Title
Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus.

Permalink
https://escholarship.org/uc/item/2c3638zx

Journal
Environmental microbiology, 11(4)

ISSN
1462-2912

Authors
Martiny, Adam C
Tai, Amos PK
Veneziano, Daniele
et al.

Publication Date
2009-04-01

DOI
10.1111/j.1462-2920.2008.01803.x

Supplemental Material
https://escholarship.org/uc/item/2c3638zx#supplemental

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed
Taxonomic resolution, ecotypes and the biogeography of Prochlorococcus

Adam C. Martiny,1,2,3,4 Amos P. K. Tai,† Daniele Veneziano,1 François Primeau2 and Sallie W. Chisholm1,**
1Department of Civil & Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.
Departments of 2Earth System Science and 3Ecology and Evolutionary Biology, University of California – Irvine, Irvine, CA 92697, USA.

Summary
In order to expand our understanding of the diversity and biogeography of Prochlorococcus ribotypes, we PCR-amplified, cloned and sequenced the 16S/23S rRNA ITS region from sites in the Atlantic and Pacific oceans. Ninety-three per cent of the ITS sequences could be assigned to existing Prochlorococcus clades, although many novel subclades were detected. We assigned the sequences to operational taxonomic units using a graduated scale of sequence identity from 80% to 99.5% and correlated Prochlorococcus diversity with respect to environmental variables and dispersal time between the sites. Dispersal time was estimated using a global ocean circulation model. The significance of specific environmental variables was dependent on the degree of sequence identity used to define a taxon: light correlates with broad-scale diversity (90% cut-off), temperature with intermediate scale (95%) whereas no correlation with phosphate was observed. Community structure was correlated with dispersal time between sample sites only when taxa were defined using the finest sequence similarity cut-off. Surprisingly, the concentration of nitrate, which cannot be used as N source by the Prochlorococcus strains in culture, explains some variation in community structure for some definitions of taxa. This study suggests that the spatial distribution of Prochlorococcus ecotypes is shaped by a hierarchy of environmental factors as well dispersal limitation.

Introduction
The marine cyanobacterium Prochlorococcus is an important contributor to global nutrient cycles, as it contributes a significant fraction of the primary production in mid-latitude oceans (Liu et al., 1997). Prochlorococcus is known to thrive under a wide range of environmental conditions, which in part may be due to the existence of several closely related (> 97% 16S rRNA similarity) but physiological and genetically distinct lineages. At broad taxonomic resolution, Prochlorococcus strains can be divided into two ecotypes, high-light (HL)-adapted and low-light (LL)-adapted (Moore et al., 1998). Most of the LL-adapted ecotypes are found in significant abundance only in deeper waters, while the HL-adapted cells typically dominate surface waters (West and Scanlan, 1999). The HL and LL groups can be further divided into at least six clades (two HL- and four LL-adapted) based on phylogenies constructed using the 16S/23S rRNA intergenic spacer sequences (ITS) (Rocap et al., 2002). At this intermediate level of taxonomic distinction, temperature and mixing intensity are important determinants of the observed distribution patterns of the different taxa (Bouman et al., 2006; Johnson et al., 2006). In contrast, nutrient concentrations do not seem to predict Prochlorococcus ecotype distributions very well (Johnson et al., 2006; Martiny et al., 2006a; Zvirgmaier et al., 2008).

The studies, upon which our understanding of Prochlorococcus diversity in the field has been established, were based on the detection of rRNA sequence diversity with PCR primers and probes designed from a collection of cultured Prochlorococcus strains and a few sequences from clone libraries constructed from wild populations of Prochlorococcus (West and Scanlan, 1999; Ahlgren et al., 2006). These methods do not capture the entire Prochlorococcus community; the number of Prochlorococcus cells counted by flow cytometry often exceeds that detected with existing molecular methods, primarily for LL-adapted cells in the deep water (Ahlgren et al., 2006; Zinser et al., 2006). This under-representation could influence our understanding of the relationship between specific environmental variables and the composition of Prochlorococcus communities. Therefore, an initial goal of
this study was to gain a more complete picture of *Prochlorococcus* diversity by analysing sequence diversity directly, by cloning and sequencing ITS regions of *Prochlorococcus* cells, from 10 stations in the Atlantic and Pacific oceans. Using this expansive data set we then asked: How is community composition, as defined at different taxonomic resolutions, related to geographic separation and environmental heterogeneity?

