, 2014): 416–420.

**As Published:** http://dx.doi.org/10.1126/science.1248575

**Publisher:** American Association for the Advancement of Science (AAAS)

**Persistent URL:** http://hdl.handle.net/1721.1/92763

**Version:** Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

# Single cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*

Nadav Kashtan[1]*, Sara E Roggensack[1], Sébastien Rodrigue[1,2], Jessie W Thompson[1], Steven J Biller[1], Allison Coe[1], Huiming Ding[1,7], Pekka Marttinen[3], Rex R Malmstrom[4], Roman Stocker[1], Michael J Follows[5], Ramunas Stepanauskas[6] & Sallie W Chisholm[1,7]*

**Affiliations:**

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

[2]Département de biologie, Université de Sherbrooke, 2500 boul. Université, Sherbrooke, Québec, J1K 2R1, Canada

[3]Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Aalto University, PO Box 15400, FI-00076 Aalto, Finland.

[4]Department of Energy Joint Genome Institute, 2800 Mitchell Dr, Walnut Creek, California 94598, USA

[5]Department of Earth, Atmosphere and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, Massachusetts 02139, USA.

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

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

*Correspondence to: Chisholm@mit.edu or nadav.kashtan@gmail.com

**Extensive genomic diversity within co-existing members of a microbial 'species' has been revealed through selected cultured isolates and metagenomic assemblies. Yet the cell-by-cell genomic composition of wild uncultured populations of co-occurring cells is largely unknown. Here, we applied large-scale single-cell genomics to study populations of the globally abundant marine cyanobacterium *Prochlorococcus*. We show that they are composed of hundreds of subpopulations with distinct 'genomic backbones' – each backbone consisting of a different set of core gene alleles linked to a small distinctive set of flexible genes. These subpopulations are estimated to have diverged at least a few million years ago suggesting ancient, stable niche-partitioning. Such a large set of coexisting subpopulations may be a general feature of free-living bacterial species with huge populations in highly-mixed habitats.**

The cyanobacterium *Prochlorococcus* is the smallest and most abundant photosynthetic cell in the oligotrophic oceans, contributing significantly to global photosynthesis (*1*). A single species by traditional measures, it can be divided into several major clades, or ecotypes, defined by the Intergenic Transcribed Spacer (ITS) region of their rRNA genes. These ecotypes are physiologically distinct (*2-4*), display distinctive seasonal, depth, and geographic patterns (*3*), and like other microorganisms (*5-10*) embody tremendous genotypic and phenotypic diversity (*4*). To begin to understand the scope and limits of ecologically meaningful diversity within the canonical *Prochlorococcus* ecotypes we examined cell-by-cell genomic diversity – within a small sample of seawater – and explored how it shifts in a dynamic environment.

We applied single cell genome sequencing (*11-14*) to wild *Prochlorococcus* cells from samples collected at the Bermuda-Atlantic Time-series Study site at three times of year (Fig. 1A, (*15*)). Since light, temperature, nutrients, and co-occurring communities change with winter deep mixing (*16*) (Fig. 1A, (*15*)), cells experience significant environmental changes over tens of generations – enough to cause shifts in abundance of ITS-defined ecotypes (*2, 15, 17*). Flow sorting and DNA amplification (*11-14*) of over a thousand co-occurring *Prochlorococcus* cells allowed us to explore the cell-by-cell genomic composition of these wild populations. We were able to identify coherent subpopulations at the whole genome level and their relationship to those defined by the ITS region, explore finely-resolved diversity patterns within and between subpopulations, and examine shifting abundances with seasonal changes in the habitat.

We first examined the population composition by sequencing the ITS of hundreds of *Prochlorococcus* cells in each sample, revealing the presence of finely resolved clusters within the broadly-defined ecotypes (Fig 1B). The populations were composed of tens to hundreds of 'nearly-identical' ITS clusters (>98% similar) within the coarse-grained ecotypes (Fig. 1B,C). The relative abundance of cells belonging to the different clusters changed with season (Fig. 1A,B,C (*15*) ), suggesting shifts in their relative fitness in response to environmental changes.

