The *Sorcerer II* Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific

Douglas B. Rusch[1*], Aaron L. Halpern[1], Granger Sutton[1], Karla B. Heidelberg[1,2], Shannon Williamson[1], Shibu Yooseph[1], Dongying Wu[1,3], Jonathan A. Eisen[1,3], Jeff M. Hoffman[1], Karin Remington[1,4], Karen Beeson[1], Bao Tran[1], Hamilton Smith[1], Holly Baden-Tillson[1], Clare Stewart[1], Joyce Thorpe[1], Jason Freeman[1], Cynthia Andrews-Pfannkoch[1], Joseph E. Venter[1], Kelvin Li[1], Saul Kravitz[1], John F. Heidelberg[1,2], Terry Utterback[1], Yu-Hui Rogers[1], Luisa I. Falcón[5], Valeria Souza[5], Germán Bonilla-Rosso[5], Luis E. Eguiarte[5], David M. Karl[6], Shubha Sathyendranath[7], Trevor Platt[7], Eldredge Bermingham[8], Victor Gallardo[9], Giselle Tamayo-Castillo[10], Michael R. Ferrari[11], Robert L. Strausberg[1], Kenneth Nealson[1,12], Robert Friedman[1], Marvin Frazier[1], J. Craig Venter[1]

1 J. Craig Venter Institute, Rockville, Maryland, United States of America, **2** Department of Biological Sciences, University of Southern California, Avalon, California, United States of America, **3** Genome Center, University of California Davis, Davis, California, United States of America, **4** Your Genome, Your World, Rockville, Maryland, United States of America, **5** Departmento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, Mexico City, Mexico, **6** Department of Oceanography, University of Hawaii, Honolulu, Hawaii, United States of America, **7** Bedford Institute of Oceanography, Dartmouth, Nova Scotia, Canada, **8** Smithsonian Tropical Research Institute, Balboa, Ancon, Republic of Panama, **9** Departamento de Oceanografía, Universidad de Concepción, Concepción, Chile, **10** Escuela de Química, Universidad de Costa Rica, San Pedro, Costa Rica, **11** Department of Environmental Sciences, Rutgers University, New Brunswick, New Jersey, United States of America, **12** Department of Earth Sciences, University of Southern California, Los Angles, California, United States of America

The world's oceans contain a complex mixture of micro-organisms that are for the most part, uncharacterized both genetically and biochemically. We report here a metagenomic study of the marine planktonic microbiota in which surface (mostly marine) water samples were analyzed as part of the *Sorcerer II* Global Ocean Sampling expedition. These samples, collected across a several-thousand km transect from the North Atlantic through the Panama Canal and ending in the South Pacific yielded an extensive dataset consisting of 7.7 million sequencing reads (6.3 billion bp). Though a few major microbial clades dominate the planktonic marine niche, the dataset contains great diversity with 85% of the assembled sequence and 57% of the unassembled data being unique at a 98% sequence identity cutoff. Using the metadata associated with each sample and sequencing library, we developed new comparative genomic and assembly methods. One comparative genomic method, termed "fragment recruitment," addressed questions of genome structure, evolution, and taxonomic or phylogenetic diversity, as well as the biochemical diversity of genes and gene families. A second method, termed "extreme assembly," made possible the assembly and reconstruction of large segments of abundant but clearly nonclonal organisms. Within all abundant populations analyzed, we found extensive intra-ribotype diversity in several forms: (1) extensive sequence variation within orthologous regions throughout a given genome; despite coverage of individual ribotypes approaching 500-fold, most individual sequencing reads are unique; (2) numerous changes in gene content some with direct adaptive implications; and (3) hypervariable genomic islands that are too variable to assemble. The intra-ribotype diversity is organized into genetically isolated populations that have overlapping but independent distributions, implying distinct environmental preference. We present novel methods for measuring the genomic similarity between metagenomic samples and show how they may be grouped into several community types. Specific functional adaptations can be identified both within individual ribotypes and across the entire community, including proteorhodopsin spectral tuning and the presence or absence of the phosphate-binding gene *PstS*.

Citation: Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The *Sorcerer II* Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. PLoS Biol 5(3): e77. doi:10.1371/journal.pbio.0050077

This article is part of the Oceanic Metagenomics collection.

Academic Editor: Nancy A. Moran, University of Arizona, United States of America

Received July 14, 2006; **Accepted** January 16, 2007; **Published** March 13, 2007

Copyright: © 2007 Rusch et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: CAMERA, Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis; GOS, Global Ocean Sampling; NCBI, National Center for Biotechnology Information

* To whom correspondence should be addressed. E-mail: DRusch@venterinstitute.org

This article is part of Global Ocean Sampling collection in *PLoS Biology*. The full collection is available online at http://collections.plos.org/plosbiology/gos-2007.php.

Author Summary

Marine microbes remain elusive and mysterious, even though they are the most abundant life form in the ocean, form the base of the marine food web, and drive energy and nutrient cycling. We know so little about the vast majority of microbes because only a small percentage can be cultivated and studied in the lab. Here we report on the Global Ocean Sampling expedition, an environmental metagenomics project that aims to shed light on the role of marine microbes by sequencing their DNA without first needing to isolate individual organisms. A total of 41 different samples were taken from a wide variety of aquatic habitats collected over 8,000 km. The resulting 7.7 million sequencing reads provide an unprecedented look at the incredible diversity and heterogeneity in naturally occurring microbial populations. We have developed new bioinformatic methods to reconstitute large portions of both cultured and uncultured microbial genomes. Organism diversity is analyzed in relation to sampling locations and environmental pressures. Taken together, these data and analyses serve as a foundation for greatly expanding our understanding of individual microbial lineages and their evolution, the nature of marine microbial communities, and how they are impacted by and impact our world.

Introduction

The concept of microbial diversity is not well defined. It can either refer to the genetic (taxonomic or phylogenetic) diversity as commonly measured by molecular genetics methods, or to the biochemical (physiological) diversity measured in the laboratory with pure or mixed cultures. However, we know surprisingly little about either the genetic or biochemical diversity of the microbial world [1], in part because so few microbes have been grown under laboratory conditions [2,3], and also because it is likely that there are immense numbers of low abundance ribotypes that have not been detected using molecular methods [4]. Our understanding of microbial physiological and biochemical diversity has come from studying the less than 1% of organisms that can be maintained in enrichments or cultivated, while our understanding of phylogenetic diversity has come from the application of molecular techniques that are limited in terms of identifying low-abundance members of the communities.

Historically, there was little distinction between genetic and biochemical diversity because our understanding of genetic diversity was based on the study of cultivated microbes. Biochemical diversity, along with a few morphological features, was used to establish genetic diversity via an approach called numerical taxonomy [5,6]. In recent years the situation has dramatically changed. The determination of genetic diversity has relied almost entirely on the use of gene amplification via PCR to conduct taxonomic environmental gene surveys. This approach requires the presence of slowly evolving, highly conserved genes that are found in otherwise very diverse organisms. For example, the gene encoding the small ribosomal subunit RNA, known as 16S, based on sedimentation coefficient, is most often used for distinguishing bacterial and archaeal species [7–10]. The 16S rRNA sequences are highly conserved and can be used as a phylogenetic marker to classify organisms and place them in evolutionary context. Organisms whose 16S sequences are at least 97% identical are commonly considered to be the same ribotype [11], otherwise referred to as species, operational taxonomic units, or phylotypes.

Although rRNA-based analysis has revolutionized our view of genetic diversity, and has allowed the analysis of a large part of the uncultivated majority, it has been less useful in predicting biochemical diversity. Furthermore, the relationship between genetic and biochemical diversity, even for cultivated microbes, is not always predictable or clear. For instance, organisms that have very similar ribotypes (97% or greater homology) may have vast differences in physiology, biochemistry, and genome content. For example, the gene complement of *Escherichia coli* O157:H7 was found to be substantially different from the K12 strain of the same species [12].

In this paper, we report the results of the first phase of the *Sorcerer II* Global Ocean Sampling (GOS) expedition, a metagenomic study designed to address questions related to genetic and biochemical microbial diversity. This survey was inspired by the British Challenger expedition that took place from 1872–1876, in which the diversity of macroscopic marine life was documented from dredged bottom samples approximately every 200 miles on a circumnavigation [13–15]. Through the substantial dataset described here, we identified 60 highly abundant ribotypes associated with the open ocean and aquatic samples. Despite this relative lack of diversity in ribotype content, we confirm and expand upon previous observations that there is tremendous within-ribotype diversity in marine microbial populations [4,7,8,16,17]. New techniques and tools were developed to make use of the sampling and sequencing metadata. These tools include: (1) the fragment recruitment tool for performing and visualizing comparative genomic analyses when a reference sequence is available; (2) new assembly techniques that use metadata to produce assemblies for uncultivated abundant microbial taxa; and (3) a whole metagenome comparison tool to compare entire samples at arbitrary degrees of genetic divergence. Although there is tremendous diversity within cultivated and uncultivated microbes alike, this diversity is organized into phylogenetically distinct groups we refer to as subtypes.

Subtypes can occupy similar environments yet remain genetically isolated from each other, suggesting that they are adapted for different environmental conditions or roles within the community. The variation between and within subtypes consists primarily of nucleotide polymorphisms but includes numerous small insertions, deletions, and hypervariable segments. Examination of the GOS data in these terms sheds light on patterns of evolution and also suggests approaches towards improving the assembly of complex metagenomic datasets. At least some of this variation can be associated with functional characters that are a direct response to the environment. More than 6.1 million proteins, including thousands of new protein families, have been annotated from this dataset (described in the accompanying paper [18]). In combination, these papers bring us closer to reconciling the genetic and biochemical disconnect and to understanding the marine microbial community.

We describe a metagenomic dataset generated from the *Sorcerer II* expedition. The GOS dataset, which includes and extends our previously published Sargasso Sea dataset [19], now encompasses a total of 41 aquatic, largely marine locations, constituting the largest metagenomic dataset yet produced with a total of ~7.7 million sequencing reads. In

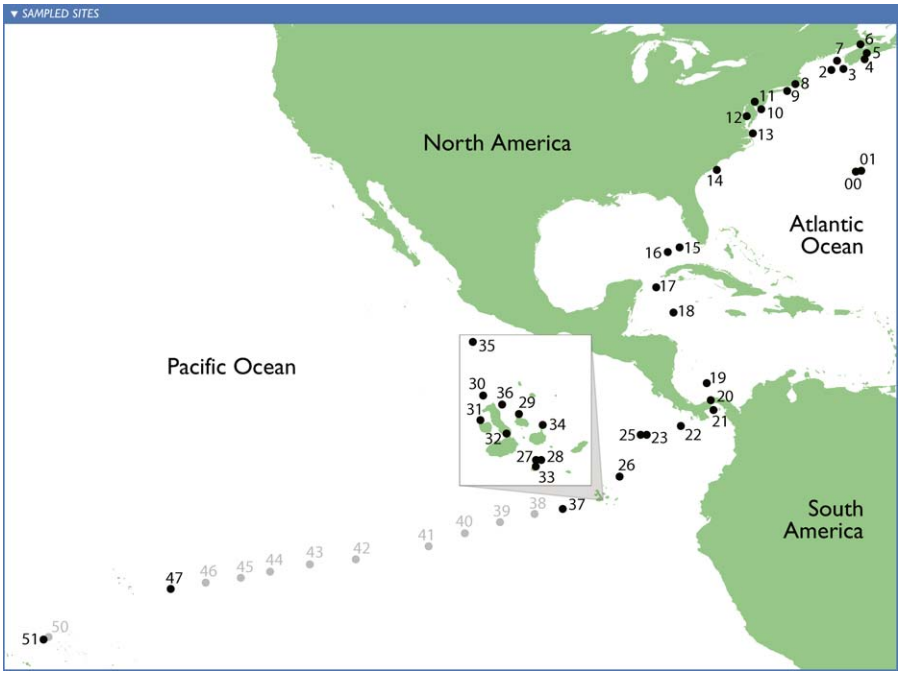


Figure 1. Sampling Sites

Microbial populations were sampled from locations in the order shown. Samples were collected at approximately 200 miles (320 km) intervals along the eastern North American coast through the Gulf of Mexico into the equatorial Pacific. Samples 00 and 01 identify sets of sites sampled as part of the Sargasso Sea pilot study [19]. Samples 27 through 36 were sampled off the Galapagos Islands (see inset). Sites shown in gray were not analyzed as part of this study.

doi:10.1371/journal.pbio.0050077.g001

the pilot Sargasso Sea study, 200 l surface seawater was filtered to isolate microorganisms for metagenomic analysis. DNA was isolated from the collected organisms, and genome shotgun sequencing methods were used to identify more than 1.2 million new genes, providing evidence for substantial microbial taxonomic diversity [19]. Several hundred new and diverse examples of the proteorhodopsin family of light-harvesting genes were identified, documenting their extensive abundance and pointing to a possible important role in energy metabolism under low-nutrient conditions. However, substantial sequence diversity resulted in only limited genome assembly. These results generated many additional questions: would the same organisms exist everywhere in the ocean, leading to improved assembly as sequence coverage increased; what was the global extent of gene and gene family diversity, and can we begin to exhaust it with a large but achievable amount of sequencing; how do regions of the ocean differ from one another; and how are different environmental pressures reflected in organisms and communities? In this paper we attempt to address these issues.

Results

Sampling and the Metagenomic Dataset

Microbial samples were collected as part of the *Sorcerer II* expedition between August 8, 2003, and May 22, 2004, by the S/V *Sorcerer II,* a 32-m sailing sloop modified for marine research. Most specimens were collected from surface water marine environments at approximately 320-km (200-mile) intervals. In all, 44 samples were obtained from 41 sites (Figure 1), covering a wide range of distinct surface marine environments as well as a few nonmarine aquatic samples for contrast (Table 1).

Several size fractions were isolated for every site (see Materials and Methods). Total DNA was extracted from one or more fractions, mostly from the 0.1–0.8-μm size range. This fraction is dominated by bacteria, whose compact genomes are particularly suitable for shotgun sequencing. Random-insert clone libraries were constructed. Depending on the uniqueness of each sampling site and initial estimates of the genetic diversity, between 44,000 and 420,000 clones per sample were end-sequenced to generate mated sequencing reads. In all, the combined dataset includes 6.25 Gbp of sequence data from 41 different locations. Many of the clone libraries were constructed with a small insert size (<2 kbp) to maximize cloning efficiency. As this often resulted in mated sequencing reads that overlapped one another, overlapping mated reads were combined, yielding a total of ~6.4 M contiguous sequences, totaling ~5.9 Gbp of nonredundant sequence. Taken together, this is the largest collection of metagenomic sequences to date, providing more than a 5-fold increase over the dataset produced from the Sargasso Sea pilot study [19] and more than a 90-fold increase over the other large marine metagenomic dataset [20].