**Results and discussion**

*Prochlorococcus* ribotype diversity

To identify the genetic diversity of *Prochlorococcus* communities, we PCR-amplified, cloned and sequenced the 16S/23S rRNA ITS region from *Prochlorococcus* from 10 sites in the Atlantic and Pacific Ocean (Fig. 1) at three depths, targeting the surface mixed layer, below mixed layer and near the bottom of the euphotic zone (Table S1). The primers were designed to capture all *Prochlorococcus* but not *Synechococcus* (see Experimental procedures), although this distinction can never be guaranteed. Phylogenies constructed with the ITS sequence relate well to those constructed with 16S rRNA sequences, but provide a higher resolution for delineating closely related genotypes (Rocap et al., 2002; Kettler et al., 2007).

Ninety-three per cent of the ITS sequences from this data set can be assigned to the described major *Prochlorococcus* clades (Fig. 2A and B). As expected, nearly all sequences collected from the surface mixed layer cluster in the HL-adapted clades eMED4 and eMIT9312. Most of the sequences that cluster with the LL-adapted clades eMIT9313, eNATL and eSS120/eMIT9211 were retrieved below the mixed layer or deep euphotic zone. Consistent with the ‘intermediate light adaptation status’ of the eNATL clade (Kettler et al., 2007; Zinser et al., 2007), the eNATL clade is densely populated by sequences collected just below the mixed layer (yellow lineages, Fig. 2B), whereas the other LL-adapted clades are not.

While the clade assignments of the sequences are consistent, for the most part, with the observations from field studies using quantitative PCR (qPCR) and probes (West and Scanlan, 1999; Zinser et al., 2007), this expanded data set reveals numerous new sublineages within each clade. Most of the sequences belonging to the LL-adapted ecotypes eMIT9313, eNATL and eSS120/eMIT9211 would not have been targeted by the previously designed primers for qPCR assays (see ‘% of clones with primer site’, Table S2 for details). This likely explains most of the gap between the total flow cytometry counts, and the cell counts determined by qPCR reported here (Table S2) and in several previous studies (Ahlgren et al., 2006; Zinser et al., 2006). Importantly, we also detected a new deeply branching group within the LL-adapted clusters, which we termed NC1 (i.e. Non-Cultured 1). Since we do not have a cultured isolate from NC1, we cannot say for sure if this group contains divinyl-chlA as the main pigment (the defining feature of prochlorophytes). However, the phylogenetic position suggests that this group is part of the *Prochlorococcus* radiation. NC1 constituted 7% of the total number of sequences, and is found almost exclusively in the deep samples, where it constitutes, on average, 23.1% of the sequences (Table S2). It is unclear if NC1 is monophyletic or constitutes multiple independent lineages, since many lineages originate from the same point in the tree in a star-like manner. This is likely due to...
Fig. 2. Phylogenetic diversity of Prochlorococcus.

A. Visualization of 16S/23S rRNA intergenic transcribed spacer (ITS) sequence distances of cultured isolates of Prochlorococcus (note log-scale) using the unrelated pair-group method with arithmetic mean (UPGMA). Bootstrap values are calculated with the phylogenetic methods neighbour joining, maximum parsimony and maximum likelihood.

B. Phylogenetic tree constructed using the ITS sequences of cultured Prochlorococcus and Synechococcus isolates (black lines), and cloned sequences from the wild obtained for this study (coloured lines). Red lines indicate surface cells, yellow ‘below mixed layer’ and blue ‘deep’ (see also Table S1). Prochlorococcus strain names are shown in black. The clades are identified using the ‘e’ terminology (grey font), which names significant clades according to their type strains in culture (Ahlgren et al., 2006). NC1 is a new group identified in this study. The tree is a PHYLIP majority consensus of bootstrap resampling (100) using FastME (Desper and Gascuel, 2002).
the limited resolution of the ITS locus for deeply branching nodes within the LL-adapted branches of the *Prochlorococcus* radiation (Rocap et al., 2002; Kettler et al., 2007). Based on their phylogeny and depth distribution, we predict cells affiliated to NC1 to be LL-adapted.

