Ecosystem-specific selection pressures revealed through comparative population genomics

Maureen L. Coleman[a] and Sallie W. Chisholm[a,b,1]

[a]Department of Civil and Environmental Engineering, and [b]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved August 24, 2010 (received for review June 30, 2010)

Bacterial populations harbor vast genetic diversity that is continually shaped by abiotic and biotic selective pressures, as well as by neutral processes. Individuals coexisting in the same geographically defined population often have significantly different gene content, but whether this variation is largely adaptive or neutral remains poorly understood. Here we quantify heterogeneity in gene content for two model marine microbes, *Prochlorococcus* and *Pelagibacter*, within and between populations in the Atlantic and Pacific Oceans, to begin to understand the selective pressures that are shaping these "population genomes." We discovered a large fraction of genes that are rare in each population, reflecting continual gene transfer and loss. Despite this high variation within each population, only a few genes significantly differ in abundance between the two biogeochemically distinct environments; nearly all of these are related to phosphorus acquisition and are enriched in the Atlantic relative to the Pacific. Moreover, P-related genes from the two sites form phylogenetically distinct clusters, whereas housekeeping genes do not, consistent with a recent spread of adaptive P-related genes in the Atlantic populations. These findings implicate phosphorus availability as the dominant selective force driving divergence between these populations, and demonstrate the promise of this approach for revealing selective agents in more complex microbial systems.

adaptation | biogeography | microbial evolution | natural selection | phosphorus limitation

Perhaps the most surprising lesson we have learned from microbial genome sequences is that closely related isolates of the same species often harbor substantially different gene complements, owing to horizontal gene transfer and gene loss (1, 2). The set of genes unique to a particular strain or subset of strains—the flexible genome—is hypothesized to be responsible for niche-specific adaptations (2, 3). However, in any given isolate, some fraction of the flexible genome might provide no fitness benefit, making it difficult to infer ecology from individual genome sequences. Moreover, snapshots of gene content in some natural populations have revealed high levels of coexisting variation (4–7), even among individuals that appear to experience similar selection pressures, suggesting that much of this variation is neutral. Alternatively, this variation could reflect adaptation by subpopulations specialized for microenvironments (8), or could reflect predation-driven frequency-dependent or diversifying selection (9). Understanding what fraction of gene content variation is in fact acted upon by selection, and what role this variation plays in ecosystem functioning, are central questions in microbial biology (10).

In the absence of selection, genes will be lost from a bacterial genome, owing to a mutational bias that favors deletions (11, 12). Therefore, genes that persist and rise to fixation in a population are inferred to be functional and to enhance organismal fitness. Likewise, genes that are differentially maintained in two populations reveal differential selection pressures acting on the two populations. Based on these principles, quantitative population genomics—in particular, the analysis of gene frequencies and patterns of sequence variation—has the power to highlight salient genetic features amid a background of continual gene transfer and gene erosion (10, 13). Thus, we can begin to quantify the roles of horizontal gene transfer and the flexible genome in microbial adaptation across environments.

Here we apply this approach to advance our understanding of evolutionary dynamics and the selective pressures facing two model microbes, the marine cyanobacterium *Prochlorococcus* and the heterotroph *Pelagibacter* (a member of the SAR11 clade; ref. 14), in two distinct oceanic regions. These microbial groups numerically dominate the biogeochemically well-characterized North Atlantic and North Pacific subtropical gyres, hereinafter referred to by the representative long-term study stations at which we sampled, the Bermuda Atlantic Time Series (BATS) station and the Hawaii Ocean Time-Series (HOT) Station ALOHA (15). Both locations are oligotrophic with similar rates of primary production and carbon export, but BATS experiences stronger seasonal mixing and nutrient supply compared with HOT (15). BATS has lower phosphate concentrations (16) but higher fluxes of dust inputs, which bring iron and other metals (17). Thus, the two sites could potentially select for a range of distinct microbial traits. Moreover, gene content at these two sites can be framed in the context of reference genomes (12 for *Prochlorococcus* and 3 for *Pelagibacter*).

Specifically, we investigated the degree of heterogeneity in gene content among individuals in marine microbial populations, how much of the flexible genome is being maintained by selection, and what the functions of these adaptive genes tell us about ecosystem-specific selective pressures. Toward this end, we quantitatively compared gene frequencies among *Prochlorococcus* and *Pelagibacter* using pyrosequenced community DNA from HOT and BATS. Because *Prochlorococcus* comprises distinct high- and low-light–adapted clades (18) whose abundance varies with depth (19, 20), we sampled three analogous depths at each site to provide a more representative picture of the entire vertically integrated (meta)population (Table S1). In doing so, we captured similar abundances of high- and low-light–adapted clades at both sites, thereby minimizing the effect of clade structure on our intersite comparison (Table S2).

Results and Discussion

To begin to understand the extent of gene content variability in natural populations, and how much of this variability is likely to be adaptive, we first quantified the occurrence of each *Prochlorococcus* flexible gene relative to the core genome (21). Genes belonging to the core genome—the 1,221 single-copy genes shared by all 12 sequenced isolates of *Prochlorococcus*—appear to be shared by nearly all *Prochlorococcus* cells in these wild pop-

Author contribution: M.L.C. and S.W.C. designed research; M.L.C. performed research; M.L.C. analyzed data; and M.L.C. and S.W.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Database deposition: The sequences reported in this paper have been deposited in the NCBI Sequence Read Archive, http://www.ncbi.nlm.nih.gov. For a list of accession numbers, see *SI Text*.