Assembly

Assembling genomic data into larger contigs and scaffolds, especially metagenomic data, can be extremely valuable, as it places individual sequencing reads into a greater genomic context. A largely contiguous sequence links genes into operons, but also permits the investigation of larger biochemical and/or physiological pathways, and also connects otherwise-anonymous sequences with highly studied "taxo-

Table 1. Sampling Locations and Environmental Data

ID	Sample Location	Country	Date, mm/dd/yy	Time	Location	Sample Depth, m	Water Depth, m	T (°C)[a]	S[b] (ppt)	Size Fraction (μm)	Habitat Type	Chl *a* Sample Month (Annual ± SE) mg/m^{-3}	Good Sequences
GS00a	Sargasso Stations 13 and 11	Bermuda (UK)	02/26/03	3:00	31°32′6″n; 63°35′42″w	5.0	>4,200	20.0 20.5	36.6	0.1–0.8	Open ocean	0.17 (0.0.9 ± 0.02)	644,551
				10:10	31°10′50″n; 64°19′27″w				36.7				
GS00b	Sargasso Stations 13 and 11	Bermuda (UK)	02/26/03	3:35	31°32′10″n; 63°35′70″w	5.0	>4,200	20.0 20.5	36.6	0.22–0.8	Open ocean	0.17 (0.0.9 ± 0.02)	317,180
				10:43	31°10′50n; 64°19′27″w				36.7				
GS00c	Sargasso Stations 3	Bermuda (UK)	02/25/03	13:00	32°09′30″n; 64°00′36″w	5.0	>4,200	19.8	36.7	0.22–0.8	Open ocean	0.17 (0.0.9 ± 0.02)	368,835
GS00d	Sargasso Stations 13	Bermuda (UK)	02/25/03	17:00	31°32′6″n; 63°35′42″w	5.0	>4,200	20.0	36.6	0.22–0.8	Open ocean	0.17 (0.0.9 ± 0.02)	332,240
GS01a	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10′00″n 64°30′00″w	5.0	>4,200	22.9	36.7	3.0–20.0	Open ocean	0.10 (0.10 ± 0.01)	142,352
GS01b	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10′00″n; 64°30′00″w	5.0	>4,201	22.9	36.7	0.8–3.0	Open ocean	0.10 (0.10 ± 0.01)	90,905
GS01c	Hydrostation S	Bermuda (UK)	05/15/03	11:40	32°10′00″n; 64°30′00″w	5.0	>4,202	22.9	36.7	0.1–0.8	Open ocean	0.1 (0.1 ± 0.01)	92,351
GS02	Gulf of Maine	USA	08/21/03	6:32	42°30′11″n; 67°14′24″w	1.0	106	18.2	29.2	0.1–0.8	Coastal	1.4 (1.12 ± 0.19)	121,590
GS03	Browns Bank, Gulf of Maine	Canada	08/21/03	11:50	42°51′10″n; 66°13′2″w	1.0	119	11.7	29.9	0.1–0.8	Coastal	1.4 (1.12 ± 0.19)	61,605
GS04	Outside Halifax, Nova Scotia	Canada	08/22/03	5:25	44°8′14″n; 63°38′40″w	2.0	142	17.3	28.3	0.1–0.8	Coastal	0.4 (0.78 ± 0.17)	52,959
GS05	Bedford Basin, Nova Scotia	Canada	08/22/03	16:21	44°41′25″n; 63°38′14″w	1.0	64	15.0	30.2	0.1–0.8	Embayment	6 (6.76 ± 0.98)	61,131
GS06	Bay of Fundy, Nova Scotia	Canada	08/23/03	10:47	45°6′42″n; 64°56′48″w	1.0	11	11.2		0.1–0.8	Estuary	2.8 (1.87 ± 0.18)	59,679
GS07	Northern Gulf of Maine	Canada	08/25/03	8:25	43°37′56″n; 66°50′50″w	1.0	139	17.9	31.7[c]	0.1–0.8	Coastal	1.4 (1.12 ± 0.19)	50,980
GS08	Newport Harbor, RI	USA	11/16/03	16:45	41°29′9″n; 71°21′4″w	1.0	12	9.4	26.5[c]	0.1–0.8	Coastal	2.2 (1.59 ± 0.17)	129,655
GS09	Block Island, NY	USA	11/17/03	10:30	41°5′28″n; 71°36′8″w	1.0	32	11.0	31.0[c]	0.1–0.8	Coastal	4.0 (2.72 ± 0.24)	79,303
GS10	Cape May, NJ	USA	11/18/03	4:30	38°56′24″n; 74°41′6″w	1.0	10	12.0	31.0[c]	0.1–0.8	Coastal	2.0 (2.75 ± 0.33)	78,304
GS11	Delaware Bay, NJ	USA	11/18/03	11:30	39°25′4″n; 75°30′15″w	1.0	8	11.0		0.1–0.8	Estuary	4.8 (9.23 ± 1.02)	124,435
GS12	Chesapeake Bay, MD	USA	12/18/03	11:32	38°56′49″n; 76°25′2″w	1.0	25	3.2	3.47[c]	0.1–0.8	Estuary	21.0 (15.0 ± 1.01)	126,162
GS13	Off Nags Head, NC	USA	12/19/03	6:28	36°0′14″n; 75°23′41″w	1.0	20	9.3		0.1–0.8	Coastal	3.0 (2.24 ± 0.25)	138,033
GS14	South of Charleston, SC	USA	12/20/03	17:12	32°30′25″n; 79°15′50″w	1.0	31	18.6		0.1–0.8	Coastal	1.70 (1.92 ± 0.25)	128,885
GS15	Off Key West, FL	USA	01/08/04	6:25	24°29′18″n; 83°4′12″w	2.0	47	25.3	36.0	0.1–0.8	Coastal	0.2 (0.27 ± 0.09)	127,362
GS16	Gulf of Mexico	USA	01/08/04	14:15	24°10′29″n; 84°20′40″w	2.0	3,333	26.4	35.8	0.1–0.8	Coastal sea	0.16 (0.11 ± 0.01)	127,122
GS17	Yucatan Channel	Mexico	01/09/04	13:47	20°31′21″n; 85°24′49″w	2.0	4,513	27.0	35.8	0.1–0.8	Open ocean	0.13 (0.09 ± 0.01)	257,581
GS18	Rosario Bank	Honduras	01/10/04	8:12	18°2′12″n; 83°47′5″w	2.0	4,470	27.4	35.4	0.1–0.8	Open ocean	0.14 (0.09 ± 0.01)	142,743
GS19	Northeast of Colón	Panama	01/12/04	9:03	10°42′59″n; 80°15′16″w	2.0	3,336	27.7	35.4	0.1–0.8	Coastal	0.23 (0.15 ± 0.02)	135,325
GS20	Lake Gatun	Panama	01/15/04	10:24	9°9′52″n; 79°50′10″w	2.0	4	28.5	0.06	0.1–0.8	Fresh water		296,355
GS21	Gulf of Panama	Panama	01/19/04	16:48	8°7′45″n; 79°41′28″w	2.0	76	27.6	30.7	0.1–0.8	Coastal	0.50 (0.73 ± 0.22)	131,798
GS22	250 miles from Panama City	Panama	01/20/04	16:39	6°29′34″n; 82°54′14″w	2.0	2,431	29.3	32.3	0.1–0.8	Open ocean	0.33 (0.28 ± 0.02)	121,662
GS23	30 miles from Cocos Island	Costa Rica	01/21/04	15:00	5°38′24″n; 86°33′55″w	2.0	1,139	28.7	32.6	0.1–0.8	Open ocean	0.07 (0.19 ± 0.02)	133,051
GS25	Dirty Rock, Cocos Island	Costa Rica	01/28/04	10:51	5°33′10″n; 87°5′16″w	1.1	30	28.3	31.4	0.8–3.0	Fringing reef	0.11 (0.19 ± 0.01)	120,671
GS26	134 miles NE of Galapagos	Ecuador	02/01/04	16:16	1°15′51″n; 90°17′42″w	2.0	2,376	27.8	32.6	0.1–0.8	Open ocean	0.22 (0.28 ± 0.02)	102,708
GS27	Devil's Crown, Floreana	Ecuador	02/04/04	11:41	1°12′58″s; 90°25′22″w	2.0	2.3	25.5	34.9	0.1–0.8	Coastal	0.40 (0.38 ± 0.03)	222,080
GS28	Coastal Floreana	Ecuador	02/04/04	15:47	1°13′1″s; 90°19′11″w	2.0	156	25.0[c]		0.1–0.8	Coastal	0.35 (0.35 ± 0.02)	189,052
GS29	North James Bay, Santigo	Ecuador	02/08/04	18:03	0°12′0″s; 90°50′7″w	2.0	12	26.2	34.5	0.1–0.8	Coastal	0.40 (0.39 ± 0.03)	131,529
GS30	Warm seep, Roca Redonda	Ecuador	02/09/04	11:42	0°16′20″n; 91°38′0″w	19.0	19	26.9		0.1–0.8	Warm seep		359,152
GS31	Upwelling, Fernandina	Ecuador	02/10/04	14:43	0°18′4″s; 91°39′6″w	12.0	19	18.6		0.1–0.8	Coastal upwelling	0.35 (0.39 ± 0.03)	436,401
GS32	Mangrove, Isabella	Ecuador	02/11/04	11:30	0°35′38″s; 91°4′10″w	0.3	0.67	25.4		0.1–0.8	Mangrove		148,018
GS33	Punta Cormorant Lagoon, Floreana	Ecuador	02/19/04	13:35	1°13′42″s; 90°25′45″w	0.2	0.33	37.6	46[c]	0.1–0.8	Hypersaline		692,255
GS34	North Seamore	Ecuador	02/19/04	17:06	0°22′59″s; 90°16′47″w	2.0	35	27.5		0.1–0.8	Coastal	0.36 (0.35 ± 0.02)	134,347
GS35	Wolf Island	Ecuador	03/01/04	16:44	1°23′21″n; 91°49′1″w	2.0	71	21.8	34.5	0.1–0.8	Coastal	0.28 (0.31 ± 0.02)	140,814
GS36	Cabo Marshall, Isabella	Ecuador	03/02/04	12:52	0°1′15″s; 91°11′52″w	2.0	67	25.8	34.6	0.1–0.8	Coastal	0.65 (0.45 ± 0.05)	77,538
GS37	Equatorial Pacific TAO Buoy	International	03/17/04	16:38	1°58′26″s; 95°0′53″w	2.0	3,334	28.8		0.1–0.8	Open ocean	0.21 (0.24 ± 0.02)	65,670
GS47	201 miles from French Polynesia	International	03/28/04	15:25	10°7′53″s; 135°26′58″w	30.0	2,400	28.6	37.3	0.1–0.8	Open ocean		66,023
GS51	Rangirora Atoll	French Polynesia	05/22/04	7:04	15°8′37″s; 147°26′6″w	1.0	10	27.3	34.2	0.1–0.8	Coral reef atoll		128,982
Total													7,697,926

[a]Temperature.
[b]Salinity.
[c]Measurements were acquired from nearby vessels and/or research stations.
doi:10.1371/journal.pbio.0050077.t001

Table 2. Summary Assembly Statistics

Category	Statistic	Value
Assembly inputs	Number of reads used for assembly	7,697,926
	Total read length (bp)	6,325,208,303
	Number of "intigs" used for assembly[a]	6,389,523
	Total intig length (bp)[a]	5,883,982,712
Assembly outputs	Number of assemblies[b]	3,081,849
	Total assembled consensus length (bp)	4,460,027,783
	Percentage of unassembled reads	53%
	Percentage of assembly at >1× coverage	15.3%
	Base pairs in contigs ≥ 10 kb	39,427,102
	Base pairs in contigs ≥ 50 kb	15,723,513
	Base pairs in scaffolds ≥ 2.5 kb (consensus bases)	458,196,599
	Base pairs in scaffolds ≥ 5 kb (consensus bases)	138,137,150
	Base pairs in scaffolds ≥ 10 kb (consensus bases)	65,238,481
	Base pairs in scaffolds ≥ 50 kb (consensus bases)	20,738,836
	Base pairs in scaffolds ≥ 100 kb (consensus bases)	16,005,244
	Base pairs in scaffolds ≥ 300 kb (consensus bases)	8,805,668
	Percentage of assembly in scaffolds ≥ 10 kb[c]	1.5%
	Length of longest contig (bp)	977,960
	Length of longest scaffold (bp)	2,097,794
	N1[d] assembly bp	15,915
	N10[d] assembly bp	2,533
	N50[d] assembly bp	1,611
	N1[d] contig bp	8,994
	N10[d] contig bp	2,447
	N50[d] contig bp	(single reads)

[a]Intigs are overlapping mated reads that have been collapsed into a single sequence as input into the assembler.
[b]Assemblies refers to the total number of scaffolds, pairs of mated nonoverlapping singletons, and singleton unmated reads.
[c]For comparison purposes, 10 kb is the average contig size predicted for 4.1× coverage for an idealized shotgun assembly of a repeat-free, clonal genome [22].
[d]N1 indicates the length of the next largest assembled sequence or contig such that 1% of the sequence data falls into longer assemblies or contigs. N10 and N50 indicate that 10% and 50% of the data fell into larger assemblies or contigs.
doi:10.1371/journal.pbio.0050077.t002

nomic markers" such as 16S or *recA*, thus clearly identifying the taxonomic group with which they are associated. The primary assembly of the combined GOS dataset was performed using the Celera Assembler [21] with modifications as previously described [19] and as given in Materials and Methods. The assembly was performed with quite stringent criteria, beginning with an overlap cutoff of 98% identity to reduce the potential for artifacts (e.g., chimeric assemblies or consensus sequences diverging substantially from the genome of any given cell). This assembly was the substrate for annotation (see the accompanying paper by Yooseph et al. [18]).

The degree of assembly of a metagenomic sample provides an indication of the diversity of the sample. A few substantial assemblies notwithstanding, the primary assembly was strikingly fragmented (Table 2). Only 9% of sequencing reads went into scaffolds longer than 10 kbp. A majority (53%) of the sequencing reads remained unassembled singletons. Scaffolds containing more than 50 kb of consensus sequence totaled 20.7 Mbp; of these, >75% were produced from a single Sargasso Sea sample and correspond to the *Burkholderia* or *Shewanella* assemblies described previously [19]. These results highlight the unusual abundance of these two organisms in a single sample, which significantly affected our expectations regarding the current dataset. Given the large size of the combined dataset and the substantial amount of sequencing performed on individual filters, the overall lack of assembly provides evidence of a high degree of diversity in surface planktonic communities. To put this in context, suppose there were a clonal organism that made up 1% of our data, or ~60 Mbp. Even a genome of 10 Mbp—enormous by bacterial standards—would be covered ~6-fold. Such data might theoretically assemble with an average contig approaching 50 kb [22]. While real assemblies generally fall short of theory for various reasons, *Shewanella* data make up <1% of the total GOS dataset, and yet most of the relevant reads assemble into scaffolds >50 kb. Thus, with few scaffolds of significant length, we could conclude that there are very few clonal organisms present at even 1% in the GOS dataset.

To investigate the nature of the implied diversity and to see whether greater assembly could be achieved, we explored several alternative approaches. Breaks in the primary assembly resulted from two factors: incomplete sequence coverage and conflicts in the data. Conflicts can break assemblies when there is no consistent way to chain together all overlapping sequencing reads. As it was possible that there would be fewer conflicts within a single sample (i.e., that diversity within a single sample would be lower), assemblies were attempted with individual samples. However, the results did not show any systematic improvements even in those samples with greater coverage (unpublished data). Upon manual inspection, most assembly-breaking conflicts were found to be local in nature. These observations suggested that reducing the degree of sequence identity required for assembly could ameliorate both factors limiting assembly: effective coverage would increase and many minor conflicts would be resolved.

