Genomic Islands and the Ecology and Evolution of *Prochlorococcus*

Maureen L. Coleman,[1] Matthew B. Sullivan,[1] Adam C. Martiny,[1] Claudia Steglich,[1]* Kerrie Barry,[2] Edward F. DeLong,[1] Sallie W. Chisholm[1]†

Prochlorococcus ecotypes are a useful system for exploring the origin and function of diversity among closely related microbes. The genetic variability between phenotypically distinct strains that differ by less that 1% in 16*S* ribosomal RNA sequences occurs mostly in genomic islands. Island genes appear to have been acquired in part by phage-mediated lateral gene transfer, and some are differentially expressed under light and nutrient stress. Furthermore, genome fragments directly recovered from ocean ecosystems indicate that these islands are variable among co-occurring *Prochlorococcus* cells. Genomic islands in this free-living photoautotroph share features with pathogenicity islands of parasitic bacteria, suggesting a general mechanism for niche differentiation in microbial species.

Closely related bacterial isolates often contain remarkable genomic diversity (*1*, *2*). Although its functional consequences have been described in a few model heterotrophic microbes (*3*), little is known about genomic microdiversity in the microbial phototrophs that dominate aquatic ecosystems. The marine cyanobacterium *Prochlorococcus* offers a useful system for studying this issue, because they are globally abundant, have very simple growth requirements, have a very compact genome [1.7 to 2.4 megabases (Mb)], and live in a well-mixed habitat. Although the latter appears to offer few opportunities for niche differentiation, *Prochlorococcus* populations consist of multiple coexisting ecotypes (*4*), whose relative abundances vary markedly along gradients of light, temperature, and nutrients (*5*–*9*). Even two high-light adapted (HL) ecotypes, whose type strains (MED4 and MIT9312) differ by only 0.8% in 16*S* ribosomal RNA (rRNA) sequence, have substantially different distributions in the wild (*5*–*9*).

Although whole-genome comparisons between the most distantly related *Prochlorococcus* isolates (97.9% 16*S* rRNA identity) have revealed the gross signatures of this niche differentiation (*10*), important insights into the evolution of diversity in this group likely lie in comparisons between very closely related strains, and between coexisting genomes from wild populations. Thus, we compared the complete genomes of the type strains, MED4 and MIT9312, that represent the two HL clades, and we analyzed genome fragments from wild cells belonging to these clades from the Atlantic and Pacific oceans.

The 1574 shared genes of MED4 and MIT9312 have conserved order and orientation, except for a large inversion around the replication terminus (Fig. 1). The average G + C content is similar in both genomes (31%), and the median sequence identity of the shared genes is 78%, surprisingly low for strains so similar at the rRNA locus (*11*). For most genes, synonymous sites are saturated and protein sequence identity is low (median 80%); this is likely a function of high mutation rates, given that HL *Prochlorococcus* lack several important DNA-repair enzymes (*10*, *12*).

The strain-specific genes between MED4 and MIT9312 (236 in MIT9312 and 139 in MED4) occur primarily (80 and 74%, respectively) in five major islands (Fig. 1). Thus, these genomes have a mosaic structure similar to that of *Escherichia coli* genomes (*1*), though on a smaller scale. The islands are located in the same position in both genomes, implying that they are hotspots for recombination, and the length of island genes is similar to the whole-genome average, suggesting that they are not degraded. We hypothesize that these islands arose via lateral gene transfer and continually undergo rearrangement, on the basis of a number of characteristics. First, three islands are associated with tRNA genes (fig. S1), which are common integration sites for mobile elements (*13*). Second, the 3′ end of tRNA-proline, which flanks ISL3 in both genomes, is repeated 13 times in MIT9312-ISL3 (Fig. 2A) and three times in MED4-ISL3 (fig. S2), suggesting repeated remodeling of this island. Third, some of the genes found in a particular island in MED4 are found in a different island in MIT9312 (Fig. 1), a rearrangement that may have been mediated by a 48–base pair sequence element we call PRE1 (*Prochlorococcus* repeat element 1; fig. S3); portions of PRE1 are repeated, almost exclusively in islands, 13 times in MED4 (fig. S2), and 9 times in MIT9312 (Fig. 2A). Finally, up to 80% of the genes in any given MIT9312 island are most similar to the genes of noncyanobacterial organisms including phage, Eukarya, and Archaea, consistent with the recent observation that horizontally acquired genomic islands reflect a gene pool that differs from that of the core genome (*14*).