[1]To whom correspondence should be addressed. E-mail: chisholm@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1009480107/-/DCSupplemental.

ulations as well, as evidenced by the single narrow peak in the distribution of relative core gene frequencies (Fig. 1). Note that this distribution for core genes is, by definition, centered around one copy per cell, but the shape of the distribution is free to vary (see *Materials and Methods*). Compared with the core genome, most flexible genes (i.e., genes found in some, but not all, cultured isolates) are rare; 974 genes at HOT and 784 genes at BATS are estimated to occur at < 0.25 copy per cell (Fig. 1). A preponderance of rare flexible genes even at a given depth (Fig. S1) indicates that light adaptation cannot solely account for the distribution shown in Fig. 1. The majority of these rare flexible genes are likely neutral and transient in the population. In addition, some may have been acquired only recently, and some encode cell surface traits that may be driven by frequency-dependent selection (9). It is also possible that some of these rare flexible genes are adaptive in microenvironments within our samples, although such microenvironments are difficult to define for a planktonic organism. In contrast, about 400 flexible genes are estimated to occur at > 0.65 copy per cell at either site, and thus are "nearly core" in these particular populations (Fig. 1 and Table S3). Most of these widespread flexible genes are prevalent at both HOT and BATS, suggesting that they confer a fitness advantage in both open ocean environments, yet nearly three-quarters of these genes have no known function. One notable exception is the urease pathway; a dozen genes encoding urea transport and metabolism are found at about one copy per cell at both sites, suggesting that urea is an important nutrient source common to both these environments.

To identify selective pressures affecting the HOT and BATS populations differentially, we compared relative gene frequencies between the two sites. Only 29 of 2,854 observed *Prochlorococcus* genes have significantly different relative frequencies between the two sites (Fig. 2*A* and Table S4). Nearly all of the 29 differentially maintained genes are enriched at BATS relative to HOT and have functions related to phosphorus (P) acquisition and metabolism, as inferred from homology to known P-uptake genes, upregulation during P-starvation conditions in cultured strains MED4 (22) and MIT9301 (Fig. S2), or colocation with such genes in the genome (Fig. 2*B*). Many of these P-related genes are present in only a small fraction of *Prochlorococcus* cells at HOT but are nearly one copy per cell at BATS (Table S4); for instance, alkaline phosphatase (*phoA*) is estimated to be present in only 3% of *Prochlorococcus* cells at HOT but in 100% of these cells at BATS. A pathway for phosphonate utilization (23) is present in an estimated 13–21% of *Prochlorococcus* cells at BATS, but in virtually none of these cells at HOT (*phnYZ*; Table S4). Thus the "population genome" (i.e., the gene inventory in the *Prochlorococcus* collective) suggests strong selection for maintenance of accessory P- acquisition genes at BATS, but not at HOT. This difference between the two populations is consistent with exceptionally low surface water phosphate concentrations and physiological indicators of microbial phosphate limitation in the western North Atlantic (16, 24–27) relative to the Pacific. What is striking, however, is that this is the sole difference in gene content that emerged, implying that phosphate scarcity has been the most persistent and influential selective force driving diversification between the HOT and BATS populations.

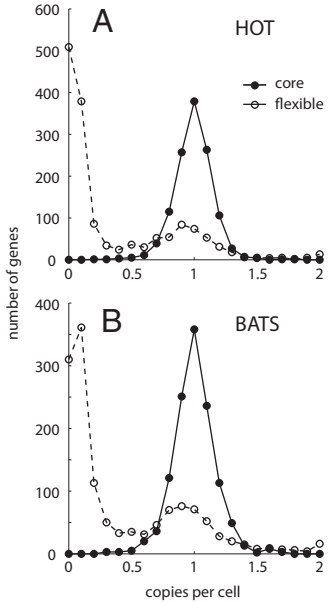

Fig. 1. Distribution of core and flexible genes (21) among *Prochlorococcus* populations at HOT (*A*) and at BATS (*B*). Core genes are defined as present in all 12 sequenced cultured isolates of *Prochlorococcus*, and the tight single-peaked distribution around the mean implies that they also are core genes and are present in one-to-one stoichiometry in these wild populations. Most flexible genes (those present in some, but not all, genomes of cultured isolates) are rare in these populations, being present in only a small proportion of the cells. The copy number per cell for each gene was estimated as described in *Materials and Methods*.

To test whether these same selective pressures have impacted the genomes of coexisting microbes in the community, we repeated the analysis with putative *Pelagibacter* (a heterotrophic bacterium) sequences (Fig. 2*C*), because this is another abundant microbe at these two sites for which reference genomes exist. Of 1,667 observed gene clusters, only 31 differ in abundance between BATS and HOT, and 29 of these are enriched at BATS (Table S5). Twenty-seven of these genes are located in one contiguous region of the genome in *Pelagibacter* strain HTCC7211, and nearly all are involved in phosphate or phosphonate metabolism (Fig. 2*D*), demonstrating that adaptation to P scarcity is a broad feature of the BATS ecosystem. Notably, however, the enriched gene set for *Pelagibacter* is distinct from that of *Prochlorococcus*. For example, genes encoding high-affinity phosphate transport (*pstSCAB*) and polyphosphate storage and breakdown (*ppk* and *ppx*) are represented equally among *Prochlorococcus* at HOT and at BATS, because they are core genes in this group (21). In contrast, the frequency of these genes varies dramatically between HOT and BATS in *Pelagibacter* (Table S5), in which these are not core genes in cultured isolates. The acquisition of beneficial nutrient-scavenging genes, along with their large cellular surface-to-volume ratio, helps explain the observation that *Prochlorococcus* and *Pelagibacter*-like bacteria can account for 90% of phosphate uptake in the North Atlantic (28).

