www.sciencemag.org/content/344/6182/416/suppl/DC1

Supplementary Materials for

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

Nadav Kashtan,* Sara E. Roggensack, Sébastien Rodrigue, Jessie W. Thompson, Steven J. Biller, Allison Coe, Huiming Ding, Pekka Marttinen, Rex R. Malmstrom, Roman Stocker, Michael J. Follows, Ramunas Stepanauskas, Sallie W. Chisholm*

*Corresponding author. E-mail: chisholm@mit.edu (S.W.C.); nadav.kashtan@gmail.com (N.K.)

Published 25 April 2014, *Science* **344**, 416 (2014)
DOI: 10.1126/science.1248575

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/344/6182/416/suppl/DC1)

Data S1 (Excel file)

Materials and Methods

Table of Contents

1. Samples

1.1 Sample details

Samples were collected from the Bermuda Atlantic Time-series Study (BATS) site (approximate 5 nautical mile radius around 31° 40′N, 64° 10′ W), see details in Table S2. 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 (*31*). Three samples were selected for analysis, each from one of three different seasons over a period of 5 months: Autumn (November 2008), Winter (February 2009) and Spring (April 2009). All samples were collected from 60m depth to ensure that they were taken from within the mixed layer.

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.

1.2 Seasonal environmental changes at BATS

Seasonal profiles of light, temperature, and nitrogen at BATS, averaged over several years, are shown in Fig. S7.

Prochlorococcus, *Synechococcus* and pico-eukaryote cell counts over the 2008-2009 seasonal cycle, determined by flow cytometry, are described in Fig. S8. The estimated abundance of total

Prochlorococcus cells in the three samples used for single cell sorting, determined by flow cytometry (mean±SE cells/mL), is listed in Table S2.

1.3 Ecotype abundance measured by qPCR

Prochlorococcus ecotypes are traditionally defined by their ITS sequences and their abundance in samples can be estimated by qPCR (*17*). Ocean water samples were collected in 2008-2009 using a Niskin rosette at 12 depths at BATS (1, 10, 20, 40, 60, 80, 100, 120, 140, 160, 180 and 200 m) and processed as previously described by Zinser et al. (*32*). The samples were analyzed on a Roche Light Cycler 480 using culture based standards and the same PCR conditions as previously described in Malmstrom et al (*17*). Abundances that fell below the lowest value of the standard curve were set to the theoretical detection limit of 0.65 cells/mL. See Fig. S9.

2. Single cell Sequencing

2.1 Construction of single amplified genome (SAG) libraries.

Single cell sorting and whole genome amplification were performed at the Bigelow Laboratory Single Cell Genomics Center (https://scgc.bigelow.org). Prior to cell sorting, the cryopreserved samples were diluted 5x with filter-sterilized and UV-treated Sargasso Sea water and then pre-screened through a 70 µm mesh-size cell strainer (BD). Cell sorting was performed with a MoFlo™ (Beckman Coulter) flow cytometer using a 488 nm argon laser for excitation, a 70 µm nozzle orifice and a CyClone™ robotic arm for droplet deposition into microplates. The cytometer was triggered on side scatter. The "purify 1 drop" mode was used for maximal sort purity, which ensures the absence of non-target particles within the target cell drop and the drops immediately surrounding the cell. *Prochlorococcus* cells were separated from other particles based on autofluorescence and light side scatter (proxy to particle size). Target cells were deposited into 384-well plates containing 600 nL per well of 1x TE buffer and then stored at -80ºC until further processing. Of the 384 wells, 315 were dedicated for single cells, 66 were used as negative controls (no droplet deposited) and three received 10 cells each (positive controls). Cells from each sample were deposited into eight 384-well plates: four of them kept as backup and four were used for whole genome amplification as described below.

Cells were lysed and their DNA denatured using cold KOH (*33*). Genomic DNA from the lysed cells was amplified using multiple displacement amplification (MDA) (*33, 34*) in 10 uL final volume. The MDA reactions contained 2 U/µL Repliphi polymerase (Epicentre), 1x reaction buffer (Epicentre), 0.4 mM each dNTP (Epicentre), 2 mM DTT (Epicentre), 50 mM random hexamers with the two 3′-terminal nucleotide bonds phosphorothioated (IDT) and 1 µM SYTO-9 (Invitrogen) (all final concentrations). The MDA reactions were incubated at 30°C for 12-16 h, then inactivated by a 15 min incubation at 65°C. Amplified genomic DNA was stored at -80°C until further processing. We refer to the MDA products originating from individual cells as single amplified genomes (SAGs) (Fig. S10).

Prior to cell sorting, the instrument and the workspace were decontaminated for DNA as previously described (*35*). High molecular weight DNA contaminants were cross-linked in all MDA reagents (*36*). Cell sorting and MDA setup were performed in a HEPA-filtered environment. As a quality control, the kinetics of all MDA reactions were monitored by measuring the SYTO-9 fluorescence using FLUOstar Omega (BMG). The critical point (Cp) was determined for each MDA reaction as the time required to produce half of the maximal

fluorescence. The Cp is inversely correlated to the amount of DNA template (*37*). The Cp values were significantly lower in 1-cell wells compared to 0-cell wells in all microplates ($p<0.001$; Wilcoxon Two Sample Test). Our previous studies and other recent publications using our single cell sequencing technique demonstrate the reliability of our methodology with insignificant levels of DNA contamination (*36, 38-42*).

2.2 ITS-rRNA screening and sequencing

ITS screen

Amplified genomic DNA was diluted 10x in UV-treated 0.2μm filtered H_2O and qPCR screened using primers (ITS-F: 5'-CCGAAGTCGTTACTYYAACCC-3', ITS-R: 5'-TCATCGCCTCTGTGTGCC-3') targeting the *Prochlorococcus* intergenic transcribed spacer (ITS) (*11*). The reaction ran using a LightCycler II 480 (Roche) and underwent 30 cycles of 95°C for 15 seconds, 55°C for 30 seconds, 72° for 45 seconds, followed by an extension at 72°C for 5 minutes and a cooling to 37°C (*11*). Each reaction contained 1.0 Units TaqB (Enzymatics), 2.0 mL diluted DNA, 0.25mM each dNTP (NEB), 0.5mM each primer, 1x buffer (12mM Tris-HCl pH 8.3, 50 mM KCl, 8 mM $MgCl_2$, 150mM trehalose, 0.2% (v/v) Tween20, 0.2 mg/ml non-acetylated BSA, 0.139X SYBR Green). Reactions were prepared using a Bio-Tek Precision 2000 Liquid Handler.

Sequencing of ITS product

Selection for sequencing was based on the kinetics from the MDA reaction; only samples which likely amplified and were confirmed as *Prochlorococcus* through the PCR screen were sent for Sanger sequencing of the ITS product. 15mL of each product were sent to MCLab (www.mclab.com) with 5mM primer (ITS-F) for purification and sequencing.

Second round MDA

Based off of the resulting ITS-sequences, 96 samples were selected to undergo a second MDA reaction in order to produce enough DNA to construct sequencing libraries. Each reaction (performed in duplicate) contained 0.63mL DNA from the first MDA reaction, 250 Units RepliPHI Phi29 DNA polymerase (Epicentre), RepliPHI Phi29 1X Reaction Buffer (40mM Tris-HCl (pH 7.5), 50mM KCl, 10mM $MgCl_2$, 5mM $(NH_4)_2SO_4$, 4mM DTT), 4mM DTT, 1mM each dNTP, and 50mM phosphorothioate-protected random hexamers (IDT). The reactions were incubated at 30°C for 12hours, then heat inactivated at 80°C for 10 minutes in a BIO-RAD C1000 Thermal Cycler (*11*).

Purification of second MDA DNA

DNA resulting from the second MDA was purified using Qiagen's QIAamp DNA Mini Kit according to the manufacturer's protocol, "Purification of REPLI-g amplified DNA." DNA yields were measured using a NanoDrop ND-1000 Spectrophotometer, and DNA from each duplicate reaction was combined in equal parts to help eliminate any bias during MDA (*11*).

Preparation of sequencing libraries

Illumina libraries were generated based on a protocol described in (*43*) using at least 2μg of purified, second round MDA product. 50mL of this DNA was sheared using 18 cycles of alternating 30 seconds ultrasonic bursts and 30 seconds pauses in a 4°C water bath , with

instrument power set to high (Bioruptor UCD-200, Diagenode). The sheared DNA was repaired at room temperature for 30 minutes using an Enzymatics End-Repair Mix.
DNA fragments were size selected with double solid phase reversible immobilization (dSPRI) (*43*) using Agencourt AMPure XP SPRI magnetic beads. In the first SPRI selection, 46.1 mL of AMPure XP beads were mixed with 50mL of DNA and incubated at room temperature for 5 minutes in a 96-well plate. The 96-well plate was placed in a magnetic holder (DynaMag-96 Side, Invitrogen), and 100mL of the supernatant was transferred to a new 96-well plate; the magnetic beads, which are bound to large fragments of DNA, are discarded. Then, 15mL of fresh AMPure XP beads were mixed with the supernatant and incubated at room temperature for 5 minutes. The plate was placed back in the magnetic holder, and the supernatant was discarded, with DNA of the desired length bound to the magnetic beads. The beads were washed twice with 150mL 70% ethanol and allowed to dry. The DNA was eluted by adding and mixing 20mL H_2O to re-suspend the magnetic beads; the mixture was incubated at room temperature for 2 minutes, then placed in the magnetic holder, where 18mL of the supernatant was recovered. This shearing and dSPRI yielded DNA fragments of approximately 420 bp (*43*).

Blunt-end DNA fragments were ligated to two distinct adapters (See Table S3); the DNA was mixed with a 5-fold molecular excess for each oligonucleotide adapter and ligated using Enzymatics Rapid Ligation kit at room temperature for 5 minutes. The newly ligated DNA was purified via SPRI selection using AMPure XP magnetic beads at a DNA/bead ratio of 0.8.

Nick translation of the DNA was performed using 5.2 units Enzymatics Manta 1.0 DNA Polymerase (exo-), 20mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10mM KCl, 2 mM $MgSO_4$, 0.1% Triton X-100, 2.6 mg/ml BSA, 0.2mM each dNTP for 25 minutes at 65°C. DNA was purified via SPRI selection at a DNA/bead ratio of 0.8.

To complete the Illumina adaptor for sequencing, to add sequencing barcodes for multiplexing, and to select fragments with one of each adaptor from the blunt ligation, the DNA fragments were PCR-amplified using KAPA SYBR FAST qPCR Kit; the reaction consisted of 1X KAPA SYBR FAST qPCR Master Mix Universal, and 0.5mM each primer (see Table S3 for oligonucleotides). The reactions were monitored in real time with a Bio-Rad CFX96, and underwent an initial denaturation at 95°C for 1 minute 30 seconds, then repeated cycles of 95°C for 3 seconds, 65°C for 20 seconds, and 72°C for 1 second. When the reactions reached late logarithmic amplification phase, they went through a final extension of 72°C for 1 minute. The reverse primer for each reaction contains a unique 6-nucleotide sequence used to barcode the libraries (Table S3). Libraries were purified using a SPRI selection with a DNA/bead ratio of 0.7.
The libraries were quantified using a BioAnalyzer (Agilent) and qPCR to determine library length and concentration.

3. ITS-rRNA population composition analysis

3.1 Composition of ITS-defined populations

The first step of our analysis (Fig. S10) used flow cytometry to identify and sort *Prochlorococcus* cells from water samples, using a gate that aimed to capture the whole *Prochlorococcus* population. These cells were sorted individually into separate wells, their DNA was MDA amplified, and the ITS region of their genomes was PCR amplified and sequenced.

These steps did not involve any selection thus the set of hundreds of ITS sequences is an unbiased representation of the population composition (w.r.t. ITS) – that spans the known ribotype diversity of *Prochlorococcus* (Fig 1B,C). We excluded from the heatmaps in Fig. 1 cells belonging to Low-Light adapted ecotypes of *Prochlorococcus* (representing only 6%-13% of the total population in the samples) because their ITS sequences are much longer (800-1000bp compared to 500-600bp for High-light ecotypes) and their exclusion made the multiple alignment much more informative. Therefore, apart from the exclusion of this small fraction of Low-Light-adapted cells (or more precisely 'long ITS' cells), we have an unbiased representation of the population.

From a total of 1596 single cell ITS-rRNA sequences (440 sequences from the autumn sample, 519 from the winter sample and 637 from the spring sample) 1381 ITS sequences remained after the removal of cells belonging to the Low-Light adapted ITS sequences (with long ITS sequences), and the removal of partial ITS sequences. Apart from the excluded cells, these 1381 sequences quantitatively represent the population composition of all small-genome *Prochlorococcus* cells in the samples. The number of sequences per sample was 399, 436 and 546 sequences of the autumn, winter and spring samples respectively. Average ITS sequence length was 550±27 bp (mean±SD) .

Sequences were multi-aligned by mafft (*44*) (http://mafft.cbrc.jp/alignment/software/), using the following command line flags: 'mafft --auto --ep 0.123'.

The ITS trees presented in Fig. 1 were generated by Matlab with 'p-distance' and 'average' linkage. Cultured cells whose ITS position is marked in the heatmaps of Fig. 1C (main text) as (*), ordered from top to bottom of each heatmap, are: NATL2A, NATL1A, MIT9515, MED4, MIT9107, MIT9302, GP2, MIT9321, MIT9201, MIT9215, MIT9312, AS9601, SB, MIT9301, MIT9314.

The 96 ITS sequences in Fig 2A were multi-aligned by Matlab with 'multialign(96-ITS, 'terminalGapAdjust', true)'.

3.2 Relative abundance of ITS-defined clusters within samples

Traditional ecotype abundance as estimated by single cell ITS sequences as well as by qPCR is summarized in Table S4. Relative abundances of the largest ITS clusters, as depicted from single cell data is summarized in Table S5. Relative abundances of the cN2 C1-C5 clades as depicted from single cell data is summarized in Table S6. To assess the standard error values presented in Tables S5 and S6, the relative abundance was calculated for each of the four 384-well plates (cells from each seasonal sample were flow-sorted into four 384-well plates; we treated each one of the four plates as a sample replication) and then bootstrapped using 1,000 resampling with repetitions.

3.3 Community comparisons between samples

Two standard methods for community comparisons were used to ask whether the *Prochlorococcus* population structure, based on ITS-rRNA sequences, differed between seasonal samples:

1) Libshuff (http://whitman.myweb.uga.edu/libshuff.html) using cutoff=0.01.
2) FastUniFrac (*45*) (http://bmf2.colorado.edu/fastunifrac/).

Pairwise comparisons of samples, by both methods, indicated the populations of any two of the three samples are significantly different (pairwise comparisons, P<0.001).

4. Sequencing and assemblies of single cell genomes

4.1 Choosing single cells for whole genome sequencing

96 single cell partial genomes were sequenced: 90 cells (30 per sample) from the cN2 'nearly identical' ITS-cluster, cN2 (Fig 1C, Fig 2) , three cells (one per sample) from cN1 and three cells (one per sample) from the c9301 ITS-cluster, as summarized in Table S7. For each time of year, cells were randomly selected from within the major ITS-ribotypes (>99% similar) within cluster cN2 (C1-C5) for whole genome sequencing, as well as one cell from each of the other two clusters (c9301-C8 and cN1-C9). Note that the relative number of cells that were selected for whole genome sequencing, per clade, does not represent their relative abundance.

4.2 Whole genome sequencing

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).

4.3 De novo assembly of single cell genomes

De novo assembly was done by clc-assembly-cell-3.2 (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 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 150bp and 1000bp. Minimal Contig size was set to 200bp (CLCbio program "clc_novo_assemble" was used with the command line flags: "-q -p fb ss 150 1000". Assembly size statistics are summarized in Fig. S11 and Table S8.

4.4 Reference-guided assembly of single cell genomes

The cN2-C1 composite genome (see section 4.6 below) was used as a reference genome. Quality trimming was done as described above in section 4.3. Paired-end reads were assembled using clc-assembly-cell-3.2 (CLCbio, http://www.clcbio.com/). Insert length was assumed to be within the range of 150bp to 1000bp (CLCbio program "clc_ref_assemble_long" was used with the command line flags: "-p fb ss 150 1000").

4.5 Genome annotation

Genome annotations of the 96 *de novo* assembled genomes, as well the cN2-C1 composite genome, were done on the RAST server (*46*). Up to 1921 open reading frames, 3 rRNA Genes (1 copy of 5S, 16S, and 23S rRNA genes) and 38 tRNAs were identified per *de novo* assembled genome. See Fig. S11 and Table S8 for assembly statistics.

4.6. Generation of a cN2-C1 composite genome sequence

Since 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 cells (according to their ITS-rRNA). These contigs were selected such that they had large enough overlaps between contigs and that they cover 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. Annotation on the RAST server identified 1971 ORFs, 3 rRNA Genes, and 37 tRNAs.

4.7 Genomic islands in a cN2-C1 composite genome

Genomic island positions were determined by these two steps:
(i) Genome alignment with previously sequenced genomes of high-light adapted *Prochlorococcus* cultures in which island regions have been identified (*4, 22*).
(ii) The gene content, in terms of core or flexible genes, was checked for each predicted island from (i). The set of core genes was defined as all the genes that are shared in the HLII genomes (equivalent to ecotype e9312) within our culture collection, as described in Section 7. If at least 66% of the genes in a predicted island consisted of non-core genes – it was determined to be a real island.
The above steps result in the identification of six island regions within these closely related clades as listed in Table S9.

5. Whole genome similarity analysis

5.1 Whole genome sequence pair-wise distance estimations

Mediator genome reference assembly

We build upon a mediator genome reference assembly approach (*18*) with few modifications that takes into account the partiality of the genomes and the absence of a representative reference genome from the cN2 ITS-rRNA cluster. This method produced a multiple alignment of the 96 partial genomes, letting us analyze the pairwise whole genome sequence variation in positions recovered on each pair of partial genomes.

Single point mutations, insertions and deletions were detected for each assembly at each recovered position on the reference genome. Pair-wise distances between any two of the 96 genomes were then estimated only for positions that had coverage ≥ 2 in both assemblies, using 'p-distance'. Insertions and deletions (Indels) were discarded from the estimation of genomic distances. One of the reasons we discarded Indels was that the number of detected Indels per genome (a few hundred) was not far from the expected number of Indels as a result of errors, as described in Section 5.8.

Note that the method we used to estimate pair-wise distances between genomes tends to somewhat underestimate the real pairwise whole-genome sequence distances, since variable positions on the genomes are less likely to be mapped to the reference genome. Island regions had a lower recovery rate – because they probably have different gene-content and are often not mapped to the composite reference genome, which represents just one arrangement of island gene content. In addition, it seemed that islands are under-represented in the *de novo* assemblies. Two possible explanations for the observed underrepresentation are: (i) a higher DNA fragility at specific sites on these regions or (ii) a higher rate of repeats that limited assembly.

Statistics of the reference-guided assemblies and the genome alignments:
Reference genome length: 1,650,354bp
Mean assembly size: 1.144M bp±0.285M bp (mean±SD). (70% of the genome).
Max assembly size: 1.570M bp (95% of the genome)
Min assembly size: 0.359M bp (22% of the genome)
Median: 1.245Mbp (75% of the genome)
Recovery percentage is estimated assuming a genome size of 1.65Mbp

Basic Statistics of the multi-alignment of the 96 single cell partial genomes:
Conserved sites: 1,193,772bp (72% of the genome)
Variable Sites: 424,125bp (26%)
Parsimony-informative sites: 259,834bp (16%)
Singleton sites: 163,260bp (10%)

5.2 ITS and whole genome tree construction
Phylogenetic trees were generated by MEGA4 (*47*). 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 (Fig. S1). The delineation of C1-C5 clades was highly robust and also observed in trees constructed from genomic position subsets (Fig. S2).

5.3 Identification of dimorphic SNPs between clades
We refer to dimorphic sites as sites that are highly different between two populations of cells and are highly conserved within each of the two populations of cells. Dimorphic sites can be detected by several methods. Here we built upon a method based on mutual information (*48*). For each pair of clades (within the five cN2 clades C1 to C5) the mutual information for each bp position along the genome was estimated. The mutual information of two discrete random variables *X* and *Y* is defined as

$$(5.1) \quad I(X,Y) = \sum_{y\in Y} P(y) \sum_{x\in X} P(x|y) \log \frac{P(x|y)}{P(x)}$$

where y is clade (e.g. *Y*=[C1,C2,…,C5]) and x is the bp value from the alphabet (e.g. *X*=['A','C','G','T']) . To correct for variations in the number of cells within each clade in the samples, we used equal weights $P(y) = 1/n$ (where n is the number of clades). In the identification of dimorphic sites we used $P(y1) = P(y1) = 1/2$. Sites with $I(X,Y) > 0.5$ and p-value<0.01, that were recovered in at least 2/3 of the cells within each of the two clades, were considered as dimorphic. To assess the significance of each position we calculated p-values by comparing to the corresponding mutual information estimation when cells are randomly assigned into clades. To generate random assignments into clades, first, a random pool of cells was created. The pool size equals the total number of cells within the two clade samples. Cells in the pool were randomly sampled from each clade, with repetition, such that each clade contributes equally. This is done to correct for differences in the number of cells within each clade. This pool is then randomly partitioned into clades, keeping the original number of cells within each clade. 10,000 random clade populations were generated and mutual information was estimated, to yield p-values. Last, a correction for multiple hypothesis comparisons was done (FDR (*49*)). Dimoprhic sites per non-overlapping 1000bp along the cN2-C1 composite genome are shown,

for all pairs within cN2-C1 to cN2-C5 clades, in Fig. S3. Note that the smaller number of cells belonging to the clades cN2-C4 and cN2-C5, in our data, limits the significance of their comparison.