It is likely that phage, which often carry host genes (*15*, *16*), mediate some of the island-associated lateral gene transfer, and the *hli* gene family in particular appears to have undergone repeated phage-host gene exchange (*16*). Of the 24 *hli* genes in MIT9312, 18 are found in the five major islands or their flanking regions. All 18 belong to the multicopy and sporadically distributed group that includes phage copies (Fig. 2A) and is well differentiated from widespread single-copy *hli* genes found in cyanobacteria (*16*). Other phagelike genes in islands include an integrase, DNA methylases, a second *phoH*, a MarR-family transcriptional regulator, a putative hemagglutinin neuraminidase, and an endonuclease (*15*), further supporting a link between phage and island dynamics.

Many island genes in the two strains appear to encode functions related to physiological stress and nutrient uptake and thus may be important in the high-light, low-nutrient surface waters dominated by HL *Prochlorococcus*. ISL2 and ISL5 in MIT9312, for example, encode 12 of the 24 *hli* genes, known to be important under a variety of stress conditions (*17*); they also encode two outer-membrane transport

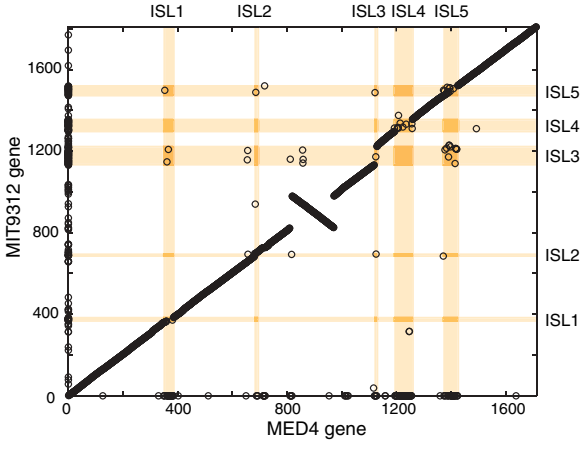

Fig. 1. Whole-genome alignment showing the positions of orthologous genes in MED4 and MIT9312. Strain-specific genes appear on the axes. The locations of five major islands defined by whole-genome alignment (*25*) are shaded.

[1]Massachusetts Institute of Technology, Department of Civil and Environmental Engineering, 15 Vassar Street, Cambridge, MA 02139, USA. [2]U.S. Department of Energy Joint Genome Institute, Production Genomics Facility, Walnut Creek, CA 94598, USA.

*Present address: University Freiburg, Department of Biology II/Experimental Bioinformatics, Schänzlestrasse 1, D-79104 Freiburg, Germany.

†To whom correspondence should be addressed. E-mail: chisholm@mit.edu

proteins; and a cyanophage-like homolog of *phoH* thought to be involved in the phosphate stress response (*15*). ISL3 in this strain contains a paralog of *psbF*, which encodes part of cytochrome b559, thought to protect against photoinhibition (*18*). Islands also contain genes involved in nutrient assimilation, including a cyanate transporter and lyase in MED4 and two transporters, for manganese/iron and amino acids, in MIT9312 (fig. S1).