Our results reveal unanticipated adaptations to phosphorus scarcity as well. Arsenate reductase and an arsenite efflux pump are enriched at BATS in *Pelagibacter* and *Prochlorococcus*, respectively (Tables S4 and S5). When cells scavenge phosphate, they sometimes take up the structurally similar arsenate ion, which is then reduced to arsenite and exported as a detoxification mechanism (29). Although average concentrations of surface inorganic arsenic (predominantly arsenate) are similar in the North Pacific and Atlantic, the arsenate:phosphate ratio is several-fold higher at BATS, where there is generally more arsenate (average, 16.3 nM) than phosphate in surface waters (30). Our population genomic analysis suggests that arsenic toxicity is an important selective force influencing *Prochlorococcus* and *Pelagibacter* at BATS, but not at HOT.

Several evolutionary scenarios could have led to this differential gene content between the two sites. If an ancestor of the BATS population had acquired genes that enhanced its fitness (e.g., alkaline phosphatase), a selective sweep could have ensued, leading to ecologically and genetically distinct HOT and BATS ecotypes (31). Alternatively, frequent gene transfer could have introduced P-acquisition genes into diverse genome backgrounds relatively recently (31). To distinguish between these scenarios, we built

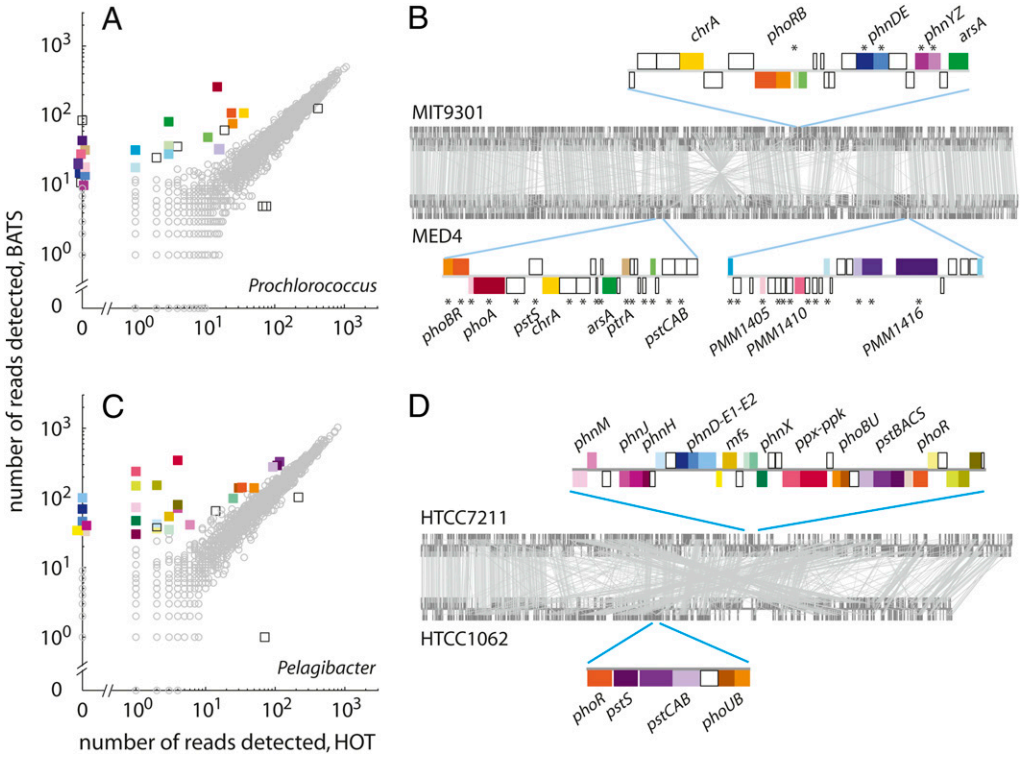

Fig. 2. Relative gene frequency among *Prochlorococcus* and *Pelagibacter* populations in the oligotrophic North Pacific (HOT) and North Atlantic (BATS) subtropical ocean gyres. (*A*) Detection of each *Prochlorococcus* gene at HOT and BATS, measured as the number of pyrosequencing reads. For each gene, the number of reads detected is proportional to the product of gene length and gene multiplicity per cell. Thus, assuming similar gene lengths at both sites, genes that fall along the diagonal trend have the same relative frequency at both sites, whereas those that fall above or below the diagonal are enriched at one site compared with the other. Squares represent significantly different frequencies between sites (G test; $P < 0.01$); circles represent nonsignificant differences. Colored squares represent genes whose chromosomal positions are depicted in *B*. (*B*) Genome comparison of two cultured isolates of *Prochlorococcus* (MED4 and MIT9301), with gray lines connecting homologous genes. Genes represented by corresponding colored squares in *A* are clustered together in a few distinct regions of the chromosome, including hypervariable genomic islands, in these isolates. Genes marked with an asterisk are up-regulated in response to P starvation (ref. 22 and Fig. S2). MED4 and MIT9301, isolated from the Mediterranean and the North Atlantic, respectively, are depicted because they carry the largest complements of phosphorus uptake genes among *Prochlorococcus* isolates (22). (*C*) Relative frequency of each *Pelagibacter* gene at HOT and BATS, as in *A*. (*D*) Genome comparison of two *Pelagibacter* strains, HTCC7211 and HTCC1062, as in *B*.

phylogenetic trees from both core genes and P-related flexible genes shared between the two populations and used Unifrac (32) to test whether the HOT- and BATS-derived alleles constituted distinct lineages. As exemplified by the *pntA* (encoding transhydrogenase) gene (Fig. 3 *A* and *B*), the housekeeping genes tested in both *Prochlorococcus* and *Pelagibacter* do not show divergence between the two populations (Table S6). Several P-uptake genes in both *Prochlorococcus* and *Pelagibacter* do show distinct HOT and BATS lineages, however (e.g., the *pstB* trees shown in Fig. 3 *C* and *D*). This phylogenetic divergence appears to be specific to P-uptake genes, and is not seen for the iron- and nitrogen-acquisition genes *fur* and *amtB* (Fig. 3 *E* and *F* and Table S6). This result is consistent with previous studies suggesting that the evolutionary history of the core genome backbone is decoupled from that of the P-associated regions (22, 33, 34). Together these results point to relatively recent gene acquisition, for instance, via phage-mediated gene transfer associated with genomic islands (35), combined with efficient selection in large populations, as the major processes underlying gene content variability.