5.4 Identification of polymorphic sites within clades

Polymorphic sites within clades were determined based on their entropy. Sites with entropy >0.5 (entropy values were calculated with log base 2) that were recovered in at least 50% of the cells within the clade, were identified as polymorphic (Fig. S4, Table S10). This is roughly equivalent to the case where at least 10% of the cells have bases other than the value of the consensus bp**.**

5.5 Dimorphic and Polymorphic sites between clades cN2-C1 and cN2-C3

We describe in detail the differences between the *cN2*-C1 and *cN2*-C3 subpopulations. These subpopulations differ in 52885 dimorphic single nucleotide polymorphisms (SNPs), which represent 3.2% of the genome (see tree in Fig. 2B). Sites at which these SNPs occur are highly conserved within clades and different between-clades. 74% and 87% of sites that are dimorphic SNPs between C1 and C3 are identical (i.e. 100% conserved) within C1 and C3 respectively. The dimorphic SNPs are scattered along the entire genome (Fig. 3A blue) except for a few regions within genomic islands where there was not enough data to detect sites (Fig. 3A). C1 and C3 dimorphic SNPs occur in 1519 genes out of 1974 genes in the genome – most of them core genes – and 8% are found in intergenic regions (cf. 9% of the genome is non-coding). Of the SNPs within coding regions 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 variation 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 – true for all pairwise comparisons within C1-C5 (Fig. S3,S4). Of the sites that are dimorphic SNPs between pairs, 77%±26% (mean±SD) are identical (i.e. 100% conserved) within clades.

5.6 Determining the set of core genes

In this study we define the set of core genes as genes that appear in all our culture collection genomes within the HLII lineage (equivalent to ecotype e9312). These include the following 13 strains: MIT9311, MIT9314, MIT9401, MIT9301, MIT9312, MIT9107, MIT9201, MIT9321, MIT9202, MIT9215, SB, GP2, and AS9601.

A set of 1463 core genes was identified by the above method (see also Additional Data file S1)

5.7 Allelic variations in core genes

Assessing genomic differentiation of genes based on F_{ST}

γ_{ST} is an equivalent measure of F_{ST} that measures genetic differentiation between subpopulations (*20*). It is widely applied to genomic data of asexual haploid organisms and is based on genomic distance between and within populations (*20, 50*). We used this estimator to assess the sequence differentiation between backbone-subpopulation across all genes (Fig. 3BC, Fig. S12). Qualitatively similar results were obtained based on amino acid sequences rather than nucleotide sequences. Interestingly, median F_{ST} of core genes was higher than that of flexible genes ($p<0.001$, Wilcoxon test), possibly due to a stronger diversifying selection or due to longer "residency" on genomes.

Assessing mutual information with more than two clades

In addition to F_{ST}, we used mutual information to assess the degree of differentiation between the fine cN2 clades C1-C5, in a similar way of the identification of dimorphic SNPs, but applied upon five instead of two subpopulations and on genes instead of bases. Average mutual information per gene within clades C1 to C5, based on nucleotide sequences, for all genes in the cN2-C1 composite genome was estimated. Only genes that appeared in at least three cells per clade, in at least three of the five clades were considered. The genome-wide mutual information for all genes was 0.0519, 0.0733 and 0.0988 (25th, 50th and 75th percentiles).

Mutual information per core gene was 0.0537, 0.0740, 0.0978 (25th, 50th and 75th percentiles). Mutual information per flexible gene was 0.0460, 0.0687, 0.1046 for these percentiles. Although core genes had higher median value, the null hypothesis that the core and flexible genes had equal median mutual information could not be rejected (Fig. S12D). There is a positive correlation between mutual information and F_{ST} ($r=0.34$, $p<0.0001$). It seems, however, that some genes with high F_{ST} have below average mutual information values. This is the case with genes that are relatively highly conserved even between backbone-subpopulations. Since F_{ST} is the ratio between inter-population diversity to whole-population diversity, F_{ST} does not depend on the absolute overall mean distance between sequences in the whole population. Mutual information however depends on the absolute overall mean distances, since it is averaged over all bases of the gene. Thus, genes can have low mutual information while high F_{ST} when absolute overall mean distance is small. Indeed there is a positive correlation between average sequence distance and mutual information ($r=0.6325$, $p<0.0001$), (Fig. S12F).

Qualitatively similar results were obtained based on amino acid sequences rather than nucleotide sequences.

Relation of the observed allelic variation to that found in genomes from cultured strains

For each core gene in the cN2-C1 composite genome three distances were estimated:
D_{RC}: mean sequence distances between previously sequenced genomes and the cN2 C1-C5 sequences.
D_{BC}: mean sequence distances between any two clades within the cN2 C1 to C5 clades.
D_{CC}: mean sequence distances within any of the cN2 C1 to C5 clades.

The mean distances over all core genes were:
$D_{RC} = 0.135 \pm 0.080$ (mean±SD)
$D_{BC} = 0.083 \pm 0.080$ (mean±SD)
$D_{CC} = 0.044 \pm 0.040$ (mean±SD)

D_{RC} was found to have a higher median than D_{BC} (Wilcoxon rank sum test, $P<10^{-10}$).
On average, D_{RC} was almost twice as large as D_{BC} (mean D_{RC}/D_{BC} ratio=~2, median=~1.8). Only 24 genes (1.6% of the core genes) had smaller D_{RC} than D_{BC}.
The above statistics indicate that the majority of core gene alleles of previously sequenced genomes are different from the same alleles found in the five cN2 C1-C5 clades.

The following set of 16 genomes, which includes all the High-Light adapted strains in our culture collection (HLI+HLII), were used to estimate D_{RC} : MIT9311, MIT9314, MIT9401, MIT9301, MIT9312, MIT9107, MIT9201, MIT9321, MIT9202, MIT9215, SB, GP2, AS9601, PMED4, P9515, RCC278.

5.8 Assessing the estimated error rates of single cell genomics

We show that the overall error rate of single cell genomics i.e. – the cumulative error of MDA errors, sequencing errors and assembly errors – is about ~0.0001 errors per bp (equivalent to ~10bases per 100Kb). This estimated error rate is based on experimental evidence as well as literature – as described in the details below. This is two orders of magnitude smaller than the average variations we observe between the *Prochlorococcus* single cells in our samples.

Control experiment based on single cell sequencing of *E. coli* K-12 EcNR2 clonal cells from a single colony.

We performed single cell whole genome amplification, sequencing, and assembly on eight replicate *E. coli* cells processed following Rinke et al (*42*). Briefly, individual cells were collected from a liquid culture inoculated from a single colony of *E. coli* EcNR2 and grown overnight in LB media at 30°C. Cells were stained with 1X SYBR Green I DNA Stain (Invitrogen) and sorted with an Influx cell sorter (BD Biosciences) based on their side scatter and DNA fluorescence (531nm) characteristics following excitation at 488nm. Cells were lysed and amplified as described in section 2, with the exception that cells were sorted into an initial volume of 0.9uL of Tris-EDTA and amplified once using a final volume of 15uL. Standard paired-end Illumina libraries (2 x 150bp) with an average insert size of 290bp were prepared from sheared amplified DNA, and sequenced on the Illumina HiSeq 2000 platform by the DOE Joint Genome Institute.

The same assembly program CLCbio version 3.2 and the same assembly command line flags (as described in section 4.4 above) were used to perform reference-guided assembly (using the *E. coli* MG1655 K-12 EcNR2 genome as a reference).

Variations between the single cell genomes and the reference genome were minimal. Average recovery was 1655±505 (mean±SD) Kb per cell (using a coverage threshold of C=2, e.g. a site is considered as recovered if it was mapped by two or more reads – identical to the threshold we used in our analysis of the *Prochlorococcus* genomes). We observed 3.7±0.7 (mean±SD) substitutions per 100Kb and 1.3±0.3 (mean±SD) insertions/deletions per 100Kb as summarized in Table S11. We then evaluated the pairwise genomic differences per bp between all cells, in exactly the same way we did for our *Prochlorococcus* cells. The average genomic difference between any two *E. coli* cells was 0.000051±0.000014 (mean±SD) which is equivalent to ~5 substitutions per 100Kb - as summarized in Table S12. A phylogenetic tree and the distribution of pairwise genetic distances are described in Fig. S13 A,B. We note that two rounds of MDA for *Prochlorococcus*, could, in the worst case, increase the error rate by a factor of two (i.e. to yield an average genomic difference of ~0.0001 per bp).

Since these cells are clonal and are assumed to have identical genomes the above results can serve as an estimation (of the upper bound) of the cumulative error rate in the overall single cell sequencing process - i.e from MDA, sequencing and assembly.

As mentioned above, this error rate is more than two orders of magnitude smaller than the variations we observe within the *Prochlorococcus* single cells in our samples. We observe, on average, 3500 bp substitutions per 100Kb between individual cell genomes, and 4700 bp substitutions per 100Kb between subpopulations. This suggests strongly that the differences we observe are biological differences and are not due to errors in MDA or single cell sequencing.

Additional evidence:

In addition to this experimental evidence, others have reported on error rates one can expect from single cell genomics, and they are very close to what we observed in the *E. coli* control experiment described above: about 10^{-4} errors per bp. More specifically, Rodrigue et al. (*11*) sequenced two putatively identical *Prochlorococcus* cells (MED4) from the same culture. The two single cells genomes were found to be different in ~20bp per 100Kb (representing error rate of ~$2x10^{-4}$). Nurk et al (*51*) estimated error rates from one *E.coli* single-cell that was illumina sequenced and assembled with the same assembly program we used (CLCbio). Their reported error rate is strikingly equivalent to the one we observed with our *E. coli* control experiment, i.e. ~5 differences and ~3 indels per 100Kb. Finally, Pamp et al. (*52*) sequenced five single-cells of the intestinal symbiont *C. arthromitus* from different fine filaments of the intestine of an individual mouse. They reported a total of 1287 base substitutions in these five 'almost identical' cells from different filaments (genomes size is ~1.5 Mb). They claim that the observed substitutions are mostly biological differences and not errors in the MDA or sequencing. In their Suppl. Mat., they make clear theoretical arguments as to why these observed differences are unlikely to be sourced from errors (see below).

Theoretical arguments

One final bit of evidence that supports the biological origin of the variation we see in wild *Prochlorococcus* is the pattern of the variation. As explained in (*52*) and in section 6.3 below, variations from errors are expected to be uniformly distributed along all genome positions and the number of mismatches per Kb should follow a Poisson distribution, and should not be clustered. Indeed the variations within the *E. coli* data follow a Poisson distribution (chi-square goodness of fit; $P<0.05$) and no apparent clustering is seen (Fig. S13 C,D,E). In addition, if error rates are so small, the errors should not overlap between different single cells. More specifically the observed distribution of the number of positions with variants that appear in one, two, three cells etc. is equivalent to a binomial distribution with the same number of variants. This is indeed the case with the *E. coli* data (1-sided Binomial test, $P<0.05$). In fact, there are 330 sites with mismatches, 328 appear in one of the eight cells and two appear in two cells. In contrast, the variations we observe in *Prochlorococcus* are significantly different: they are clustered and correlated. This is described in detail in section 6 where we talk about signatures of selection.

Closest single-cell genomes in our samples and detection limit

The closest pairs of individual cell genomes within our samples differ in between 300 to 500 bp substitutions (20-30 bp substitutions per 100Kbp). This rate of substitution is slightly higher than the expected error rate of 10^{-4} errors per bp. The *distribution* of substitutions along the genome in these pairs, however, indicates that the differences are not all from errors and that at least part of them are real (Their frequency per 1Kb has higher variance than a similar distribution expected from random errors, Two sample F-test, $P<0.001$, see also 6.2 below).

Given the combination of the vast diversity and the strong physical mixing (described in detail in SM section 10), it may not be that surprising that within the extent of our sampling, of few hundreds cells per sample, we did not detect cells with identical genomes.

6. Signatures of selection

6.1 Overview

Is the differentiation between genomic backbones we have observed in *Prochlorococcus* a product of selection? To try to answer this question, we compared the observed sequence variation patterns to those obtained from coalescent simulations of a neutrally evolving genome sequence (*53-55*) (assuming a single constant-sized population - a reasonable assumption for *Prochlorococcus* evolution, See section 6.3 below) and find that the distributions of both dimorphic and polymorphic SNPs are qualitatively and quantitatively different from those simulated (Section 6.3), indicating that selection likely acted differently on different genomic regions. In addition, in contrast to simulations of neutral evolution, both SNPs classes tend to cluster in the genome (Fig. S16), possibly due to co-selection of gene cassettes and/or adaptive hitchhiking (*54*). Notably, the observed per-gene F_{ST} distribution is significantly different from those obtained by simulation (*54*) (Kolmogorov-Smirnov test, $p<10^{-10}$). It has significantly more genes with very high F_{ST} ($F_{ST}>0.95$) – likely a result of diversifying selection, and a long tail of genes with low F_{ST} ($F_{ST}<0.5$) – at least in part reflecting negative selection (Fig. 3B). An additional signature of selection is the presence of highly polymorphic genomic regions within one backbone that are highly conserved in all other backbones - again in contrast to simulations - possibly indicating clade-specific selection pressures (or their absence within a specific clade) (Fig. S4). Lastly, we applied a complementary method (*56*) that is free of the demographic assumptions made in the coalescent simulations, and find that different functional classes of single nucleotide sites, e.g. intergenic and genic positions, have statistically different genome-wide F_{ST} distributions – indicating that selection acts differently on each of these classes of nucleotides (*56*) (Fig. S17 and section 6.4 below).

6.2 Coalescent simulations of neutral evolution

We used coalescent simulations (*53, 57*) to obtain genome-wide distributions of dimorphic SNPs between clades, polymorphic SNPs within clades and population differentiation between clades (F_{ST}) under a selectively neutral model, following Akey et al (*54*), with adjustments to fit to *Prochlorococcus* evolution. The coalescent simulates the evolution of the largest ecotype e9312, which account for almost 90% of *Prochlorococcus* cells in the upper 200m of the Atlantic Ocean. We assumed a single population, a constant population size and no recombinations. The simulated genomes were identical in length to the cN2-C1 composite genome (1,650,354 bp) and with a similar GC content (32%). Sample size was 96 – as in our real sample (i.e. 96 single cell genomes).The scaled mutation rate (Θ) was determined such that the simulated average pair-wise sequence distance (D_{all}) between genomes was similar to the observed D_{all} (see Fig. S14).

To compare population differentiation between the simulated genomes and the observed genomes, we clustered the resulted simulated genomes based on pair-wise genomic distance, to obtain the same number of clusters (subpopulations) as in our real data (Fig. S15). The simulated genomes were clustered to no more than seven clusters and then the five largest clusters were considered for the analysis (for a comparison to the observed results of the five cN2 clades C1 to

C5). Polymorphic and dimorphic sites were determined in the same way as described for the observed data. To compute the genome-wide per gene distribution of F_{ST}, the positions of genes on the chromosome in the simulated genomes were identical to their positions in the cN2-C1 composite genome.

A Θ value of 0.05 yielded average π values that are similar to the observed data (Fig. S14). This choice of Θ yielded nearly maximal levels of median genome-wide F_{ST} values (gene-by-gene F_{ST}). In fact, as can be seen in Fig. S14, there is no choice of Θ that yield median F_{ST} values as high as in the observed real data.

We note that the estimated Θ value, to yield comparable average pair-wise genomic distances to the observed one, was much smaller than our estimation of Θ in *Prochlorococcus* based on the consensus population size. See more in Section 11.
A Θ value of 0.05, which produced the same average pair-wise genetic distance as in our observed data, correspond to $\sim 2.5 \cdot 10^8$ generations to the most recent common simulated ancestor (assuming $\mu = 10^{-10}$ mutations per base per generation). This is the mean number of generations that, under a neutral evolution, results in the same amount of nucleotide diversity as in our observed data. This number of generations is equivalent to $\sim 10^6$ years (assuming a generation time of ~1day).

For the above coalescent simulations we used the Richard Hudson "MS" software tool (*58*) (http://home.uchicago.edu/rhudson1/source/mksamples.html) to generate random genealogies. We then used the molecular sequence simulator software tool "seq_gen" (*59*) to generate the neutrally evolved genome sequences (using the command line flags ' –mHKY -f 0.34 0.16 0.16 0.34'; the second flag was used to yield mean GC-content of 32%).

We compared between simulated and observed distributions as follows:
Dimorphic SNPs density profile along the chromosome was calculated per 1000bp (non-overlapping windows). The distribution of dimorphic SNPs for each pair of clades (clusters), for each simulation was then evaluated (similar to Fig. 3A in the main text). The empirical distributions of dimorphic SNPs density were found to be similar to Poisson distributions (chi-square goodness of fit; $P<0.001$). The observed distributions did not resemble Poisson distributions and had much higher variance (Two sample F-test, $P<0.001$). Comparisons between the observed and simulated genome-wide gene-by-gene F_{ST} distributions were performed by the Kolmogorov-Smirnov test, all yielded $p<10^{-10}$. See also Fig. S15, S16.

6.3 Comparison of F_{ST} distributions of different single-nucleotide genomic classes
We built upon the analysis of Barreiro et al (*56*) and examined the genome-wide distributions of different genomic positions: Intergenic sites (i.e. non-coding sites), Genic sites (i.e. coding sites), codon first base, second base and third base. The distributions were found to be significantly different (Fig. S17). The fraction of positions with very low F_{ST} values ($F_{ST} <0.05$) was significantly different between all pairs of classes (chi-square, $P<0.0001$). The fraction of genic positions with very low F_{ST} was smaller than the corresponding fraction of intergenic positions. This may be interpreted as negative selection globally reducing population differentiations at genic regions, especially at the first and second codon bases. On the other hand the fraction of

positions with very high F_{ST} (F_{ST} >0.95) was significantly different between all pairs of nucleotide classes (chi-square, P<0.0001).

6.4 Additional notes on identifying signatures of selection

Most classic population genetics theories have not been developed with huge mostly-asexual populations, with a large population-scaled mutation rate $N \cdot \mu >> 1$ (here N is the census populations size, not the 'bounded' estimated-from-data effective population size), in mind, or to deal with adaptation from standing genetic variation (see also section 11). Our choice of methods for identifying signatures of selection was carefully determined. We chose not to use methods based on dN/dS, as we believe their interpretation could be questionable in the context of our data (*60*) due to two reasons: (i) It is not clear if cells in our samples represent a single population or evolutionary independent lineages for the sake of the dN/dS statistical analysis; this may have significant implications on the interpretation of the analysis - as nicely explained by Kryazhimskiy and Plotkin (*60*). (ii) Synonymous substitutions might not be neutral in *Prochlorococcus* evolution (or at least not all of them can be considered as neutral). For example, synonymous substitutions may influence internal codon preference or may affect DNA or RNA structural properties (*61, 62*). Since even weak fitness differentials may play a role in *Prochlorococcus* evolution (as proposed in the main text) we think dN/dS methods could be misleading here. We hope that the growing availability of single cell genomics data will invite the development of a population genetics theory that will be more directly applicable to free-living bacterial species.

7. Ortholog clustering and gene content analysis

Clusters of Orthologous Genes

Genes were classified into Clusters of Orthologous Genes (COGs) using the pipeline described in (*63*). Genes from previously sequenced *Prochlorococcus* as well as all genes from the 96 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. We note that due to the partiality of the genomes, the high number of contigs (resulting in many partial sequences of genes), and the high sequence diversity, these clusters are not perfect and we had to manually check any result that were based on the gene clustering analysis.

Detection of differential gene sets between backbone-subpopulations (or clades)

Genes that were candidates to be "differential" between clades (i.e. appear in one or more clades but absent from the other clades) were selected by the following 3 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 HLII core genes (genes that appear in all the culture HLII genomes).
   (ii) Genes that appear in less than 3 cells in total or in more than 50 cells in total.

3. Apply ‘hierarchical clustering’ to genes according to their presence/absence in the 96 partial single cell genomes.

Steps 1 and 2 resulted in a set of 404 genes. The genes were then clustered using standard hierarchical clustering, using ‘hamming distance’ and ‘complete’ linkage in Matlab (Fig. S5). Based on this clustering analysis and on multiple alignment of the annotated partial genomes, gene cassettes (or part of cassettes) that were present in one or more clades but absent from the others were identified (Table 1, Table S1).

Detection of gene cassettes shared by a few closely related cells (subclades) within backbone-subpopulations

Several gene cassettes were found to be shared by a small number of closely related cells within backbone-subpopulations, and not by other cells. For example a cassette with Type II secretion and type IV pilus genes was identified in 8 cells forming a subclade within cN2-C1 (named C1a; include the cells:518D8, 527P5, 528K19, 521B10, 521O20, 519O11, 527L16, 495N16) see Table S13 and Fig. S5. Interestingly one of the cells in the C1a subclade (518D8) was flow-sorted attached to a gamma-proteo bacterium (the two cells were physically attached and DNA from both cells was amplified and sequenced).
Other examples of gene cassettes associated with specific subclades are listed in Table S13.