Accordingly, we produced a series of assemblies based on 98%, 94%, 90%, 85%, and 80% identity overlaps for two subsets of the GOS dataset, again using the Celera Assembler. Assembly lengths increased as the overlap cutoff decreased from 98% to 94% to 90%, and then leveled off or even dropped as stringency was reduced below 90% (Table 3). Although larger assemblies could be generated using lower identity overlaps, significant numbers of overlaps satisfying the chosen percent identity cutoff still went unused in each assembly. This is consistent with a high rate of conflicting overlaps and in turn diagnostic of significant polymorphism.

In mammalian sequencing projects the use of larger insert libraries is critical to producing larger assemblies because of their ability to span repeats or local polymorphic regions [23]. The shotgun sequencing libraries from the GOS filters were typically constructed from inserts shorter than 2 kb. Longer plasmid libraries were attempted but were much less stable. We obtained paired-end sequences from 21,419 fosmid clones (average insert size, 36 kb; [24,25]) from the 0.1-micron fraction of GS-33. The effect of these long mate pairs on the GS-33 assembly was quite dramatic, particularly at high stringency (e.g., improving the largest scaffold from 70 kb to 1,247 kb and the largest contig from 70 kb to 427 kb). At least for GS-33 this suggests that many of the polymorphisms affect small, localized regions of the genome that can be spanned using larger inserts. This degree of improvement may be greater than what could be expected in general, as the diversity of GS-33 is by far the lowest of any of the currently

Table 3. Evaluation of Alternative Assembly Methods

Dataset	Type	Percent Identity	Base Pairs in 10 k Contigs	Base Pairs in 100 k Contigs
GS33 plasmids	WGS[a]	98	13,669,678	0
		94	19,536,324	2,749,543
		90	20,996,826	3,729,765
		85	20,327,989	3,505,324
		80	19,245,637	4,195,959
	E-asm[b]	98	22,000,579	5,604,857
		94	22,781,462	7,302,801
		90	22,702,764	7,600,441
		85	22,570,933	7,937,079
		80	20,335,558	4,779,684
GS33 with fosmids	WGS[a]	98	15,031,557	1,306,992
		94	22,310,335	4,449,710
		90	22,944,278	5,585,959
		85	22,251,738	5,485,013
		80	21,088,975	5,684,925
GS17,18,23,26	WGS[a]	98	185,058	0
		94	5,422,366	213,755
		90	10,694,783	373,822
		85	11,514,421	800,290
		80	9,004,221	879,401
	E-asm[b]	98	2,047,524	0
		94	10,668,547	1,184,881
		90	15,215,981	2,634,227
		85	15,786,515	3,132,152
		80	13,767,929	2,942,160
Combined GOS	WGS[a]	98	39,427,102	11,488,828
		94	98,887,937	12,376,236
	E-asm[b]	98	91,526,091	16,444,304
		94	163,612,717	25,564,163
		90	186,614,813	28,752,198
		85	181,887,218	27,154,335
		80	161,160,091	23,794,832

[a]Whole-genome shotgun (WGS) assembly performed with the Celera Assembler.
[b]Assemblies performed using extreme assembly approach (E-asm).
doi:10.1371/journal.pbio.0050077.t003

sequenced GOS samples, yet it clearly indicates the utility of including larger insert libraries for assembly.

Fragment Recruitment

In the absence of substantial assembly, direct comparison of the GOS sequencing data to the genomes of sequenced microbes is an alternative way of providing context, and also allows for exploration of genetic variation and diversity. A large and growing set of microbial genomes are available from the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov). At the time of this study, we used 334 finished and 250 draft microbial genomes as references for comparison with the GOS sequencing reads. Comparisons were carried out in nucleotide-space using the sequence alignment tool BLAST [26]. BLAST parameters were designed to be extremely lenient so as to detect even distant similarities (as low as 55% identity). A large proportion of the GOS reads, 70% in all, aligned to one or more genomes under these conditions. However, many of the alignments were of low identity and used only a portion of the entire read. Such low-quality hits may reflect distant evolutionary relationships, and therefore less information is gained based on the context of the alignment. More stringent criteria could be imposed requiring that the reads be aligned over nearly their entire length without any large gaps. Using this stringent criterion only about 30% of the reads aligned to any of the 584 reference genomes. We refer to these fully aligned reads as "recruited reads." Recruited reads are far more likely to be from microbes closely related to the reference sequence (same species) than are partial alignments. Despite the large number of microbial genomes currently available, including a large number of marine microbes, these results indicate that a substantial majority of GOS reads cannot be specifically related to available microbial genomes.

The amount and distribution of reads recruited to any given genome provides an indication of the abundance of closely related organisms. Only genomes from the five bacterial genera *Prochlorococcus, Synechococcus, Pelagibacter, Shewanella,* and *Burkholderia* yielded substantial and uniform recruitment of GOS fragments over most of a reference genome (Table 4). These genera include multiple reference genomes, and we observed significant differences in recruitment patterns even between organisms belonging to the same species (Figure 2A–2I). Three genera, *Pelagibacter* (Figure 2A), *Prochlorococcus* (Figure 2B–2F), and *Synechococcus* (Figure 2G–2I), were found abundantly in a wide range of samples and together accounted for roughly 50% of all the recruited reads (though only ~15% of all GOS sequencing reads). By contrast, although every genome tested recruited some GOS reads, most recruited only a small number, and these reads clustered at lower identity to locations corresponding to large highly conserved genes (for typical examples see Figure 2E–2F). We refer to this pattern as nonspecific recruitment as it reflects taxonomically nonspecific signals, with the reads in

Table 4. Microbial Genera that Recruited the Bulk of the GOS Reads

Genus	Read Count			Best Strain		
	All Reads	80%+[a]	90%+[b]	All Reads	80%+[a]	90%+[b]
Pelagibacter	922,677	195,539	36,965	HTCC1062	HTCC1062	HTCC1062
Prochlorococcus	208,999	159,102	84,325	MIT9312	MIT9312	MIT9312
Synechococcus	60,650	26,365	21,594	CC9902	RS9917	RS9917
Burkholderia	151,123	108,610	93,081	383	383	383
Shewanella	59,086	34,138	27,693	MR-1	MR-1	MR-1
Remaining	43,244	2,367	564	*Buchnera aphidicola* Str. Sg	*Buchnera aphidicola* Str. APS	*Alteromonas macleodii*

[a]Reads aligned at or above 80% identity over the entire length of the read.
[b]Reads aligned at or above 90% identity over the entire length of the read.
doi:10.1371/journal.pbio.0050077.t004

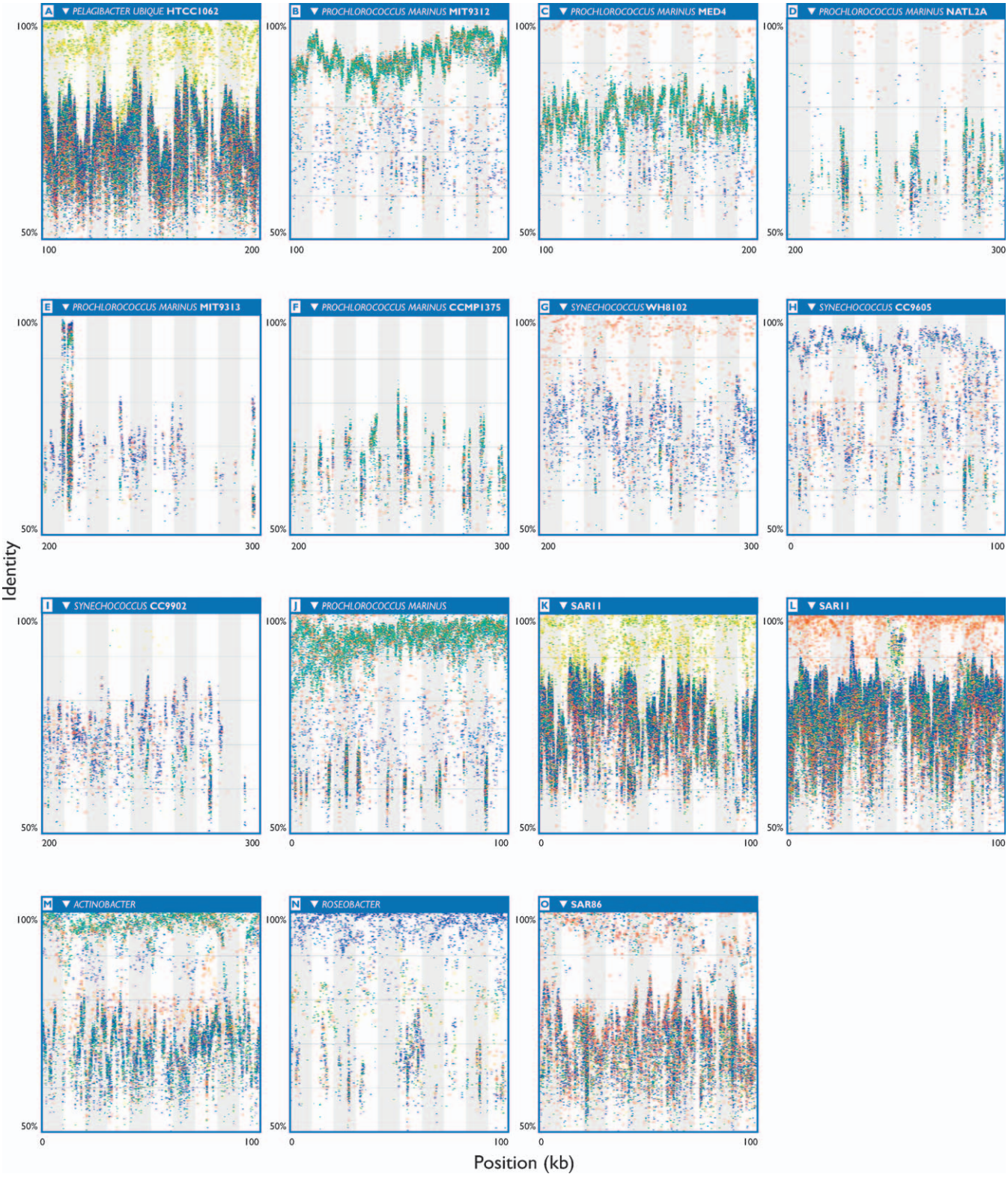


Figure 2. Fragment Recruitment Plots

The horizontal axis of each panel corresponds to a 100-kb segment of genomic sequence from the indicated reference microbial genome. The vertical axis indicates the sequence identity of an alignment between a GOS sequence and the reference genomic sequence. The identity ranges from 100% (top) to 50% (bottom). Individual GOS sequencing reads were colored to reflect the sample from which they were isolated. Geographically nearby samples have similar colors (see Poster S1 for key). Each organism shows a distinct pattern of recruitment reflecting its origin and relationship to the environmental data collected during the course of this study.

(A) *P. ubique* HTCC1062 recruits the greatest density of GOS sequences of any genome examined to date. The GOS sequences show geographic

stratification into bands, with sequences from temperate water samples off the North American coast having the highest identity (yellow to yellow-green colors). At lower identity, sequences from all the marine environments could be aligned to HTCC1062.
(B) *P. marinus* MIT9312 recruits a large number of GOS sequences into a single band that zigzags between 85%–95% identity on average. These sequences are largely derived from warm water samples in the Gulf of Mexico and eastern Pacific (green to greenish-blue reads).
(C) *P. marinus* MED4 recruits largely the same set of reads as MIT9312 (B) though the sequences that form the zigzag recruit at a substantially lower identity. A small number of sequences from the Sargasso Sea samples (red) are found at high identity.
(D) *P. marinus* NATL2A recruits far fewer sequences than any of the preceding panels. Like MED4, a small number of high-identity sequences were recruited from the Sargasso samples.
(E) *P. marinus* MIT9313 is a deep-water low-light–adapted strain of *Prochlorococcus*. GOS sequences were recruited almost exclusively at low identity in vertical stacks that correspond to the locations of conserved genes. On the left side of this panel is a very distinctive pattern of recruitment that corresponds to the highly conserved 16S and 23S mRNA gene operon.
(F) *P. marinus* CCMP1375, another deep-water low-light–adapted strain, does not recruit GOS sequences at high identity. Only stacks of sequences are seen corresponding to the location of conserved genes.
(G) *Synechococcus* WH8102 recruits a modest number of high-identity sequences primarily from the Sargasso Sea samples. A large number of moderate identity matches from the Pacific and hypersaline lagoon (GS33) samples are also visible.
(H) *Synechococcus* CC9605 recruits largely the same sequences as does *Synechococcus* WH8102, but was isolated from Pacific waters. GOS sequences from some of the Pacific samples recruit at high identity, while sequences from the Sargasso and hypersaline lagoon (bluish-purple) were recruited at moderate identities.
(I) *Synechococcus* CC9902 is distantly related to either of the preceding *Synechococcus* strains. While this strain also recruits largely the same sequences as the WH8102 and CC9902 strains, they recruit at significantly lower identity.
(J–O) Fragment recruitment plots to extreme assemblies seeded with phylogenetically informative sequences. Using this approach it is not only possible to assemble contigs with strong similarities to known genomes but to identify contigs from previously uncultured genomes. In each case a 100-kb segment from an extreme assembly is shown. Each plot shows a distinct pattern of recruitment that distinguishes the panels from each other.
(J) Seeded from a *Prochlorococcus marinus*-related sequence, this contig recruits a broad swath of GOS sequences that correspond to the GOS sequences that form the zigzag on *P. marinus* MIT9312 recruitment plots (see [B] or Poster S1 for comparison).
(K–L) Seeded from SAR11 clones, these contigs show significant synteny to the known *P. ubique* HTCC1062 genome. (K) is strikingly similar to previous recruitment plots to the HTCC1062 genome (see [A] or Poster S1). In contrast, (L) identifies a different strain that recruits high-identity GOS sequences primarily from the Sargasso Sea samples (red).
(M–O) These three panels show recruitment plots to contigs belonging to the uncultured *Actinobacter, Roseobacter,* and SAR86 lineages.
doi:10.1371/journal.pbio.0050077.g002

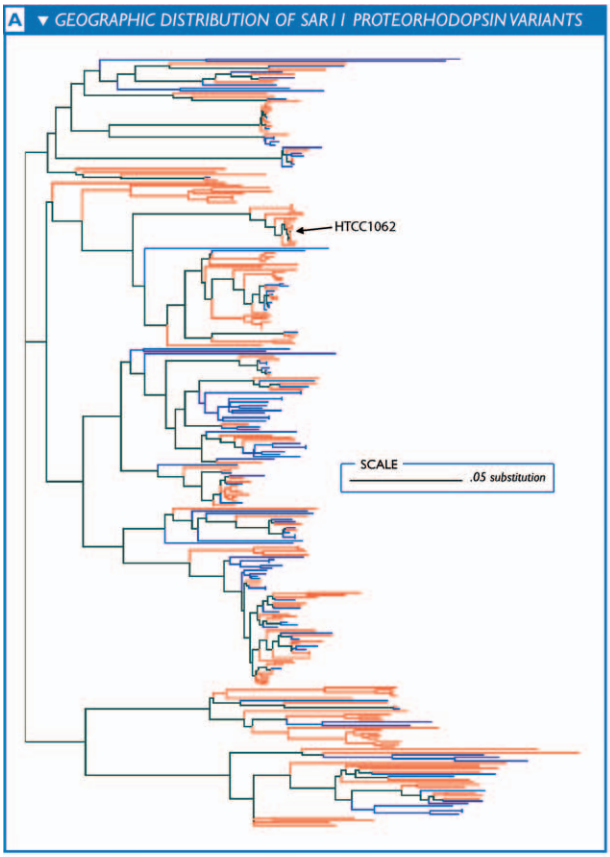
question often recruiting to distantly related sets of genomes. Most microbial genomes, including many of the marine microbes (e.g., the ubiquitous genus *Vibrio*), demonstrated this nonspecific pattern of recruitment.