To study the fine-scale genomic variation and to compare it with the ITS-defined clusters, we sequenced the partial genomes (representing on average 70% of the total genome) of 90 individual cells (30 per sample) from the largest 'nearly identical' ITS-cluster, cN2 (Fig 1C, Fig 2), as well as six cells from two other clusters, cN1 and c9301. For each time of year, cells were randomly selected for genome sequencing from within the major ITS-ribotypes (>99% similar) within cluster cN2 (C1-C5) (30 cells), as well as from c9301-C8 and cN1-C9 (one cell each) as detailed in (*15*). We analyzed between-cell variation in the partial genomes recovered, using a modified mediator genome reference assembly approach (*15, 18*). The topology of the ITS and genomic trees were highly congruent (Fig. 2), indicating that ITS sequences can serve as a proxy for genome sequences in *Prochlorococcus* at a much finer level of resolution than previously demonstrated (*4, 19*). The genomic data further revealed that the largest cluster cN2 is divided into five major clades (C1-C5, Fig. 2) and a few additional minor clades represented by only one cell each. The delineation of C1-C5 clades was highly robust and also observed in trees constructed from genomic position subsets (Figs. S1,S2).

To explore the evolutionary forces that shaped the cN2 C1-C5 clades, we examined differences in nucleotide sequences within and between clades. For example, the C1 and C3 subpopulations (Fig. 2B) differ in 52,885 dimorphic single nucleotide polymorphisms (SNPs), which represent 3.2% of their genomes (Fig 3A blue). The C1,C3 dimorphic SNPs are scattered across the genomes, occurring in 1519 out of 1974 genes – most of them core genes; 8% of these SNPs are found in intergenic regions (9% of the genome is non-coding). Of the intragenic SNPs, 37% are non-synonymous, thus affecting the amino-acid sequences of the proteins they encode. In contrast to the scattered nature of the sequence variation between the C1 and C3 clades, the polymorphism within them is confined to a few regions of the genome (Fig. 3A black) indicating that most regions along the genome are conserved within clades and different between them (*15)* – true for all pairwise comparisons within C1-C5 (Figs. S3,S4).

This emerging pattern was further supported by a standard measure for genetic differentiation between populations, $F_{ST}$ (*20*), applied at gene-by-gene resolution to the 5 cN2 clades, C1-C5 (Fig. 3B,C) . Seventy five percent of the core genes had high $F_{ST}$ values (greater than >0.8), (Fig. 3B,C, (*15*)), meaning different clades contained significantly different alleles. Some of the differentiated core genes have functions involved in the interaction between the cell and environmental stimuli (e.g. transporters, and genes that affect oxidative stress responses, and cell surface biosynthesis/modification; Additional data file S1) – i.e. they are not all simply "house-keeping genes" that control central metabolism. For example, alleles of phosphoglucosamine mutase, which is involved in the biosynthesis of outer membrane lipopolysaccharides (*21*) differ by an average of 10% of their amino acid sequences (Fig. 3C), with substitutions in the hydrophilic center of the enzyme (*21*) possibly affecting its specificity and kinetics.

We next asked whether different clade-subpopulations carry distinct sets of flexible genes. Using *de novo* assemblies to capture regions unmapped by the reference assemblies (*15*) we found that each subpopulation carries a small set of distinct genes – typically in the form of cassettes within genomic islands (Table 1). Remarkably, cassettes containing genes in the glycosyltransferase family account for much of the gene content variation between these clade-subpopulations (Table 1, Table S1). The gene content in these cassettes suggests involvement in outer membrane modifications, possibly affecting phage attachment (*22*), recognition by grazers (*23*), cell-to-cell communication or interactions with other bacteria (*24*).

We conclude that these clade-subpopulations have distinct 'genomic backbones' (and are henceforth referred to as 'backbone-subpopulations') consisting of highly conserved (within subpopulation) alleles of the majority of core genes, and a small distinct set of flexible genes that is linked with a particular backbone. This co-variation between the core alleles and flexible gene content, and its fine scale resolution, represents a new dimension of micro-diversity within wild *Prochlorococcus* populations. It is noteworthy that similar patterns have been identified in selected cultured isolates and metagenomic assemblies within coexisting members of a few other microbial 'species' with very different ecologies (*5-10, 25*), suggesting that differentiated genomic backbones may be a feature of diverse types of microbial populations.

At a finer resolution of diversity we observed that cells within the five cN2 backbone-subpopulations differ by 19,000 nucleotide positions on average, in comparison to 77,000 positions between backbone-subpopulations (equivalent to 1.2% and 4.7% of the genome, respectively) (Fig. 2B). The most similar pairs of individual cell genomes in our samples differ in a few hundred bp (close to the detection limit when one considers single cell processing and sequencing error (*15*)); we infer that some likely have identical gene content (*15*). Except from these few pairs, each cell carries at least one gene cassette not found in any other. In some cases a few closely related cells (a subclade) within backbones share a 'unique' gene cassette. Among these genes are, again, glycosyltransferase genes, as well as transporters and genes involved in nucleotide binding and processing. In a few cases cells from different backbone-subpopulations carry similar flexible gene cassettes (e.g. high-light related genes (Table 1) and phosphonate related genes), demonstrating the combinatorial nature of backbones and flexible genes.