In addition to genes involved in potentially growth-limiting processes, islands also contain genes that could play a role in selective mortality. ISL4 in both MED4 and MIT9312 encodes proteins involved in cell surface modification, including biosynthesis of lipopolysaccharide, a common phage receptor (*19*) (fig. S1). Phages are important agents of mortality in the oceans (*20*), and thus cell surface properties are likely under strong selection.

Clearly, for island genes to influence a cell's fitness, they must be expressed. When MED4 cells are starved for phosphorus, nine ISL5 genes are differentially expressed, nearly all of unknown function (table S1). When cells are shifted to high light, 38 island genes are differentially expressed, including seven *hli* genes (table S1) that in *Synechocystis* encode proteins that accumulate when cells absorb excess excitation energy (e.g., under high light, nutrient limitation, and low temperatures) (*17*). Thus, 26% of all MED4 island genes are differentially expressed under P starvation or high-light stress; only one of these is differentially expressed under both conditions (conserved hypothetical gene PMM1416), suggesting that island genes contribute to specific stress responses.

The genome variation within the eMIT9312 clade [sensu (*7*)] was examined in wild populations of *Prochlorococcus* by aligning short genome fragments from the Sargasso Sea (*21*), where this clade dominates (*7*), against the MIT9312 genome (Fig. 2B). Nearly constant coverage was observed, confirming a stable core genome, except for notable gaps at ISL1, ISL3, and ISL4. This finding indicates that very few wild sequences match genes in these islands, and it supports the hypothesis that these regions are hypervariable in HL *Prochlorococcus* genomes. In contrast, genes belonging to ISL2 and ISL5 are relatively well represented in the Sargasso Sea data set (Fig. 2B, fig. S2). In MED4 and MIT9312, these islands contain about half of the *hli* genes, lack the tRNA genes implicated in integration of mobile elements, and contain a smaller fraction of noncyanobacterial genes than do the other islands. This finding suggests that the genes in these islands have become fixed in this wild population.

Examination of 36 large genome fragments (1.1 Mb total sequence; median size 34 kb) (table S2) from the Hawaii Ocean Time-Series Station (*22*) further confirms that a stable core genome surrounds islands of variability, because most fragments showed remarkable conservation of gene content and order with respect to the MED4 and MIT9312 genomes. Thirty-four of the 36 fragments were more similar to MIT9312 than to MED4; two contained rRNA operons, confirming their phylogenetic affiliation with the eMIT9312 clade (fig. S4). The eMIT9312 fragments have about 90% identity with the MIT9312 genome and about 80% with MED4 (Table 1). Collectively, these results suggest that the wild eMIT9312 population is a coherent group identifiable by sequence similarity in the absence of an rRNA operon (*11*). eMIT9312 genome fragments from this wild population are more similar to each other than to the genome of the type strain MIT9312 (isolated from the Atlantic Ocean), but still share only 93% average sequence identity (Table 1), indicating high coexisting diversity in core genes.

Five eMIT9312 genome fragments from the Hawaii sample border the major islands defined above. About 60% of the genes in these islands have no ortholog in either MED4 or MIT9312, and two fragments border ISL1, yet their gene content is largely different from each other and from the MIT9312 and MED4 genomes (fig. S5). Indeed, a third of the island genes in these two fragments are novel, i.e., have no detectable homologs, implying that cells have access to a large novel gene pool in the oceans (*14*). Like the islands in the MED4 and MIT9312 genomes, these two fragments contain signatures of mobility, including duplicated tRNA genes, copies of the repeat PRE1, and an integrase gene. This reveals that islands are dynamic even within a single ecotype clade as we have defined it.