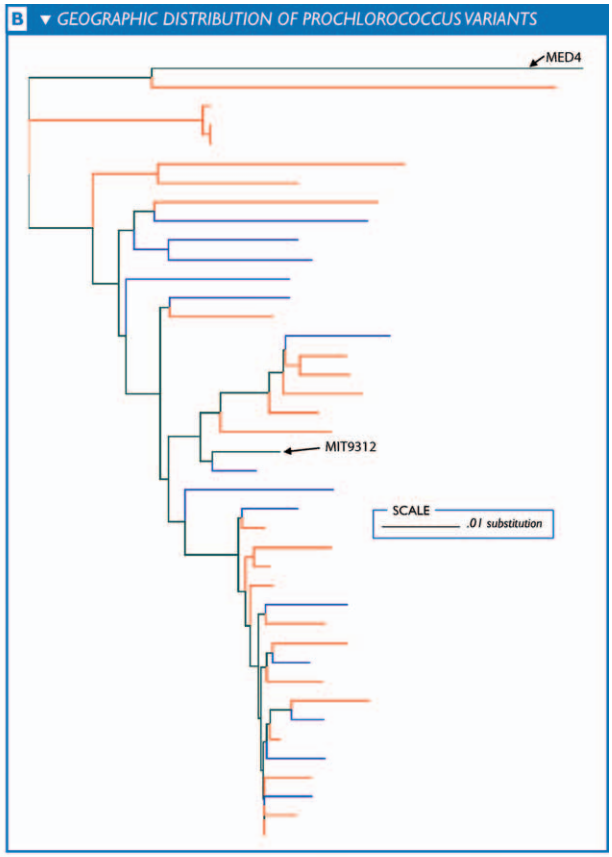
The relationship between the similarity of an individual sequencing read to a given genome and the sample from which the read was isolated can provide insight into the structure, evolution, and geographic distribution of microbial populations. These relationships were assessed by constructing a "percent identity plot" [27] in which the alignment of a read to a reference sequence is shown as a bar whose horizontal position indicates location on the reference and whose vertical position indicates the percent identity of the alignment. We colored the plotted reads according to the samples to which they belonged, thus indirectly representing various forms of metadata (geographic, environmental, and laboratory variables). We refer to these plots that incorporate metadata as fragment recruitment plots. Fragment recruitment plots of GOS sequences recruited to the entire genomes of *Pelagibacter ubique* HTCC1062, *Prochlorococcus marinus* MIT9312, and *Synechococcus* WH8102 are presented in Poster S1.

Within-Ribotype Population Structure and Variation

Characteristic patterns of recruitment emerged from each of these abundant marine microbes consisting of horizontal bands made up of large numbers of GOS reads. These bands seem constrained to a relatively narrow range of identities that tile continuously (or at least uniformly, in the case when abundance/coverage is lower) along ~90% of the reference sequence. The uninterrupted tiling indicates that environmental genomes are largely syntenic with the reference genomes. Multiple bands, distinguished by degree of similarity to the reference and by sample makeup, may arise on a single reference (Poster S1D and S1F). Each of these bands appears to represent a distinct, closely related population we refer to as a subtype. In some cases, an abundant subtype is highly similar to the reference genome, as is the case for *P. marinus* MIT9312 (Poster S1) and *Synechococcus* RS9917 (unpublished data). *P. ubique* HTCC1062 and other *Synechococcus* strains like WH8102 show more complicated banding patterns (Poster S1D and S1F) because of the presence of multiple subtypes that produce complex often overlapping bands in the plots. Though the recruitment patterns can be quite complex they are also remarkably consistent over much of the reference genome. In these more complicated recruitment plots, such as the one for *P. ubique* HTCC1062, individual bands can show sudden shifts in identity or disappear altogether, producing a gap in recruitment that appears to be specific to that band (see *P. ubique* recruitment plots on Poster S1B and S1E, and specifically between 130–140 kb). Finally, phylogenetic analysis indicates that separate bands are indeed evolutionarily distinct at randomly selected locations along the genome.

The amount of sequence variation within a given band cannot be reliably determined from the fragment recruitment plots themselves. To examine this variation, we produced multiple sequence alignments and phylogenies of reads that recruited to several randomly chosen intervals along given reference genomes to show that there can be considerable within-subtype variation (Figure 3A–3B). For example, within the primary band found in recruitment plots to *P. marinus* MIT9312, individual pairs of overlapping reads typically differ on average between 3%–5% at the nucleotide level (depending on exact location in the genome). Very few reads that recruited to MIT9312 have perfect (mismatch-free) overlaps with any other read or to MIT9312, despite ~100-fold coverage. While many of these differences are silent (i.e., do not change amino acid sequences), there is still considerable variation at the protein level (unpublished data). The amount of variation within subtypes is so great that it is likely that no two sequenced cells contained identical genomes.


A
GEOGRAPHIC DISTRIBUTION OF SAR11 PROTEORHODOPSIN VARIANTS
HTCC1062
SCALE
.05 substitution


B
GEOGRAPHIC DISTRIBUTION OF PROCHLOROCOCCUS VARIANTS
MED4
MIT9312
SCALE
.01 substitution

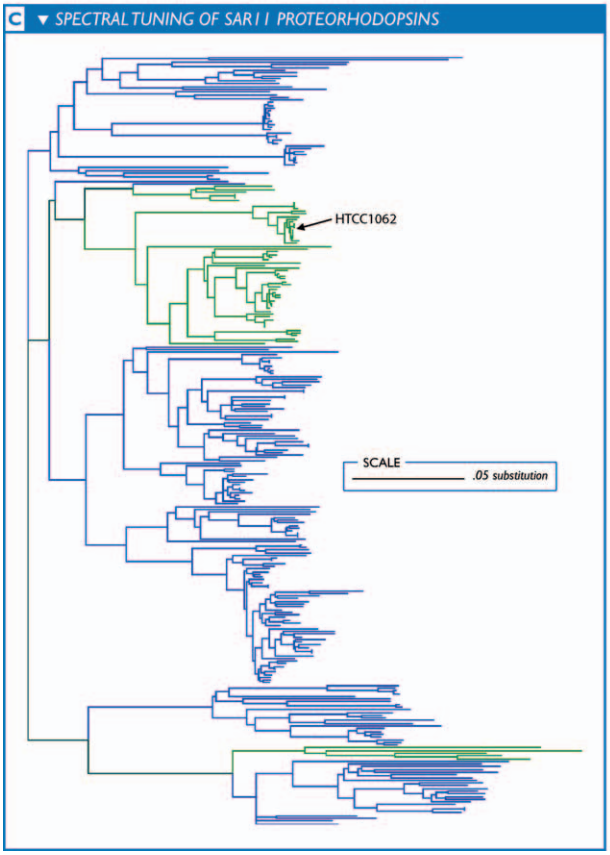

C
SPECTRAL TUNING OF SAR11 PROTEORHODOPSINS
HTCC1062
SCALE
.05 substitution

Figure 3. Population Structure and Variation as Revealed by Phylogeny

Phylogenies were produced using neighbor-joining. There is significant within-clade variation as well as an absence of strong geographic structure to variants of SAR11 (*P. ubique* HTCC1062) and *P. marinus* MIT9312. Similar reads are not necessarily from similar locations, and reads from similar locations are not necessarily similar.

(A) Geographic distribution of SAR11 proteorhodopsin variants. Keys to coloration: blue, Pacific; pink, Atlantic.

(B) Geographic distribution of *Prochlorococcus* variants. Keys to coloration: blue, Pacific; pink, Atlantic.

(C) Origins of spectral tuning of SAR11 proteorhodopsins. Reads are colored according to whether they contain the L (green) or Q (blue) variant at the spectral tuning residue described in the text. The selection of tuning residue is lineage restricted, but each variant must have arisen on two separate occasions.

doi:10.1371/journal.pbio.0050077.g003

Identifying Genomic Structural Variation with Metagenomic Data

Variation in genome structure in the form of rearrangements, duplications, insertions, or deletions of stretches of DNA can also be explored via fragment recruitment. The use of mated sequencing reads (pairs of reads from opposite ends of a clone insert) provides a powerful tool for assessing structural differences between the reference and the environmental sequences. The cloning and sequencing process determines the orientation and approximate distance between two mated sequencing reads. Genomic structural variation can be inferred when these are at odds with the way in which the reads are recruited to a reference sequence. Relative location and orientation of mated sequences provide a form of metadata that can be used to color-code a fragment recruitment plot (Figure 4). This makes it possible to visually identify and classify structural differences and similarities between the reference and the environmental sequences (Figure 5). For the abundant marine microbes, a high proportion of mated reads in the "good" category (i.e., in the proper orientation and at the correct distance) show that synteny is conserved for a large portion of the microbial population. The strongest signals of structural differences typically reflect a variant specific to the reference genome and not found in the environmental data. In conjunction with the requirement that reads be recruited over their entire length without interruption, recruitment plots result in pronounced recruitment gaps at locations where there is a break in synteny. Other rearrangements can be partially present or penetrant in the environmental data and thus may not generate obvious recruitment gaps. However, given sufficient coverage, breaks in synteny should be clearly identifiable using the recruitment metadata based on the presence of "missing" mates (i.e., the mated sequencing read that was recruited but whose mate failed to recruit; Figure 4). The ratio of missing mates to "good" mates determines how penetrant the rearrangement is in the environmental population.

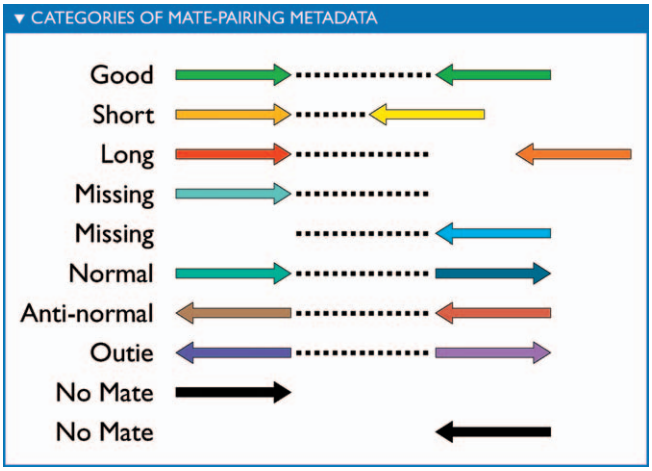


Figure 4. Categories of Recruitment Metadata

The recruitment metadata distinguishes eight different general categories based on the relative placement of paired end sequencing reads (mated reads) when recruited to a reference sequence in comparison to their known orientation and separation on the clone from which they were derived. Assuming orientation is correct, two mated reads can be recruited closer together, further apart, or within expected distances given the size of the clone from which the sequences were derived. These sequences are categorized as "short," "long," or "good," respectively. Alternately, the mated reads may be recruited in a misoriented fashion, which trumps issues of separation. These reads can be categorized as "normal," "anti-normal," or "outie." In addition, there are two other categories. "No mate" indicates that no mated read was available for recruitment, possibly due to sequencing error. Perhaps most useful of any of the recruitment categories, "missing" mates indicate that while a mated sequence was available, it was not recruited to the reference. "Missing" mates identify breaks in synteny between the environmental data and the reference sequence.

doi:10.1371/journal.pbio.0050077.g004

In theory, all genome structure variations that are large enough to prevent recruitment can be detected, and all such rearrangements will be associated with missing mates. Depending on the type of rearrangement present other recruitment metadata categories will be present near the rearrangements' endpoints. This makes it possible to distinguish among insertions, deletions, translocations, inversions, and inverted translocations directly from the recruitment plots. Examples of the patterns associated with different rearrangements are presented in Figure 5. This provides a rapid and easy visual method for exploring structural variation between natural populations and sequenced representatives (Poster S1A and S1B).

Genomic Structural Variation in Abundant Marine Microbes

Variation in genome structure potentially results in functional differences. Of particular interest are those differences between sequenced (reference) microbes and environmental populations. These differences can indicate how representative a cultivated microbe might be and shed light on the evolutionary forces driving change in microbial populations. Fragment recruitment in conjunction with the mate metadata helped us to identify both the consistent and the rare structural differences between the genomes of microbial populations in the GOS data and their closest sequenced relatives. Our analysis has thus far been confined to the three microbial genera that were widespread in the GOS dataset as represented by the finished genomes of *P. marinus* MIT9312, *P. ubique* HTCC1062, and to a lesser extent *Synechococcus* WH8102. Each of these genomes is characterized by large and small segments where little or no fragment recruitment took place. We refer to these segments as "gaps." These gaps

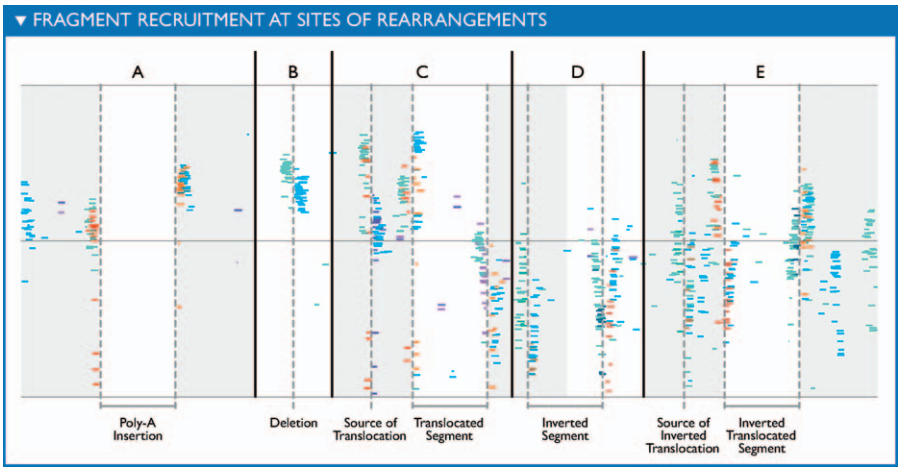


Figure 5. Fragment Recruitment at Sites of Rearrangements

Environmental sequences recruited near breaks in synteny have characteristic patterns of recruitment metadata. Indeed, each of five basic rearrangements (i.e., insertion, deletion, translocation, inversion, and inverted translocation) produced a distinct pattern when examining the recruitment metadata. Here, example recruitment plots for each type of rearrangement have been artificially generated. The "good" and "no mate" categories have been suppressed. In each case, breaks in synteny are marked by the presence of stacks of "missing" mate reads. The presence or absence of other categories distinguishes each type of rearrangement from the others.

doi:10.1371/journal.pbio.0050077.g005

represent reference-specific differences that are not found in the environmental populations rather than a cloning bias that identifies genes or gene segments that are toxic or unclonable in *E. coli*. The presence of missing mates flanking these gaps indicates that the associated clones do exist, and therefore that cloning issues are not a viable explanation for the absence of recruited reads. Although the reference-specific differences are quite apparent due to the recruitment gaps they generate, there are also sporadic rearrangements associated with single clones, mostly resulting from small insertions or deletions.