Over time, acquired genes can become core-like not only in their frequency in the population (i.e., one copy per cell), but also in their arrangement within individual genomes. For example, genes first acquired in hypervariable genomic islands can eventually migrate to more stable regions of the chromosome where insertions and deletions are less likely (36). Phosphate-acquisition genes are clustered in hypervariable regions in the genomes of isolates of both *Prochlorococcus* (22, 35) (Fig. 2*B*) and *Pelagibacter* (37) (Fig. 2*D*). To assess how conserved the gene neighborhoods surrounding accessory P-uptake genes have become within natural populations at BATS, we used small- insert clone libraries, constructed from the same DNA used for pyrosequencing (see *Materials and Methods*). We identified clones in which at least one read matched a *Prochlorococcus* or *Pelagibacter* BATS-enriched gene (Tables S4 and S5). In 22% of such *Prochlorococcus* clones ($n = 307$), but in 40% of such *Pelagibacter* clones ($n = 554$), the paired-end read matches another BATS-enriched gene, suggesting that these genes are more often clustered on the chromosome in *Pelagibacter* than in *Prochlorococcus*. Furthermore, we counted the number of unique clusters of orthologous groups of proteins (COGs) adjacent to each BATS-enriched gene in *Prochlorococcus* and *Pelagibacter* clones, and found a greater diversity of adjacent COGs in *Prochlorococcus* (Fig. S3). Notably, core genes and BATS-enriched genes in *Pelagibacter* show similar levels of local gene order conservation, whereas in *Prochlorococcus*, local gene order is significantly more variable surrounding BATS-enriched genes compared with core genes (Fig. S3). Together, these results suggest that accessory P-uptake genes have become core-like in BATS populations of *Pelagibacter* in both frequency and arrangement, perhaps because *Pelagibacter* has experienced P limitation for a longer time; indeed, despite its smaller cell and genome size, *Pelagibacter* likely has a higher per-cell quota for P than *Prochlorococcus*, owing to its dependence on phospholipids (24).

Our analysis reveals that P availability is the primary selective force driving genome divergence between these two ocean regions.

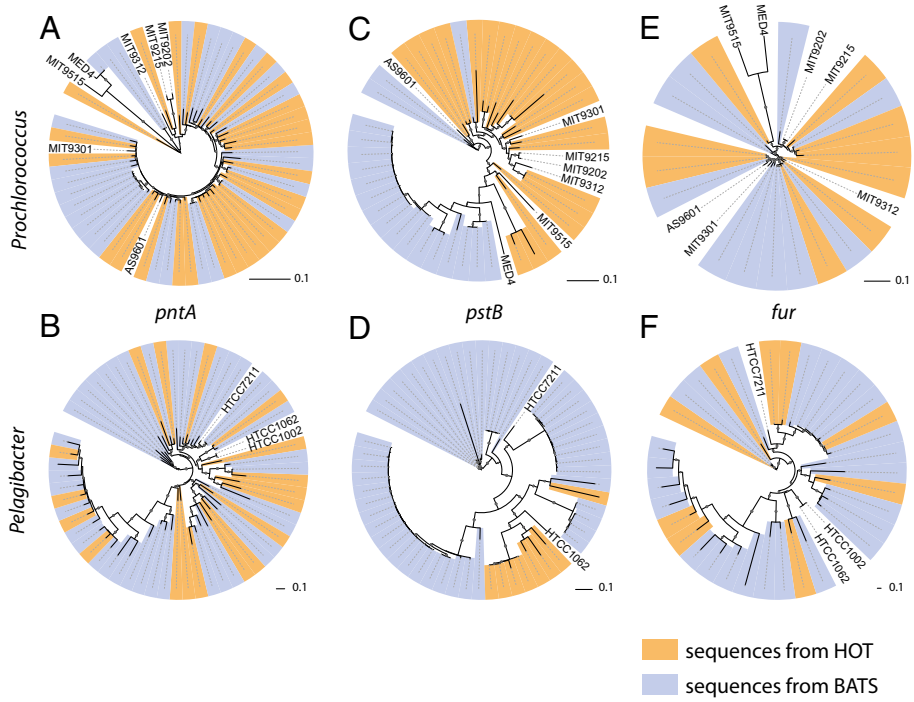

Fig. 3. Three examples of phylogenetic patterns observed in *Prochlorococcus* and *Pelagibacter* genes, from shotgun sequences sampled at HOT (orange leaf coloring) and BATS (blue leaf coloring). (*A* and *B*) *pntA*, a housekeeping gene found in similar abundance at HOT and at BATS, is not phylogenetically distinct between the sites. (*C* and *D*) *pstB*, a gene involved in phosphate transport, is phylogenetically distinct between the two sites, reflecting recent recombination and/or selection. (*E* and *F*) *fur*, a gene involved in regulation of iron metabolism, is not phylogenetically distinct between the sites. For clarity, sequence identifiers have been omitted except in the case of cultured isolates.