Predictions of the position on the genome of the differential genes (Table 1, Table S1)

To predict the likely position of genes we performed the following steps:

1. The *de novo* assembled partial genome of each cell was aligned to the cN2-C1 consensus genome using mauve (*64*).
2. For each gene for each cell
    a. If the contig is aligned then get position from alignment.
    b. Else (not aligned)
    c. Try to find if other parts of the contigs are aligned
    d. If yes, then use these as anchors and predict gene position by extrapolation.
    e. If not, gene position is not predicted.

Predictions of gene-content similarity between pairs of nearest-neighbor cells

Pairs of annotated single cell nearest-neighbors genomes were aligned using mauve (*64*). Pairs were determined as non-identical if there is a difference in gene content in aligned contigs. When a whole contig was not aligned to any contigs in the other cell, no violation to the identical gene-content test were considered. Because these are partial genomes, pairs of cells could only be predicted to have identical gene-content. Three pairs of sister cells were determined as possibly having identical gene content: (i) B241_529J11_C1 and B245a_518E10_C1 (see also Table S13); (ii) B241_527L16_C1 and B243_495N16_C1; (iii) B245a_521N3_C1 and B241_528N8_C1. Note that the pair-wise genomic distances of these pairs was among the smallest in our dataset (between 300 to 500 substitutions across the entire genome, based on the mediator reference-guided assembly method described in section 5.1).

Genome synteny

The exact nature of genome synteny could not be determined since the genomes are partial and each is composed of hundreds of contigs. Never the less, multiple alignment of the partial

genomes tells us that they are broadly syntenous although the synteny is often broken within genomic islands (Fig. S21).

8. Genomic comparison of populations between samples

The whole genome population differentiation estimator F_{ST} (*20*) did not show population differentiation between cN2-C1 populations between samples when applied on whole genome distances (pairwise comparison, P>0.05). We could not assess this test to the other clades (i.e. C2 to C5), due to the small number of sequenced-cells from each one of them.

A weak signal of changes in allele frequencies, over the seasons, was observed in a small number of genes (Fig. S18). This could hint at selection for specific alleles of these genes. As discussed later in Section 13 we predict that a change in allele frequency is a possible adaptation strategy over ecological timescales.

9. Estimating the number of backbone subpopulations and their relative abundances

99% clusters were the best match to backbone-subpopulation clusters defined with whole genomes

Clusters of ITS-rRNA at the level of 99% sequence similarity were the best match to backbone-subpopulation clusters defined with whole genomes. ITS-rRNA clusters at the level of 99% sequence similarity were decided by Mothur (*65*), with manual verification within the cN2 C1-C5 backbones. To assess standard errors of 99%-ITS clusters abundance, the relative abundance was calculated for each of the four 384-well plates (cells from each sample were flow-sorted into four 384-well plates, thus we treated each of the four plates as a sample replication) and then bootstrapped by 1,000 resampling from each plate with replacement (Fig. 4A in the main text). To assess whether a backbone subpopulation significantly changed in abundance between samples, samples where pooled and then randomly assigned into samples (10,000 times) to compute p-values. Multiple hypothesis correction was done by FDR (α=0.05). To assess the significance of differences in the relative abundance changes among samples between each pair of backbone subpopulations we normalized the relative abundance by the mean relative abundance among samples and define the distance D_{ij} between normalized relative abundance profiles of subpopulation i and subpopulation j as:

$$(9.1) \quad D_{ij} = \sqrt{\sum_{k=1}^{S}(P_{ik} - P_{jk})^2}$$

where P_{ik} is the normalized relative abundance of subpopulation i in sample k and S is the number of samples. To assess the significance (p-values) of the pairwise distances D_{ij} we compared this measure to the same measure but with random assignments to samples, in a similar way as described above (10,000 times). Multiple hypothesis correction was done by FDR (at α=0.05). 17 out of 55 pairs within the 11 largest backbone- subpopulations (whose profiles presented in Fig. 4A main text) were found to have significantly different abundance profiles over the seasons (reflected by significantly higher D_{ij} than expected with randomized samples). The equivalent mean number of pairs with different abundance dynamics expected by chance is <1.

The rarefaction curves in Fig. 4B in the main text were generated by Mothur at the level of ITS-rRNA sequence similarity of 99%.

10. Estimation of the population size of *Prochlorococcus* that becomes well-mixed within ecologically relevant time scales

Estimating the distance between two 'just-divided' daughter cells over time

Prochlorococcus cells are non-motile, neutrally buoyant, and do not form aggregates. Therefore, the dispersal of single cells is dictated by Brownian motion at early times (order of seconds) and by ambient fluid motion at longer times. Here we focus on the latter, since the former only dominates for very short timescales. Importantly, only relative fluid motion (i.e., differences in fluid velocity) matters for dispersal (*66*), because a uniform flow transports cells without changing their relative distances. A dominant source of relative fluid motion in the ocean is turbulence, which entails velocity differences between different points in space. Since non-motile, neutrally buoyant cells cannot move relative to the fluid, those same velocity differences govern the separation between cells. As a consequence, any two cells tend, on average, to separate over time.

A fundamental length scale in turbulence is the Kolmogorov scale (*67*), η, the scale at which the kinetic energy transferred down from larger scales by inertia balances the dissipation of energy by viscous forces. Typical values of the Kolmogorov scale in the upper ocean are η = 1-5 mm. At scales smaller than η, turbulence reduces to laminar shear, where the velocity difference between two points, u_S, simply increases linearly with their separation distance, d, and with the magnitude of the fluid velocity gradient, γ, as (*68*):

(10.1) $u_S = 0.42\gamma d = 0.42(\varepsilon/\nu)^{1/2} d$ (for $d \ll \eta$),

where $\nu = 10^{-6}$ m^{2}s^{-1} is the kinematic viscosity of water and $\gamma = (\varepsilon/\nu)^{1/2}$ is the Kolmogorov shear rate. At separation distances d larger than η, the velocity difference between two points, u_L, scales with the one third power of the energy dissipation rate, as (*68*):

(10.2) $u_L = 1.37(\varepsilon d)^{1/3}$ (for $d \gg \eta$) .

We can use these expressions for the separation velocities to compute the separation distance of two *Prochlorococcus* cells over time. First we ask how much time it takes for two 'just-divided' cells to be at a distance greater than η (see also ref. (*69*)). This time is obtained by integrating the inverse of the velocity in Equation 10.1 over the separation distance, from the initial separation distance (taken to be the cell diameter of *Prochlorococcus*, D = 0.6 μm) to the Kolmogorov length scale:

(10.3) $T_S = \int_D^\eta \frac{\mathrm{d}x}{u_S(x)} = \int_D^\eta \frac{\mathrm{d}x}{0.42(\varepsilon/\nu)^{0.5} x} = \frac{1}{0.42(\varepsilon/\nu)^{0.5}} \ln(\frac{\eta}{D})$

Typical values of ε in the ocean range from 10^{-8} m^{2}s^{-3} below the mixed layer to 10^{-6} m^{2}s^{-3} within the mixed layer. For these values, one obtains $\eta = (\nu^3/\varepsilon)^{1/4}$ = 3.2 mm and 1 mm, respectively, resulting in T_S = 204 s below the mixed layer and T_S = 18 s within the mixed layer. Therefore, two 'just-divided' cells will be separated by a distance larger than the Kolmogorov

scale within a few minutes at most, and from that point on, the distance between them is prescribed by Equation 10.2.

We can apply the same approach to compute the separation time once cells are in the second regime (separation distance larger than η), obtaining:

(10.4) $T_L = \int_{\eta}^{L} \frac{dx}{u_L(x)} = \int_{\eta}^{L} \frac{dx}{1.37(\varepsilon x)^{1/3}} = \frac{1}{2/3}\frac{1}{1.37\varepsilon^{1/3}}\left(L^{2/3} - \eta^{2/3}\right) = \frac{L^{2/3}-\eta^{2/3}}{0.91\varepsilon^{1/3}}$,

which can be solved to obtain the separation distance L after a given time T_L:

(10.5) $L = (0.91\varepsilon^{1/3}T_L + \eta^{2/3})^{3/2} \sim 0.87\varepsilon^{1/2}{T_L}^{3/2}$ (for $L >> \eta$)

For $\varepsilon = 10^{-8}$ $m^2 s^{-1}$, the separation distance estimated by this approach will be L ~ 19 m after T_L = 1 hour; L ~ 2.2 km after T_L = 1 day; and L ~ 41 km after T_L = 1 week.

At long times, these values of L might be an overestimate of the actual separation distance and the actual values of L may be an order of magnitude smaller. This can be seen by considering results from tracer dispersal experiments in the ocean. Because *Prochlorococcus* cells can be assumed to behave as passive tracers (they are non-motile and neutrally buoyant), we can empirically estimate their dispersal also by using dispersion coefficients obtained experimentally for patches of tracers injected in the ocean. This is done through estimation of the variance σ_{rc}^2 of the patch radius (a proxy for the area of the patch) after a time t following point-injection of the tracer (*70-73*). Observations resulted in the empirical relation $\sigma_{rc}^2 = 0.0108t^{2.34}$, where σ_{rc} is in cm and t is in seconds (see Fig. 1 in ref. (*70*)). This relation yields $\sigma_{rc} = 6$ km after a time t = 1 week.

Therefore, both estimates indicate that two 'just-divided' *Prochlorococcus* cells will be separated by at least a few kilometers over the course of a week, for typical turbulence conditions in the upper ocean. We conclude that a conservative estimate is that over the course of one week, cells in a horizontal area of 3 km by 3 km in the upper ocean are well mixed.

To summarize, in characteristics upper ocean water' just-divided' *Prochlorococcus* cells will not be within the same milliliter of water within minutes, will be tens of meters apart within one hour, and will be a few kilometers away within a week.

Vertical dispersion

The above analysis addresses horizontal separation. The vertical separation d between cells after a time t can be estimated from the relation $d = (2K_V t)^{1/2}$, using empirically measured values of the vertical dispersion coefficient, K_V. Within the mixed layer, typical values of the vertical dispersion coefficient are on the order of $K_V = 10^{-2}$ m^2 s^{-1}. Over the course of one week, this results in a vertical separation of ~100 meters, implying that mixing spans the entire depth of typical mixed layers. Below the mixed layer, $K_V = 10^{-5}$ m^2 s^{-1} and the separation distance over one week is ~3.5 m.

Estimating the census population size within a well-mixed water parcel

A conservative estimation is that the population in a water parcel of 3km x 3km x 3m can be considered as well-mixed over a week time, which translates to a total of more than 10^{17} *Prochlorococcus* cells per such water parcel. To get lower bounds of the number of cells of a backbone-subpopulation in such a water parcel, let us assume there are hundreds of backbone-subpopulations with a minimal relative abundance of 10^{-4} (equivalent to assuming there is at least one cell from each subpopulation per 1mL). Thus there are at least $3x10^{13}$ cells from each backbone-subpopulation in such a water parcel. For the more abundant clades and within the mixed layer, these numbers are probably larger ($>10^{15}$ cells).

Estimating lower bounds of evolutionary relevant census population size (N)
Prochlorococcus populations are stable, with minimal annual and inter-annual density $>10^4$ cells/ml. Assuming no significant population bottlenecks, we can estimate N to be equal to the minimal annual census population size. We note that High-Light adapted populations spend part of the year below the mixed layer (in summer when the mixed layer is very shallow). At these times the populations are more stratified. Thus, mixed water parcels in the summer are effectively smaller in the vertical dimension than in the winter time, when they are effectively mixed over ecological time scales, across the whole mixed layer. Since we are interested in a conservative estimation of N, we consider the 'below mixed layer' vertical range of ~3m as well-mixed. The smaller vertical range is however compensated, to some degree, by a larger census population size in summer ($\sim10^5$ cells/ml as opposed to $\sim10^4$ in winter). To conclude, a conservative estimate is that of $N > 3x10^{13}$ cells for most High-Light-adapted clades, in particular the ones investigated in this study.

11. Estimating 'effective population size' and its evolutionary consequences

Why is it hard to estimate effective population size for *Prochlorococcus* populations?
The 'effective population size' (*74*), N_e, is defined as the size of an imaginary, theoretically ideal population affected by genetic drift at the same rate per generation as the population being studied. Estimating N_e is hard in general and is even harder in the case of *Prochlorococcus*. There are two main reasons for this difficulty:

1) The huge census population size suggests a very large N_e. So large that N_e may well be much larger than the number of generations to the most recent common ancestor (MRCA) of all High-Light adapted *Prochlorococcus* cells in the oceans. Estimation of N_e is commonly done using coalescent theory (*53*). In a coalescent the mean number of generations to MRCA is $\sim N_e$. If the number of generations back to the MRCA is much fewer than N_e generations, then the coalescent does not describe the situation well. Let us assume the MRCA of all High-Light adapted cells was alive 100 Million years ago (a reasonable assumption). This is equivalent to $\sim 2 \cdot 10^{10}$ generations, assuming ~200 generations per year. In a situation where $N_e > 2 \cdot 10^{10}$ the estimation of N_e from a coalescent will yield a smaller N_e than the real one.
2) It is reasonable to believe that due to the large N_e and a streamlined genome there are very few truly neutral positions on the genome (since a large N_e result in selection of even weak fitness differentials). Synonymous sites are commonly used for the estimation of N_e, but they are unlikely to be truly neutral (see section 6.4). Thus, using nucleotide diversity (π) based on synonymous sites, likely underestimates π. It could be that the real 'neutral' divergence is in fact saturated. Saturation is a situation where most neutral positions have mutated more than once. The nucleotide diversity at equilibrium depends on what assumptions are made, and

can range anywhere between π ~0.1 to 0.5 depending on codon bias, GC content, amino-acid content and other factors (*75*).

We estimate below lower bounds of N_e based on π. Since π values may be close to realistic saturation values, N_e is likely larger than could be estimated from nucleotide divergence. We believe the lower bounds may be even a few orders of magnitude smaller than the real N_e.

Estimating lower bounds to N_e based on nucleotide diversity (π)

A common method is to estimate N_e based on 'neutral' nucleotide diversity (π). In the absence of a better way to estimate N_e from the genome sequences, we estimate here lower bounds to N_e based on nucleotides divergence of non-conserved third codon positions (167562 positions) as an approximation for synonymous sites. These are likely not 'true' neutral positions and thus, π values from 'true' neutral sites, are probably higher. Assuming a constant population size and a known constant mutation rate (another assumption that has to be taken with care) one can estimate N_e by $N_e = \frac{1.5\pi}{\mu(3-4\pi)}$ as described in Lynch and Conery (*76*). Using this approach we get π=0.216 which gives us $N_e\mu = \frac{1.5\pi}{3.5\pi} = 0.1516$. Assuming $\mu = 10^{-10}$ mutations per bp per generation we get lower bounds of N_e= ~$1.5 \cdot 10^9$. Note that this is a lower bound of N_e for the e9312 ecotype. Lower bounds for the whole *Prochlorococcus* species should be larger.

A reasonable estimation of the real N_e

Several factors are known to decrease the ratio between N_e and the census population size (*77*) including: large variation in offspring number, age and stage structure, and factors common in sexually reproducing organisms (e.g. division into two sexes, inbreeding). Since the above factors do not play role in the evolution of *Prochlorococcus*, and because it is also reasonable to assume no major bottlenecks in population size (though this is hard to prove as we simply do not know) it is realistic that N_e values are in fact much closer to the census population size than to the lower bounds calculated from the data. It is thus reasonable to assume that N_e of each backbone-subpopulation is much closer to the census population size of at least 10^{13} cells than to the lower bounds estimated from nucleotide divergence.

We suggest that *Prochlorococcus* is likely the organism with the largest N_e on the planet.

Evolutionary consequences of a very large N_e

The huge N_e together with a mutation rate (μ) that is commensurate with other bacteria (*78*) (~10^{-10} mutations per bp per generation), and a streamlined genomes size (1.5 to 1.8Mbp) imply that adaptation mostly occurs from standing genetic variation (*27, 79*), is not mutation limited (*80*), and is probably characterized by "soft genetic sweeps" (*80, 81*) and clonal interference (*82-84*) in which independently generated adaptive mutations rise in frequency simultaneously. Future work is required to better understand the exact mechanisms and timescales of adaptive evolution in wild *Prochlorococcus* populations.

12. Homologous recombination

BratNextGen software (http://www.helsinki.fi/bsg/software/BRAT-NextGen/) was used with default settings to detect recombination (*85*) from the 96 reference-guided assemblies. The learning algorithm was run for 20 iterations, and the statistical significance of the recombinations

(p=0.05) was determined using permutation sampling with 50 replicate analyses, which were run in parallel on a computing cluster. This approach has previously been used to detect recombinations in *Staphylococcus aureus* data (*86*) and in (*85*) it was shown to yield almost identical results with the analysis of the *Streptococcus pneumonia* data from (*87*).

On average a total of 13737±14000 bp (mean±SD) per single cell genome were predicted to be acquired by recombinations, reflected in 9.3±2.5 (mean±SD) recombined stretches of DNA, (Fig. S19).

Only a small fraction of the dimorphic SNPs between pairs of the five cN2 C1-C5 clades, coincides with positions detected as recombined. For example, only 15% (2028 bases) of the 13,437 dimorphic SNPs between C1 and C2 coincide with a position detected as recombined in at least one cell in C1 or C2 population samples; 6.7% (3580 bases) of the 52,885 dimorphic SNPs between C1 and C3; and 4.1% (1520 bases) of the 36874 dimorphic SNPs between C2 and C3 (Fig. S20). Thus, the majority of the observed dimorphic SNPs likely originated by mutation and not recombination. As a comparison, in (*87*) a total of 57736 SNPs were identified in 240 *Streptococcus pneumoniae* isolates, 50720 (88%) of which were predicted to be introduced by 702 recombination events.

Only a small fraction of polymorphic sites within clades are identified as recombined (see Table S10); therefore, homologous recombination does not seem to be the main mechanism to explain the cohesion of backbone-subpopulations.

13. Estimation of lower bounds of adaptation times

To estimate lower bounds of adaptation times we assume a simple logistic growth model (*88*) with some maximum carrying capacity. We assume the population is composed of a wild type and a mutant. The relative abundance of the mutant at time t is p. The change of p over time, assuming the mutant has a fitness advantage s is described by:

(13.1) $\frac{dP}{dt} = sp(1-p)$

The solution for this equation is:

(13.2) $P(t) = \frac{P_0}{P_0+(1-P_0)e^{-st}}$

We can estimate the time it takes a mutant with initial relative frequency P_0 to reach a significant fraction of the population (say 50%), and get:

(13.3) $T_{50} = -\frac{1}{S} ln \frac{P_0}{1-P_0}$

Lower bounds for the time of establishment of new *de novo* mutations

A new *de novo* mutation that did not exist in the population has initial frequency $P_0 = \frac{1}{N_e}$ assuming a conservative value of $N_e = 10^{13}$ will estimate $T_{50} \sim -\frac{1}{S} lnP_0 = -\frac{1}{S} ln 10^{13} \sim -30\frac{1}{S}$. That means that it takes a new mutant with selection advantage of 10% (which is a huge selection advantage) around 300 generations (>1 year) to reach 50% of the population. With more realistic s values of ~1% the establishment takes 3000 generations (>10 years).

Note these estimations are lower bounds for the estimation of time to reach 50%, because (i) we assume there are only two equi-fitness types in the population while in real *Prochlorococcus* populations there are many more equi-fitness types (ii) Conditions are assumed to not change after time t=0. (iii) We assume no other mutations are introduced after time t>0.
We therefore conclude that a new *de novo* mutation is unlikely to be established over ecological timescales (e.g. over seasons - tens of generations).

Lower bounds for the time of establishment of new acquired gene or a gene cassette
Let us assum the gene is acquired by just one cell in the population, and that it confers a selective advantage s. This case is equivalent to the behavior of a *de novo* mutation (assuming it is not rapidly transferred horizontally to other cells in the population). Thus lower bounds for s=1% is ~10 years.

Lower bounds for the time of establishment of a standing mutation
Assuming an initial frequency of a standing mutation is $P_0 = \frac{1}{\mu}$ we get an establishment time of $T_{50} = -\frac{1}{S} ln \frac{P_0}{1-P_0} = -\frac{1}{S} ln \frac{\frac{1}{\mu}}{1-\frac{1}{\mu}} \sim \frac{1}{S} ln\mu \sim 23\frac{1}{S}$ assuming a mutation rate of $\mu = 10^{-10}$ mutation per bp per generation. This is not very different from the T_{50} of the establishment of a *de novo* mutation. Unless a standing mutation has a very strong selective advantage it will take at least hundreds of generations to establish (>1 year).

A possible strategy for adaptation over seasonal timescales
An important consequence of the above analysis is that only mutations with a large initial frequency P_0 can be established over seasonal timescales of tens of generations assuming ecologically realistic s values. For example if $P_0 = 0.1$ and s=0.1, we get $T_{50} \sim 20$ generations. This suggests a design principle that allows fast response of populations to environmental changes that occur within tens of generations – through shifts in allele frequency. This principle is also valid if the selected entities are clades instead of alleles. Thus, populations can respond to rapid environmental changes, such as seasonal changes, through shifts in the relative abundance of clades that have different fitness in different seasons. Our data suggest that adaptation over seasonal timescales in *Prochlorococcus* is mainly achieved through such shifts in the relative abundance of clades – as observed in the change in the relative abundance of backbone-subpopulations over the seasons. There is only a weak signal of a change in allele frequency in a few genes between seasons (see Fig. S18 and section 8 above).

14. Estimation of backbone-subpopulations divergence times
Estimating divergence time for prokaryotes can be challenging. Here we try to give rough estimation of the divergence times between the backbone-subpopulations we observed in our data.