Careful examination of the unrecruited mates of the reads flanking the gaps allowed us to identify, characterize, and quantify specific differences between the reference genome and their environmental relatives. The results of this analysis for *P. ubique* and *P. marinus* have been summarized in Table 5. With few exceptions, small gaps resulted from the insertion or deletion of only a few genes. Many of the genes associated with these small insertions and deletions have no annotated function. In some cases the insertions display a degree of variability such that different sets of genes are found at these locations within a portion of the population. In contrast, many of the larger gaps are extremely variable to the extent that every clone contains a completely unrelated or highly divergent sequence when compared to the reference or to other clones associated with that gap. These segments are hypervariable and change much more rapidly than would be expected given the variation in the rest of the genome. Sites containing a hypervariable segment nearly always contained some insert. We identified two exceptions both associated with *P. ubique*. The first is approximately located at the 166-kb position in the *P. ubique* HTCC1062 genome. Though no large gap is present, the mated reads indicate that under many circumstances a highly variable insert is often present. The second is a gap on HTCC1062 that appears between 50 and 90 kb. This gap appears to be less variable than other hypervariable segments and is occasionally absent based on the large numbers of flanking long mated reads (Poster S1A). Interestingly, the long mated reads around this gap seem to be disproportionately from the Sargasso Sea samples, suggesting that this segment may be linked to geographic and/or environmental factors. Thus, hypervariable segments are highly variable even within the same sample, can on occasion be unoccupied, and the variation, or lack thereof, can be sample dependent.

Hypervariable segments have been seen previously in a wide range of microbes, including *P. marinus* [28], but their precise source and functional role, especially in an environmental context, remains a matter of ongoing research. For clues to these issues we examined the genes associated with the missing mates flanking these segments and the nucleotide composition of the gapped sequences in the reference genomes. In some rare cases the genes identified on reads that should have recruited within a hypervariable gap were highly similar to known viral genes. For example, a viral integrase was associated with the *P. ubique* HTCC1062 hypervariable gap between 516 and 561 kb. However, in the majority of cases the genes associated with these gaps were uncharacterized, either bearing no similarity to known genes or resembling genes of unknown function. If these genes were indeed acquired through horizontal transfer then we might expect that they would have obvious compositional biases. Oligonucleotide frequencies along the *P. ubique* HTCC1062 and *Synechococcus* WH8102 genomes are quite different in the large recruitment gaps in comparison to the well-represented portions of the genome (Poster S1). Surprisingly, this was less true for *P. marinus* MIT9312, where the gaps have been linked to phage activity [28]. These results suggest that these hypervariable segments of the genome are widespread among marine microbial populations, and that they are the product of horizontal transfer events perhaps mediated by phage or transposable elements. These results are consistent with and expand upon the hypothesis put forward by Coleman et al. [28] suggesting that these segments are phage mediated, and conflicts with initial claims that the HTCC1062 genome was devoid of genes acquired by horizontal transfer [29].

Table 5. Atypical Segments in *P. marinus* MIT9312 and *P. ubique* HTCC1062 (SAR11)

Reference Genome	Begin[a]	End[b]	Size, bp	Type of Variant[c]	Description
MIT9312	36,401	38,311	2,132	Variable deletion	12 out of 66 clones support simple deletion. Remaining clones show considerable sequence variation amongst themselves.
MIT9312	124,448	125,219	771	Variable insertion	Associated with ASN tRNA gene; in the environment, half the reads identify pair of small inserts with no similarity to known genes; the other half point to small and large inserts of undetermined nature.
MIT9312	233,826	233,910	84	Insertion	All clones support small insert (270 bp) with no clear sequence similarity to known genes or sequences.
MIT9312	243,296	245,115	424	Variable deletion	24 of 42 clones support simple deletion of hypothetical protein. 13 support slightly larger deletion. Remaining not clearly resolved.
MIT9312	296,818	300,888	4,344	Variable deletion	44 clones support deletion of 3,070 bp segment containing 4 genes (3 hypothetical; 1 carbamoyltransferase). 17 clones support alternative sequences with little or no similarity to each other.
gMIT9312	342,404	342,662	326	Variable insert	In environment is 93% chance of finding deoxyribodopyrimiden photolyase with 7% chance of finding an ABC type Fe3+ siderophore transport system permease component.
MIT9312	345,933	365,351	19,418	Hypervariable	Very limited similarity among clones indicates that this is a hypervariable segment. Note that it is closely associated with a site-specific integrase/recombinase.
MIT9312	551,347	552,025	678	Deletion	Two small deletions within a hypothetical protein.
MIT9312	617,914	621,556	3,642	Hypervariable	About 50% of the clones support small deletions among several hypothetical proteins. Remaining support significant variability suggesting hypervariable segment. At least two of the missed mates contain integrase-like genes.
MIT9312	646,340	652,375	6,035	Deletion/Hypervariable	Majority of clones support simple deletion of eight hypothetical and hypothetical genes. Small number of clones indicate this may by hypervariable as well.
MIT9312	655,241	655,800	559	Deletion	Small deletion between hypothetical proteins.
MIT9312	665,000	678000	12,000	Deletions	Complex set of deletions and replacements that vary with geographic location.
MIT9312	665,824	666,380	556	Insertion	Small inserted hypothetical protein.
MIT9312	670,747	671,933	1,186	Deletion	Deletes hypothetical gene.
MIT9312	736,266	736,289	23	Insertion	All clones support insertion of fructose-bisphosphate aldolase and fructose-1,6-bisphosphate aldolase.
MIT9312	762,156	762,717	561	Deletion	Small deletion between hypothetical proteins.
MIT9312	779,006	779,309	303	Insertion	Small hypothetical protein inserted.
MIT9312	874,349	874,913	564	Insertion	Small insertion including gene with similarity to RNA-dependent RNA-polymerase.
MIT9312	943,389	946,997	3,608	Variable	Several small changes.
MIT9312	1,043,129	1,131,874	88,745	Hypervariable	
MIT9312	1,140,922	1,141,412	490	Insertion	Small insertion.
MIT9312	1,144,307	1,144,790	483	Insertion	Small insertion of several genes.
MIT9312	1,155,123	1,156,440	1,317	Deletion	Deletes high-light–inducible protein.
MIT9312	1,172,609	1,177,292	4,683	Variable	Several genes have been replaced or deleted.
MIT9312	1,202,643	1,274,335	71,692	Hypervariable	
MIT9312	1,288,481	1,290,367	1,886	Variable deletion	Deletes polysaccharide export-related periplasmic protein (28 out of 55). Other deletions are variable and may include replacement with alternate sequences.
MIT9312	1,323,606	1,324,523	917	Deletion	Environmental sequences lack a small hypothetical protein.
MIT9312	1,369,637	1,369,996	359	Insertion	NAD-dependent DNA ligase absent from MIT9312; has possible paralog
MIT9312	1,381,273	1,382,049	776	Deletion	Small insert in MIT9312 not present in environment
MIT9312	1,384,664	1,385,110	446	Deletion	Deletes delta(12)-fatty acid dehydrogenase and replaces gene with small (~100 bp) sequence. There is some variation in the exact location and replacement sequence.
MIT9312	1,388,430	1,389,718	1,288	Replacement	Segment between two high-light–inducible proteins swapped for different sequence with no similarity.
MIT9312	1,392,865	1,392,976	111	Replacement	Small replacement deletes hypothetical gene and replaces it with small, unknown sequence. There is some variation in the precise boundaries of the deletion and in the replacement sequences.
MIT9312	1,397,696	1,420,005	22,309	Hypervariable	
MIT9312	1,486,145	1,487,971	231	Variable	Small insertion of dolichyl-phosphate-mannose-protein mannosyltransferase; alternately deletes glycosyl transferase (8 out of 56).
MIT9312	1,519,810	1,520,860	1,050	Variable deletion	Deletes a single hypothetical protein; about half of the deletions contain variable sequences of unknown origin.
MIT9312	1,568,049	1,569,121	1,072	Replacement	Typically 928-bp portion of MIT9312 replaced by 175-bp stretch in environment; some small amount of variation in environmental replacement sequence (11 out of 51).
HTCC1062	50,555	93,942	43,387	Variable deletion	Low recruitment segment containing many hypothetical, transporter, and secretion genes.
HTCC1062	146,074	146,415	341	Deletion	Often deleted segment containing DoxD-like and ferredoxin dependent glutamate synthase peptide.
HTCC1062	166,600	166,700	100	Variable replacement	GOS sequences indicate that variable blocks of genes are frequently inserted here.
HTCC1062	308,720	309,633	913	Deletion	Potential sulfotransferase domain deleted.
HTCC1062	339,545	339,951	406	Deletion	Deletes a predicted O-linked N-acetylglucosamine transferase.
HTCC1062	385,348	386,224	876	Deletion	SAM-dependent methyltransferase deleted.

Table 5. Continued.

Reference Genome	Begin[a]	End[b]	Size, bp	Type of Variant[c]	Description
HTCC1062	441,074	441,152	78	Deletion	Deletes possible methyltransferase FkbM.
HTCC1062	516,041	561,604	45,563	Variable replacement	Several high identity "missed" mates match phage genes, including an integrase.
HTCC1062	660,413	660,978	565	Deletion	Deletes hypothetical protein.
HTCC1062	675,141	676,399	1,258	Deletion	Deletes hypothetical protein.
HTCC1062	766,022	768,263	2,241	Deletion	Deletes four hypothetical proteins.
HTCC1062	814,015	816,386	2,371	Deletion	Deletes steroid monoxygenase and short-chain dehydrogenase.
HTCC1062	893,450	922,604	29,154	Deletion	Deletes large segment including a 7317-aa hypothetical gene.
HTCC1062	941,203	942,403	1,200	Deletion	Deletes hypothetical protein.
HTCC1062	991,461	997,861	6,400	Deletion	Deletes several hypotheticals and mix of other genes; adjacent to recombinase.
HTCC1062	1,117,126	1,143,483	26,357	Deletion	Large number of hypothetical and transporters that are deleted.
HTCC1062	1,160,908	1,166,589	5,681	Deletion	Deletes a small cluster of peptides with various functions.
HTCC1062	1,188,887	1,189,231	344	Deletion	Deletes portion of winged helix DNA-binding protein but inserts sequences with similarity to gi|71082757 sodium bile symporter family protein found in large gap between 50555 and 93942.

[a]Begin indicates the approximate bp position which marks the beginning of the gap in recruitment.
[b]End indicates the approximate bp position which marks the ending of the gap in recruitment.
[c]The type of change indicates what would have to happen to the reference genome to produce the sequences seen in the environment (e.g., a deletion indicates that the indicated portion of the reference would have to be deleted to generate the variant(s) seen in the environment).
doi:10.1371/journal.pbio.0050077.t005

Though insertions and deletions accounted for many of the obvious regions of structural variation, we also looked for rearrangements. The high levels of local synteny associated with *P. ubique* and *P. marinus* suggested that large-scale rearrangements were rare in these populations. To investigate this hypothesis we used the recruitment data to examine how frequently rearrangements besides insertions and deletions could be identified. We looked for rearrangements consisting of large (greater than 50 kb) inversions and translocations associated with *P. marinus;* however, we did not identify any such rearrangements that consistently distinguished environmental populations from sequenced cultivars. Rare inversions and translocations were identified in the dominant subtype associated with MIT9312 (Table 6). Based on the amount of sequence that contributed to the analysis, we estimate that one inversion or translocation will be observed for every 2.6 Mbp of sequence examined (less than once per *P. marinus* genome).

A further observation concerns the uniformity along a genome of the evolutionary history among and within subtypes. For instance, the similarity between GOS reads and *P. marinus* MIT9312 is typically 85%–95%, while the similarity between MIT9312 and *P. marinus* MED4 is generally ~10% lower. However, there are several instances where the divergence of MIT9312 and MED4 abruptly decreases to no more than that between the GOS sequences and MIT9312 (Poster S1G). These results are consistent either with horizontal transfer (recombination) or with inhomogeneous

Table 6. Six Large-Scale Translocations and Inversions Were Identified in the Abundant *P. marinus* Subtype

Group	Low Genome Begin	Low Genome End	High Genome Begin	High Genome End	Read ID	Low Read Begin	Low Read Breakpoint	Low Read Strand	High Read Breakpoint	High Read End	High Read Strand	Read Length	Inversion	Sample
1	34,428	34,904	1,536,417	1,536,774	1092255385627	0	476	1	477	834	1	834	No	15
2	607,778	608,467	1,131,368	1,131,621	1093017685727	270	959	0	0	251	1	959	Yes	18
3	618,997	619,375	1,172,217	1,172,728	1092963065572	19	397	1	425	939	1	939	No	17
3	618,997	619,372	1,172,203	1,172,728	1095433012642	0	375	1	403	931	1	941	No	26
3	618,997	619,251	1,172,216	1,172,728	1091140913752	2	256	1	284	799	1	799	No	19
3	618,997	619,331	1,172,280	1,172,728	1641121	2	337	1	365	816	1	816	No	00d
3	619,007	619,375	1,172,223	1,172,728	1092963490951	19	387	1	425	933	1	933	No	17
4	652,933	653,483	1,369,774	1,370,077	200560	325	875	0	0	303	1	875	Yes	00a
4	652,979	653,350	1,369,774	1,370,200	1492647	0	371	1	439	865	0	867	Yes	00d
4	652,979	653,496	1,369,774	1,370,086	1092256128910	10	527	1	595	905	0	906	Yes	15
4	652,979	653,496	1,369,774	1,370,061	1092405979387	14	531	1	599	886	0	886	Yes	25
4	652,979	653,353	1,369,774	1,370,132	1093017637883	2	376	1	444	802	0	802	Yes	18
4	652,980	653,496	1,369,774	1,370,111	1092343389654	6	522	1	591	928	0	928	Yes	25
5	1,049,484	1,049,793	1,172,301	1,172,782	1400802	518	827	0	1	483	0	827	No	00d
6	1,219,993	1,220,519	1,485,834	1,486,222	1092256207885	2	528	1	532	919	1	919	No	17
6	1,219,993	1,220,496	1,485,925	1,486,222	380485	300	803	0	0	296	0	803	No	00a
6	1,219,993	1,220,477	1,485,782	1,486,204	1484583	0	484	1	506	927	1	927	No	00d
6	1,220,005	1,220,388	1,485,733	1,486,216	1682914	507	890	0	2	484	0	913	No	00d

doi:10.1371/journal.pbio.0050077.t006

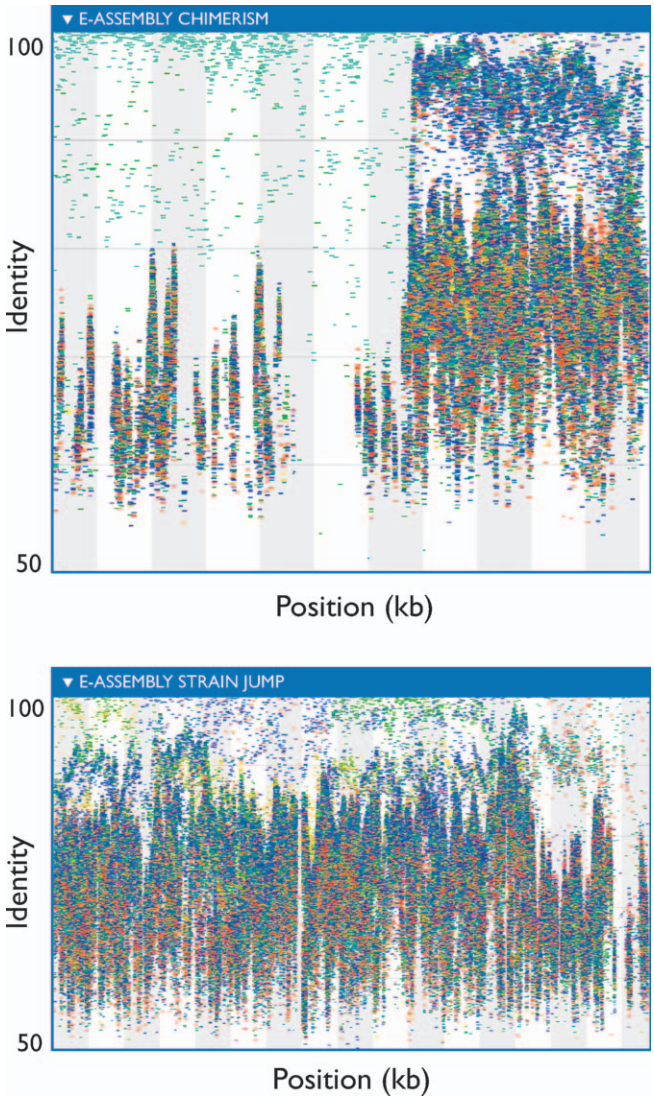


Figure 6. Examples of Chimeric Extreme Assemblies

(A) Fragment recruitment to an extreme assembly contig indicates the assembly is chimeric between two organisms, based on dramatic shifts in density of recruitment, level of conservation, and sample distribution.