One observation that stimulated this work is the dramatic difference in distribution and abundance of the two HL *Prochlorococcus* ecotype clusters (*5*–*9*), as defined by their rRNA internal transcribed spacer (ITS) sequence similarity. Although strains belonging to these two clusters have different island gene content, so do cells from field populations that belong to a single cluster. Therefore, other genomic features are likely to be important in explaining niche differentiation between eMED4 and eMIT9312 cells in the wild. Differential temperature adaptation, for example, which is thought to be an important determinant of ecotype distribu-

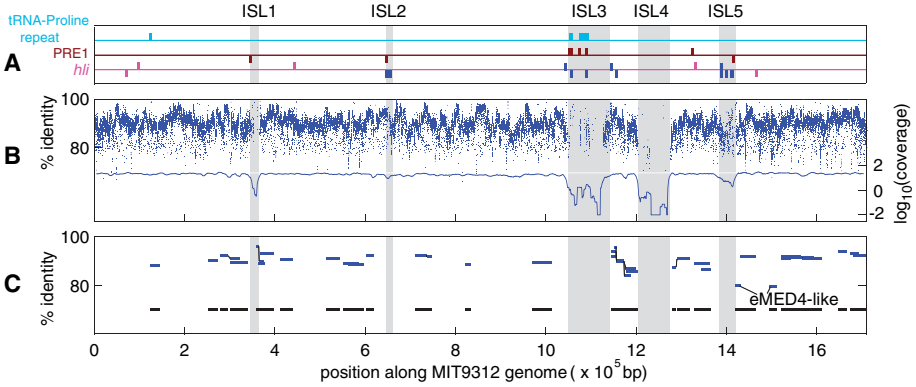

Fig. 2. Features of genomic islands (shaded) in the *Prochlorococcus* strain MIT9312 genome compared with wild sequences from the Atlantic and Pacific Oceans. (**A**) Locations of repetitive elements and *hli* genes in MIT9312, shown above or below the horizontal line for the forward or reverse strand, respectively. *hli* genes shown in pink belong to the single-copy conserved group and those shown in blue belong to the multicopy phage-encoded group (*16*). (**B**) Percent identity of Sargasso Sea shotgun database sequences (*21*) aligned to MIT9312 (top, left axis) and average coverage in the database of a given position in the MIT9312 genome (bottom, right axis). Log_{10}(coverage) is set to −2 when coverage equals 0. (**C**) Genomic locations and percent identity of wild genome fragments (eMIT9312-like unless noted) aligned to MIT9312. Where the alignment is interrupted, a black line connects aligned segments of a single fragment. Fragments are projected down to 70% horizontal to visualize total coverage.

Table 1. Median pairwise percent identities, for all orthologous gene pairs and for large aligned regions >4 kb (*25*). Numbers in parentheses indicate the number of orthologous gene pairs from which the median was calculated.

	MED4-MIT9312	MED4-eMIT9312 fragments	MIT9312-eMIT9312 fragments	Overlapping fragments
Orthologs (nucleotides)	78.4 (1574)	79.5 (1063)	90.6 (1092)	93.2 (434)
Orthologs (amino acids)	80.0 (1574)	82.4 (1063)	92.9 (1092)	95.2 (434)
Large aligned regions	79.0	79.9	90.7	92.6

tions (*5*), can be achieved through sequence (*23*) or regulatory (*24*) changes in the core genome. Nonetheless, given their prevalence, mobility, and expression under relevant conditions, islands likely play a role in adaptation, but on shorter time scales, or more local spatial scales, in the context of large populations that harbor substantial genomic variability.

Thus, although streamlined for life in the oligotrophic oceans, the genomes of HL *Prochlorococcus* are not static. Cell-to-cell genome variability is concentrated in islands containing genes that are differentially expressed under stresses typical of oceanic environments. Just as pathogenicity islands alter the host specificity and virulence of pathogenic bacteria (*3*), genomic islands in *Prochlorococcus* may contribute to niche differentiation in the surface oceans. Although other factors, such as small insertions and deletions, substitutions in homologous proteins, and differential regulation are important contributors to diversity, the prevalence of genomic islands and their features argue that these also play an influential role. We postulate that lateral gene transfer in genomic islands is an important mechanism for local specialization in the oceans. If true, genomic islands of natural taxa should contain genes that are ecologically important in a given environment, regardless of the core genome phylogeny. Testing this hypothesis will not only advance our understanding of microbial diversity in the ocean, but also contribute to a unified understanding of genomic evolutionary mechanisms and their impact on microbial ecology.