Estimation based on sequence divergence along branches
In the absence of 'true' neutral positions at hand we based our analysis of sequence divergence on non-conserved third base codons (as described in section 11). The cN2 C1-C5 clades show divergence of d=~0.2 substitutions per bp (excluding cN2-C2 that is more closely related to cN2-C1). Since these values are in a range that could be within saturation (*75*) we can only estimate

lower bounds of divergence times. Saturation is a situation where too much time has passed from divergence and most neutral positions have mutated more than once.

Assuming a constant mutation rate of 10^{-10} mutations per base-pair per generation (*78*) it is possible to estimate the total branch length between two leaves in a phylogenetic tree by $T = d/2\mu$ where d is the estimated number of mutations that have occurred on the branches, μ is the mutation rate per bp per generation and T is the number of generations from the most recent common ancestor. Based on these assumptions the cN2-C1 to cN2-C5 clades likely diverged at least a few million years ago.

Comparison of divergence rates with other organisms

This is another useful method for the estimation of divergence times. The average rate of sequence divergence at synonymous sites in homologous protein coding regions between *E. coli* and *S. enterica* (*89, 90*) were estimated at 0.9% per million years. If divergence rates within *Prochlorococcus* are similar it would indicate the cN2 C1-C5 clades shared a common ancestor at least 10 million years ago. Note that the estimated number of generations per year for both *E. coli* and *Prochlorococcus* is very similar (100-300 per year) as does the mutation rate (μ) (*90*).

Cytochrome C amino acid substitutions, which are often used as a molecular clock, estimate the divergence of the cN2 C1-C5 clades could have been even earlier. For example the number of amino acid substitutions between cN2-C1 and cN2-C3 Cytochrome C (10% amino acids substitutions) is about the same as the number of Cytochrome C substitutions between human and horse (estimated to have diverged between 100-160 million years ago). *Prochlorococcus* proteins have been shown to evolve faster than other organisms though (*91*).

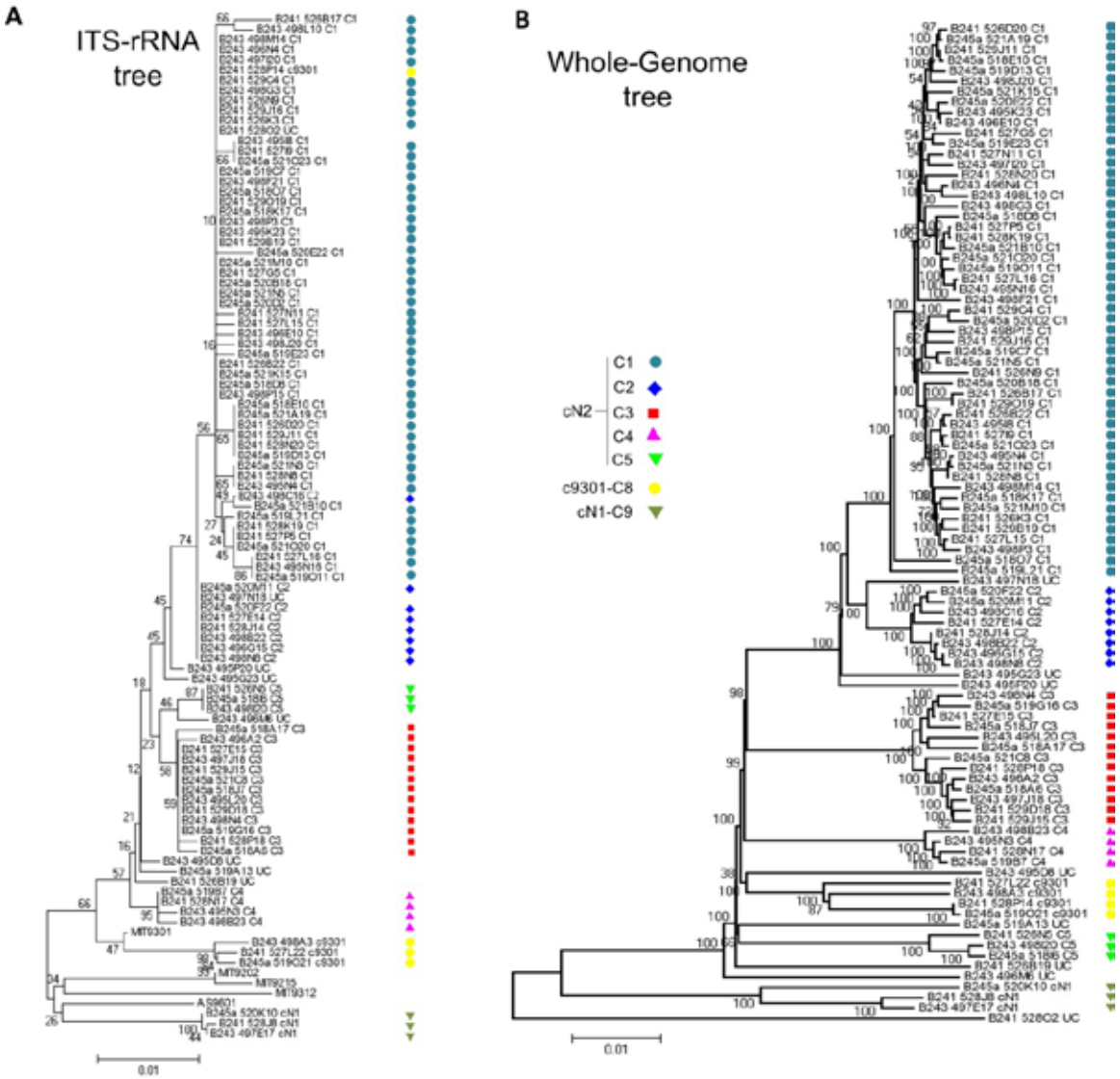

Fig. S1. Bootstrap values of the ITS-rRNA tree (A) and whole-genome tree (B) of the 96 sequenced single cells. Trees are neighbor joining with 'p-distance' (proportion of nucleotide differences). ITS sequences from cultured representatives of the same ecotype are also included. Numbers near internal nodes are bootstrap values. Trees were constructed by MEGA4.

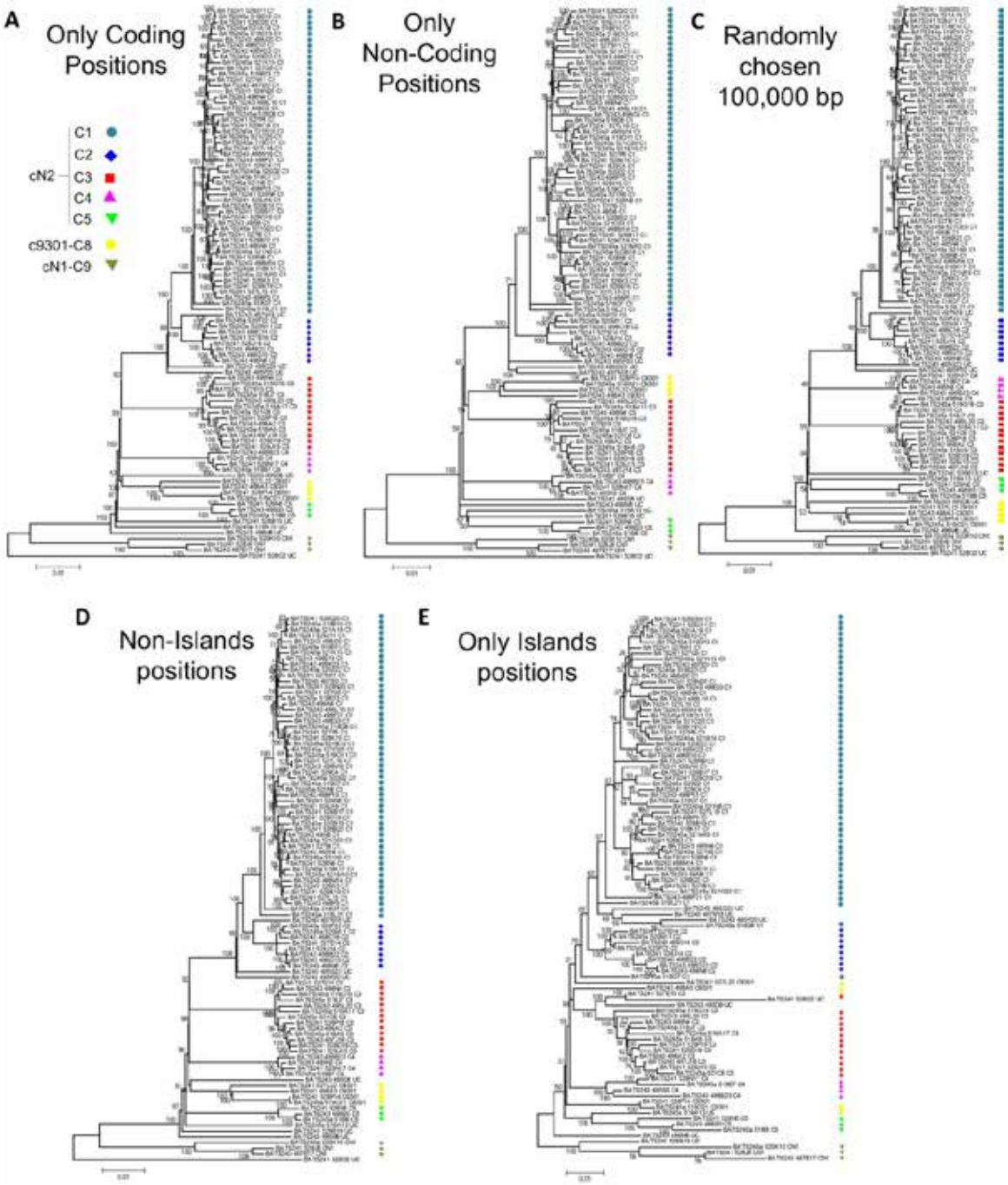

Fig. S2. Phylogenetic tree of the 96 single cells based on different classes of genomic positions. (**A**) Coding positions (1,491,155 bp). (**B**) Non-coding positions (159,199 bp). (**C**) Randomly chosen 100 Kbp. (**D**) Positions excluding genomics islands (1,433,955 bp). (**E**) Positions within genomic islands (216,399 bp). Trees are neighbor joining using p-distance. Numbers near internal nodes are bootstrap values. Trees were constructed by MEGA4.

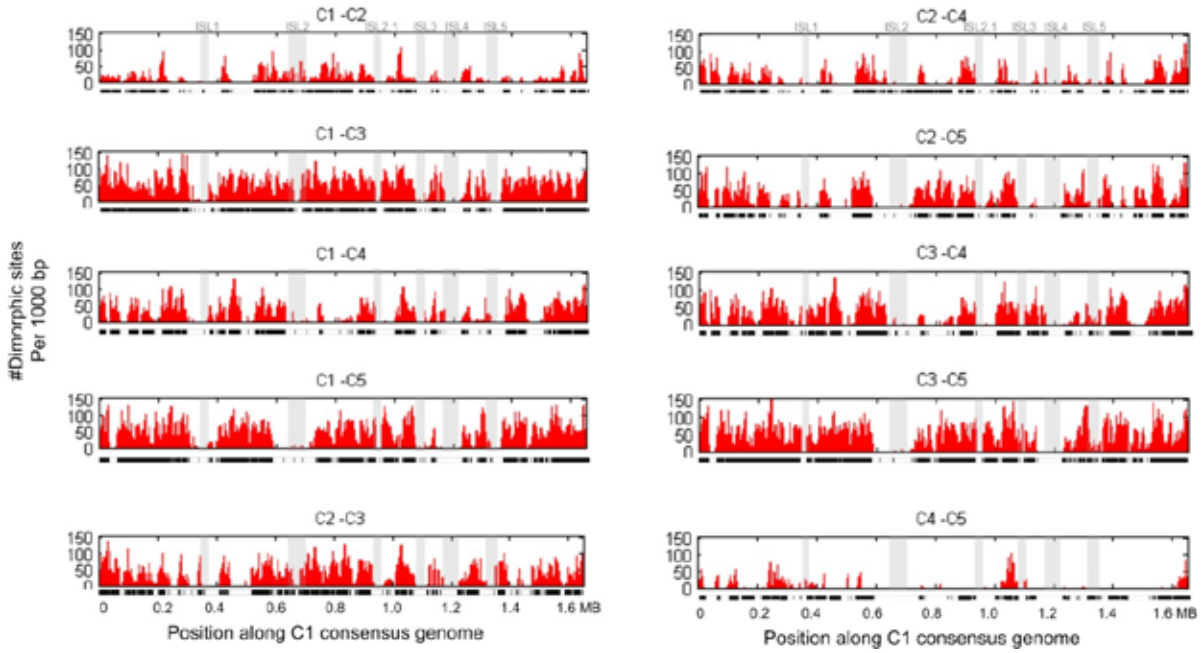

Fig. S3. Abundance of dimorphic sites, per non-overlapping 1000bp, between all pairs of the five clades within the cN2 ITS-cluster. Black/white stripes below each graph indicate positions with sufficient data to support the dimorphic site analysis (red). Gray bars represent genomic islands as defined in section 4.7.

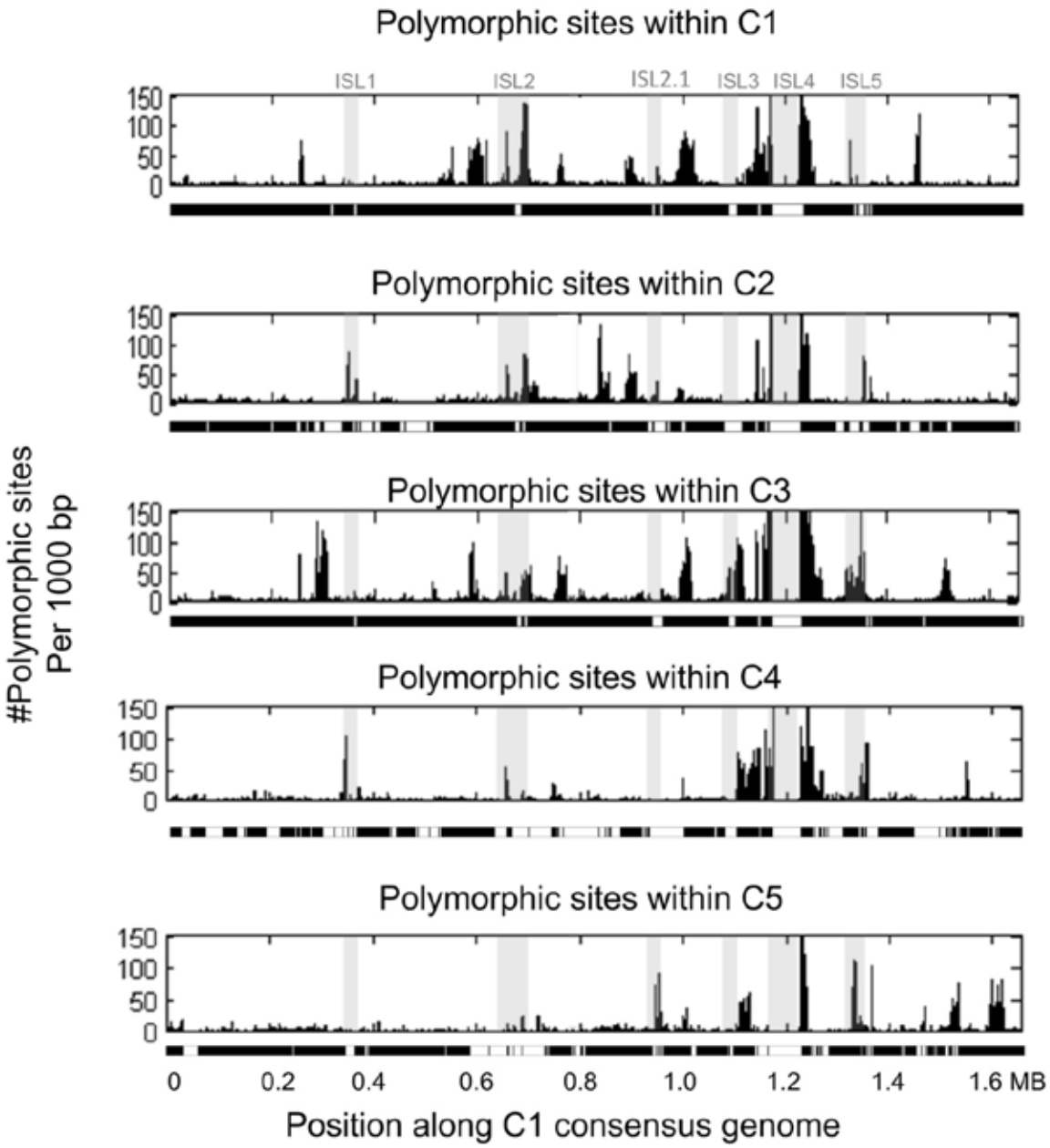

Fig. S4. Abundance of polymorphic sites, per non-overlapping 1000bp, within clades cN2 C1-C5. Black/white stripes below each graph indicate positions with sufficient data to support the polymorphic site analysis (black). Gray bars represent genomic islands as defined in section 4.7.

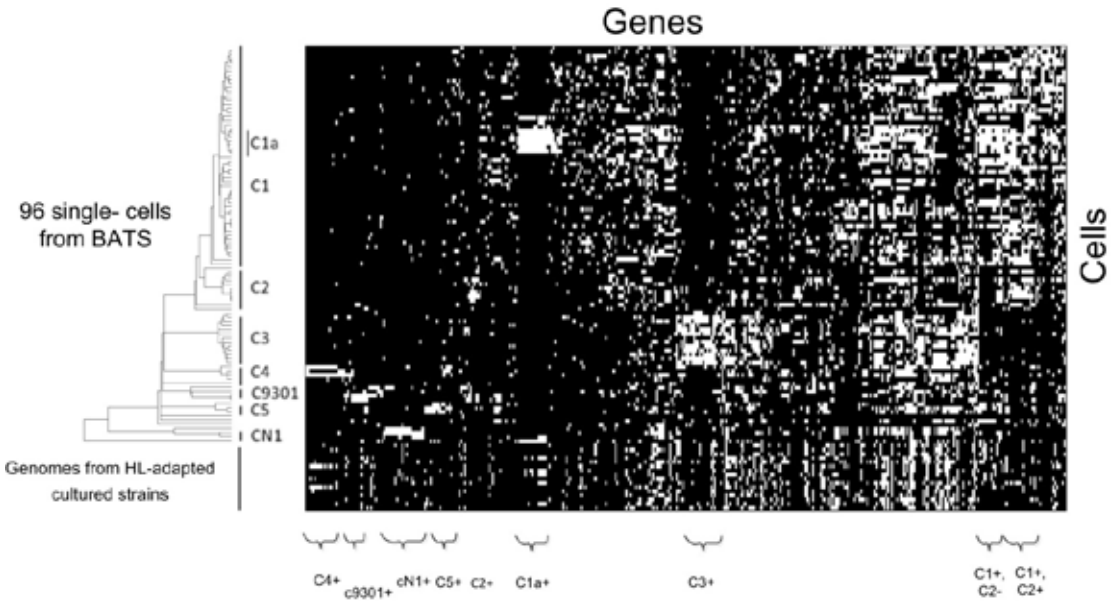

Fig. S5. **Differential gene sets between clades.** Each column is a gene. Each row represents a single cell. The order of the single cells is according to the leaf order of the whole genome phylogenetic tree. Matrix representation: Each white/black dot represents the existence/absence of a gene in the partial genome of a single cell. Note that since these 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. Also note that the order of the genes in columns does 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 sets of genes indicate genes that are differentially abundant in a pattern associated with a particular clade or clades.