If backbone-subpopulations have differential fitness we would expect their relative abundance to change with changing environmental conditions (Fig 1). Accordingly, the majority of the largest subpopulations exhibited significant seasonal abundance variation (Fig. 4A) higher than expected by chance (*15*). This is consistent with the hypothesis that this reflects selection, but more data is needed to draw that conclusion. Backbone-subpopulations maintain their genomic composition between seasons (tested for C1, (*15*)), which we would expect, as establishment of new mutations and acquisition/loss of genes is not likely to be in play on these timescales (*15*).

The congruency of genomic and ITS phylogenies in *Prochlorococcus* at both coarse (*4, 19*) and fine resolution (Fig. 2), suggests that ITS-ribotype clusters coincide, in most cases, with distinct genomic backbones (*15*). This allowed us to estimate the number of coexisting backbone-subpopulations in our samples through rarefaction analysis, revealing at least hundreds of coexisting subpopulations with distinct backbones (Fig. 4B) in each sample. These backbone-subpopulations are estimated to have diverged at least a few million years ago (*15*) suggesting ancient, stable niche-partitioning. That they have different alleles of 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.

Enormous population sizes and immense physical mixing likely played a role in the evolution of diverse genomic backbones in *Prochlorococcus*. A simple fluid mechanics model bridging the micron and km-scale for a 'typical' ocean suggests that 'just-divided' cells will be centimeters apart within minutes, tens of meters apart within an hour, and a few kilometers apart within a week (*15*). Thus *Prochlorococcus* populations are expected to be well-mixed over large water parcels (~10km$^2$ area x 3m depth) on ecologically relevant time scales (~1 week) (*15*). This mixing and a stable collective *Prochlorococcus* population density of $10^7$-$10^8$ cells L$^{-1}$ (*17*) make the size of each backbone-subpopulation in such parcels enormous (>$10^{13}$ cells, (*15*)). The effective population size is arguably close to this census population size (*15*), implying that

*Prochlorococcus* evolution is governed by selection, not genetic drift (based on population genetics theory (*26*)). Consistent with this argument, the difference in the observed $F_{ST}$ distribution from that estimated for no-selection (Fig. 3B ), provides further evidence that the differentiation of genomic backbones in *Prochlorococcus* is a product of selection (*15*).

The correlation between phylogeny and flexible gene content (Table 1, Tables S1,S13, and Fig. S5) leads us to propose that the emergence of a new backbone is initiated by the acquisition of a beneficial flexible gene cassette followed by slow fine-adjustment of the core gene alleles to the new niche dimension afforded by the acquired cassette. Indeed, given the huge effective population size, even extremely weak fitness differentials among alleles (*27*) can facilitate fine-adjustment of core genes (*15*) over the millions of years of evolution after divergence.

The diverse set of hundreds of subpopulations with distinct genomic backbones likely plays an important role in the dynamic stability of the *Prochlorococcus* 'collective' in the global oceans (Fig. S6). Small fitness differentials, niche differentiation, and selective phage and grazer predation, in the context of temporal and spatial environmental variation, helps explain their co-existence (*28, 29*). On seasonal timescales, the abundance of the *Prochlorococcus* 'collective' maintains relatively stable through temporal and local adjustments in the relative abundance of backbone-subpopulations (Fig. 1C, Fig 4A, Fig. S6D). On longer timescales (decades to millions of years), the collective may respond to shifting selective pressures through exchange of gene cassettes between and within backbone-subpopulations, and through the evolution of the backbones themselves. The coherence of the collective population holds as long as subpopulations do not diverge to the point where they are no longer able to exchange flexible genes and backbone extinction and emergence rates are relatively balanced. If *Prochlorococcus* backbone-subpopulations were designated as distinct species (*30*) it would imply that the global collective is an assortment of thousands of species. It is likely that such a large set of coexisting subpopulations with distinct genomic backbones is a characteristic feature of all free-living bacterial species with very large population sizes living in highly mixed habitats.

**References and Notes:**

1. F. Partensky, W. R. Hess, D. Vaulot, *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiology and Molecular Biology Reviews* **63**, 106 (1999).