It is important to note, however, that we sampled both sites only in October. Unlike HOT, where the water column is stably stratified year-round, BATS experiences a pronounced seasonal cycle. At BATS, deep mixing events every winter supply nutrients that fuel spring phytoplankton blooms, and as these blooms draw down nutrient stocks and thermal stratification intensifies, the surface layer becomes progressively more oligotrophic (38). Our October sample was collected near the end of summer stratification at BATS, when *Prochlorococcus* are typically most abundant (39, 40). If we sampled in winter or spring, would we still find P availability to be the predominant selective pressure distinguishing HOT and BATS? We suspect that the answer is yes, at least for *Prochlorococcus*, because the increased P supply at BATS is accompanied by an increased P demand from faster-growing phytoplankton groups (*Synechococcus* and eukaryotes); the abundance of *Prochlorococcus* is actually lowest at BATS in late winter and early spring and increases as conditions become more stratified and oligotrophic (39, 40). To properly compare the population genomes at HOT and BATS in winter or spring, we also would need to control for the ecotype composition of the water column. Whereas we captured similar ecotype abundance profiles at both sites in October (Table S2), we would expect to capture very different profiles in winter or spring, owing to the stronger seasonal succession of ecotypes at BATS relative to HOT (40). These dynamics, and similar ecotype oscillations within the SAR11 clade at BATS (41), underscore the need for even finer-scale phylogenetic and temporal resolution in future population genomics studies.

Here we have used a population-genomic approach, unbiased by a priori assumptions about the relative strengths of selective factors, to compare two distinct ocean regions. This approach also could reveal adaptations in response to environmental change over time in a single habitat. For instance, recently it has been suggested that the inorganic P inventory at HOT is decreasing, and if this trend toward P limitation continues, the community will likely experience a shift toward organisms with lower P requirements and alternative P-acquisition mechanisms (42, 43). Based on our results, *Prochlorococcus* and *Pelagibacter* have not yet faced prolonged P limitation at HOT, at least not relative to their Atlantic counterparts. Over time, repeated sampling of these population genomes will reveal whether selection is indeed shaping P-related gene content and gene sequence at HOT and will help constrain the time scales of evolutionary change in the oceans. Moreover, these types of analyses will likely reveal unanticipated evolutionary pressures being "felt" by the microbial community. In this way, population genomics of ocean microbes not only is a powerful tool for diagnosing environmental change, but also can illuminate the fundamental evolutionary processes underlying biological organization (44).

Materials and Methods

Seawater was collected in October 2006 from HOT Station ALOHA (22°44′N, 158°2′W) and BATS (31°40′N, 64°10′W) from three depths at each site: in the mixed layer, just below the mixed layer, and in the deep chlorophyll maximum (corresponding to 20/50/100 m at BATS and 25/75/110 m at HOT). Sampling, DNA extraction, and sequencing were performed as described previously (45). Pyrosequencing reads were mapped to taxa using BLASTX against NCBI-nr, with a minimum bit score of 40. *Prochlorococcus* reads were mapped to gene clusters using BLASTN against 12 reference genomes, and were assigned to a gene cluster only if the top three hits (or all hits if fewer than three total) belonged to the same gene cluster, with bit score > 40 and alignment length > 40. *Pelagibacter* reads were mapped to gene clusters if the top three hits (or all hits if fewer than three total) belonged to the same gene cluster with bit score > 40; if there was a large bit score dropoff (>30) between hits, then the lower hits were not considered. Gene clusters for *Prochlorococcus* have been described by Kettler et al. (21); clusters for *Pelagibacter* were constructed by reciprocal best BLASTP hits using available genomes (*Pelagibacter* HTCC1002, HTCC1062, and HTCC7211). Gene frequencies were compared using the G test and Bonferroni-corrected for multiple comparisons (46).

Estimating Multiplicity per Cell. The number of DNA fragments, n_i, observed for a gene i of length L_i and average multiplicity per genome, m_i, in a sample of N_{Pro} sequences is expected to be binomially distributed,

$$n_i(L_i, m_i) \sim binomial\left(N_{Pro}, \frac{L_i m_i}{\sum L_i m_i}\right).$$

Thus, each sequencing read that we sample can be classified as either a "success" or "failure;" either it belongs to gene i or it does not. The probability of success (i.e., that the read came from a given gene i) depends on the length of the gene; longer genes will be detected more frequently. Indeed, using 1,221 single-copy core genes (i.e., $m_i = 1$, defined based on isolate genomes; ref. 21), we found a tight linear relationship between gene length and the number of sequencing reads detected ($r^2 = 0.95$ for HOT and 0.95 for BATS), and also found that the number of reads detected for nearly all core genes fell within the confidence intervals predicted by the binomial distribution assuming that $m_i = 1$. From this relationship, based on all 1,221 known single-copy genes, we then inferred gene multiplicity per cell for the "flexible" genome (*sensu* ref. 21); this approach is more robust than normalizing to a single core gene. Multiplicity per cell for gene i was calculated as

$$m_i = \frac{n_i}{bL_i},$$

where n_i is the number of sequence reads mapped to gene i, L_i is the gene length, and b is the slope of the length-versus-reads detected relationship for core genes. Confidence intervals for m_i were estimated using the binomial distribution. This approach was taken for *Pelagibacter* estimates of multiplicity per genome as well, using its respective core genome.