In order to address the correspondence between *Prochlorococcus* community composition and environmental variation, we needed to understand to what degree the representation of different ITS sequences in our clone libraries is a measure of their actual relative abundance in *Prochlorococcus* populations in the field. In other words, is a bias introduced when we amplify and clone the sequences? To answer this question, we measured the abundance of cells belonging to specific clades using the standard qPCR primers used in the past, and compared this with the relative frequency of sequences containing no more than one mismatch to the qPCR primers times the total abundance measured with flow cytometry. These two measures of abundance correspond particularly well for ecotypes that are present at higher abundances (Fig. 3), whereas we observed higher variance among samples with low abundance of *Prochlorococcus* (Fig. 2A). When a taxon is defined more stringently, i.e. when cells with at least 97% or 99.5% sequence identity are assigned to the same taxon, the number of OTUs increases significantly in most samples (Fig. 4B and C) and some rarefaction curves are not close to saturating. Many of these narrower taxa were only observed once (Fig. 4D). While we have clearly not sampled all the fine-scale diversity, we have good coverage of most samples at the 97% or lower cut-off levels. Sequencing more clones, however, would likely not change our observation of higher richness in the deep water samples compared with surface samples (Shaw et al., 2008). We observed an interesting trend whereby the population is large close to the sunlit surface but only contains a few taxa at broad cut-offs. In contrast, fewer *Prochlorococcus* cells live deeper in the photic zone but these cells belong to many divergent types. This trend is largely a reflection of the more similar HL-adapted lineages compared with the

**Extent of sampling coverage**

In preparation for analysis, we resolved sequences into taxa using a sliding scale of ITS sequence similarity cut-offs ranging from $\geq 80\%$ to $\geq 99.5\%$ similarity, thereby creating a range of phylogenetic distances applied in defining a taxon, or operational taxonomic unit (OTU). Representative rarefaction distances for the $\geq 90\%$, $\geq 97\%$ and $\geq 99.5\%$ cut-off definitions are shown in Fig. 4A–C. When a taxon is defined broadly, i.e. as all cells with 90% or more sequence similarity, there are only 1–5 OTU in the surface and mid-depth samples at most stations (Fig. 4A; Table S3). All but two of the deepest water samples had a dramatically higher number of OTUs than did the shallower depths (Table S3), reflecting the more divergent LL-adapted clades of *Prochlorococcus* (Fig. 2A). When a taxon is defined more stringently, i.e. when cells with at least 97% or 99.5% sequence identity are assigned to the same taxon, the number of OTUs increases significantly in most samples (Fig. 4B and C) and some rarefaction curves are not close to saturating. Many of these narrower taxa were only observed once (Fig. 4D). While we have clearly not sampled all the fine-scale diversity, we have good coverage of most samples at the 97% or lower cut-off levels. Sequencing more clones, however, would likely not change our observation of higher richness in the deep water samples compared with surface samples (Shaw et al., 2008). We observed an interesting trend whereby the population is large close to the sunlit surface but only contains a few taxa at broad cut-offs. In contrast, fewer *Prochlorococcus* cells live deeper in the photic zone but these cells belong to many divergent types. This trend is largely a reflection of the more similar HL-adapted lineages compared with the
more divergent LL types. This may be driven by increased growth rate and predation in the surface waters potentially leading to more rapid purging of less competitive types.