2. L. R. Moore, G. Rocap, S. W. Chisholm, Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**, 464 (1998).
3. Z. I. Johnson *et al.*, Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737 (2006).
4. G. C. Kettler *et al.*, Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* **3**, e231 (2007).
5. J. Grote *et al.*, Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio* **3**, (2012).
6. D. E. Hunt *et al.*, Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**, 1081 (2008).
7. S. L. Simmons *et al.*, Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol* **6**, e177 (2008).
8. H. Cadillo-Quiroz *et al.*, Patterns of gene flow define species of thermophilic Archaea. *PLoS Biol* **10**, e1001265 (2012).
9. A. Gonzaga *et al.*, Polyclonality of concurrent natural populations of *Alteromonas macleodii*. *Genome biology and evolution* **4**, 1360 (2012).
10. R. T. Papke *et al.*, Searching for species in haloarchaea. *Proceedings of the National Academy of Sciences* **104**, 14092 (2007).
11. S. Rodrigue *et al.*, Whole genome amplification and de novo assembly of single bacterial cells. *PLoS One* **4**, e6864 (2009).
12. T. Kalisky, P. Blainey, S. R. Quake, Genomic analysis at the single-cell level. *Annual Review of Genetics* **45**, 431 (2011).
13. R. Stepanauskas, Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* **15**, 613 (2012).
14. R. S. Lasken, Genomic sequencing of uncultured microorganisms from single cells. *Nature reviews microbiology* **10**, 631 (2012).
15. Materials and methods are available as supplementary material on Science Online.
16. A. F. Michaels *et al.*, Seasonal patterns of ocean biogeochemistry at the U.S. JGOFS Bermuda Atlantic Time-series Study site. *Deep-Sea Research (Part 1, Oceanographic Research Papers)* **41**, 1013 (1994).
17. R. R. Malmstrom *et al.*, Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *Isme J* **4**, 1252 (2010).
18. O. Wurtzel, M. Dori-Bachash, S. Pietrokovski, E. Jurkevitch, R. Sorek, Mutation detection with next-generation resequencing through a mediator genome. *PLoS One* **5**, e15628 (2010).
19. M. Mühling, On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. *Environ Microbiol* **14**, 567 (2012).
20. M. Nei, Evolution of human races at the gene level. *Human genetics, part A: the unfolding genome. Alan R. Liss, New York*, 167 (1982).
21. R. Mehra-Chaudhary, J. Mick, L. J. Beamer, Crystal structure of Bacillus anthracis phosphoglucosamine mutase, an enzyme in the peptidoglycan biosynthetic pathway. *J Bacteriol* **193**, 4081 (2011).
22. S. Avrani, O. Wurtzel, I. Sharon, R. Sorek, D. Lindell, Genomic island variability facilitates Prochlorococcus-virus coexistence. *Nature* **474**, 604 (2011).
23. J. Pernthaler, Predation on prokaryotes in the water column and its ecological implications. *Nature reviews microbiology* **3**, 537 (2005).
24. F. Malfatti, F. Azam, Atomic force microscopy reveals microscale networks and possible symbioses among pelagic marine bacteria. *Aquatic Microbial Ecology* **58**, 1 (2010).
25. U. Dobrindt, B. Hochhut, U. Hentschel, J. Hacker, Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews* **2**, 414 (2004).
26. J. F. Crow, M. Kimura, *An introduction to population genetics theory*. An introduction to population genetics theory. (Harper & Row, 1970).
27. R. D. Barrett, D. Schluter, Adaptation from standing genetic variation. *Trends Ecol Evol* **23**, 38 (2008).
28. A. D. Barton, S. Dutkiewicz, G. Flierl, J. Bragg, M. J. Follows, Patterns of diversity in marine phytoplankton. *Science* **327**, 1509 (2010).
29. F. Rodriguez-Valera *et al.*, Explaining microbial population genomics through phage predation. *Nature reviews microbiology* **7**, 828 (2009).
30. C. C. Thompson *et al.*, Genomic Taxonomy of the Genus *Prochlorococcus*. *Microb Ecol* **66**, 752 (2013).

**Acknowledgments:** We thank S. Itzkovitz, P.H.R. Calil, D. Sher, R. Milo, P.M. Berube, A.P. Yelton, R. Braakman and particularly M.F. Polz for comments on the manuscript. We thank the Bermuda Atlantic time-series Study for sample collection and the BioMicroCenter facility at MIT for their contributions to the generation of genomic data. N.K. acknowledges the Rothschild Foundation (Yad Hanadiv) and the NOAA 'Climate and Global Change' Postdoctoral Research Fellowships. This work was supported 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 US Department of Energy-GTL and the Gordon and Betty Moore Foundation Marine Microbiology Initiative, to R.S. from the NSF Biological Oceanography Section, and to R.R.M from the US Department of Energy (contract No. DE-AC02-05CH11231)

Genomic data have been deposited in NCBI GenBank under accession numbers KJ477896 to KJ479276 and JFKN000000000 to JFOE00000000. Additional data files have been deposited to Dryad (doi: 10.5061/dryad.9r0p6).