References and Notes

1. R. A. Welch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 17020 (2002).
2. J. R. Thompson *et al.*, *Science* **307**, 1311 (2005).
3. J. Hacker, J. B. Kaper, *Annu. Rev. Microbiol.* **54**, 641 (2000).
4. L. R. Moore, G. Rocap, S. W. Chisholm, *Nature* **393**, 464 (1998).
5. Z. I. Johnson *et al.*, *Science* **311**, 1737 (2006).
6. E. R. Zinser *et al.*, *Appl. Environ. Microbiol.* **72**, 723 (2006).
7. N. Ahlgren, G. Rocap, S. W. Chisholm, *Environ. Microbiol.* **8**, 441 (2006).
8. N. J. West, D. J. Scanlan, *Appl. Environ. Microbiol.* **65**, 2585 (1999).
9. N. J. West *et al.*, *Microbiology* **147**, 1731 (2001).
10. G. Rocap *et al.*, *Nature* **424**, 1042 (2003).
11. K. T. Konstantinidis, J. M. Tiedje, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 2567 (2005).
12. A. Dufresne, L. Garczarek, F. Partensky, *Genome Biol.* **6**, R14 (2005).
13. W. D. Reiter, P. Palm, S. Yeats, *Nucleic Acids Res.* **17**, 1907 (1989).
14. W. W. Hsiao *et al.*, *PLoS Genet.* **1**, e62 (2005).
15. M. B. Sullivan, M. L. Coleman, P. Weigele, F. Rohwer, S. W. Chisholm, *PLoS Biol.* **3**, e144 (2005).
16. D. Lindell *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11013 (2004).
17. Q. He, N. Dolganov, O. Bjorkman, A. R. Grossman, *J. Biol. Chem.* **276**, 306 (2001).
18. D. H. Stewart, G. W. Brudvig, *Biochim. Biophys. Acta* **1367**, 63 (1998).
19. A. Wright, M. McConnell, S. Kanegasaki, in *Virus Receptors*, L. L. Randall, L. Philipson, Eds. (Chapman and Hall, New York, 1980), pp. 27–57.
20. J. A. Fuhrman, *Nature* **399**, 541 (1999).
21. J. C. Venter *et al.*, *Science* **304**, 66 (2004).
22. E. F. DeLong *et al.*, *Science* **311**, 496 (2006).
23. G. N. Somero, *Annu. Rev. Physiol.* **57**, 43 (1995).
24. M. M. Riehle, A. F. Bennett, R. E. Lenski, A. D. Long, *Physiol. Genomics* **14**, 47 (2003).
25. Materials and methods are available as supporting material on *Science* Online.

26. We thank T. Rector, N. Hausman, and R. Steen for Affymetrix microarray processing; M. Polz for helpful discussions; and D. Lindell and A. Tolonen for comments on the manuscript. This work was supported by grants from NSF Biological Oceanography (S.W.C.) and Microbial Observatory (E.F.D.) Programs, the U.S. Department of Energy (DOE) GTL Program (to S.W.C. and G. Church), and the Gordon and Betty Moore Foundation (S.W.C. and E.F.D.). Sequencing support came from the DOE Microbial Genomics Program (E.F.D.) and DOE GTL and Community Sequencing Program (S.W.C.), conducted at the DOE Joint Genome Institute. Sequences are available in GenBank: BX548174 (MED4 genome), CP000111 (MIT9312 genome), and DQ366711 to DQ366746 (environmental genome fragments).

Supporting Online Material

www.sciencemag.org/cgi/content/full/311/5768/1768/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S3
References

31 October 2005; accepted 17 February 2006
10.1126/science.1122050