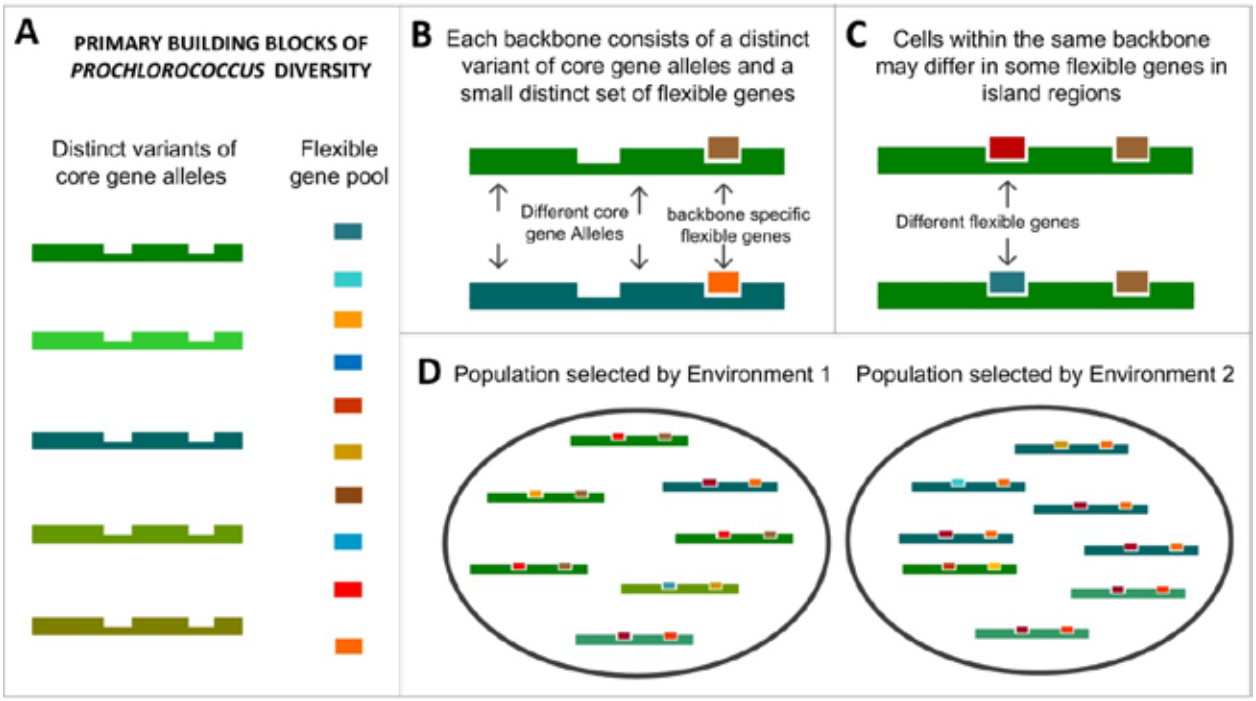

Fig. S6. Schematic of fundamental components of the genomic backbones that define *Prochlorococcus* subpopulations. (**A**) The building blocks of *Prochlorococcus* diversity include hundreds of variants with distinct core gene alleles (shades of green) – produced by selection – and a pool of thousands of flexible gene cassettes. Both contribute to niche differentiation. (**B**) Each backbone is characterized by different alleles of core genes and a small distinct set of the same flexible genes. (**C**) Cells within a backbone-subpopulation – i.e. with shared backbones – are still observed to carry a few different environment-specific genes within genomic islands, contributing an additional level of variability. (**D**) The composition of local populations is fine-tuned to local conditions by adjustment of the relative abundance of hundreds of backbone-subpopulations, reflecting their slightly different fitness, as well as variability in the genes they carry from the flexible gene pool.

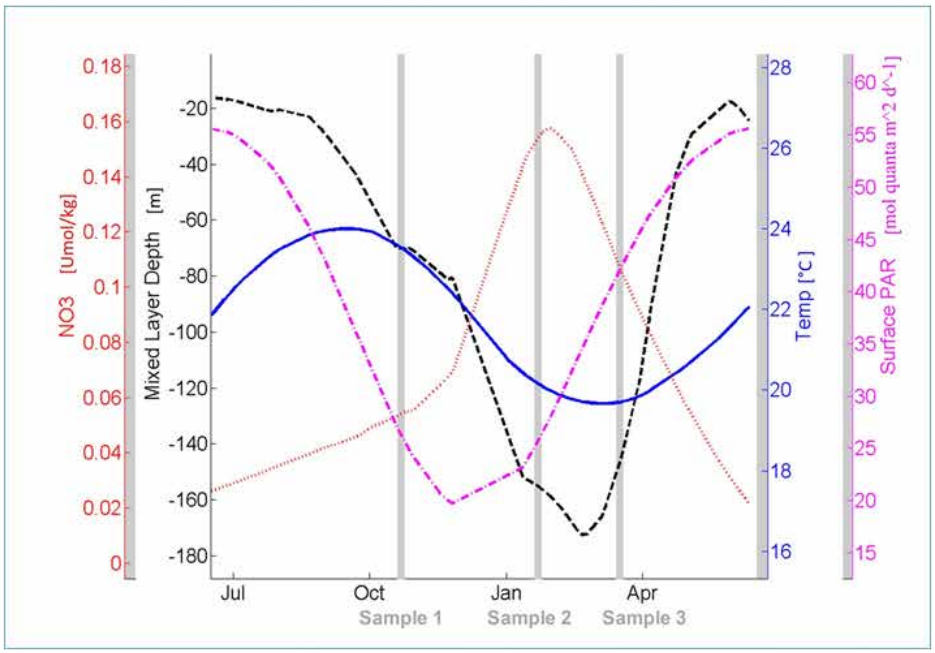

Fig. S7. Average seasonal profiles at the Bermuda Atlantic Time-series Study (BATS) site indicating conditions when the three samples used in this study were collected. Shown are profiles of water temperature, surface light, nitrate+nitrite (NO_3+NO_2) and mixed layer depth. The graphs are smoothed curves (smoothed in a similar manner as in (*17*) of average mixed layer depth, mean temperature in the top 100m, mean surface PAR (Photosynthetically Active Radiation)* (mol quanta m^{-2} d^{-1}) and mean NO_3+NO_2 concentration (µmol/kg) at the top 100m, over 10 years (1999-2009).* Light is averaged over the years (2004-2009). Data from http://bats.bios.edu/.

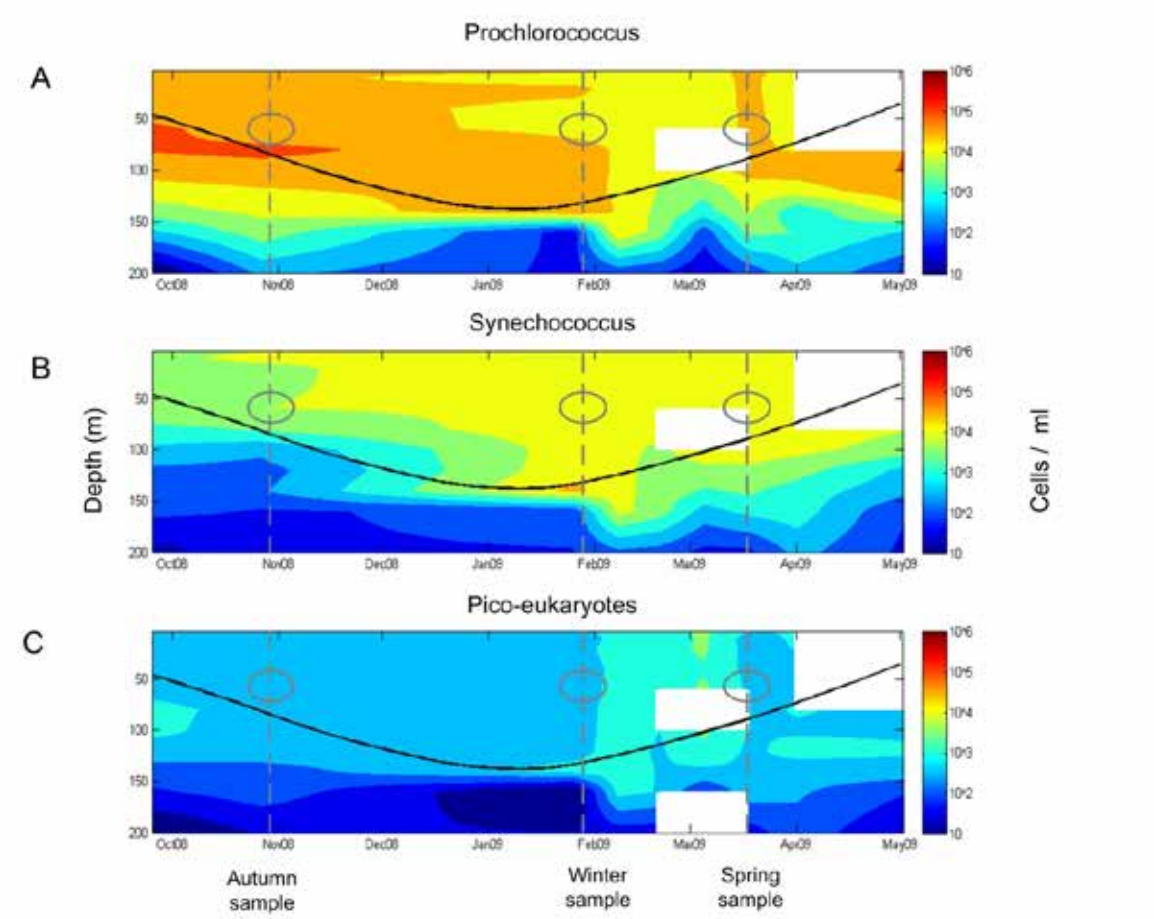

Fig. S8. *Prochlorococcus, Synechococcus* and pico-eukaryote abundance at Bermuda-Atlantic Time-series Study (BATS) site indicating conditions when the three samples used in this study were collected. Shown are species abundance over 2008-2009 seasons. Data from http://bats.bios.edu/. Samples in the current study are marked as ellipses. Black solid line marks the mixed layer depth.

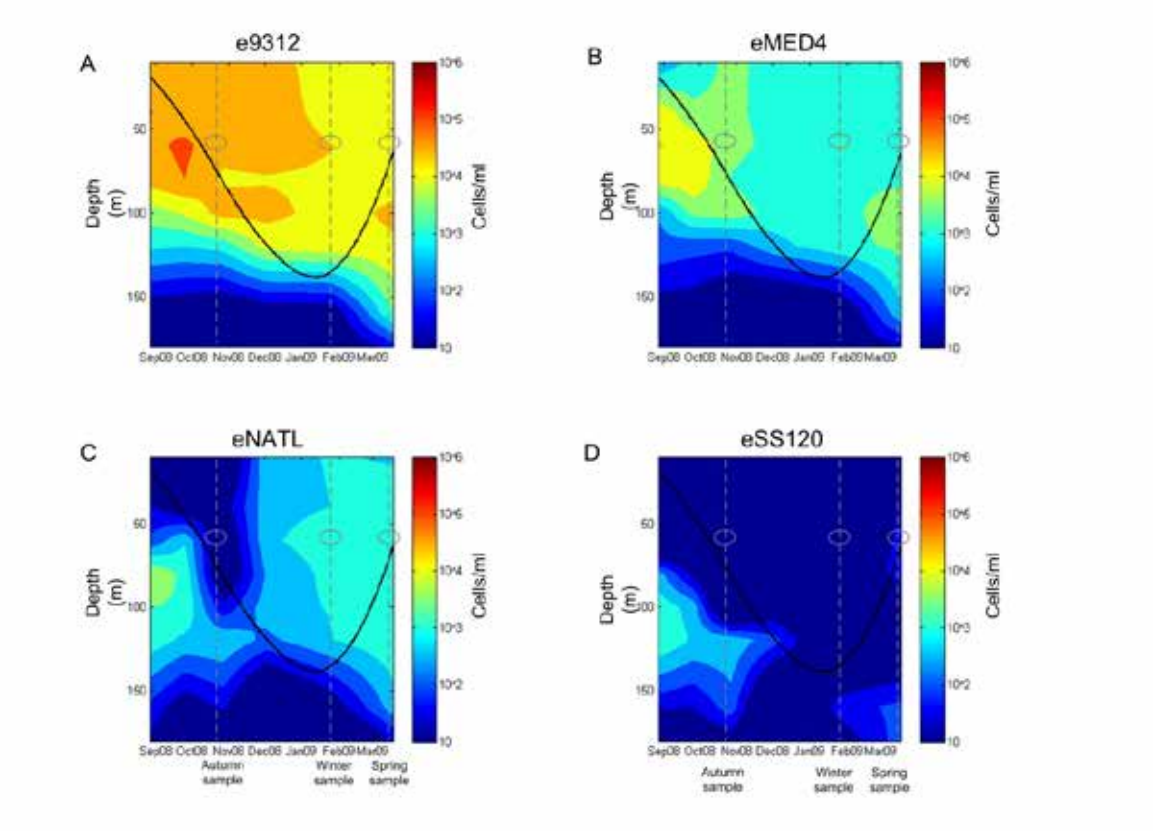

Fig. S9. *Prochlorococcus* traditional ecotype abundance over 2008-2009 seasons at Bermuda-Atlantic Time-series Study (BATS) site. Ecotype abundances are determined by qPCR. Samples in the current study are marked as ellipses. Black solid line marks the mixed layer depth.

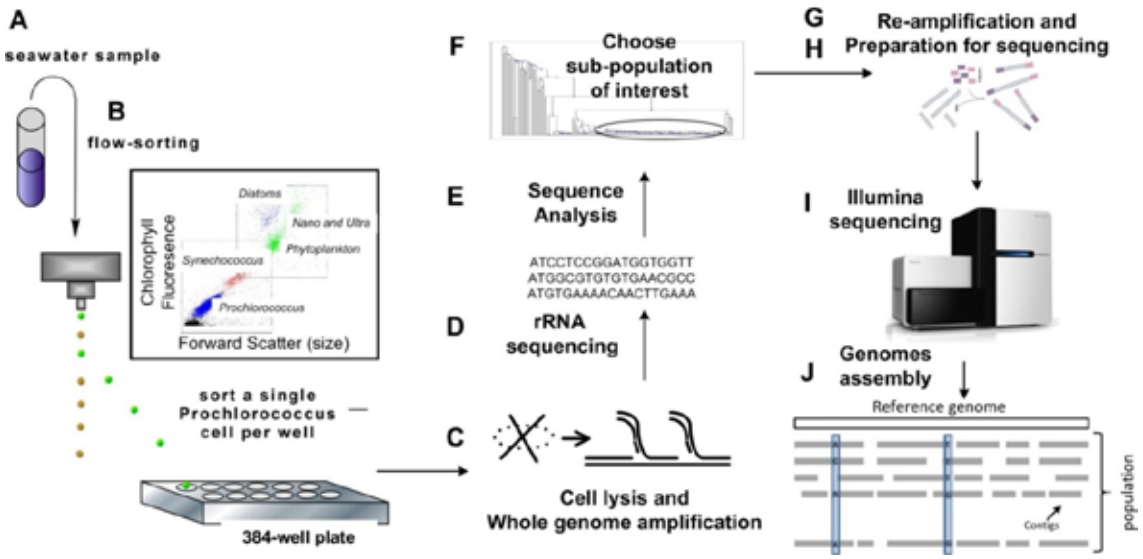

Fig. S10. A schematic representation of the single cell pipeline applied in this study. (**A**) Sea-water samples are collected. (**B**) *Prochlorococcus* is identified through flow cytometry based on light-scatter and autofluorescence, and sorted into 384-well plates (one cell per well). (**C**) Whole Genome Amplification is performed using Multiple Displacement Amplification (MDA). (**D**) Single cell Amplified Genomes (SAGs) are screened for the genetic marker(s) of choice (in this study the ITS region of the rRNA operon) using PCR followed by sequencing. (**E**) Population structure is analyzed based on the ITS sequences, using multiple sequence alignment followed by phylogenetic analysis. (**F**) Candidate cells for genome sequencing are selected. (**G,H**) A second amplification (using MDA) is performed on the selected SAGs, to obtain DNA for sequencing. (**I**) Barcoded DNA libraries are created and sequenced using Illumina technology. (**J**) *De novo* assembly or referenced guided assembly of the sequence reads into genomes, followed by genetic analysis of the population.

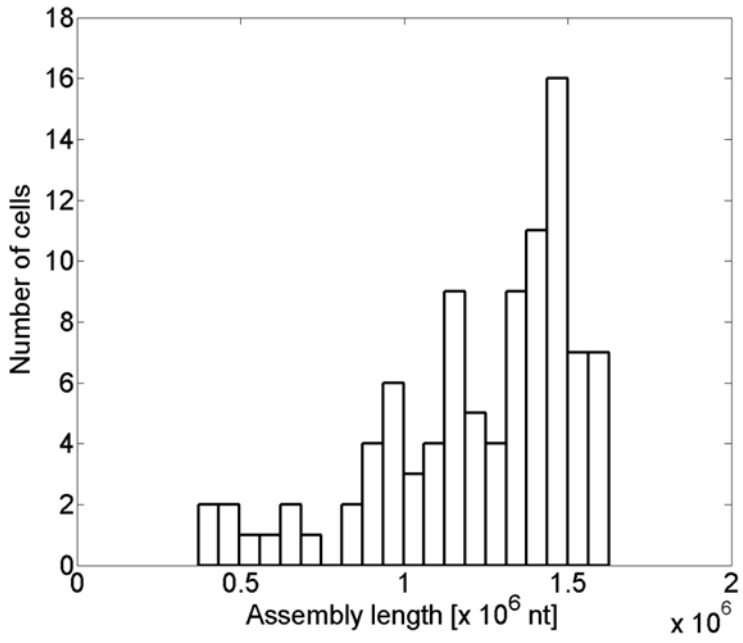

Fig. S11. Histogram of the *de novo* assembly lengths of the 96 partial single cell genomes. The median length is ~1.3 million bp - equivalent to 78% of the estimated complete genome size of ~1.65 million bp. nt = nucleotides.

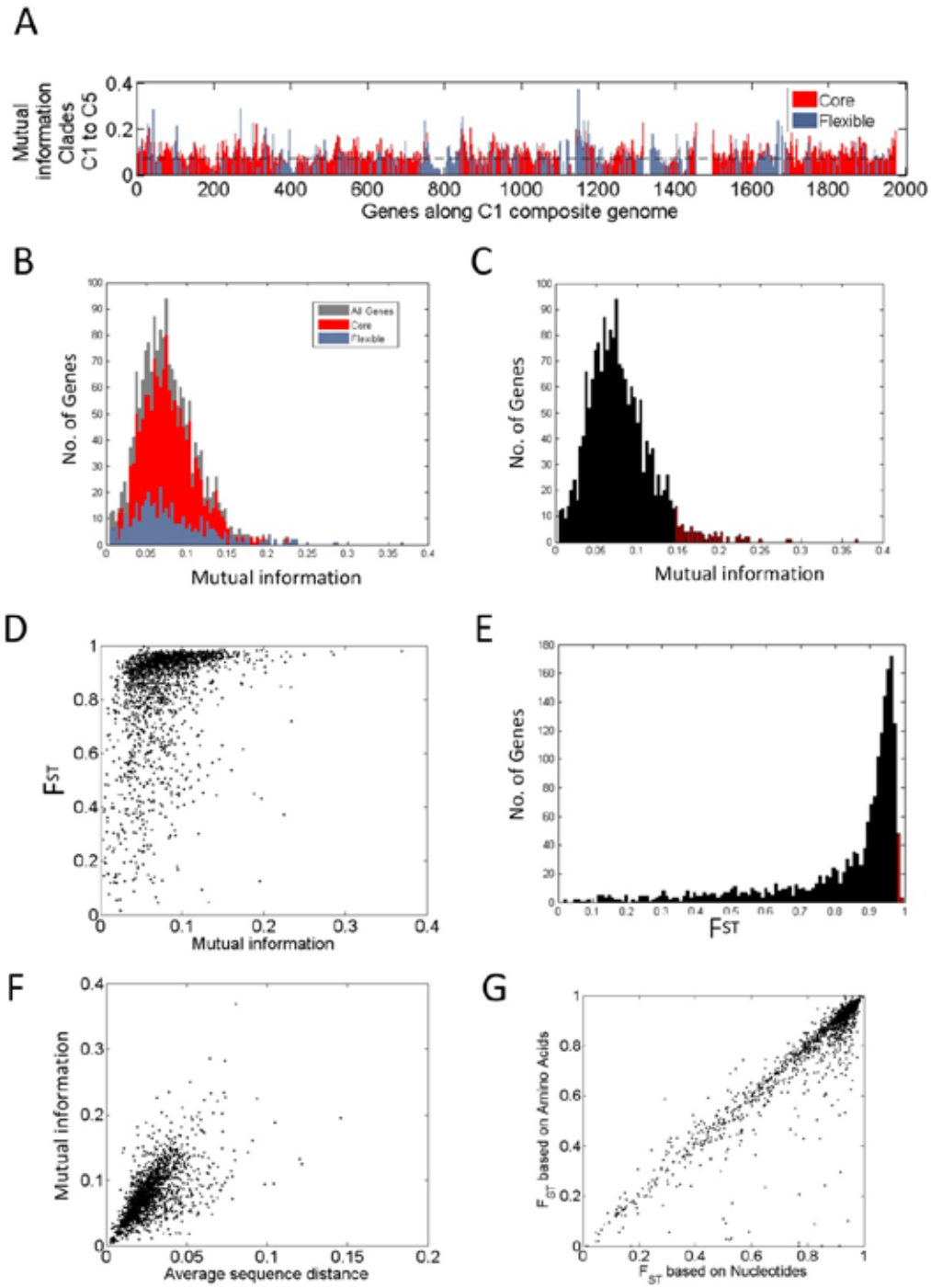

Fig. S12. Genetic differentiation of genes among clades cN2-C1 to cN2-C5. (**A**) Mutual information of genes (based on nucleotide sequences). (**B**) Distribution of mutual information values among core and flexible genes. (**C**) Highest 5% mutual information values. (**D**) F_{ST} vs. mutual information**.** (**E**) Highest 5% F_{ST} values**.** (**F**) Mutual information vs. Average sequence distance. (**G**) F_{ST} values based on amino-acid protein sequences vs. same values based on gene nucleotide sequences.