Biogeography of Prochlorococcus

Next, we examined the variation of Prochlorococcus communities in different oceanic regions. We first compared the composition at the Hawaii Ocean Time-Series (HOT) and the Bermuda Atlantic Time-Series (BATS) sites (station 1A, 1B, 2A and 2B, three depths from each station). At the 90% sequence similarity cut-off, 84% of the sequences fall within a taxon that is present at both HOT and BATS, whereas only 2% at HOT and 0% at BATS are found in more than one sample but exclusively at one site. At the 97% or 99.5% cut-off, the number of shared sequences drops to 59% and 9% respectively. However, the number of multiple-sample, single-site sequences increases to 13% and 7% (HOT and BATS) at the 97% cut-off and 29% and 18% at the 99.5% cut-off. Thus, the number of shared sequences increases between these two oceanic regions at broader definitions of a taxon whereas we find more unique types in a given region at finer cut-offs. This suggests the presence of microdiversity specific for a given region (i.e. non-random distribution) whereas broader phylogenetic groups within Prochlorococcus are cosmopolitan (i.e. as a result of similar environmental conditions).

To investigate this further, we analysed the degree of similarity of community structure between all samples as a function of the time it would take a hypothetical cell to be physically transported from one site to another, i.e. dispersal time. Similarity between communities was estimated using the Bray–Curtis index. The dispersal time between each sampling site was estimated by a global ocean circulation model (Fig. S1) (Primeau, 2005). We then asked whether any of the differences between communities could be explained by variation in dispersal time alone using a Mantel test (Mantel, 1967) (Table S4). Community composition was correlated to dispersal time only when taxa were defined at the 99.5% cut-off – i.e. when very fine-scale diversity was resolved – suggesting that local speciation and dispersal limitation establish endemic microdiverse Prochlorococcus populations. Due to the large population size and relatively short generation time, cells may evolve faster than ocean currents can mix them and this may maintain local microdiversity.
Influence of environmental factors on Prochlorococcus assemblage structure

We next explored the influence of environmental variation on the biogeography of Prochlorococcus by applying canonical correspondence analysis (CCA) (ter Braak, 1986) (see Experimental procedures). The environmental variables used in the analysis were light intensity, temperature, nitrate and phosphate concentration (Fig. 5; Table S1) – all important physical and chemical regulators of the growth of marine phytoplankton. When Prochlorococcus is divided into clusters of 97% or higher ITS sequence similarity, nitrate concentration explains a significant part of the genetic variation (Fig. 5). However, when clustering is based on taxa with 95% or more similarity, temperature and light become significant factors. Finally, when Prochlorococcus is divided into large phylogenetic groups (approximately 90% or more similarity), light is the most important descriptor. Variation in phosphate concentration does not correlate with composition at any cut-off, consistent with previous studies showing that there is no relationship between ribotype and genome composition with respect to P-acquisition genes (Martiny et al., 2006a). Despite the significant correlation between environmental variables and community composition, we also see significant unexplained variance, highlighting our incomplete understanding of the factors that structure this system, including selection by grazers and phage.

We also investigated the relative influence of environmental factors on the distribution of specific pairs of major phylogenetic clades (Fig. S2) as clusters with comparable levels of similarity may be associated with different environmental variables. (The analyses in Fig. 5 consider all clades at a specified similarity cut-off level and therefore return only overall average effects.) In congruence with the CCA analysis, we observed that light explained most of the variation of the distribution of HL versus LL clades, whereas temperature was the most important descriptor of the difference in distribution of the two HL clades. Furthermore, nitrate was moderate descriptor of the distribution of many clades.

Given that no cultured Prochlorococcus strains can grow on nitrate as an N source (Moore et al., 2002), it is perhaps surprising that nitrate concentration explains some of the variation in Prochlorococcus community composition (Fig. 5). There is evidence from field studies, however, that Prochlorococcus cells in the wild can use nitrate (Casey et al., 2007), and this could explain this correlation. Considering all the genetic diversity we have observed in this study, there is no reason to believe that the culture collections represent all phenotypic diversity. Furthermore, culture collections are dominated by HL-adapted Prochlorococcus ecotypes, which may be less likely to retain this trait because nitrate is commonly most abundant in deeper waters. Alternatively, the correlation between nitrate and Prochlorococcus diversity could be indirect and reflect abundance of competitors like picoeukaryotes and Synechococcus that can use nitrate.

