Ecosystem-specific selection pressures revealed through comparative population genomics

The MIT Faculty has made this article openly available. *Please share* how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1073/pnas.1009480107">http://dx.doi.org/10.1073/pnas.1009480107</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>National Academy of Sciences (U.S.)</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Wed Sep 12 17:39:21 EDT 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/77562">http://hdl.handle.net/1721.1/77562</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher’s policy and may be subject to US copyright law. Please refer to the publisher’s site for terms of use.</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td></td>
</tr>
</tbody>
</table>
Ecosystem-specific selection pressures revealed through comparative population genomics

Maureen L. Coleman* and Sallie W. Chisholm**

*Department of Civil and Environmental Engineering, and **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 so doing, 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-
Materials and Methods

and strain HTCC7211, and nearly all are

Pelagibacter 18635 (in 0.25 copy per cell (Fig. 1). A pre-
cells at BATS, but in 13–21% of Prochlorococcus cells at BATS, but in virtually none of these cells at BATS. A pathway for phosphonate utilization (23) is present in an estimated 5% of Prochlorococcus cells 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.

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. 2C), 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. 2D), 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 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. 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 single-peak 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.
Table S6 and Coleman and Chisholm. Prochlorococcus (encoding trans-clones (Pelagibacter = 307), but in 40% of populations in the oligotrophic North Pacific, show similar levels of local gene order conservation, Pelagibacter do not show divergence gene at HOT and BATS, measured as the number of pyrosequencing reads. For each gene, the isolates (22). (Fig. 3 trees shown in Fig. 3 and www.pnas.org/cgi/doi/10.1073/pnas.1009480107. 

Pelagibacter do show distinct HOT and BATS, as in the largest complements of phosphorus uptake genes among P starvation (ref. 22 and 37) (Fig. 2 < than in clones, in both frequency and arrangement, perhaps because 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. 2B) and Pelagibacter (37) (Fig. 2D). 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. 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 r.

Fig. S3 was not mentioned in the text.

Fig. S3 was not mentioned in the text.
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 BLAST 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 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 \( N_{\text{read}} \) sequences is expected to be binomially distributed,

\[
n_i (L_i, m_i) \sim \text{binomial} \left( N_{\text{read}} \right) \frac{L_i}{L_i + L_i}.
\]

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 \((\rho = 0.95 \text{ for HOT and 0.95 for BAT5})\), 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

\[
m_i = \frac{n_i L_i}{N_{\text{read}}},
\]

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 more confidently 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 BAT5 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 BAT5 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 identified as Prochlorococcus than as Pelagibacter.

Synteny Analysis. From these same clone libraries, we identified clones in which at least one read matched a BAT5-enriched gene from Prochlorococcus or Pelagibacter, and clones that at least one read matched an otherwise conserved gene from either taxon. We then compared the diversity of the paired-end reads associated with BAT5-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 BAT5-enriched gene as a measure of gene order conservation (Fig. 53). We used the \( \chi^2 \) goodness-of-fit test to evaluate whether the BAT5-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 HTCC2711 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, the mit9301 (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 rpB 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).

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 N2, and stored at −80 °C. RNA was extracted using the mirVana kit (Ambion), 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 rpB 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 Steph Schuster and Ed DeLong and members of their laboratories for pyrosequencing the BAT5216 and HOT186 samples, respectively; the HOT and BAT5 teams and the captain and crew of the R/V Kilo Moana and R/V Atlantic Explorer; Matt Sullivan, Jay McCauley, Suzanne Kern, Sarah Bagby, and Yannine 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 Waldhuber, Vanja Klepac-Ceraj, Jesse Sha- piro, 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.).

4. Simmons SL, et al. (2008) Population genomic analysis of strain variation in Leptospirillum/gsp strain MIT9301 for 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 identified as Prochlorococcus than as Pelagibacter.