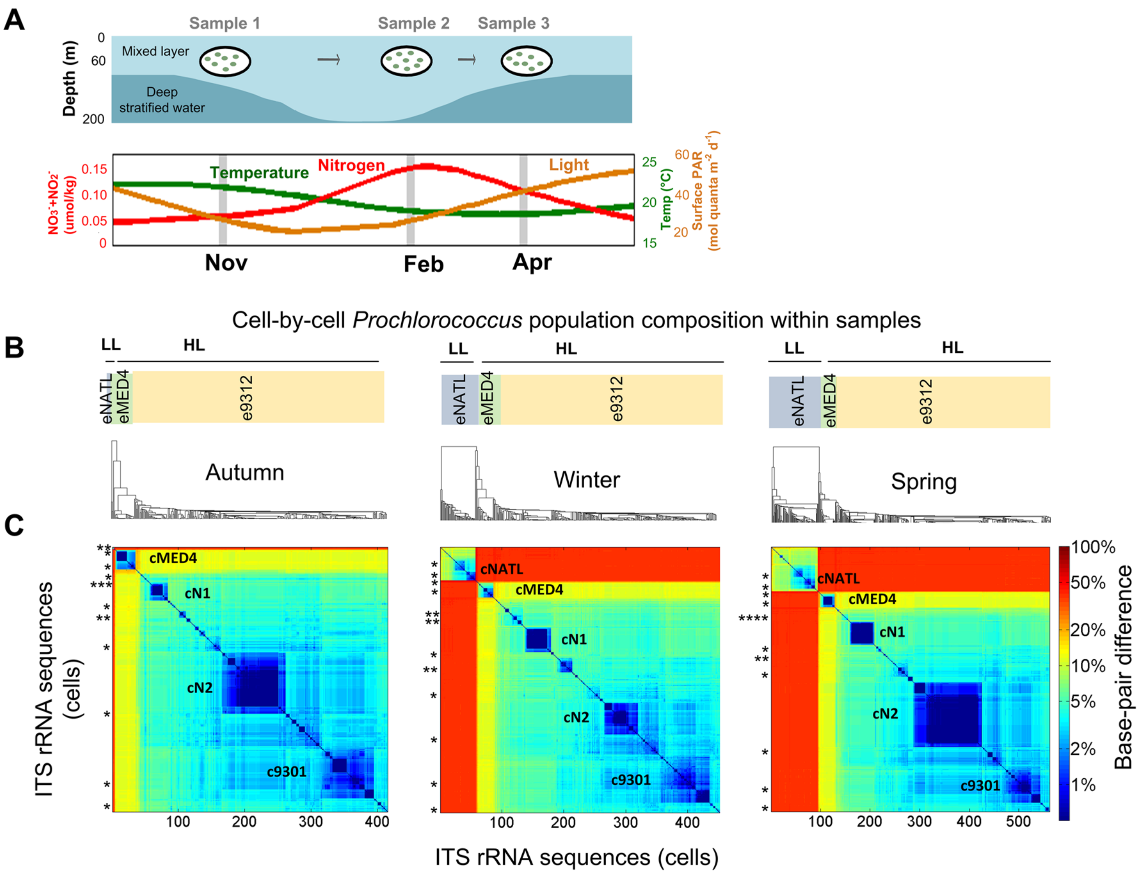

**Fig. 1**. **Cell-by-cell *Prochlorococcus* population composition in samples from 3 times of year at the Bermuda Atlantic Time-series Study (BATS) site.** Cells were collected within the mixed-layer at 60m depth in November 2008, February 2009 and April 2009, see (*15*) (**A**) Schematic of seasonal dynamics at BATS, and sampling design. (top): A typical mixed layer depth profile and context of our three samples. (bottom): Typical average dynamics of light (smoothed mean surface PAR [Photosynthetically Active Radiation] over 2004-2009), temperature, and nitrogen (within mixed layer, averaged over 1999-2009) experienced by cells (*15*). Winter deep mixing brings cold nutrient-rich water to the surface. (**B**) Phylogenetic trees from pairwise genetic distances of ITS-rRNA sequences of individual cells from each sample (based on multiple alignment, (*15*)). The relevant sub-tree range of the 'known' ecotypes (*2*) are marked above each tree if cells belonging to that ecotype were found, as is the division into Low-Light adapted (LL) and High-Light adapted (HL) groups (*2*) . (**C**) Heatmaps describing the pairwise distance matrix between ITS-rRNA sequences of individual cells from each sample. Rows and columns are arranged according to the order of leaves of the trees shown in **B**. 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. ITS sequences from cultured isolates with completely sequenced genomes are marked as (*) centered on the relevant line. Names of the largest clusters are marked in bold (e.g. cN2).