Validation of Taxon Mapping. An obvious limitation of our approach is that we can never know whether a pyrosequencing read with strong similarity to a particular organism's genome was truly derived from a cell of that organism, or whether it resides in the genome of another organism due to horizontal gene transfer. To address this issue, we used paired-end reads from small-insert shotgun clone libraries constructed from the same DNA used for pyrosequencing. Clone libraries were constructed and sequenced by the Joint Genome Institute according to standard production protocols. We identified putative *Prochlorococcus* clones, in which at least one of the two paired-end reads matched *Prochlorococcus* as the best hit, and then identified the best-hit taxon for its paired end using BLASTN against a custom database of marine microbial genomes. If both paired-end reads match *Prochlorococcus*, we can be more confident that the clone came from a *Prochlorococcus* cell, although multigene horizontal transfer events are also possible. Only 9.3% of the putative *Prochlorococcus* clones at BATS and 8.6% of those at HOT matched *Prochlorococcus* on one end and a different taxon on the other end. We repeated the analysis for putative *Pelagibacter* clones and found that 18.2% of clones at BATS and 21.6% at HOT matched *Pelagibacter* on one end and a different taxon on the other end. Clones in which only one of two paired reads matched *Prochlorococcus* (or *Pelagibacter*) are still quite likely to have derived from a *Prochlorococcus* (or *Pelagibacter*) cell, given the high abundance of *Prochlorococcus* (and *Pelagibacter*) in these communities. There are more reference genomes available for *Prochlorococcus* than for *Pelagibacter* (12 vs. 3), providing a more comprehensive (although still incomplete) picture of the *Prochlorococcus* pan-genome; this larger set of reference sequences likely explains why paired-end reads were more often coidentified as *Prochlorococcus* than as *Pelagibacter*.

Synteny Analysis. From these same clone libraries, we identified clones in which at least one read matched a BATS-enriched gene from *Prochlorococcus* or *Pelagibacter*, and clones in which at least one read matched a core gene from either taxon. We then compared the diversity of the paired-end reads associated with BATS-enriched genes and with core genes. We used rpsblast (47) to identify the best COG match for each paired-end read, and counted the number of unique COGs associated with each core gene and each BATS-enriched gene as a measure of gene order conservation (Fig. S3). We used the χ^2 goodness-of-fit test to evaluate whether the BATS-enriched sample came from the null distribution specified by the core genes (Poisson).

Phylogenetic Analyses. We constructed phylogenetic trees using small-insert shotgun clone sequences. Environmental gene sequences were aligned to the reference strain MIT9301 for *Prochlorococcus* and strain HTCC7211 for *Pelagibacter* using BlastAlignP (48), allowing a maximum of 70% gaps in the alignment due to the variable length and location of the environmental clone sequences. For *Prochlorococcus*, alignments were pruned to include only sequences most similar to high-light–adapted isolates, for reasons given below. Trees were constructed using PhyML (49) using the HKY85 model, with 10 different random starting trees, 4 gamma rate categories, and 0 invariant sites and parameters estimated from the data, with SH-like aLRT branch supports. Maximum likelihood is the most accurate method when using gene fragments and is the least sensitive to missing data (50). These trees were then used as input for UniFrac (32). The pruning to high-light adapted *Prochlorococcus* was done because long branches in the LL clades, which were detected only occasionally (the high-light–adapted clades eMIT9312 and eMED4 account for about 85% of cells at both sites; Table S2), can strongly influence the UniFrac test (e.g., if a LL clone was detected at HOT but not at BATS). Both the UniFrac significance test and the P test were performed for the pairwise HOT versus BATS comparison with 100 replicates.

Gene Expression. To measure gene expression during P starvation, strain MIT9301 was grown to midexponential phase in Pro99 medium, harvested at 8,000 g, washed twice with −P Pro99 (no added phosphate), and resuspended in the same. Duplicate cultures, grown in continuous light at 20 μmol photon·m^{-2}·s^{-1} and 21 °C, were analyzed. At each time point, an aliquot of cells was harvested at 10,000 g, resuspended in storage buffer [200 mM sucrose, 10 mM sodium acetate (pH 5.2), and 5 mM EDTA], frozen in liquid N_2, and stored at −80 °C. RNA was extracted using the *mir*Vana kit (Ambion) and Dnase-treated using Turbo DNA-*free* (Ambion). RNA (2–10 ng per reaction) was reverse-transcribed using SuperScript II (Invitrogen) and gene-specific primers. Transcripts of each gene were quantified using the QuantiTect SYBR Green Kit (Qiagen) and normalized to *rnpB* transcripts. The five genes tested in MIT9301 (*P9301_12441*, *12511*, *12521*, *12551*, and *12561*) were chosen because they are absent from strain MED4, for which whole genome expression data under P starvation are available (22).

ACKNOWLEDGMENTS. We thank Stephan Schuster and Ed DeLong and members of their laboratories for pyrosequencing the BATS216 and HOT186 samples, respectively; the HOT and BATS teams and the captain and crew of the R/V *Kilo Moana* and R/V *Atlantic Explorer*; Matt Sullivan, Jay McCarren, Suzanne Kern, Sarah Bagby, and Yanmei Shi for help with sample collection and processing; Scott Chilton for laboratory assistance; Daniele Veneziano for advice on statistical analyses; Kerrie Barry and the Joint Genome Institute for shotgun sequencing; and Jake Waldbauer, Vanja Klepac-Ceraj, Jesse Shapiro, and Eric Alm for comments and discussion. This work was supported in part by the Gordon and Betty Moore Foundation, the National Science Foundation (NSF), the U.S. Department of Energy, and an NSF Graduate Research Fellowship (to M.L.C.).