(B) Fragment recruitment to a SAR11-related extreme assembly. Changes in color, density, and vertical location toward the top of the figure indicate transitions among multiple subtypes of SAR11.

doi:10.1371/journal.pbio.0050077.g006

selectional pressures. Similar patterns are present in the two high-identity subtypes seen on the *P. ubique* HTCC1062 genome (Poster S1D). Other regions show local increases in similarity between MIT9312 and the dominant subtype that are not reflected in the MIT9312/MED4 divergence (e.g., near positions 50 kb, 288 kb, 730 kb, 850 kb, and 954 kb on MIT9312; also see Poster S1G). These latter regions might reflect either regions of homogenizing recombination or regions of higher levels of purifying selection. However, the lengths of the intervals (several are 10 kb or more) are longer than any single gene and correspond to genes that are not extremely conserved over greater taxonomic distances (in contrast to the ribosomal RNA operon). Equally, if widespread horizontal transfer of an advantageous segment explains these intervals, the transfers occurred long enough ago for appreciable variation to accumulate (unpublished data).

Extreme Assembly of Uncultivated Populations

The analyses described above have been confined to those organisms with representatives in culture and for which genomes were readily available. Producing assemblies for other abundant but uncultivated microbial genera would provide valuable physiological and biochemical information that could eventually lead to the cultivation of these organisms, help elucidate their role in the marine community, and allow similar analyses of their evolution and variation such as those performed on sequenced organisms. Previous assembly efforts and the fragment recruitments plots showed that there is considerable and in many cases conflicting variation among related organisms. Such variation is known to disrupt whole-genome assemblers. This led us to try an assembly approach that aggressively resolves conflicts. We call this approach "extreme assembly" (see Materials and Methods). This approach currently does not make use of mate-pairing data and, therefore produces only contigs, not scaffolded sequences. Using this approach, contigs as large as 900 kb could be aligned almost in their entirety to the *P. marinus* MIT9312 and *P. ubique* HTCC1062 genomes (Figure 2J–2L). Consistent patterns of fragment recruitment (see below) generally provided evidence of the correctness of contigs belonging to otherwise-unsequenced organisms. Accordingly, large contigs from these alternate assemblies were used to investigate genetic and geographic population structure, as described below. However, the more aggressive assemblies demonstrably suffered from higher rates of assembly artifacts, including chimerism and false consensus sequences (Figure 6). Thus, the more stringent primary assembly was employed for most assembly-based analyses, as manual curation was not practical.

As just noted, many of the large contigs produced by the more aggressive assembly methods described above did not align to any great degree with known genomes. Some could be tentatively classified based on contained 16S sequences, but the potential for computationally generated chimerism within the rRNA operon is sufficiently high that inspection of the assembly or other means of confirming such classifications is essential. An alternative to an unguided assembly that facilitates the association of assemblies with known organisms is to start from seed fragments that can be identified as belonging to a particular taxonomic group. We employed fragments outside the ribosomal RNA operon that were mated to a 16S-containing read, limiting extension to the direction away from the 16S operon. This produced contigs of 100 kb or more for several of the ribotypes that were abundant in the GOS dataset. When evaluated via fragment recruitment (Figure 2M–2O), these assemblies revealed patterns analogous to those seen for the sequenced genomes described above: multiple subtypes could be distinguished along the assembly, differing in similarity to the reference sequence and sample distribution, with occasional gaps. Hypervariable segments by definition were not represented in these assemblies, but they may help explain the termination of the extreme assemblies for *P. marinus* and SAR11 and provide a plausible explanation for termination of assemblies of the other deeply sampled populations as well.

This directed approach to assembly can also be used to investigate variation within a group of related organisms (e.g., a 16S ribotype). We explored the potential to assemble

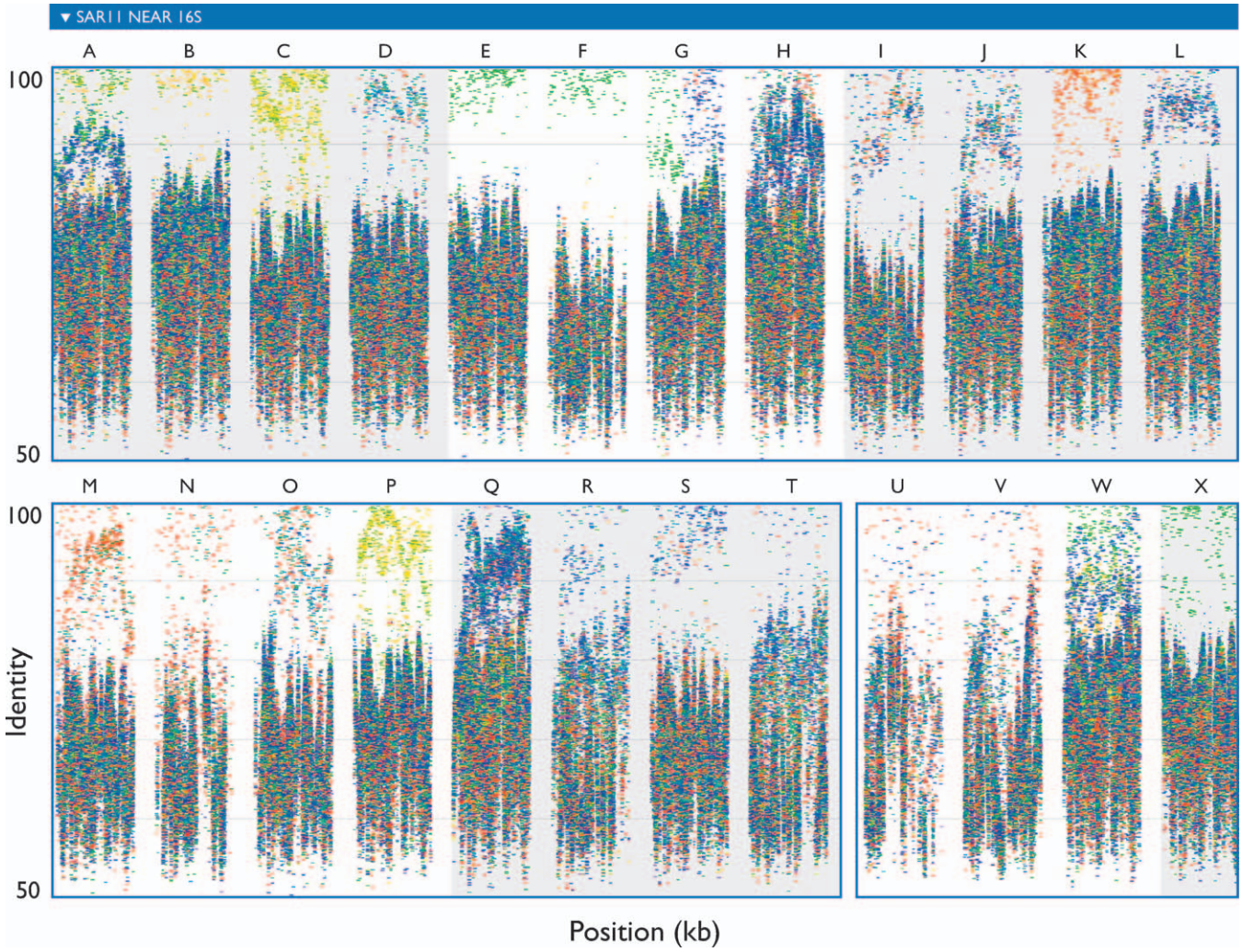


Figure 7. Fragment Recruitment Plots to 20-kb Segments of SAR11-Like Contigs Show That Many SAR11 Subtypes, with Distinct Distributions, Can Be Separated by Extreme Assembly

Each segment is constructed of a unique set of GOS sequencing reads (i.e., no read was used in more than one segment). Segments are arbitrarily labeled (A–X) for reference in Figure 8.

doi:10.1371/journal.pbio.0050077.g007

distinct subtypes of SAR11 by repeatedly seeding extreme assembly with fragments mated to a SAR11-like 16S sequence. Figure 7 compares the first 20 kb from each of 24 independent assemblies. Eighteen of these segments could be aligned full-length to a portion of the HTCC1062 genome just upstream of 16S, while six appeared to reflect rearrangements relative to HTCC1062. The rearranged segments were associated with more divergent 16S sequences (8%–14% diverged from the 16S of HTCC1062), while those without rearrangements corresponded to less divergent 16S (averaging less than 3% different from HTCC1062). In each segment, many reads were recruited above 90% identity, but different samples dominated different assemblies. Phylogenetic trees support the inference of evolutionarily distinct subtypes with distinctive sample distributions (Figure 8).

Taxonomic Diversity

Environmental surveys provide a cultivation-independent means to examine the diversity and complexity of an environmental sample and serve as a basis to compare the populations between different samples. Typically, these surveys use PCR to amplify ubiquitous but slowly evolving genes such as the 16S rRNA or *recA* genes. These in turn can be used to distinguish microbial populations. Since PCR can introduce various biases, we identified 16S genes directly from the primary GOS assembly. In total, 4,125 distinct full-length or partial 16S were identified. Clustering of these sequences at 97% identity gave a total of 811 distinct ribotypes. Nearly half (48%) of the GOS ribotypes and 88% of the GOS 16S sequences were assigned to ribotypes previously deposited in public databases. That is, more than half the ribotypes in the GOS dataset were found to be novel at what is typically considered the species level [30]. The overall taxonomic distribution of the GOS ribotypes sampled by shotgun sequencing is consistent with previously published PCR based studies of marine environments (Table 7) [31]. A smaller amount (16%) of GOS ribotypes and 3.4% of the GOS 16S sequences diverged by more than 10% from any publicly available 16S sequence, thus being novel to at least the family level.

A census of microbial ribotypes allows us to identify the abundant microbial lineages and estimate their contribution

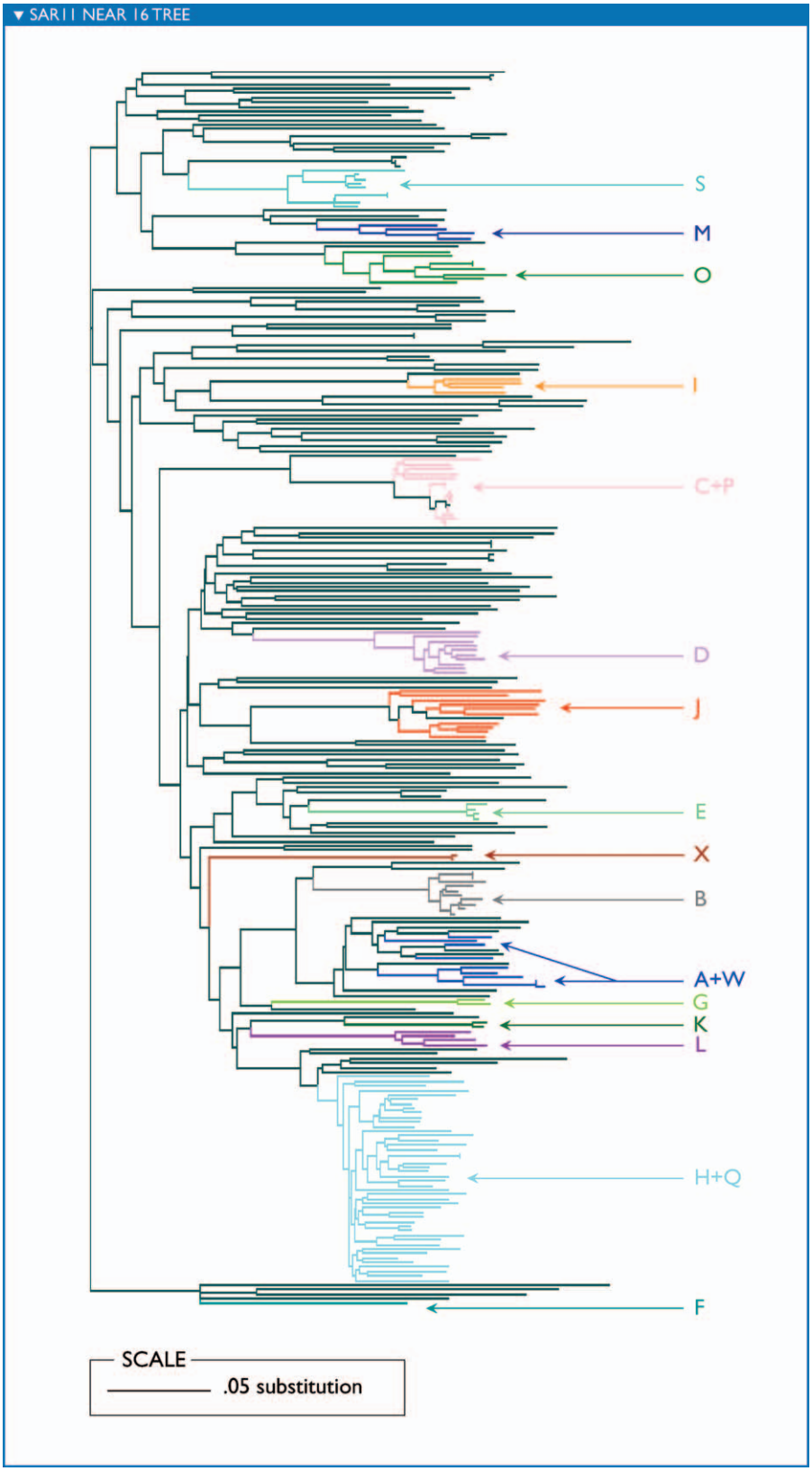


Figure 8. Phylogeny of GOS Reads Aligning to *P. ubique* HTCC1062 Upstream of 16S Gene Indicates That the Extreme Assemblies in Figure 7 Correspond to Monophyletic Subtypes

Coloring of branches indicates that the corresponding reads align at >90% identity to the extreme assembly segments shown in Figure 7; colored labels (A–X) correspond to the labels in Figure 7, indicating the segment or segments to which reads aligned.

doi:10.1371/journal.pbio.0050077.g008

Table 7. Taxonomic Makeup of GOS Samples Based on 16S Data from Shotgun Sequencing

Phylum or Class	Fraction[a]
Alpha *Proteobacteria*	0.32
Unclassified *Proteobacteria*	0.155
Gamma *Proteobacteria*	0.132
Bacteroidetes	0.13
Cyanobacteria	0.079
Firmicutes	0.075
Actinobacteria	0.046
Marine Group A	0.022
Beta *Proteobacteria*	0.017
OP11	0.008
Unclassified *Bacteria*	0.008
Delta *Proteobacteria*	0.005
Planctomycetes	0.002
Epsilon *Proteobacteria*	0.001

[a]Values shown are averages over all samples.
doi:10.1371/journal.pbio.0050077.t007

to the GOS dataset. Of the 811 ribotypes, 60 contain more than 8-fold coverage of the 16S gene (Table 8); jointly, these 60 ribotypes accounted for 73% of all the 16S sequence data. All but one of the 60 have been detected previously, yet only a few are represented by close relatives with complete or nearly complete genome sequencing projects (see Fragment Recruitment for further details). Several other abundant 16S sequences belong to well-known environmental ribotypes that do not have cultivated representatives (e.g., SAR86, *Roseobacter* NAC-1–2, and branches of SAR11 other than those containing *P. ubique*). Interestingly, archaea are nearly absent from the list of dominant organisms in these near-surface samples.