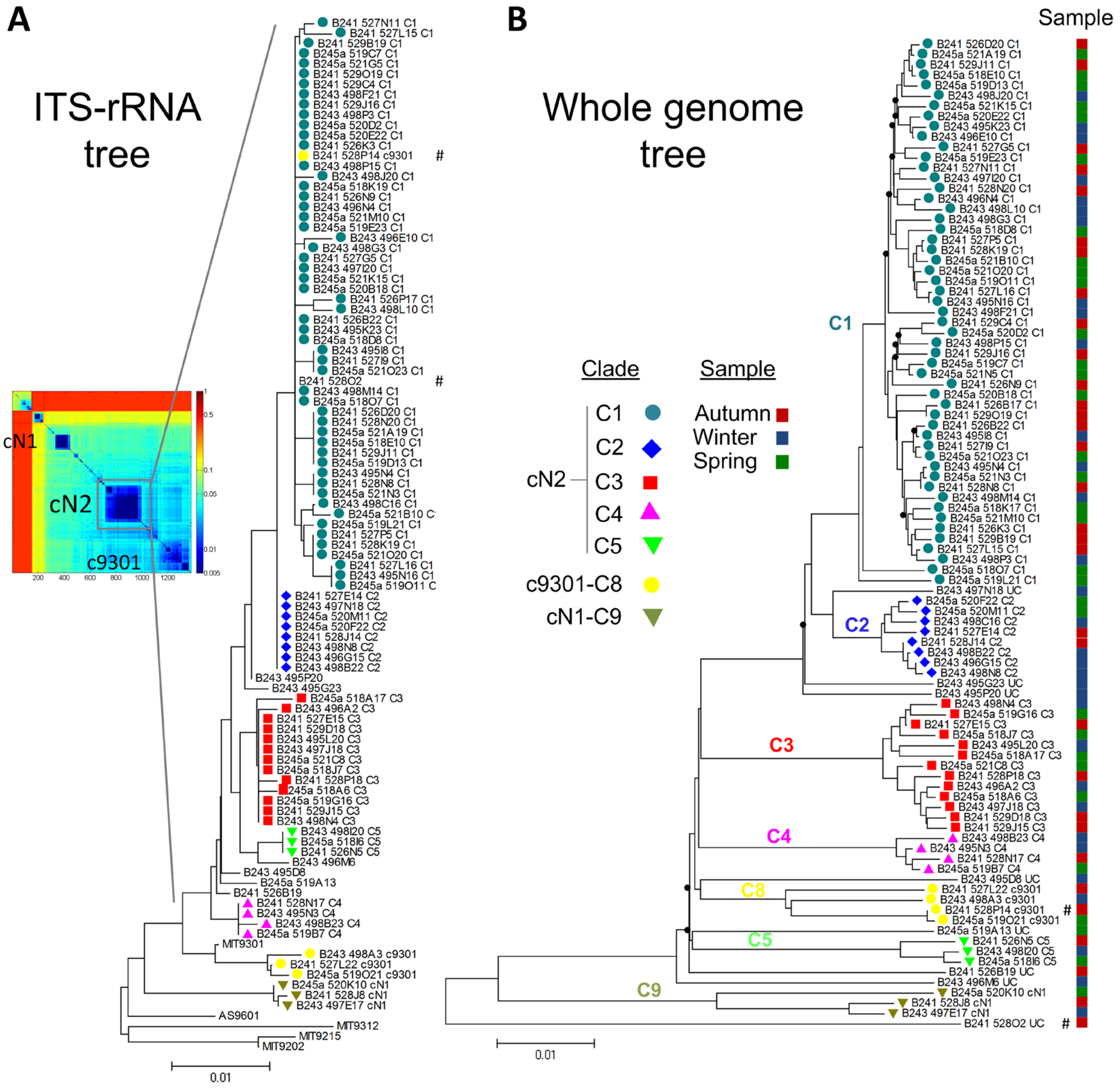

**Fig. 2. ITS-rRNA sequence and whole-genome neighbor-joining phylogenetic trees at a fine resolution of diversity.** (**A**) Phylogenetic tree based on ITS-rRNA sequences of 96 single cells (90 cN2-ribotypes, three cN1-ribotypes, and three c9301-ribotypes) as well as additional five High-Light adapted cultured strains. (**B**) Phylogenetic tree of the 96 single cells based on whole genome sequences. The colored symbols to the left of the leaf labels in **A** and **B** represent the different clades depicted from the deep branches observed in the whole genome tree. The sample origin of each cell is marked with red, blue and green squares (autumn, winter and spring respectively) on the right. Distance units are base substitutions per site (see scale bar, (*15*)). Bootstrap values <80 are marked as black dots on the internal nodes in **B** (Fig. S1). Cells marked with (#) fall into an ITS-clade that differs from the genome-defined clade. Neighbor-joining trees in A and B were constructed using p-distance.