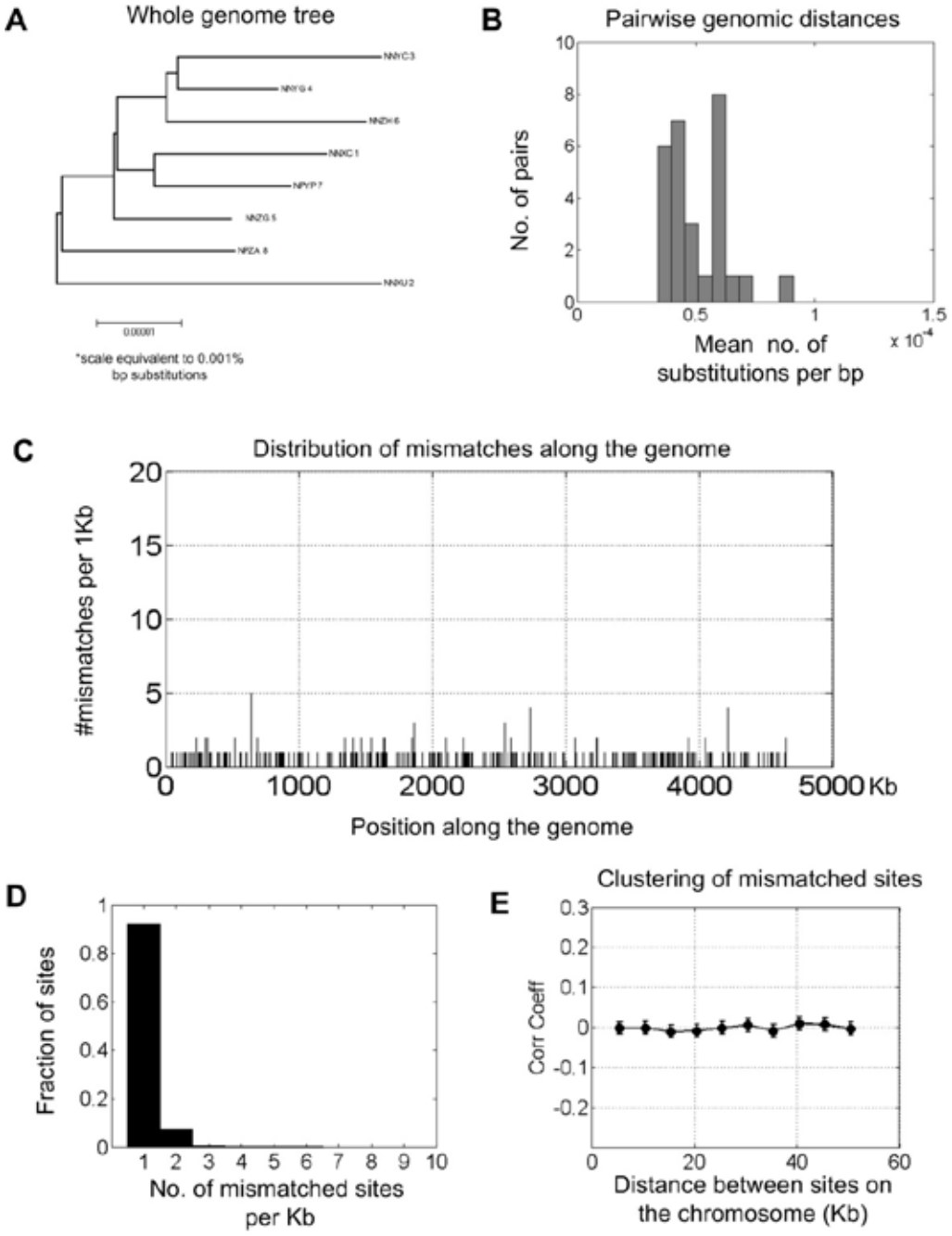

Fig. S13. Estimation of the error rate from single cell genomics based on a control experiment with eight clonal *E. coli* single cell genomes. (A) Whole-genome phylogentic tree. Neighbor joining with p-distance (computed in a similar manner to Fig. 2B in the main text) (B) Distribution of estimated pairwise genomic distances (#substitutions per bp). (C) Distribution of mismatches along the reference genome (per 1Kb). (D) Abundance distributions of sites with mismatches along the genome - similar to what expected by a Poisson distribution (E) Correlation coefficient between the abundance of sites with mismatches and distance between sites on the chromosome – indicating no apparent clustering.

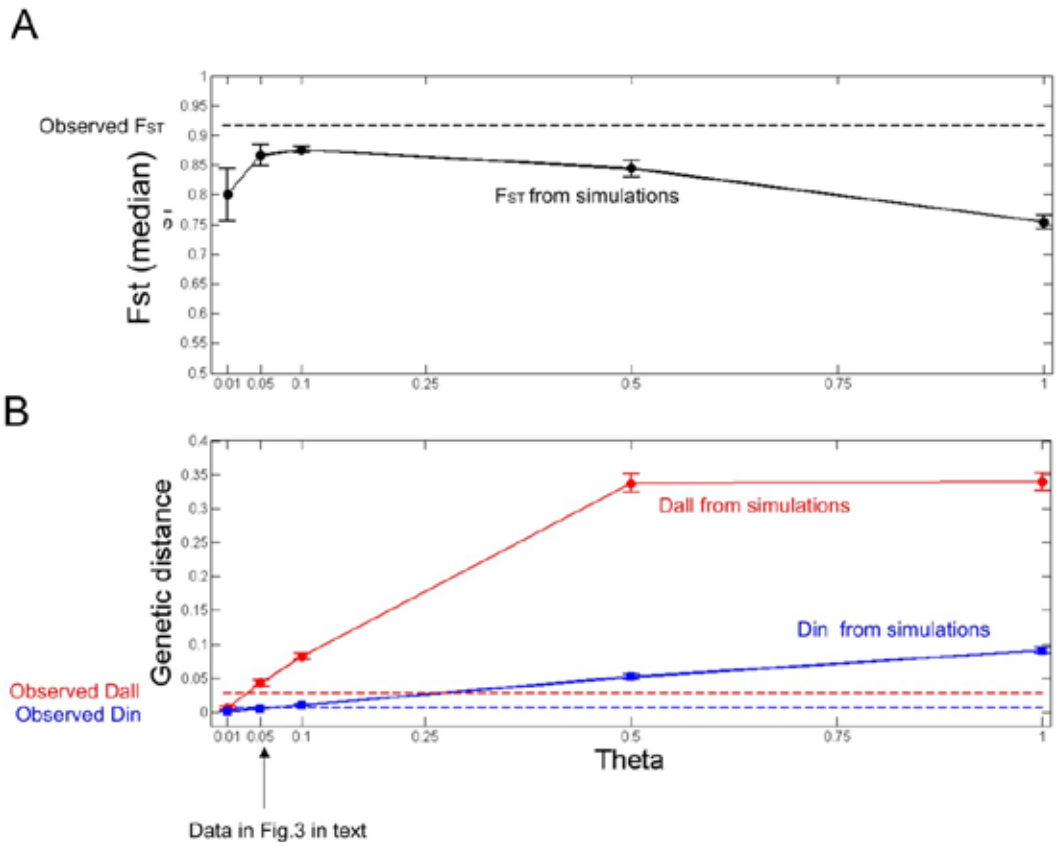

Fig. S14. Coalescent simulations. (**A**) Median F_{ST} values vs. Θ. Error bars are SE from 5 simulations. Dashed line is the observed median F_{ST} in our real genomic data. (**B**) Same as in **A** but for median genomic distance between all genomes (D_{all}) and median genomic distance within backbone-subpopulations (D_{in}). Dashed lines are the observed corresponding distances in our real data. The simulation data in Fig. 3B and Fig. S15 is for Θ=0.05, empirically found to yield the closest values of D_{all} and D_{in} to those of the real data. Note that no choice of Θ, in the tested range, reaches the observed median F_{ST}. Θ values larger than 1 are expected to yield even smaller median values of F_{ST} than those of Θ<1.

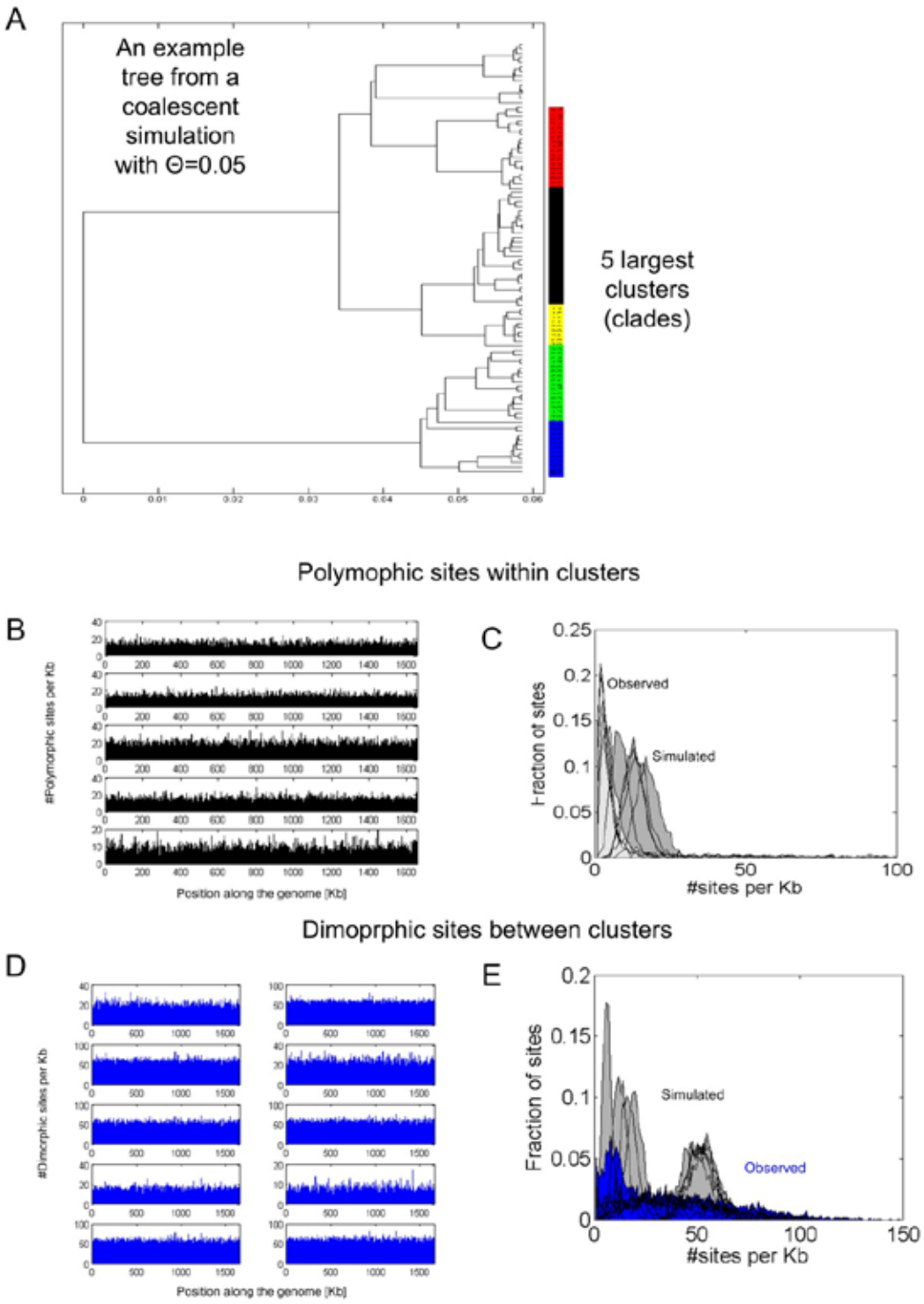

Fig. S15. A typical coalescent simulation of neutral evolution with Θ=0.05. This choice of Θ yielded similar average pairwise genetic distance to the observed ones (Fig. 2B). The simulations results in Fig. 3B and Fig. S15 are from this specific simulation. (**A**) The resulted tree. Different colors mark the 5 clusters identified. (**B**) Polymorphic sites within the five clades. (**C**) Abundance distributions of polymorphic sites along the genomes. (**D,E**), same as for **B** and **C** but for dimorphic sites between clusters.

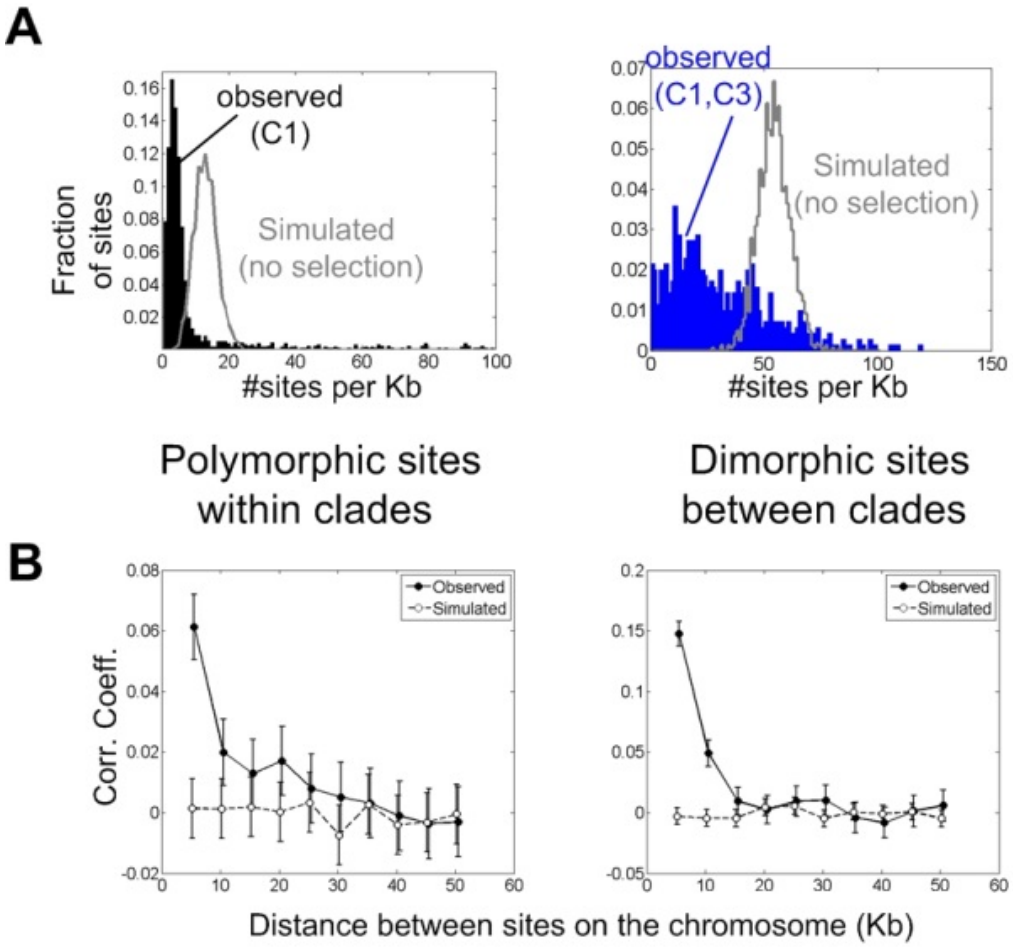

Fig. S16. Observed and simulated (no selection) abundance distributions and correlations coefficient between sites. Dimorphic sites (per non-overlapping 1000bp) between clades cN2-C1 and cN2-C3. (A) Abundance distributions of polymorphic and dimorphic sites along the genomes within and between C1 and C3, as well as typical distributions from coalescent simulations of neutral evolution (See section 6). (B) Correlation coefficient between sites abundance and distance between sites on the chromosome – indicating clustering that is not observed in coalescent simulations.

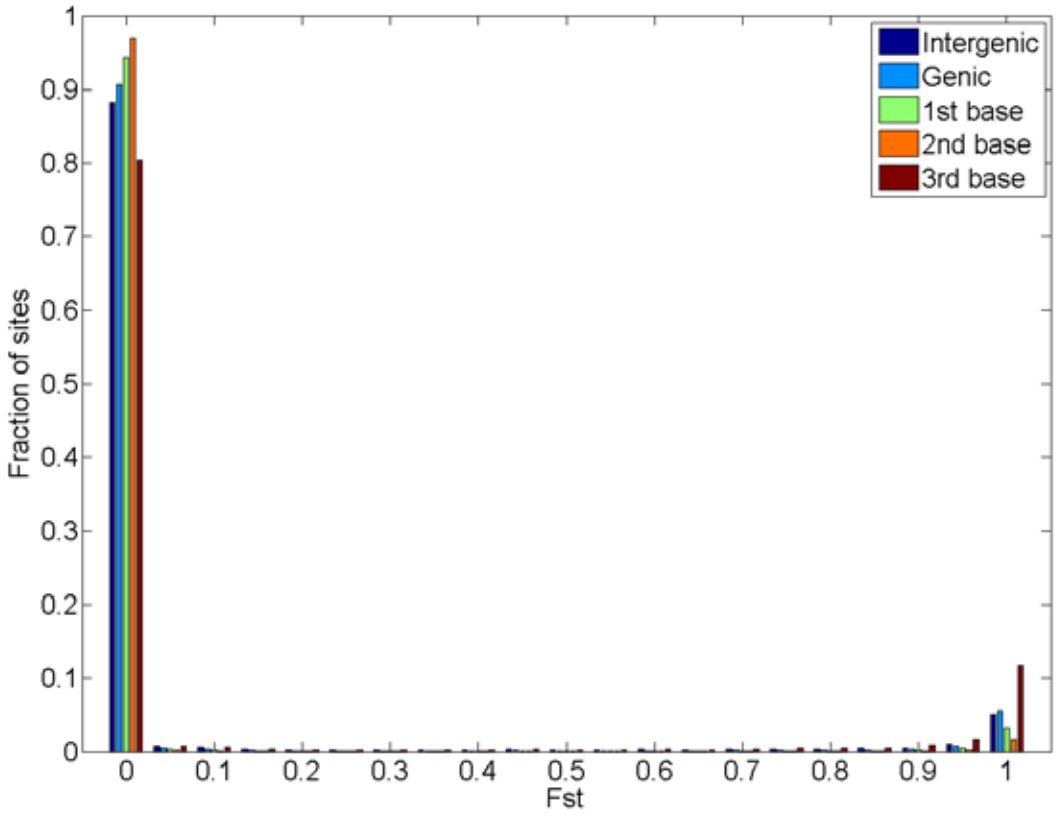

Fig. S17. F_{ST} distributions of different functional classes of single nucleotides. Classes are: Intergenic positions, Genic positions, and 1st, 2nd and 3rd codon bases. The fraction of positions with very low F_{ST} (<0.05) was significantly different between all pairs of nucleotide classes. The fraction of positions with very high F_{ST} (>0.95) was also different between all pairs of nucleotide classes.

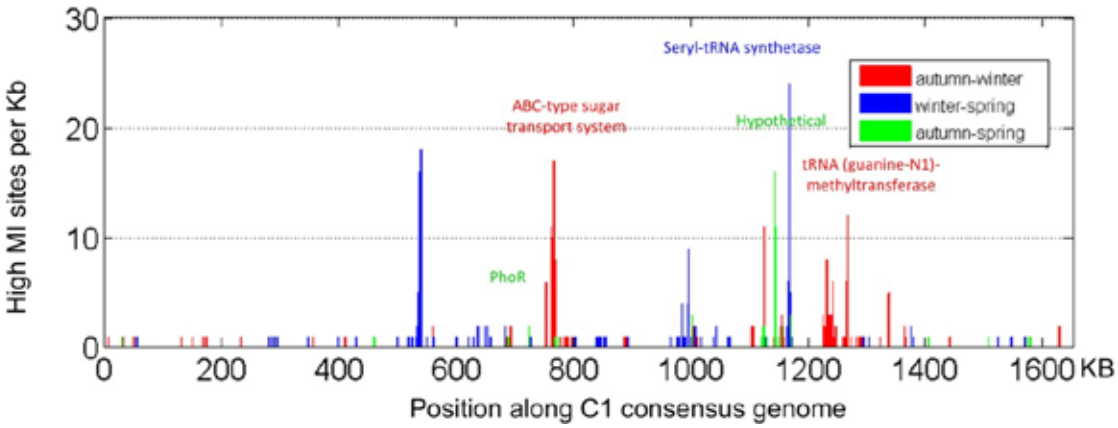

Fig. S18. Changes in allele frequency within cells belonging to the cN2-C1 clade between seasonal samples. Shown are sites with significantly high mutual information positions ($P<0.01$). A few genes with such changes are marked. Note these sites are not dimorphic but are sites with a significant change in allele frequency (e.g. from 100% 'A's in one season to 60% 'A's and 40% 'C's in the other season).

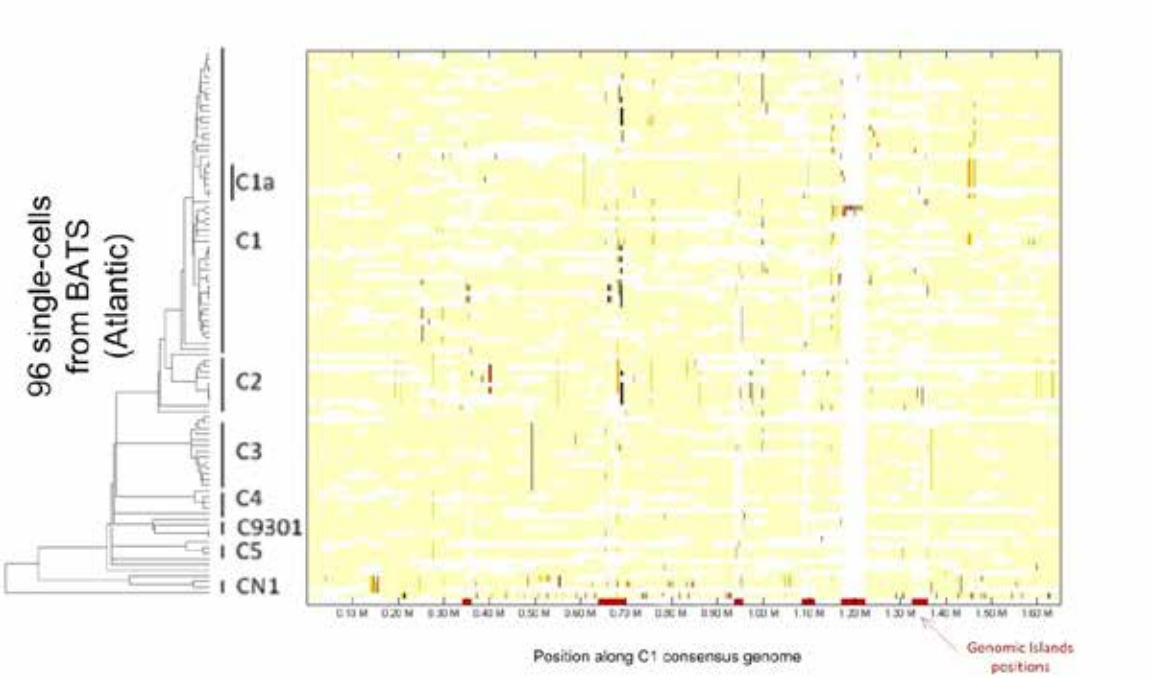

Fig. S19. Predicted Homologous Recombination (HR) within the 96 single cells. Each row represents a single cell genome. Yellow/white represents covered/missing site of each specific position in each partial genome. Other colors represent stretches of DNA predicted to be acquired through HR. Similar colors within the same position represent highly similar blocks (likely of same origin). Last row, stretches in red indicate the location of genomic islands. HR was predicted using the BratNextGen tool.