1. Welch RA, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci USA* 99:17020–17024.
2. Konstantinidis KT, Tiedje JM (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* 102:2567–2572.
3. Hacker J, Carniel E (2001) Ecological fitness, genomic islands and bacterial pathogenicity: A Darwinian view of the evolution of microbes. *EMBO Rep* 2:376–381.
4. Simmons SL, et al. (2008) Population genomic analysis of strain variation in *Leptospirillum* group II bacteria involved in acid mine drainage formation. *PLoS Biol* 6: e177.
5. Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5:e77.
6. Martín-Cuadrado AB, et al. (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE* 2:e914.
7. Cuadros-Orellana S, et al. (2007) Genomic plasticity in prokaryotes: The case of the square haloarchaeon. *ISME J* 1:235–245.
8. Hunt DE, et al. (2008) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* 320:1081–1085.
9. Wildschutte H, Wolfe DM, Tamewitz A, Lawrence JG (2004) Protozoan predation, diversifying selection, and the evolution of antigenic diversity in *Salmonella*. *Proc Natl Acad Sci USA* 101:10644–10649.
10. Wilmes P, Simmons SL, Denef VJ, Banfield JF (2009) The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev* 33:109–132.
11. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* 17:589–596.
12. Kuo C-H, Ochman H (2009) The fate of new bacterial genes. *FEMS Microbiol Rev* 33: 38–43.
13. Whitaker RJ, Banfield JF (2006) Population genomics in natural microbial communities. *Trends Ecol Evol* 21:508–516.
14. Morris RM, et al. (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* 420:806–810.
15. Karl DM, et al. (2001) Building the long-term picture: The U.S. JGOFS time-series programs. *Oceanography* 14:6–17.
16. Wu JF, Sunda W, Boyle EA, Karl DM (2000) Phosphate depletion in the western North Atlantic Ocean. *Science* 289:759–762.

17. Jickells TD, et al. (2005) Global iron connections between desert dust, ocean biogeochemistry, and climate. *Science* 308:67–71.
18. Moore LR, Rocap G, Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* 393:464–467.
19. West NJ, Scanlan DJ (1999) Niche-partitioning of *Prochlorococcus* populations in a stratified water column in the eastern North Atlantic Ocean. *Appl Environ Microbiol* 65:2585–2591.
20. Zinser ER, et al. (2006) *Prochlorococcus* ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. *Appl Environ Microbiol* 72:723–732.
21. Kettler GC, et al. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3:e231.
22. Martiny AC, Coleman ML, Chisholm SW (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc Natl Acad Sci USA* 103:12552–12557.
23. Martinez A, Tyson GW, Delong EF (2010) Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ Microbiol* 12:222–238.
24. Van Mooy BAS, et al. (2009) Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. *Nature* 458:69–72.
25. Ammerman JW, Hood RR, Case DA, Cotner JB (2003) Phosphorus deficiency in the Atlantic: An emerging paradigm in oceanography. *Eos Trans AGU*, 84: 10.1029/2003EO180001.
26. Bjorkman K, Thomson-Bulldis AL, Karl DM (2000) Phosphorus dynamics in the North Pacific subtropical gyre. *Aquat Microb Ecol* 22:185–198.
27. Cotner JB, Ammerman JW, Peele ER, Bentzen E (1997) Phosphorus-limited bacterioplankton growth in the Sargasso Sea. *Aquat Microb Ecol* 13:141–149.
28. Zubkov MV, et al. (2007) Microbial control of phosphate in the nutrient-depleted North Atlantic subtropical gyre. *Environ Microbiol* 9:2079–2089.
29. Sanders JG, Windom HL (1980) The uptake and reduction of arsenic species by marine algae. *Estuar Coast Mar Sci* 10:555–567.
30. Cutter GA, Cutter LS (1995) Behavior of dissolved antimony, arsenic, and selenium in the Atlantic Ocean. *Mar Chem* 49:295–306.
31. Cohan FM (2006) Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc Lond B Biol Sci* 361:1985–1996.
32. Lozupone C, Hamady M, Knight R (2006) UniFrac: An online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics* 7: article no. 371.
33. Martiny AC, Huang Y, Li W (2009) Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* 11:1340–1347.
34. Martiny AC, Tai AP, Veneziano D, Primeau F, Chisholm SW (2009) Taxonomic resolution, ecotypes and the biogeography of *Prochlorococcus*. *Environ Microbiol* 11: 823–832.
35. Coleman ML, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311:1768–1770.
36. Lawrence JG, Hendrickson H (2005) Genome evolution in bacteria: Order beneath chaos. *Curr Opin Microbiol* 8:572–578.
37. Wilhelm LJ, Tripp HJ, Givan SA, Smith DP, Giovannoni SJ (2007) Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct* 2: article no. 27.
38. Steinberg D, et al. (2001) 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 Top Stud Oceanogr* 48:1405–1447.
39. DuRand MD, Olson RJ, Chisholm SW (2001) Phytoplankton population dynamics at the Bermuda Atlantic Time-series station in the Sargasso Sea. *Deep Sea Res Part II Top Stud Oceanogr* 48:1983–2003.
40. Malmstrom RR, et al. Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J*, 10.1038/ismej.2010.60.
41. Carlson CA, et al. (2009) Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* 3:283–295.
42. Karl DM (2007) *Manual of Environmental Microbiology*, ed Hurst CJ (ASM Press, Washington, DC), pp 523–539.
43. Karl DM (2007) Microbial oceanography: Paradigms, processes and promise. *Nat Rev Microbiol* 5:759–769.
44. Woese CR, Goldenfeld N (2009) How the microbial world saved evolution from the Scylla of molecular biology and the Charybdis of the modern synthesis. *Microbiol Mol Biol Rev* 73:14–21.
45. Frias-Lopez J, et al. (2008) Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* 105:3805–3810.
46. Sokal RR, Rohlf FJ (1995) *Biometry: The Principles and Practice of Statistics in Biological Research* (Freeman, New York).
47. Marchler-Bauer A, Bryant SH (2004) CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res* 32(Web Server issue):W327–W331.
48. Belshaw R, Katzourakis A (2005) BlastAlign: A program that uses blast to align problematic nucleotide sequences. *Bioinformatics* 21:122–123.
49. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
50. Hartmann S, Vision TJ (2008) Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? *BMC Evol Biol* 8:95.