The distribution of these ribotypes reveals distinct microbial communities (Figure 9 and Table 8). Only a handful of the ribotypes appear to be ubiquitously abundant; these are dominated by relatives of SAR11 and SAR86. Many of the ribotypes that are dominant in one or more samples appear to reside in one of three separable marine surface habitats. For example, several SAR11, SAR86, and alpha *Proteobacteria*, as well as an *Acidimicrobidae* group, are widespread in the surface waters, while a second niche delineated by tropical samples contains several different SAR86, *Synechococcus* and *Prochlorococcus* (both cyanobacterial groups), and a *Rhodospirillaceae* group. Other ribotypes related to *Roseobacter* RCA, SAR11, and gamma *Proteobacteria* are abundant in the temperate samples but were not observed in the tropical or Sargasso samples. Not surprisingly, samples taken from nonmarine environments (GS33, GS20, GS32), estuaries (GS11, GS12), and larger-sized fraction filters (GS01a, GS01b, GS25) have distinguishing ribotypes. Furthermore, as the complete genomes of these dominant members are obtained, the capabilities responsible for their abundances may well lend insight into the community metabolism in various oceanic niches.

Sample Comparisons

The most common approach for comparing the microbial community composition across samples has been to examine the ribotypes present as indicated by 16S rRNA genes or by analyzing the less-conserved ITS located between the 16S and 23S gene sequences [7,8,16,17]. However, a clear observation emerging from the fragment recruitment views was that the reference ribotypes recruit multiple subtypes, and that these subtypes were distributed unequally among samples (Figures 2, 7, 8; Poster S1D, S1F, and S1I).

We developed a method to assess the genetic similarity between two samples that potentially makes use of all portions of a genome, not just the 16S rRNA region. This similarity measure is assembly independent; under certain circumstances, it is equivalent to an estimate of the fraction of sequence from one sample that could be considered to be in the other sample. Whole-metagenomic similarities were computed for all pairs of samples. Results are presented for comparisons at $\geq$98% and 90% identity. No universal cutoff consistently divides sequences into natural subsets, but the 98% identity cutoff provides a relatively high degree of resolution, while the 90% cutoff appears to be a reasonable heuristic for defining subtypes. For instance, a 90% cutoff treats most of the reads specifically recruited to *P. marinus* MIT9312 as similar (those more similar to MED4 notably excepted), while reasonably separating clades of SAR11 (Figures 7 and 8). Reads with no qualifying overlap alignment to any other read in a pair of samples are uninformative for this analysis, as they correspond to lineages that were so lightly sequenced that their presence in one sample and absence in another may be a matter of chance. For the 90% cutoff, 38% of the sequence reads contributed to the analysis. The resulting similarities reveal clear and consistent groupings of samples, as well as the outlier status of certain samples (Figures 10 and 11).

The broadest contrast was between samples that could be loosely labeled "tropical" (including samples from the Sargasso Sea [GS00b, GS00c, GS00d] and samples that are temperate by the formal definition but under the influence of the Gulf Stream [GS14, GS15]) and "temperate." Further subgroups can be identified within each of these categories, as indicated in Figures 10 and 11. In some cases, these groupings were composed of samples taken from different ocean basins during different legs of the expedition. A few pairs of samples with strikingly high similarity were observed, including GS17 and GS18, GS23 and GS26, GS27 and GS28, and GS00b and GS00d. In each case, these pairs of samples were collected from consecutive or nearly consecutive samples. However, the same could be said of many other pairs of samples that do not show this same degree of similarity. Indeed, geographically and temporally separated samples taken in the Atlantic (GS17, GS18) and Pacific (GS23, GS26) during separate legs of the expedition are more similar to one another than were most pairs of consecutive samples. The samples with least similarity to any other sample were from unique habitats. Thus, similarity cannot be attributed to geographic separation alone.

The groupings described above can be reconstructed from taxonomically distinct subsets of the data. Specifically, the major groups of samples visible in Figure 10 were reproduced when sample similarities were determined based only on fragments recruiting to *P. ubique* HTCC1062 (unpublished data). Likewise, the same groupings were observed when the fragments recruiting to either HTCC1062 or *P. marinus* MIT9312, or both, were excluded from the calculations (unpublished data). Thus, the factors influencing sample similarities do not appear to rely solely on the most abundant

Table 8. Most Abundant Ribotypes (97% Identity Clusters)

Ribotype Classification[a]	Depth of Coverage[b]	Range	Number of Matching GenBank Entries[c]
SAR11 Surface 1	581	Widespread	100+
SAR11 Surface 2	182	Sargasso and GS31[d]	100+
Burkholderia	139	00a	100+
Acidomicrobidae type a	133	Tropical and Sargasso[d]	77
Prochlorococcus	112	Tropical and Sargasso	76
SAR11 Surface 3	109	Widespread	63
SAR86-like type a	108	Widespread	88
Shewanella	80	00a	49
Synechococcus	59	Tropical[d]	100+
Rhodospirillaceae	50	Tropical and Sargasso[d]	15
SAR86-like type b	47	Hypersaline pond[d]	24
SAR86-like type c	47	Tropical and Sargasso	75
Chlorobi-like	40	Hypersaline pond	1
Alpha *Proteobacteria* type a	38	Widespread	10
Roseobacter type a	35	Tropical and Sargasso[d]	42
Cellulomonadaceae type a	34	Hypersaline pond	12
SAR86-like type d	28	Widespread	30
Alpha *Proteobacteria* type b	27	Widespread	43
SAR86-like type e	26	Widespread	71
Cytophaga type a	24	Tropical and Sargasso	21
Bacteroidetes type a	23	Widespread	17
Bdellovibrionales type a	21	Tropical and Sargasso	36
Acidomicrobidae type b	21	Temperate	41
SAR116-like	21	Widespread	28
Marine Group A type a	20	Tropical and Sargasso[d]	9
Remotely SAR11-like type a	19	Sargasso and tropical	12
Frankineae type a	19	Fresh and estuary	6
Frankineae type b	18	Hypersaline pond[d]	71
SAR86-like type f	18	Tropical	15
SAR86-like type g	17	Tropical	13
Remotely SAR11-like type b	16	Tropical and Sargasso	4
Gamma *Proteobacteria* type a	16	Sargasso and GS14	18
Microbacteriaceae	15	Hypersaline pond	1
SAR102/122-like	15	Tropical and Sargasso	15
SAR86-like type h	15	Tropical and Sargasso	27
Bacteroidetes type b	14	Tropical[d]	14
Remotely SAR11-like type c	14	Tropical and Sargasso	18
SAR86-like type i	14	Sargasso[d]	13
Rhodobium-like type a	14	Sargasso and GS31[d]	9
Marine Group A type b	13	Tropical and Sargasso[d]	28
Gamma *Proteobacteria* type b	12	S. temperate and GS31	22
Oceanospirillaceae	12	Mangrove	1
Gamma *Proteobacteria* type c	12	Widespread	14
SAR11-like type a	12	Temperate	17
SAR11-like type b	11	Fresh and estuary	11
SAR11-like type c	11	Sargasso and GS31[d]	12
SAR86-like type j	11	Tropical and Sargasso	1
Roseobacter Algicola	11	Widespread	20
Remotely SAR102/122-like	11	Hypersaline pond	0
Frankineae type c	10	Fresh and estuary	11
Rhodobium-like type b	10	Widespread	9
Roseobacter RCA	10	Temperate	39
Remotely SAR11-like type d	10	Tropical	32
Acidobacteria	9	Fresh	4
Remotely SAR11-like type e	9	Tropical and Sargasso	8
Frankineae type d	9	Estuary and fresh	89
SAR11-like type d	9	Sargasso and GS31	4
Methylophilus	9	Temperate[d]	41
Archaea C1 C1a	8	Sargasso and GS31	2
Cytophaga type b	8	Widespread	9

[a]Taxonomic classifications based on Hugenholtz ARB database. Labels indicate the most specific taxonomic assignment that could be confidently assigned to each ribotype. "Type a," "type b," etc., used to arbitrarily discriminate separate 97% ribotypes that would otherwise be given the same name.
[b]Note that the 16S rRNA gene can be multicopy.
[c]Matching GenBank entries required full-length matches at $\geq$ 98% identity.
[d]Less than 1× coverage outside described range.
doi:10.1371/journal.pbio.0050077.t008

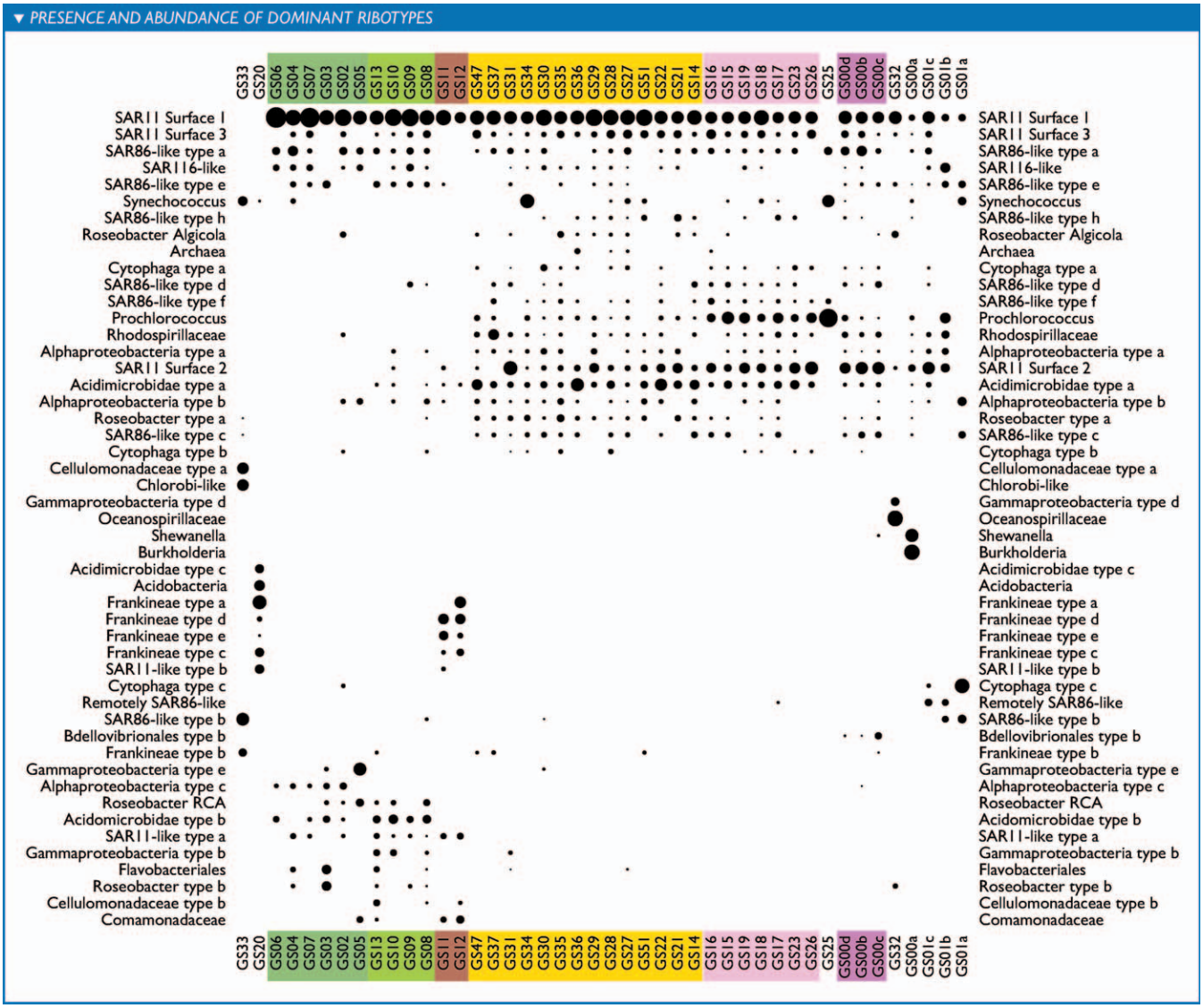


Figure 9. Presence and Abundance of Dominant Ribotypes

The relative abundance of various ribotypes (rows) in each filter (columns) is represented by the area of the corresponding spot (if any). The listed ribotypes each satisfied the following criteria in at least one filter: the ribotype was among the five most abundant ribotypes detected in the shotgun data, and was represented by at least three sequencing reads. Relative abundance is based on the total number of 16S sequences in a given filter. Order and grouping of filters is based on the clustering of genomic similarity shown in Figure 11. Ribotype order was determined based on similarity of sample distribution. A marked contrast between temperate and tropical groups is visible. Estuarine samples GS11 and GS12 contained a mix of ribotypes seen in freshwater and temperate marine samples, while samples from nonmarine habitats or larger filter sizes were pronounced outliers. The presence of large amounts of *Burkholderia* and *Shewanella* in one Sargasso Sea sample (GS00a) makes this sample look much less like other Sargasso and tropical marine samples than it otherwise would. Note that 16S is not a measure of cell abundance since 16S genes can be multicopy.

doi:10.1371/journal.pbio.0050077.g009

organisms but rather are reflected in multiple microbial lineages.