Environmental parameters also vary with distance. Thus, the correlation between dispersal time and fine-scale diversity described in the previous section could be a result of co-variation with an environmental factor. To address this, we removed the variance of each environmental factor using a partial Mantel test, and then compared dispersal time and composition (Table S4). The conclusions were the same. There was only a significant relationship between dispersal time and community structure when taxa were defined with a 99.5% cut-off.

General conclusions

From this and previous studies, a more coherent picture of the ecology, evolution and community structure of Prochlorococcus assemblages is emerging. It is already well established that the group can be divided into two major phylogenetic groups – one HL- and one LL-adapted (Moore et al., 1998) (Fig. 2A), and the HL-adapted group has two subgroups (eMIT9312 and eMED4) that have different temperature optima, and are distributed differently along temperature gradients (Johnson et al., 2006). Furthermore, some of the LL-adapted strains can use nitrate, whereas none has been shown to use nitrite (Moore et al., 2002). There is evidence, however, that some Prochlorococcus cells in the wild utilize nitrate (Casey et al., 2007). In addition, several studies suggest...
that the P-acquisition system in Prochlorococcus is not related to phylogeny, but rather the ocean of origin of the cell in question (Martiny et al., 2006a; Rusch et al., 2007). Thus, taxonomy should have no predictive power with regard to this environmental variable.

The results of this study are consistent with this growing understanding of Prochlorococcus ecology and evolution, but take it one step further. That is, consistent with earlier studies, we found that light intensity was an important determinant of community structure when taxa are defined with broad resolution, temperature becomes important at intermediate level resolution and that P-availability does not explain community structure, regardless of taxonomic resolution. In addition, we found that nitrate availability explains some of the variability in Prochlorococcus community structure when taxa are defined with fine resolution, which is consistent with the evidence that some cells in the wild have the capability to utilize nitrate. The consistency of all of these observations, given the diversity of approaches used for analysis, is remarkable. We also found that dispersal time was correlated with assemblage composition when it was analysed using finely resolved taxonomic resolution. This suggests that cosmopolitan ecotypes inhabit local communities, which likely can recruit ecotypes from the entire metacommunity due to rapid dispersal and thereby belong to one ‘province’ (sensu Martiny et al., 2006b). However, due to the large population size and short generation time, Prochlorococcus cells may evolve faster than ocean currents can mix them, which results in locally distinct microdiversity.

The hierarchical association between environmental gradients and phylogenetic depth we observed may be shaped by the rate of adaptation of cells to a new environment. If lateral gene transfer (LGT) of a single gene or a few point mutations can alter a functional trait, shifts in phenotype may be common and the phylogenetic depth of an ecotype is likely to be shallow. At the other extreme would be the case of light adaptation, which involves a large number of interacting proteins and may not easily change (Shi et al., 2005), thus we see deeply branching sequence clusters with respect to this trait (Kettler et al., 2007). Thus, an ecotype is likely to encompass sequence clusters of various sizes and this results in different patterns of Prochlorococcus biogeography depending on how you define a taxon. Putative nitrate utilizing Prochlorococcus are a case in point: if indeed our observed relationship between nitrate availability and Prochlorococcus community structure reflects the presence of microdiverse clusters of yet-to-be cultured Prochlorococcus lineages that can utilize nitrate, this would be consistent with this hierarchical model. Nitrate assimilation in cyanobacteria involves nitrate and nitrite reductase, a transporter and multiple genes responsible for the biosynthesis of a cofactor (molybdopterin) (Flores et al., 2005). Thus, this trait might be expected to be more complex than phosphate acquisition, but less complex than light adaptation, and therefore be associated with specific phylogenetic clades at relatively fine resolution.

**Experimental procedures**

### Sampling and sequence analysis

We used 100 ml of water samples collected from the HOT, BATS and AMT13 (Johnson et al., 2006) cruise at three depths (approximately 25, 75 and 150 m, see Table S1) and DNA was extracted as described elsewhere (Zinder et al., 2006). Primers targeting 16S rRNA (2F: GAAGTCGT TACTYAAACC) and 23S rRNA (3R: TCATCGCTCTTGT GTGCC) were designed to amplify the ITS region targeting all Prochlorococcus lineages and a few Synechococcus based on rRNA sequence information from cultured isolates (Rocap et al., 2002). The Prochlorococcus specificity of the primers was furthermore independently verified using previously published environmental sequence libraries (Rocap et al., 2002; Venter et al., 2004; Brown et al., 2005). Sequences that clearly clustered phylogenetically within the marine Synechococcus clade were removed from the analysis. All others were retained.