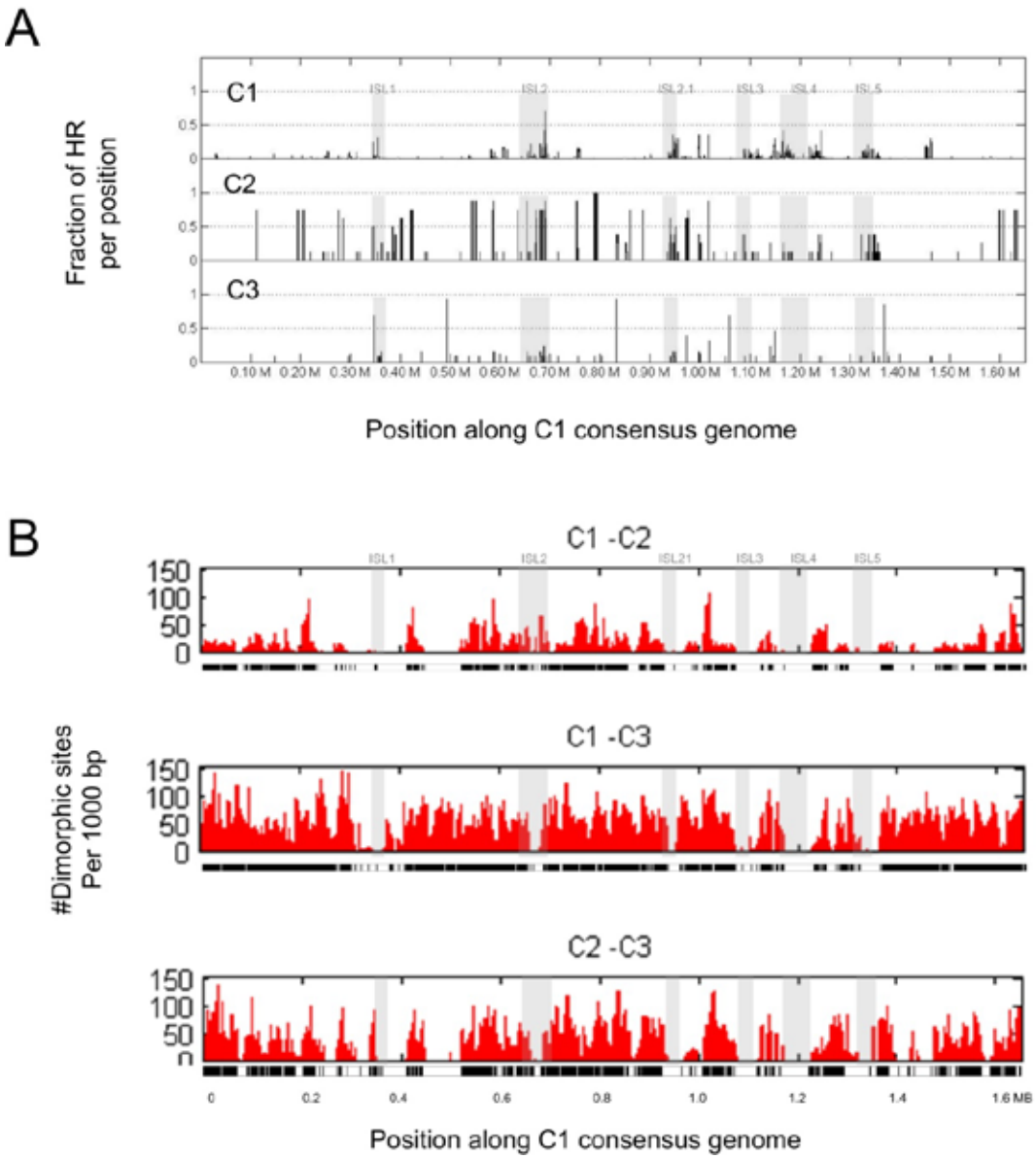

Fig. S20. **Homologous recombination does not explain dimorphic SNPs.** (**A**) Fraction of detected recombined sites within clades C1,C2 and C3 (cN2), per non-overlapping 1Kb (**B**) Dimorphic sites between pairs within clades C1,C2,C3.

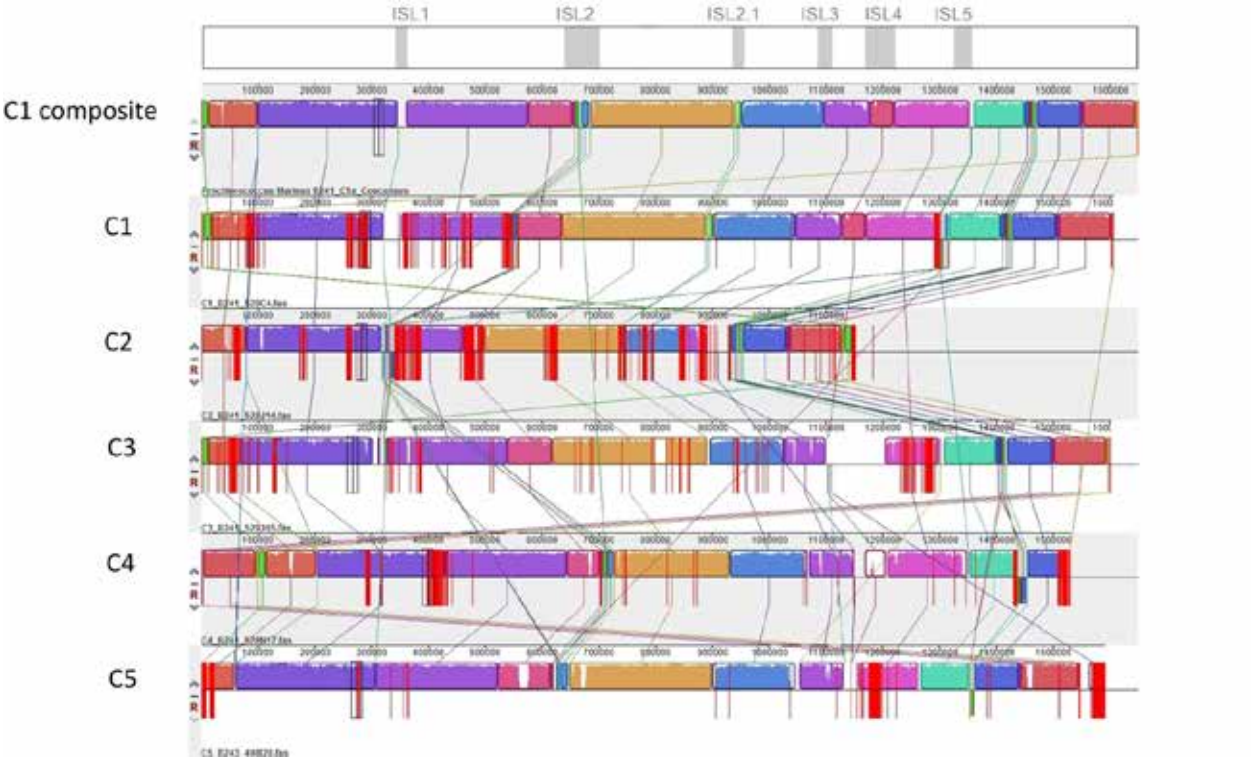

Fig. S21. Genome synteny between clades cN2-C1 to cN2-C5. Shown is a multiple alignment of representative single cell partial genomes from each of the clades cN2-C1 to cN2-C5. Each clade is represented by one cell. Alignment was done by Mauve (*64*). The top genome is the cN2-C1 composite genome that serve as a reference, with the islands locations marked in gray above. The aligned genomes are from *de novo* assemblies. The different colored blocks are "Locally co-linear blocks" (LCBs) which are conserved segments that appear to be internally free from genome rearrangements.

Table S1. Flexible gene cassettes associated with different genomic backbones

Clades	Cassette ID	COG ID	Description	Position
cN2-C1, cN2-C4	CST_I	17430	hypothetical protein	Island 2.1
		82	Possible Cytochrome oxidase c subunit VIb	
		5925	hypothetical protein	
		100193	high light inducible protein	
cN2-C1, cN2-C2	CST_II	11507	Glycosyltransferase of PMT family	Island 4
		5069	Glycosyltransferase	
		2779	Sugar transferase	
		3653	ABC-type multidrug transport system ATPase and permease components	
		6172	glycosyl transferase; group 1	
		14302	UDP-galactopyranose mutase (EC 5.4.99.9)	
		4701	predicted protein	
cN2-C3	CST_III	51079	possible Glycosyl transferase	Island 1
		59087	Glycosyl transferase family 11	
cN2-C4	CST_IV	299	UDP-glucose dehydrogenase (EC 1.1.1.22)	Cassette island 4
		1614	Glucose-1-phosphate thymidylyltransferase (EC 2.7.7.24)	
		3209	dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)	
		67595	hypothetical protein	
		53203	hypothetical protein	
		61572	HlpA protein	
		68307	putative glycosyltransferase	
		65350	hypothetical protein	
		59677	glycosyltransferase	
		3155	UDP-N-acetylmuramyl pentapeptide phosphotransferase/UDP-N-acetylglucosamine-1-phosphat transferase	
		56016	hypothetical protein	
		52032	CpsL	
		65878	Asparagine synthetase [glutamine-	

			hydrolyzing] (EC 6.3.5.4)	
		411	UDP-N-acetylglucosamine 4,6-dehydratase (EC 4.2.1.-)	
cN2-C5	CST_IV	61789	glycosyltransferase, group 1	Cassette Island 4
		48	UDP-glucose 4-epimerase (EC 5.1.3.2)	
		45361	UDP-glucose dehydrogenase (EC 1.1.1.22)	Cassette Island 4
		72971	hypothetical protein	
		67514	Glycosyltransferase	
c9301-C8	CST_VI	60774	conserved hypothetical protein	Island 1
		66999	type II DNA modification methyltransferase	
		66324	ulcer associated adenine specific DNA methyltransferas	
		70558	hypothetical protein	
cN1-C9	CST_IX	60426	putative rieske (2Fe-2S) family protein	Island 5
		35	Urea carboxylase-related ABC transporter, ATPase protein	
		59708	Urea carboxylase-related ABC transporter, permease protein	
		30352	Urea carboxylase-related ABC transporter, periplasmic substrate-binding protein	
		50117	[NiFe] hydrogenase nickel incorporation-associated protein HypB	
		19523	[NiFe] hydrogenase nickel incorporation protein HypA	
		62244	Agmatinase (EC 3.5.3.11)	
	CST_VII	27390	Repeats containing protein	Island 4
		?	hypothetical protein	
		64707	Glycosyl transferase, group 1	
		1744	Mannose-1-phosphate guanylyltransferase (GDP) (EC2.7.7.22)	
		?	hypothetical protein	
		57933	Glycosyltransferase	
		13831	UDP-N-acetylglucosamine 2-epimerase (EC 5.1.3.14)	
	CST_VIII	29029	conserved hypothetical membrane protein	Island 4
		1754	Glycosyltransferase	
		57082	hypothetical protein	

Table S2. Collected sample details.

Sample	Date	Name	Cruise	Depth	Cells/ml (mean±SE)
1	Nov 8th 2008	‘autumn sample’	BATS 241	60m	41350±750
2	Feb 8th 2009	‘winter sample’	BATS 243	60m	33100±800
3	Apr 1st 2009	‘spring sample’	BATS 245a	60m	33000±1350

Table S3. Adapters and primers for Illumina libraries.

Oligonucleotide for making adapters (no barcode in the insert)	
IGA-A0-down	AGA TCG GAA GAG CGT CGT GTA GGG AAA GAG TGT AC/3AmM/
IGA-A0-up	/5AmMC6/ACA CTC TTT CCC TAC ACG ACG CTC TTC CGA TCT
IGA-PE-B0-down	/5AmMC6/CTC GGC ATT CCT GCT GAA CCG CTC TTC CGA TCT
IGA-PE-B0-up	AGA TCG GAA GAG CGG TTC AGC AGG AAT GCC GAG /3AmM/
Oligonucleotide for PCR amplification	
IGA-PCR-PE-F	AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC T
Illumina sequencing barcoded primers (barcodes are in bold and the reverse complemented sequence is obtained)	
IGA-RACE-PCR-R64-b19	CAA GCA GAA GAC GGC ATA CGA GAT **CAGCTG** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b40	CAA GCA GAA GAC GGC ATA CGA GAT **TGAAGC** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b15	CAA GCA GAA GAC GGC ATA CGA GAT **GCACAT** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b11	CAA GCA GAA GAC GGC ATA CGA GAT **TCCCCT** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b35	CAA GCA GAA GAC GGC ATA CGA GAT **CTCCTC** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b61	CAA GCA GAA GAC GGC ATA CGA GAT **AACTAA** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b8	CAA GCA GAA GAC GGC ATA CGA GAT **TAGAGT** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b44	CAA GCA GAA GAC GGC ATA CGA GAT **GGTACC** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b54	CAA GCA GAA GAC GGC ATA CGA GAT **CTTGGA** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b29	CAA GCA GAA GAC GGC ATA CGA GAT **AGTTAG** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b49	CAA GCA GAA GAC GGC ATA CGA GAT **TAATTA** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b30	CAA GCA GAA GAC GGC ATA CGA GAT **TCTGAG** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b47	CAA GCA GAA GAC GGC ATA CGA GAT **GTGCAC** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b26	CAA GCA GAA GAC GGC ATA CGA GAT **ACAGCG** CGG TCT CGG CAT TCC TGC TGA AC
IGA-RACE-PCR-R64-b9	CAA GCA GAA GAC GGC ATA CGA GAT **CTCTCT** CGG TCT CGG CAT TCC TGC TGA AC

IGA-RACE-PCR-R64-b51	CAA GCA GAA GAC GGC ATA CGA GAT **CGACTA** CGG TCT CGG CAT TCC TGC TGA AC

Table S4. Traditional ecotype abundance as estimated by single cell ITS-rRNA and qPCR.

		Autumn sample		**Winter sample**		**Spring sample**	
Ecotype		Single cell ITS	qPCR ecotypes	Single cell ITS	qPCR ecotypes	Single cell ITS	qPCR ecotypes
e9312	#cells:	35000±1000	46200±1400	23200±1000	32000±600	22400±1000	26100±4400
	Relative abundance:	**92% ± 1%**	**90% ± 3%**	**81% ± 3%**	**86% ± 2%**	**78% ± 3%**	**85% ± 14%**
eMED4	#cells:	2700±100	4800±400	1500±100	3200±100	1450±100	3000±500
	Relative abundance:	**7% ± 1%**	**9% ± 1%**	**6% ± 2%**	**9% ± 1%**	**5% ± 1%**	**10% ± 2%**
eNATL	#cells:	100±100	5±5	3500±200	2000±200	4700±200	1600±200
	Relative abundance:	**<%1**	**<0.1%**	**7% ± 1%**	**5% ± 1%**	**8% ± 1%**	**5% ± 1 %**
eSS120	#cells:	NA	65±10	NA	35±30	NA	50±10
	Relative abundance:	**NA**	**<0.1%**	**NA**	**<0.1%**	**NA**	**<0.1%**

Table S5. Relative abundance of ITS-clusters as depicted from single cell data (Percent of whole population).

ITS cluster	Autumn sample	Winter sample	Spring sample
cNATL	0.3% ± 0.3%	7% ± 1.1%	8% ± 1%
cMED4	7.1% ± 0.3%	3.6% ± 0.8%	4% ± 0.6%
cN1	6.6% ± 0.2%	9.7% ± 2.4%	9.2% ± 0.3%
cN2	23.5% ± 1.6%	11.5% ± 2.2%	24.8% ± 1.9%
c9301	20.6% ± 2.2%	17.5% ± 1.9%	13.9% ± 1.6%

Table S6. Relative abundance of cN2 C1-C5 clades. Percent of whole population (mean±SE).

cN2 clade	Autumn sample	Winter sample	Spring sample
C1	14.4% ± 1.6%	3.3% ± 0.8%	17.8% ± 1.6%
C2	1.5% ± 0.9%	0.9% ± 0.3%	1.3% ± 0.4%
C3	2.9% ± 0.8%	2.8% ± 0.7%	4.3% ± 1.1%
C4	0.8% ± 0.5%	1.5% ± 0.2%	0.3% ± 0.2%
C5	0.4% ± 0.4%	0.4% ± 0.2%	0.2% ± 0.2%

Table S7. Number of whole-genome sequenced single cells within ITS-clusters and clades

ITS-cluster	Clade	Autumn sample	Winter sample	Spring sample
cN2	C1	19	14	20
	C2	2	4	2
	C3	4	4	5
	C4	1	2	1
	C5	1	1	1
	Other	2	5	1
c9301	C8	2	1	1
cN1	C9	1	1	1
Total		32	32	32

Table S8. *De novo* assembly statistics. Genomes were *de novo* assembled using CLCbio assembler. A median assembly size of 1.3 million bp reflects a median genome recovery of ~78% (assuming a complete genome size of 1.65 million bp).

		Percentiles	
	25%	**50%**	**75%**
Assembly size (million bp)	1.1	1.3	1.5
No. of contigs	180	280	350
N50 (bp)	50,000	75,000	115,000
Average contig length (bp)	3300	4500	6300
Largest contig (bp)	110,000	190,000	290,000

Table S9. Genomic islands

Island	Position on cN2-C1 composite genome	No. of genes	No. of non-core genes
ISL1	341529-361790	33	28
ISL2	639080-700682	104	73
ISL2.1	936188-956506	37	32
ISL3	1085348-1113669	64	44
ISL4	1170430-1222632	43	38
ISL5	1325898-1359593	68	49

Table S10. Polymorphic sites (bp) within clades

	Shared polymorphic positions between clades				Total No. of Polymorphic positions within clades	Clade-Unique Polymorphic sites	Putatively recombined positions within-clade	Polymorphic and putatively recombined positions
Clade	**C2**	**C3**	**C4**	**C5**				
C1	2531	3376	1417	943	14295	8907	206341	4416
C2		1777	982	582	10285	6799	65749	1763
C3			1748	902	18643	13512	31162	604
C4				446	8695	5812	7989	159
C5					8448	6776	17022	330

Table S11. Estimation of the number of substitutions and insertions/deletions of clonal *E. coli* single cell genomes (per 100Kb) with respect to a reference genome. SAG = single amplified genome.

	SAG	Substitutions	Insertions/Deletions	Sites Recovered (Kb)	Substitutions per 100Kb	Indels per 100Kb
1	**NNXC**	28	15	844	3.3	1.8
2	**NNXU**	74	26	1391	5.3	1.9
3	**NNYC**	49	20	1460	3.3	1.3
4	**NNYG**	75	28	2309	3.2	1.2
5	**NNZG**	71	25	2215	3.2	1.1
6	**NNZH**	58	19	1573	3.5	1.1
7	**NPYP**	78	24	2045	3.8	1.1
8	**NPZA**	54	11	1306	4.1	0.8
	Mean±SD			**1655±505**	**3.7±0.7**	**1.3±0.3**

Table S12. Estimation of pairwise differences between clonal *E. coli* single cell genomes (per 100Kb). SAG = single amplified genome.