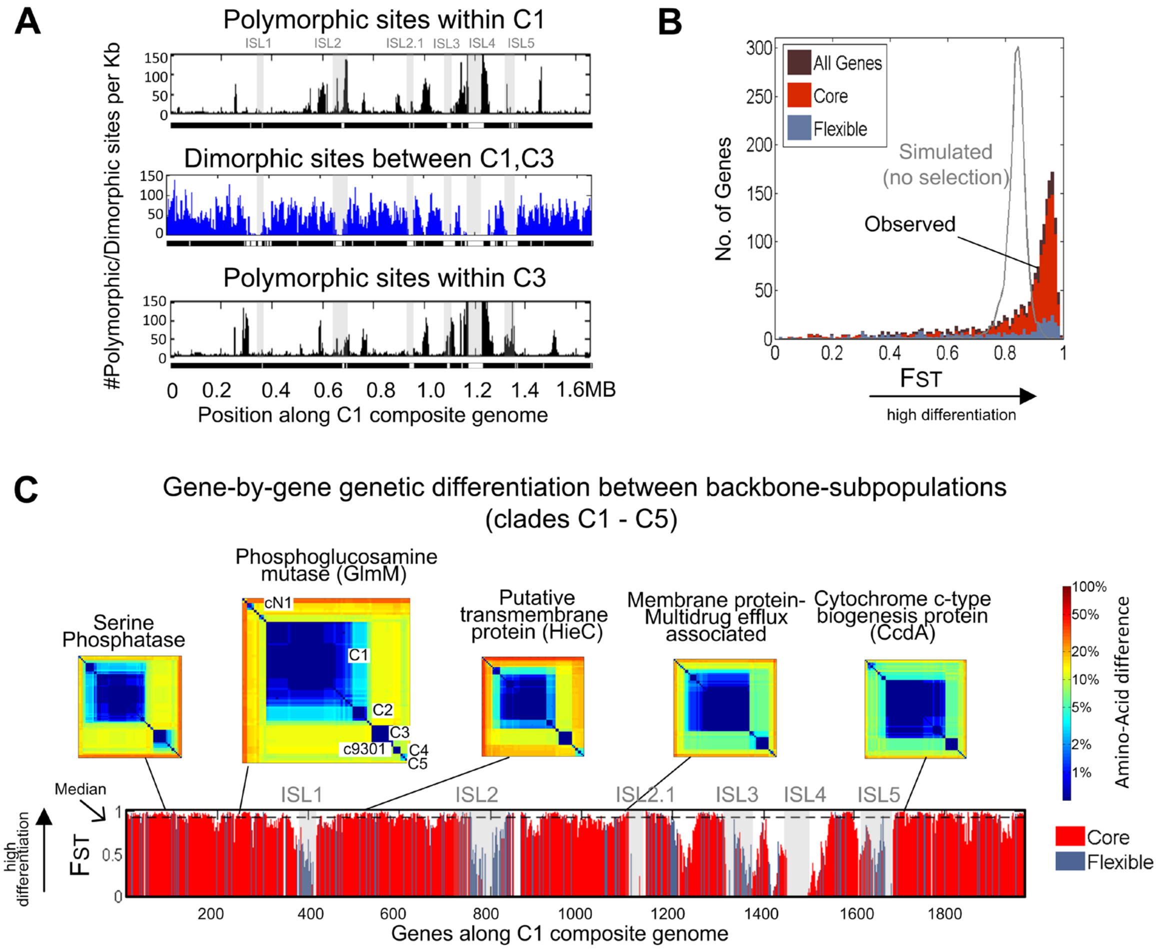

**Fig. 3. Evidence for distinct genomic backbones defining *Prochlorococcus* subpopulations.** (**A**) Polymorphic sites within the cN2 clades C1 and C3 (black) and dimorphic sites between the two clades (blue) (*15*). The black-striped line below each bar-graph marks positions with sufficient data for evaluation of site statistics. Genomic islands (Table S9) are shaded gray. (**B**) Genome-wide distributions of $F_{ST}$ of all genes in the cN2-C1 composite genome as computed for the five cN2 clades C1-C5, based on nucleotide sequences. Also shown is a representative $F_{ST}$ distribution from coalescent simulations of neutral evolution (*15*). Genes with high $F_{ST}$ exhibit higher sequence variation between the clades than within the clades. (**C**) Gene-by-gene profile of genetic differentiation between backbone-subpopulations ($F_{ST}$). $F_{ST}$ is estimated by $\gamma_{ST}$ (*20*). Heatmaps above are displayed for a few core genes with high $F_{ST}$. Each heatmap shows the percentage of amino acid sequence substitutions between single cells as well as cultured High-Light adapted cells.