Initially, the number of PCR cycles necessary for a very faint band on an agarose gel was determined in order to apply a minimum number of PCR cycles (20–30 cycles). We amplified 10 parallel reactions for each sample followed by a three-cycle reconditioning step (Thompson et al., 2002) using an annealing temperature of 62°C. The amplified DNA was pooled, gel-purified and cloned using TOPO TA cloning ver. 4.0 and 75 inserts from each sample were sequenced at Genaissance Pharmaceuticals (CT, USA). All base calls were manually verified using Sequencer (ver. 4.6 Gene Codes, MI, USA) and chimeric sequences were identified using Bellerophon (Huber et al., 2004).

To avoid analysing artificially introduced genetic diversity in our samples, we determined the sequence error introduced during PCR amplification and sequencing. This was done in three ways: (i) we amplified and sequenced two known ITS sequences along with each of the samples (Prochlorococcus MED4), (ii) we re-sequenced 300 fragments, and (iii) we identified base variation in two highly conserved tRNA sequences inserted in the ITS fragment. The latter yields an overestimation of error due to natural base variation in some Prochlorococcus ecotypes, e.g. ENATL sequences (Rocap et al., 2002). The three methods gave an error frequency of $1.5 \times 10^{-3}$, $2.7 \times 10^{-3}$ and $2.9 \times 10^{-3}$ respectively. Based on this estimation, we only interpret sequence variation higher than 0.5%.

Sequences were aligned using a custom database in ARB (Ludwig et al., 2004). A sequence alignment excluding the two internal tRNAs was exported and a distance matrix was generated using logdet correction. Neighbour-joining, maximum parsimony and maximum likelihood bootstrapping and consensus trees were calculated using PHYLIP (Felsenstein, 2006). Sequence differences of Prochlorococcus strains were visualized on a log-scale using unweighted pair group method with arithmetic mean (UPGMA) in Matlab (Mathworks, MA, USA) (Fig. 2A). The phylogenetic position
of all sequences was determined using a distance-based method [FastME (Desper and Gascuel, 2002)] and visualized in HyperTree (Bingham and Sudarsanam, 2000) (Fig. 2B).

To evaluate potential PCR bias in our clone library, we compared the abundance of specific clades in the seawater sample measured by qPCR with the frequency of sequences targeted by these primers in each clone library (see Table S2 for the details and the data underlying this calculation). Data from BATS were excluded from this analysis because flow cytometry samples were not available for these samples. We scored a sequence as targeted by qPCR if both priming sites contained one or less mismatch per primer (Whiley and Sloots, 2005). Since the eMIT9312 primers targets a region outside the ITS locus, we assigned sequences as eMIT9312 based on their phylogenetic affiliation. We then calculated the abundance of each clade in our libraries (as defined by the qPCR primers) by multiplying the frequency of sequences containing the primer target sites multiplied by the total abundance of Prochlorococcus measured with flow cytometry. It is important to note that there is variation in the number of rRNA (and ITS) copies among bacteria, including Prochlorococcus. Of all the Prochlorococcus genomes sequenced to date, however, only two from the eMIT9313 clade have two rRNA operons; all other Prochlorococcus genomes have only one. Obviously we do not know how many operons novel lineages contain, including the potential for variation within the major clades. Given this uncertainty of rRNA copies in each cell, we simplified our analysis by assuming one copy per cell for all cells. However, if we allowed for two copies in eMIT9313, this would not change our conclusion that abundance measured by qPCR is well correlated to abundance inferred from clone libraries. Clades present in a sample at abundance less than 0.9% were excluded from the analysis, since there was less than 50% chance of detecting them in our clone library when 75 sequences were analysed. Sequences from this study are submitted to GenBank under Accession Nos FJ179678–FJ181998.