	SAG (cell)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	**NNXC**								
2	**NNXU**	6.2							
3	**NNYC**	6.2	9.1						
4	**NNYG**	4.3	6.2	3.6					
5	**NNZG**	3.4	6.1	4.2	3.6				
6	**NNZH**	6	7.1	4.8	3.8	4.4			
7	**NPYP**	3.7	6.7	4.3	4.0	4.0	4.7		
8	**NPZA**	4.9	6.0	5.6	4.3	3.9	5.8	5.7	
	Mean±SD	5.1±**1.4**							

Table S13. Examples of gene cassettes shared by a few closely related cells (subclades) within backbone-subpopulations

Clade	Cells in Subclade	No. of genes	Genes (partial list)	System/Function	Position
cN2-C1	518D8, 527P5, 528K19, 521B10, 521O20, 519O11, 527L16, 495N16	21	twin-arginine translocation pathway signal sequence; Leader peptidase (Prepilin peptidase) (EC3.4.23.43); general secretion pathway protein H; possible general (type II) secretion pathway protein D precursor, Type IV fimbrial assembly; ATPase PilB, Twitching motility protein PilT; Type II secretory pathway, component PulF / Type IV fimbrial assembly protein PilC;	Type II secretion and type IV pilus	Island 2
cN2-C1	495N4, 528N8, 521N3	4	Methyltransferase FkbM; Glucose-1-phosphate thymidylyltransferase (EC 2.7.7.24);	nucleotide sugar precursor synthesis	Island 4
cN2-C1	529J11 518E10	40	polysaccharide export-related periplasmic protein; Arabinose 5-phosphate isomerase (EC 5.3.1.13); Asparagine synthetase [glutamine-hydrolyzing] (EC 6.3.5.4); glycosyl transferase; Glucose-1-phosphate cytidylyltransferase (EC2.7.7.33); Bacterial sugar transferase	Polysaccharide biosynthesis and export	Island 4
cN2-C2	498B22, 498N8, 496G15	12	Possible Natural resistance-associated macrophage Protein (Nramp); high light inducible protein-like; possible Ribosomal RNA adenine dimethylase;	Membrane surface modification (possibly related to phage resistance)	Island 5

			Putative phosphatase		

Additional file Data S1

Gene-by-gene F_{ST} values for all genes in the cN2-C1 composite genome (Excel table).

References and Notes

1. F. Partensky, W. R. Hess, D. Vaulot, *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **63**, 106–127 (1999). Medline

2. L. R. Moore, G. Rocap, S. W. Chisholm, Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* **393**, 464–467 (1998). Medline doi:10.1038/30965

3. Z. I. Johnson, E. R. Zinser, A. Coe, N. P. McNulty, E. M. Woodward, S. W. Chisholm, Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* **311**, 1737–1740 (2006). doi:10.1126/science.1118052

4. G. C. Kettler, A. C. Martiny, K. Huang, J. Zucker, M. L. Coleman, S. Rodrigue, F. Chen, A. Lapidus, S. Ferriera, J. Johnson, C. Steglich, G. M. Church, P. Richardson, S. W. Chisholm, Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLOS Genet.* **3**, e231 (2007). Medline doi:10.1371/journal.pgen.0030231

5. J. Grote, J. C. Thrash, M. J. Huggett, Z. C. Landry, P. Carini, S. J. Giovannoni, M. S. Rappé, Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio.* **3**, e00252-12 (2012). Medline doi:10.1128/mBio.00252-12

6. D. E. Hunt, L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, M. F. Polz, Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**, 1081–1085 (2008). doi:10.1126/science.1157890

7. S. L. Simmons, G. Dibartolo, V. J. Denef, D. S. Goltsman, M. P. Thelen, J. F. Banfield, Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLOS Biol.* **6**, e177 (2008). Medline doi:10.1371/journal.pbio.0060177

8. H. Cadillo-Quiroz, X. Didelot, N. L. Held, A. Herrera, A. Darling, M. L. Reno, D. J. Krause, R. J. Whitaker, Patterns of gene flow define species of thermophilic Archaea. *PLOS Biol.* **10**, e1001265 (2012). Medline doi:10.1371/journal.pbio.1001265

9. A. Gonzaga, A. B. Martin-Cuadrado, M. López-Pérez, C. Megumi Mizuno, I. García-Heredia, N. E. Kimes, P. Lopez-García, D. Moreira, D. Ussery, M. Zaballos, R. Ghai, F. Rodriguez-Valera, Polyclonality of concurrent natural populations of *Alteromonas macleodii*. *Genome Biol. Evol.* **4**, 1360–1374 (2012). Medline doi:10.1093/gbe/evs112

10. R. T. Papke, O. Zhaxybayeva, E. J. Feil, K. Sommerfeld, D. Muise, W. F. Doolittle, Searching for species in haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14092–14097 (2007). Medline doi:10.1073/pnas.0706358104

11. S. Rodrigue, R. R. Malmstrom, A. M. Berlin, B. W. Birren, M. R. Henn, S. W. Chisholm, Whole genome amplification and de novo assembly of single bacterial cells. *PLOS ONE* **4**, e6864 (2009). Medline doi:10.1371/journal.pone.0006864

12. T. Kalisky, P. Blainey, S. R. Quake, Genomic analysis at the single-cell level. *Annu. Rev. Genet.* **45**, 431–445 (2011). Medline doi:10.1146/annurev-genet-102209-163607

13. R. Stepanauskas, Single cell genomics: An individual look at microbes. *Curr. Opin. Microbiol.* **15**, 613–620 (2012). Medline doi:10.1016/j.mib.2012.09.001

14. R. S. Lasken, Genomic sequencing of uncultured microorganisms from single cells. *Nat. Rev. Microbiol.* **10**, 631–640 (2012). Medline doi:10.1038/nrmicro2857

15. Materials and methods are available as supplementary materials on *Science* Online.

16. A. F. Michaels, A. H. Knap, R. L. Dow, K. Gundersen, R. J. Johnson, J. Sorensen, A. Close, G. A. Knauer, S. E. Lohrenz, V. A. Asper, M. Tuel, R. Bidigare, Seasonal patterns of ocean biogeochemistry at the U.S. JGOFS Bermuda Atlantic Time-series Study site. *Deep Sea Res. Part I* **41**, 1013–1038 (1994). doi:10.1016/0967-0637(94)90016-7

17. R. R. Malmstrom, A. Coe, G. C. Kettler, A. C. Martiny, J. Frias-Lopez, E. R. Zinser, S. W. Chisholm, Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J.* **4**, 1252–1264 (2010). Medline doi:10.1038/ismej.2010.60

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). Medline doi:10.1371/journal.pone.0015628

19. M. Mühling, On the culture-independent assessment of the diversity and distribution of *Prochlorococcus*. *Environ. Microbiol.* **14**, 567–579 (2012). Medline doi:10.1111/j.1462-2920.2011.02589.x

20. M. Nei, "Evolution of human races at the gene level." in *Human Genetics, Part A: The Unfolding Genome*, B. Bonné-Tamir, T. Cohen, R. M. Goodman, Eds. (Alan R. Liss, New York, 1982), p. 167.

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**, 4081–4087 (2011). Medline doi:10.1128/JB.00418-11

22. S. Avrani, O. Wurtzel, I. Sharon, R. Sorek, D. Lindell, Genomic island variability facilitates *Prochlorococcus*-virus coexistence. *Nature* **474**, 604–608 (2011). Medline doi:10.1038/nature10172

23. J. Pernthaler, Predation on prokaryotes in the water column and its ecological implications. *Nat. Rev. Microbiol.* **3**, 537–546 (2005). Medline doi:10.1038/nrmicro1180

24. F. Malfatti, F. Azam, Atomic force microscopy reveals microscale networks and possible symbioses among pelagic marine bacteria. *Aquat. Microb. Ecol.* **58**, 1–14 (2009). doi:10.3354/ame01355

25. U. Dobrindt, B. Hochhut, U. Hentschel, J. Hacker, Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* **2**, 414–424 (2004). Medline doi:10.1038/nrmicro884

26. J. F. Crow, M. Kimura, *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).

27. R. D. Barrett, D. Schluter, Adaptation from standing genetic variation. *Trends Ecol. Evol.* **23**, 38–44 (2008). Medline doi:10.1016/j.tree.2007.09.008

28. A. D. Barton, S. Dutkiewicz, G. Flierl, J. Bragg, M. J. Follows, Patterns of diversity in marine phytoplankton. *Science* **327**, 1509–1511 (2010). doi:10.1126/science.1184961

29. F. Rodriguez-Valera, A. B. Martin-Cuadrado, B. Rodriguez-Brito, L. Pasić, T. F. Thingstad, F. Rohwer, A. Mira, Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009). Medline doi:10.1038/nrmicro2235

30. C. C. Thompson, G. G. Silva, N. M. Vieira, R. Edwards, A. C. Vicente, F. L. Thompson, Genomic taxonomy of the genus *Prochlorococcus*. *Microb. Ecol.* **66**, 752–762 (2013). Medline doi:10.1007/s00248-013-0270-8

31. D. K. Steinberg, C. A. Carlson, N. R. Bates, R. J. Johnson, A. F. Michaels, A. H. Knap, Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): A decade-scale look at ocean biology and biogeochemistry. *Deep Sea Res. Part II* **48**, 1405–1447 (2001). doi:10.1016/S0967-0645(00)00148-X

32. E. R. Zinser, A. Coe, Z. I. Johnson, A. C. Martiny, N. J. Fuller, D. J. Scanlan, S. W. Chisholm, Prochlorococcus ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl. Environ. Microbiol.* **72**, 723–732 (2006). Medline doi:10.1128/AEM.72.1.723-732.2006

33. A. Raghunathan, H. R. Ferguson Jr., C. J. Bornarth, W. Song, M. Driscoll, R. S. Lasken, Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71**, 3342–3347 (2005). Medline doi:10.1128/AEM.71.6.3342-3347.2005

34. F. B. Dean, S. Hosono, L. Fang, X. Wu, A. F. Faruqi, P. Bray-Ward, Z. Sun, Q. Zong, Y. Du, J. Du, M. Driscoll, W. Song, S. F. Kingsmore, M. Egholm, R. S. Lasken, Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5261–5266 (2002). Medline doi:10.1073/pnas.082089499

35. R. Stepanauskas, M. E. Sieracki, Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9052–9057 (2007). Medline doi:10.1073/pnas.0700496104

36. T. Woyke, A. Sczyrba, J. Lee, C. Rinke, D. Tighe, S. Clingenpeel, R. Malmstrom, R. Stepanauskas, J. F. Cheng, Decontamination of MDA reagents for single cell whole genome amplification. *PLOS ONE* **6**, e26161 (2011). Medline doi:10.1371/journal.pone.0026161

37. K. Zhang, A. C. Martiny, N. B. Reppas, K. W. Barry, J. Malek, S. W. Chisholm, G. M. Church, Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.* **24**, 680–686 (2006). Medline doi:10.1038/nbt1214

38. T. Woyke, G. Xie, A. Copeland, J. M. González, C. Han, H. Kiss, J. H. Saw, P. Senin, C. Yang, S. Chatterji, J. F. Cheng, J. A. Eisen, M. E. Sieracki, R. Stepanauskas, Assembling the marine metagenome, one cell at a time. *PLOS ONE* **4**, e5299 (2009). Medline doi:10.1371/journal.pone.0005299

39. B. K. Swan, B. Tupper, A. Sczyrba, F. M. Lauro, M. Martinez-Garcia, J. M. González, H. Luo, J. J. Wright, Z. C. Landry, N. W. Hanson, B. P. Thompson, N. J. Poulton, P. Schwientek, S. G. Acinas, S. J. Giovannoni, M. A. Moran, S. J. Hallam, R. Cavicchioli, T. Woyke, R. Stepanauskas, Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 11463–11468 (2013). Medline doi:10.1073/pnas.1304246110

40. B. K. Swan, M. Martinez-Garcia, C. M. Preston, A. Sczyrba, T. Woyke, D. Lamy, T. Reinthaler, N. J. Poulton, E. D. Masland, M. L. Gomez, M. E. Sieracki, E. F. DeLong, G. J. Herndl, R. Stepanauskas, Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**, 1296–1300 (2011). doi:10.1126/science.1203690

41. K. G. Lloyd, L. Schreiber, D. G. Petersen, K. U. Kjeldsen, M. A. Lever, A. D. Steen, R. Stepanauskas, M. Richter, S. Kleindienst, S. Lenk, A. Schramm, B. B. Jørgensen, Predominant Archaea in marine sediments degrade detrital proteins. *Nature* **496**, 215–218 (2013). Medline doi:10.1038/nature12033

42. C. Rinke, P. Schwientek, A. Sczyrba, N. N. Ivanova, I. J. Anderson, J. F. Cheng, A. Darling, S. Malfatti, B. K. Swan, E. A. Gies, J. A. Dodsworth, B. P. Hedlund, G. Tsiamis, S. M. Sievert, W. T. Liu, J. A. Eisen, S. J. Hallam, N. C. Kyrpides, R. Stepanauskas, E. M. Rubin, P. Hugenholtz, T. Woyke, Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437 (2013). Medline doi:10.1038/nature12352

43. S. Rodrigue, A. C. Materna, S. C. Timberlake, M. C. Blackburn, R. R. Malmstrom, E. J. Alm, S. W. Chisholm, Unlocking short read sequencing for metagenomics. *PLOS ONE* **5**, e11840 (2010). Medline doi:10.1371/journal.pone.0011840

44. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002). Medline doi:10.1093/nar/gkf436

45. M. Hamady, C. Lozupone, R. Knight, Fast UniFrac: Facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* **4**, 17–27 (2010). Medline doi:10.1038/ismej.2009.97

46. R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, O. Zagnitko, The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008). Medline doi:10.1186/1471-2164-9-75

47. K. Tamura, J. Dudley, M. Nei, S. Kumar, MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* **24**, 1596–1599 (2007). Medline doi:10.1093/molbev/msm092

48. T. M. Cover, J. A. Thomas, "Entropy, relative entropy and mutual information." in *Elements of Information Theory* (Wiley, Hoboken, NJ, 1991), p. 12.

49. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

50. R. R. Hudson, D. D. Boos, N. L. Kaplan, A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* **9**, 138–151 (1992). Medline

51. S. Nurk, A. Bankevich, D. Antipov, A. A. Gurevich, A. Korobeynikov, A. Lapidus, A. D. Prjibelski, A. Pyshkin, A. Sirotkin, Y. Sirotkin, R. Stepanauskas, S. R. Clingenpeel, T. Woyke, J. S. McLean, R. Lasken, G. Tesler, M. A. Alekseyev, P. A. Pevzner,

Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *J. Comput. Biol.* **20**, 714–737 (2013). Medline doi:10.1089/cmb.2013.0084

52. S. J. Pamp, E. D. Harrington, S. R. Quake, D. A. Relman, P. C. Blainey, Single-cell sequencing provides clues about the host interactions of segmented filamentous bacteria (SFB). *Genome Res.* **22**, 1107–1119 (2012). Medline doi:10.1101/gr.131482.111

53. J. F. Kingman, The coalescent. *Stochastic Process. Appl.* **13**, 235–248 (1982). doi:10.1016/0304-4149(82)90011-4

54. J. M. Akey, G. Zhang, K. Zhang, L. Jin, M. D. Shriver, Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002). Medline doi:10.1101/gr.631202

55. R. Nielsen, Molecular signatures of natural selection. *Annu. Rev. Genet.* **39**, 197–218 (2005). Medline doi:10.1146/annurev.genet.39.073003.112420

56. L. B. Barreiro, G. Laval, H. Quach, E. Patin, L. Quintana-Murci, Natural selection has driven population differentiation in modern humans. *Nat. Genet.* **40**, 340–345 (2008). Medline doi:10.1038/ng.78

57. R. R. Hudson, "Gene genealogies and the coalescent process." in *Oxford Surveys in Evolutionary Biology*, D. Futuyma, J. Antonovics, Eds. (Oxford Univ. Press, New York, 1990), vol. 7, p. 44.

58. R. R. Hudson, Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002). Medline doi:10.1093/bioinformatics/18.2.337

59. A. Rambaut, N. C. Grassly, Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**, 235–238 (1997). Medline

60. S. Kryazhimskiy, J. B. Plotkin, The population genetics of dN/dS. *PLOS Genet.* **4**, e1000304 (2008). Medline doi:10.1371/journal.pgen.1000304

61. H. Akashi, Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139**, 1067–1076 (1995). Medline

62. J. V. Chamary, L. D. Hurst, Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals. *Genome Biol.* **6**, R75 (2005). Medline doi:10.1186/gb-2005-6-9-r75

63. L. Kelly, K. H. Huang, H. Ding, S. W. Chisholm, ProPortal: A resource for integrated systems biology of *Prochlorococcus* and its phage. *Nucleic Acids Res.* **40**, D632–D640 (2012). Medline doi:10.1093/nar/gkr1022

64. A. E. Darling, B. Mau, N. T. Perna, progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLOS ONE* **5**, e11147 (2010). Medline doi:10.1371/journal.pone.0011147

65. P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, C. F. Weber, Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial

communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009). Medline doi:10.1128/AEM.01541-09

66. J. S. Guasto, R. Rusconi, R. Stocker, Fluid mechanics of planktonic microorganisms. *Annu. Rev. Fluid Mech.* **44**, 373–400 (2012). doi:10.1146/annurev-fluid-120710-101156

67. M. T. Landahl, *Turbulence and Random Processes in Fluid Mechanics* (Cambridge Univ. Press, Cambridge, 1992).

68. P. Hill, A. Nowell, P. Jumars, Encounter rate by turbulent shear of particles similar in diameter to the Kolmogorov scale. *J. Mar. Res.* **50**, 643–668 (1992). doi:10.1357/002224092784797539

69. T. Kiorboe, *A Mechanistic Approach to Plankton Ecology* (Princeton Univ. Press, Princeton, NJ, 2008).

70. A. Okubo, Oceanic diffusion diagrams. *Deep Sea Res. Oceanogr. Abstr.* **18**, 789–802 (1971).

71. A. Okubo, S. A. Levin, *Diffusion and Ecological Problems: Modern Perspectives* (Springer, New York, 2001), vol. 14.

72. T. M. Powell, A. Okubo, Turbulence, diffusion and patchiness in the sea. *Philos. Trans. R. Soc. London Ser. B Biol. Sci.* **343**, 11–18 (1994). doi:10.1098/rstb.1994.0002

73. J. R. Ledwell, A. J. Watson, C. S. Law, Mixing of a tracer in the pycnocline. *J. Geophys. Res. Oceans* **103**, 21499–21529(1998). doi:10.1029/98JC01738

74. S. Wright, Evolution in Mendelian populations. *Genetics* **16**, 97–159 (1931). Medline

75. J. M. Smith, N. H. Smith, Synonymous nucleotide divergence: What is “saturation”? *Genetics* **142**, 1033–1036 (1996). Medline

76. M. Lynch, J. S. Conery, The origins of genome complexity. *Science* **302**, 1401–1404 (2003). doi:10.1126/science.1089370

77. B. Charlesworth, Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205 (2009). Medline doi:10.1038/nrg2526

78. M. S. Osburne, B. M. Holmbeck, A. Coe, S. W. Chisholm, The spontaneous mutation frequencies of *Prochlorococcus* strains are commensurate with those of other bacteria. *Environ. Microbiol. Rep.* **3**, 744–749 (2011). Medline doi:10.1111/j.1758-2229.2011.00293.x

79. H. A. Orr, A. J. Betancourt, Haldane’s sieve and adaptation from the standing genetic variation. *Genetics* **157**, 875–884 (2001). Medline

80. T. Karasov, P. W. Messer, D. A. Petrov, Evidence that adaptation in *Drosophila* is not limited by mutation at single sites. *PLOS Genet.* **6**, e1000924 (2010). Medline doi:10.1371/journal.pgen.1000924

81. J. Hermisson, P. S. Pennings, Soft sweeps: Molecular population genetics of adaptation from standing genetic variation. *Genetics* **169**, 2335–2352 (2005). Medline doi:10.1534/genetics.104.036947

82. J. A. G. de Visser, D. E. Rozen, Clonal interference and the periodic selection of new beneficial mutations in *Escherichia coli*. *Genetics* **172**, 2093–2100 (2006). Medline doi:10.1534/genetics.105.052373

83. M.-C. Lee, C. J. Marx, Synchronous waves of failed soft sweeps in the laboratory: Remarkably rampant clonal interference of alleles at a single locus. *Genetics* **193**, 943–952 (2013). Medline doi:10.1534/genetics.112.148502

84. S.-C. Park, J. Krug, Clonal interference in large populations. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 18135–18140 (2007). Medline doi:10.1073/pnas.0705778104

85. P. Marttinen, W. P. Hanage, N. J. Croucher, T. R. Connor, S. R. Harris, S. D. Bentley, J. Corander, Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40**, e6 (2012). Medline doi:10.1093/nar/gkr928

86. S. Castillo-Ramírez, J. Corander, P. Marttinen, M. Aldeljawi, W. P. Hanage, H. Westh, K. Boye, Z. Gulay, S. D. Bentley, J. Parkhill, M. T. Holden, E. J. Feil, Phylogeographic variation in recombination rates within a global clone of methicillin-resistant *Staphylococcus aureus*. *Genome Biol.* **13**, R126 (2012). Medline doi:10.1186/gb-2012-13-12-r126

87. N. J. Croucher, S. R. Harris, C. Fraser, M. A. Quail, J. Burton, M. van der Linden, L. McGee, A. von Gottberg, J. H. Song, K. S. Ko, B. Pichon, S. Baker, C. M. Parry, L. M. Lambertsen, D. Shahinas, D. R. Pillai, T. J. Mitchell, G. Dougan, A. Tomasz, K. P. Klugman, J. Parkhill, W. P. Hanage, S. D. Bentley, Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011). doi:10.1126/science.1198545

88. A. Tsoularis, J. Wallace, Analysis of logistic growth models. *Math. Biosci.* **179**, 21–55 (2002). Medline doi:10.1016/S0025-5564(02)00096-2

89. H. Ochman, A. C. Wilson, Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26**, 74–86 (1987). Medline doi:10.1007/BF02111283

90. H. Ochman, S. Elwyn, N. A. Moran, Calibrating bacterial evolution. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12638–12643 (1999). Medline doi:10.1073/pnas.96.22.12638

91. A. Dufresne, L. Garczarek, F. Partensky, Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol.* **6**, R14 (2005). Medline doi:10.1186/gb-2005-6-2-r14