Supporting Information

Coleman and Chisholm 10.1073/pnas.1009480107

SI Text

The pyrosequencing reads have been deposited in the NCBI Sequence Read Archive, http://www.ncbi.nlm.nih.gov, under the following accession numbers: SRX007372 (HOT186 25m), SRX007369 (HOT186 75m), SRX007370 (HOT186 110m), SRX008032 (BATS216 20m), SRX008033 (BATS216 50m), SRX008035 (BATS216 100m). Shotgun sequence libraries have been deposited in the NCBI Trace Archive with the following trace identifiers: 2243992048-2244135183 (BATS216 20m), 2244135184-2244269775 (BATS216 50m), 2281908320-2281966591 (HOT186 25m), 2281966591-2282003263 and 2282006336-2282042335 (HOT186 75m).

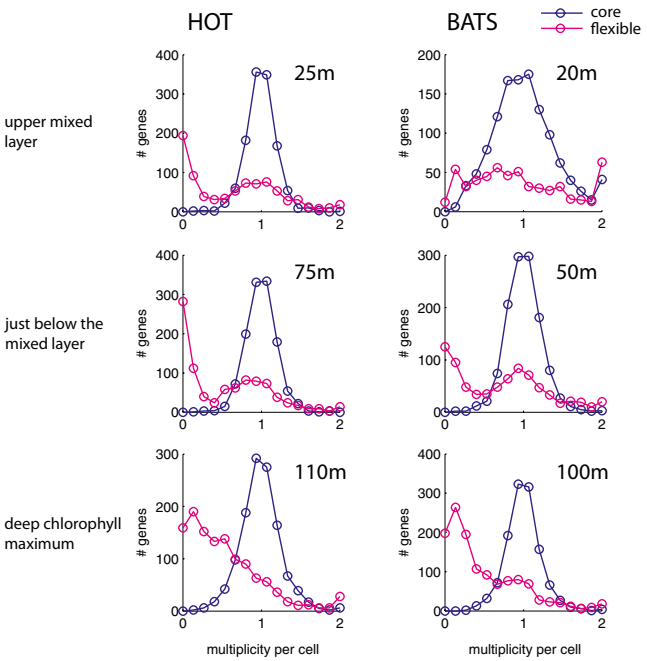

Fig. S1. Copies per *Prochlorococcus* cell of core and flexible genes at each depth. Core genes, defined using whole genome sequences from cultured isolates, also are one copy per cell in these populations, whereas many flexible genes are rare (present in only a subset of cells) at any given depth. Note that the 110-m sample from HOT and the 20-m sample from BATS each had significantly fewer *Prochlorococcus* reads (Table S1), resulting in the unusual shape of the flexible gene distribution.

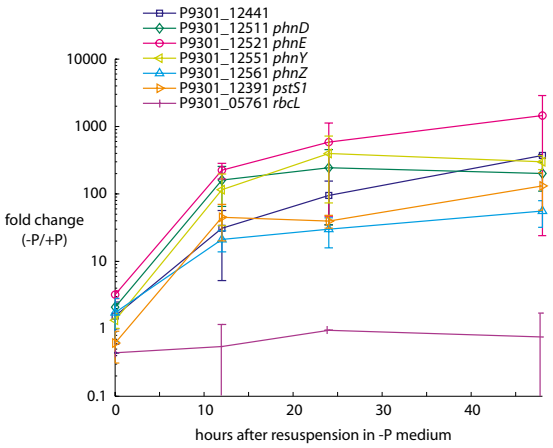

Fig. S2. Expression of phosphonate utilization genes in *Prochlorococcus* strain MIT9301 under P-starvation conditions. Two control genes—*pstS*, encoding the substrate-binding protein for phosphate, and *rbcL*, encoding the large subunit of RuBisCO—are plotted for comparison. Exponentially growing cells in P-replete media were resuspended just before time 0 in a medium with no added phosphate. For each gene, expression is normalized to that of the control gene, *rnpB*. Error bars show 2 SDs of two biological replicates. The locations of these genes in the genomes are denoted with an asterisk in Fig. 2*B*.

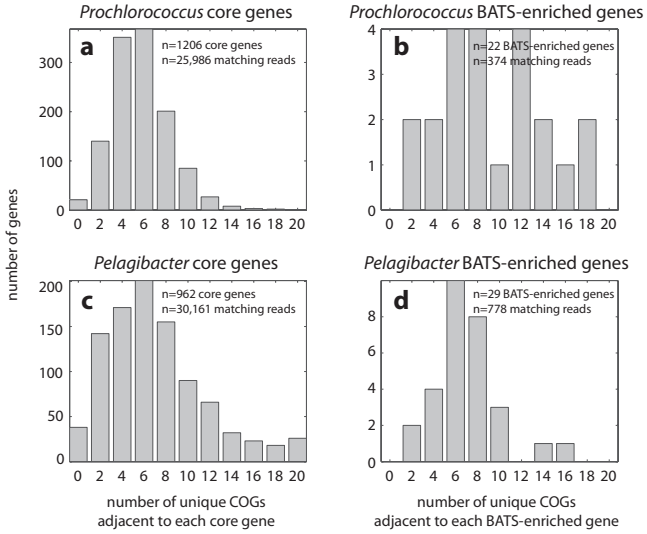

Fig. S3. Gene order variability surrounding core genes and BATS-enriched genes. We identified small-insert clones in which one read matched either a core gene (*A* and *C*) or a BATS-enriched gene (*B* and *D*). For each gene, we asked how many unique COGs are found adjacent to this gene of interest in the clone library. The distribution of COGs adjacent to *Prochlorococcus* BATS-enriched genes (*B*) is significantly different from that of COGs adjacent to *Prochlorococcus* core genes (*A*; χ^2 goodness-of-fit test, $P = 0.002$). The two distributions are not significantly different for *Pelagibacter* (*C* and *D*; $P = 0.6$).

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (DOC)
Table S4 (DOC)
Table S5 (DOC)
Table S6 (DOC)