**Dispersal time analysis**

The dispersal time between each sample site was estimated with a global ocean circulation model (Primeau, 2005). The model has 29 depth layers and an approximate horizontal resolution of 3.75° × 3.75°. Unit pulses of tracer were injected at each site and allowed to disperse freely to all other sites. The dispersal time was then defined as the time at which the tracer concentration reached 10%, 25% and 50% of their long time asymptotic concentration (Fig. S1). For sites located in the same grid cell, we used sample time difference as estimation of dispersal time. Samples at the same station but different depths were assigned a dispersal time of 0.1 years.

The dispersal time between stations was estimated from the Green function (G) for the advection-diffusion equation governing the evolution of passive tracers in the ocean:

\[
\frac{\partial}{\partial t} G(t, r, r') + \nabla \cdot (u - K V) G(t, r, r') = V \delta(t) \delta(r - r')
\]

\[
\nabla G(t, r, r') = 0
\]

\[
G(0, r, r') = 0
\]

In the above equation, u is the annually average fluid velocity field, and K is the annually averaged subgrid scale eddy diffusivity tensor. The Green function, \(G(t, r, r')\), is the response at the sample station (Fig. 1), r, due to a Dirac-delta function pulse of tracer at the source point, \(r'\) scaled by the total ocean volume, \(V\). As defined, the Green function is dimensionless and has a long time asymptotic value of one. For each site location, \(r, G(t, r, r')\) was computed as a function of t and r. The time, \(t(t, r')\), at which \(G(t, r, r')\) reached the threshold value, was recorded for each pair of site.

To solve the above partial differential equation, we used a finite difference approximation in both space and time. For the time discretization we used a second-order accurate Crank Nicholson scheme. For the space discretization we evaluated three different schemes: (i) a second-order accurate centred difference scheme, (ii) a first-order accurate upwind scheme for the first year and then switched to the second-order accurate centred difference scheme for the rest of the integration, and (iii) the first-order accurate upwind scheme for the first 10 years and then switched to the second-order accurate scheme for the rest of the integration. In the presence of sharp tracer gradients such as those that occur in the early part of the integration, the centred difference scheme can produce negative tracer concentrations that are difficult to interpret. The first-order upwind scheme, while being less accurate, does not produce unphysical negative tracer concentrations.

**Genetic composition analysis**

Individual taxa (or OTUs) at specific sequence similarity cut-offs were assigned with DOTUR (Schloss and Handelsman, 2005). The community composition of each sample at a given cut-off was defined as the specific occurrence (i.e. number of sequences) assigned to each taxon. Rarefaction curves (Coleman analytical method) and community indices were calculated using EstimateS (Colwell, 2008).

The difference in community composition between the samples was found using a square-root transformed Bray–Curts sample similarity matrix determined in PRIMER v5 (PrimerE, UK). Per cent surface light intensity, temperature, phosphate and nitrate concentrations are compiled from AMT cruise data (Johnson et al., 2006), HOT (Karl, 1999) and BATS (Steinberg et al., 2001) (see Table S1 for values). In cases where values of zero were reported among environmental variables, we re-assigned the values to the minimum non-zero value. Mantel and partial mantel tests of dispersal time (log transformed), environmental variables (log transformed) and community composition were performed in R using the ecostat package (10 000 permutations). We also tested the effect of non-transformed and the commonly used non-linear log(x + 1) transformation but saw no significant change in the correlation patterns. The variance contribution of individual environmental factors on Prochlorococcus community composition was determined with partial canonical correspondence analysis (forward selection, \(\alpha = 0.05\) and 9999 perturbations) using CANOCO (ver. 4.0, Microcomputer Power, NY, USA) (ter Braak, 1986). Identification of taxa with significant correspondence to elevated nitrate concentration was performed using CANOCO ordination plots visualized in CanoDraw.
The influence of each environmental factor on specific phylogenetic clades is also measured by the R-value ($R^2$ being the fraction of each environmental parameter’s variance explained by the classification) (Fig. S2).