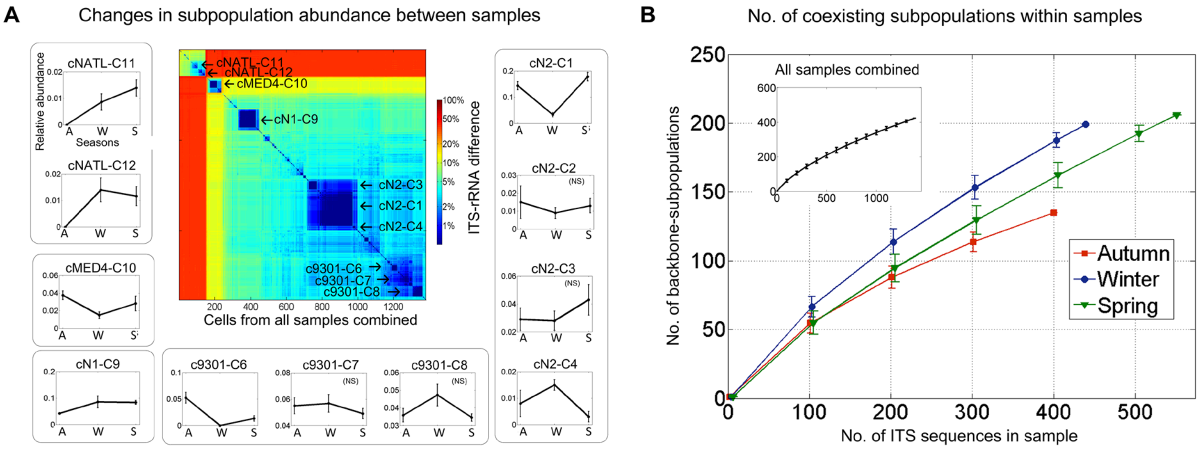

**Fig. 4. Abundance profiles of backbone-subpopulations and the estimated number of coexisting subpopulations within samples**. (**A**) Relative abundance profiles of the 11 largest backbone-subpopulations in our samples within the ITS-clusters cNATL, cMED4, cN1, cN2 and c9301 . A= autumn, W= winter, and S= spring.  Backbone names are marked near the relevant cluster on the ITS heatmap. Backbone-subpopulations were predicted by 99% ITS similarity for the full set of 1381 ITS sequences. Error bars represent standard errors. (NS) – no significant changes between seasons (FDR, α=0.05) (*15*). (**B**) Rarefaction curves estimating the number of coexisting backbone-subpopulations within samples (*15*). Backbones predicted as in (A). Error bars represent 95% confidence intervals. **Inset**, rarefaction curve of all samples combined.

**Table 1.** Flexible gene cassettes associated with different cN2 backbone-subpopulations highlighting gene content that may contribute to ecological differentiation. (GT) Glycosyltransferase; (ABC-T) ABC transporter; (HLIP) Highlight-inducible protein; (CO) Cytochrome oxidase c subunit VIb. (HlpA) outer membrane protein (CpsL) polysaccharide biosynthesis protein

| Clade | Cassette ID | Position | Total # genes in cassette | Selected gene annotations* | Cassette function |
|---|---|---|---|---|---|
| **cN2-C1** | CST_I | Island 2.1 | 4 | HLIP, CO | UV-Protection? |
| | CST_II | Island 4 | 7 | 3GT, ABC-T | Outer membrane modification |
| **cN2-C2** | CST_II | Island 4 | 7 | 3GT, ABC-T | Outer membrane modification |
| **cN2-C3** | CST_III | Island 1 | 2 | 2GT | Outer membrane modification |
| **cN2-C4** | CST_I | Island 2.1 | 4 | HLIP, CO | UV-Protection? |
| | CST_IV | Island 4 | 14 | 3GT, HlpA, CpsL, | Outer membrane modification |
| **cN2-C5** | CST_V | Island 4 | 5 | 2GT | Outer membrane modification |

* Numbers before gene annotation refer to number of that type of gene. A complete list of the genes in each cassette is described in Table S1 (*15*).

**Supplementary Materials:**

Materials and Methods

Figs. S1-S21

Tables S1-S13

Additional data file S1

References (*31*-*92*)

**Supplementary Materials:**

Are submitted as a separate file that includes: Materials and Methods, Supplementary Figures, Supplementary Tables and Supplementary Data Files.