Acknowledgements

We thank Ed Delong and Dennis Ryan for help with phylogenetic analysis. We also thank Jennifer Hughes Martin, Jorge Frias-Lopez, Rick Follis, John Avise, Steven Allison, Rex Malmstrom and Jason Bragg for many helpful discussions throughout this work, Daniel Sher, Sarah Bagby and Katya Frois-Moniz for detailed comments on the manuscript, Jed Fuhrman and Mark Brown for early access to unpublished ITS sequences and two anonymous reviewers for many helpful comments. Research was supported in part by NSF Biological Oceanography, NSF C-MORE, DOE GTL programme, the Gordon and Betty Moore Foundation (to S.W.C.), University of California, and the Danish National Science Foundation (to A.C.M.). Also, this study was supported by the UK Natural Environment Research Council through the Atlantic Meridional Transect consortium (NER/ O/ S2001/00680). This is contribution number 173 of the AMT programme.

References


Shaw, A.K., Halpern, A.L., Beeson, K., Tran, B., Venter, J.C., and Martiny, J.B. (2008) It’s all relative: ranking the diver-


**Fig. S2.** Influence of specific environmental variables on the distribution of specific phylogenetic nodes. The $R$-value is estimated as the difference in explanatory power of an environmental variable on the distribution of combined node or the two splitting lineages. For eMIT9313 and NC1, $R$ is calculated as the difference in explanatory power between each node and all *Prochlorococcus*.

**Table S1.** Location, sampling date and corresponding environmental values of samples analysed.

**Table S2.** Details of the data set that was used in this article. Total *Prochlorococcus* numbers were measured by flow cytometry (FCM) to identify the total number of cells in the population in each sample. The ecotype clades are as defined by Ahlgren and colleagues (2006). The abundance of each is shown as measured by previously designed qPCR primers (Zinser *et al.*, 2006). Also shown is the per cent of the clones sequenced in this study that would be captured by those primers, and the per cent of those clones that are taxonomically affiliated with the clade, regardless of whether they have primer sites or not. NC1 is the per cent of the clones that are affiliated taxonomically with the new group (NC1) identified in Fig. 2. These data were used to determine whether the relative frequency of the clones that contained primer sites matched that of the relative abundance of ecotypes measured by qPCR (Fig. 3). That is, to determine if there was a bias in the collection of the cloned sequences used in this study. Seventy-five clones were sequenced at each station except #7, where only 50 were analysed.

**Table S3.** Characterization of the *Prochlorococcus* assemblage in each sample as a function of three definitions of a taxon: $\geq 90\%$, $\geq 97\%$ and $\geq 99.5\%$ sequence similarity. The richness (# OTUs) and evenness of community composition is shown as a function of taxon definition.

**Table S4.** Influence of dispersal time on *Prochlorococcus* community composition analysed for three different sequence divergence definitions of a taxonomic unit. Mantel tests were used to detect variables significantly correlated with *Prochlorococcus* community composition without considering co-variables, and partial Mantel tests were used to remove the effects of each of the potential co-variables individually. Variables with significant Mantel $r$ values explain some of the variation in community structure ($P$-values less than 0.05 are deemed significant, bold). Also shown is the degree to which this is influenced by individual co-variables. Dispersal time is calculated as the time it takes for a tracer added at one sample site to reach 50% of maximum concentration at the other sites in a global ocean circulation model simulation. *Prochlorococcus* community composition is based on an operational taxonomic unit assignment of each clone using a specific rRNA sequence similarity cut-off, followed by a square-root transformed Bray-Curtis similarity index between each sample (see materials and methods for details).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

**Supporting information**

Additional Supporting Information may be found in the online version of this article:

**Fig. S1.** Estimation of dispersal time between each sample site using passive tracer injection. Panels 1–8 represent each site where the tracer was injected and the curve represents the tracer concentration at the other sites as a function of time. The dispersal time was recorded when the tracer concentration reached a threshold of 10%, 25% and 50% of the final concentration (arrows on y-axis). Tracer dispersal during the first 10 years was estimated using a first-order upwind scheme, followed by a second-order scheme for the remaining time.