It is tempting to view the groups of similar samples as constituting community types. Sample similarities based on genomic sequences correlated significantly with differences in the environmental parameters (Table 1), particularly water temperature and salinity (unpublished data). Samples that are very similar to each other had relatively small differences in temperature and salinity. However, not all samples that had similar temperature and salinity had high community similarities. Water depth, primary productivity, fresh water input, proximity to land, and filter size appeared consistent with the observed groupings. Other factors such as nutrients and light for phototrophs and fixed carbon/energy for chemotrophs may ultimately prove better predictors, but these results demonstrate the potential of using metagenomic data to tease out such relationships.

Examining the groupings in Figure 11 in light of habitat and physical characteristics, the following may be observed. The first two samples, a hypersaline pond in the Galapagos Islands (GS33) and the freshwater Lake Gatun in the Panama Canal (GS20) are quite distinct from the rest. Salinity—both higher and lower than the remaining coastal and ocean samples—is the simplest explanation.

Twelve samples form a strong temperate cluster as seen in the similarity matrix of Figure 11 as a darker square bounded by GS06 and GS12. Embedded within the temperate cluster are three subclusters. The first subcluster includes five samples from Nova Scotia through the Gulf of Maine. This is followed by a subcluster of four samples between Rhode Island and North Carolina. The northern subcluster was sampled in August, the southern subcluster in November and

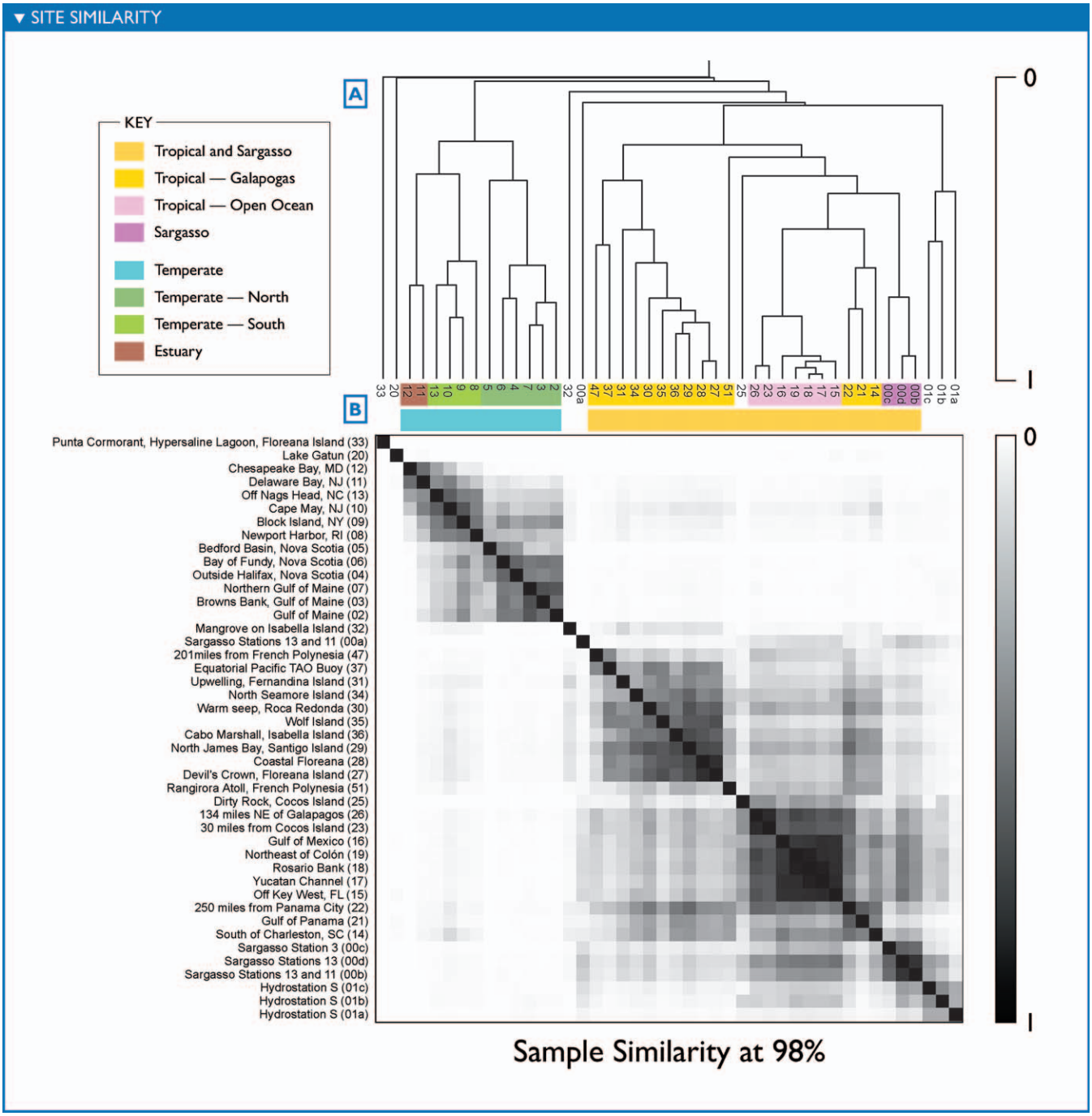


Figure 10. Similarity between Samples in Terms of Shared Genomic Content

Genomic similarity, as described in the text, is an estimate of the amount of the genetic material in two filters that is "the same" at a given percent identity cutoff—not the amount of sequence in common in a finite dataset, but rather in the total set of organisms present on each filter. Similarities are shown for 98% identity.

(A) Hierarchical clustering of samples based on pairwise similarities.

(B) Pairwise similarities between samples, represented as a symmetric matrix of grayscale intensities; a darker cell in the matrix indicates greater similarity between the samples corresponding to the row and column, with row and column ordering as in (A). Groupings of similar filters appear as subtrees in (A) and as squares consisting of two or more adjacent rows and columns with darker shading. Colored bars highlight groups of samples described in the text; labels are approximate characterizations rather than being strictly true of every sample in a group.

doi:10.1371/journal.pbio.0050077.g010

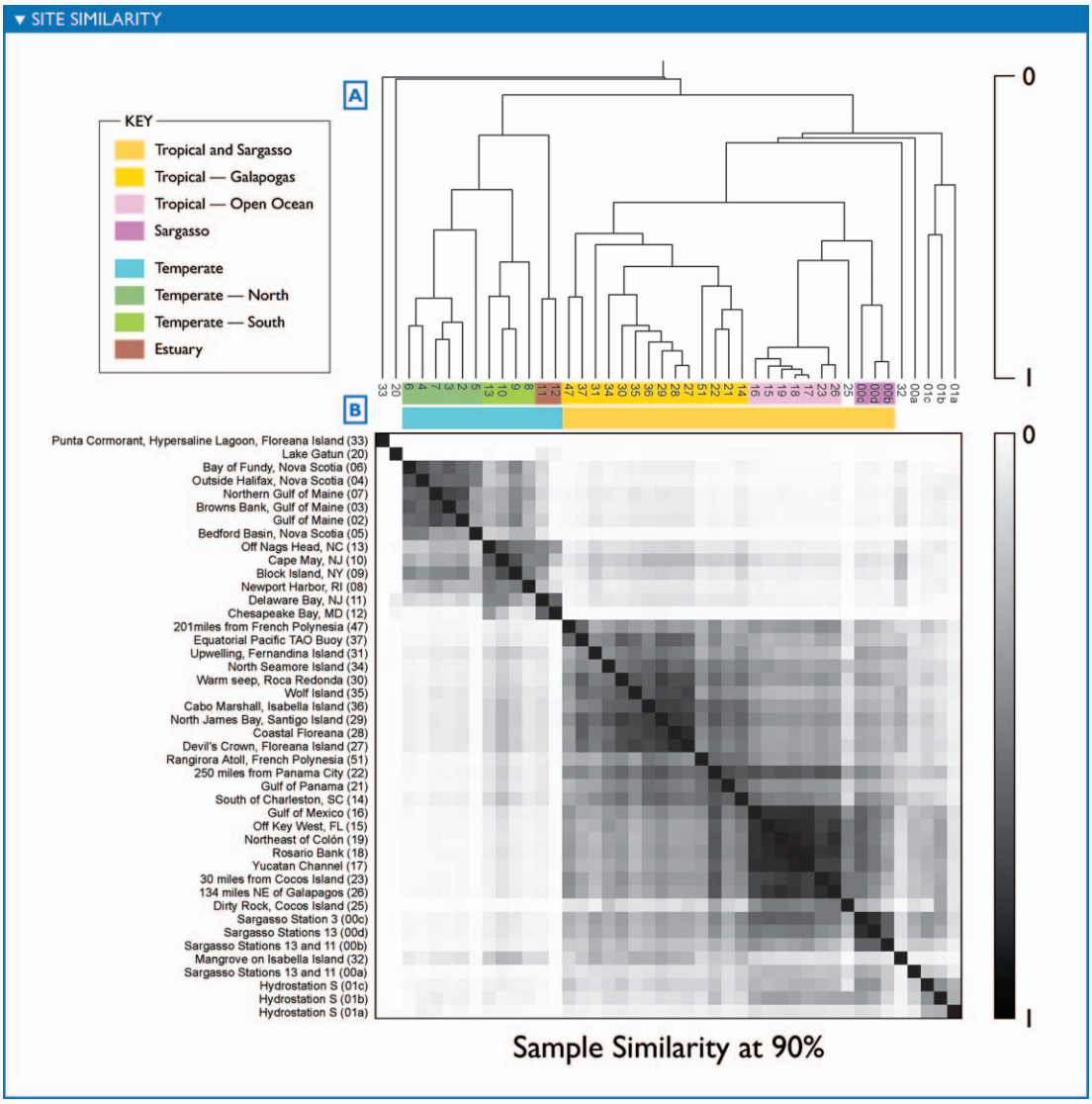


Figure 11. Sample Similarity at 90% Identity

Similarity between samples in terms of shared genomic content similar to Figure 10, except that the plots were done using a 90% identity cutoff that has proven reasonable for separating some moderately diverged subtypes

doi:10.1371/journal.pbio.0050077.g011

December. Though all samples were collected in the top few meters, the southern samples were in shallower waters, 10 to 30 m deep, whereas most of the northern samples were in waters greater than 100 m deep. Monthly average estimates of chlorophyll *a* concentrations were typically higher in the southern samples as well (Table 1). All of these factors—temperature, system primary production, and depth of the sampled water body—likely contribute to the differences in microbial community composition that result in the two well-defined clusters. The final temperate subgroup includes two estuaries, Chesapeake Bay (GS12) and Delaware Bay (GS11), distinguished by their lower salinity and higher productivity. However, GS11 is markedly similar not only to GS12 but also to coastal samples, whereas the latter appears much more unique. Interestingly, the Bay of Fundy estuary sample (GS06) clearly did not group with the two other estuaries, but rather with the northern subgroup, perhaps reflecting differences in the rate or degree of mixing at the sampling site.

Continuing to the right and downward in Figure 11, one can see a large cluster of 25 samples from the tropics and Sargasso Sea, bounded by GS47 and GS00b. This can be further subdivided into several subclusters. The first subcluster (a square bounded by GS47 and GS14) includes 14 samples, about half of which were from the Galapagos. The second distinct subcluster (a square bounded by GS16 and GS26) includes seven samples from Key West, Florida, in the Atlantic Ocean to a sample close to the Galapagos Islands in the Pacific Ocean. Loosely associated with this subcluster is a sample from a larger filter size taken en route to the Galapagos (GS25). The remaining samples group weakly with the tropical cluster. GS32 was taken in a coastal mangrove in the Galapagos. The thick organic sediment at a depth of less than a meter is the likely cause for it being unlike the other samples. Sample 00a was from the Sargasso Sea and contained a large fraction of sequence reads from apparently clonal *Burkholderia* and *Shewanella* species that are atypical. When this

Table 9. Relative Abundance of TIGRFAMs Associated with a Specific Sample

TIGRFAM	Number of Peptides	Sample[a]	Relative Abundance[b]	Major Category	Minor Category	Description
TIGR01526	131	GS01a	3.4			Nicotinamide-nucleotide adenylyltransferase
TIGR00661	214	GS01a	2.6	Hypothetical	Conserved	Conserved hypothetical protein
TIGR01833	135	GS01a	2			Hydroxymethylglutaryl-CoA synthase
TIGR01408	267	GS01b	4.5			Ubiquitin-activating enzyme E1
TIGR01678	144	GS01b	2.6			Sugar 1,4-lactone oxidases
TIGR01879	758	GS01b	2.5			Amidase, hydantoinase/carbamoylase family
TIGR00890	131	GS01b	2.4	Transport	Carbohydrates,	Oxalate/Formate Antiporter
TIGR01767	112	GS01b	2.4			5-methylthioribose kinase
TIGR00101	495	GS01b	2.3	Central	Nitrogen	Urease accessory protein UreG
TIGR01659	186	GS01b	2.2			Sex-lethal family splicing factor
TIGR00313	455	GS01b	2.1	Biosynthesis	Heme,	Cobyric acid synthase CobQ
TIGR00317	306	GS01b	2.1	Biosynthesis	Heme,	Cobalamin 5′-phosphate synthase
TIGR00601	186	GS01b	2.1	DNA	DNA	UV excision repair protein Rad23
TIGR01792	904	GS01b	2.1	Central	Nitrogen	Urease, alpha subunit
TIGR00749	483	GS01b	2	Energy	Glycolysis/gluconeogenesis	Glucokinase
TIGR01001	485	GS01b	2	Amino	Aspartate	Homoserine O-succinyltransferase
TIGR02238	165	GS32	5.1			Meiotic recombinase Dmc1
TIGR02239	159	GS32	3.5			DNA repair protein RAD51
TIGR02232	140	GS32	2.6			*Myxococcus* cysteine-rich repeat
TIGR02153	212	GS32	2.5	Protein	tRNA	Glutamyl-tRNA(Gln) amidotransferase, subunit D
TIGR00519	289	GS32	2.4			L-asparaginases, type I
TIGR00288	248	GS32	2	Hypothetical	Conserved	Conserved hypothetical protein TIGR00288
TIGR01681	136	GS32	2			HAD-superfamily phosphatase, subfamily IIIC
TIGR02236	143	GS32	2	DNA	DNA	DNA repair and recombination protein RadA
TIGR00028	110	GS33	14.7			*Mycobacterium tuberculosis* PIN domain family
TIGR01550	131	GS33	9.1	Unknown	General	Death-on-curing family protein
TIGR01552	200	GS33	5.3	Mobile	Other	Prevent-host-death family protein
TIGR00143	151	GS33	4.6	Protein	Protein	[NiFe] hydrogenase maturation protein HypF
TIGR01641	131	GS33	4.3	Mobile	Prophage	Phage putative head morphogenesis protein, SPP1 gp7 family
TIGR01710	1,926	GS33	3.9	Cellular	Pathogenesis	General secretion pathway protein G
TIGR01539	251	GS33	3.6	Mobile	Prophage	Phage portal protein, lambda family
TIGR00016	217	GS33	3.5	Energy	Fermentation	Acetate kinase
TIGR00942	117	GS33	3.5	Transport	Cations	Multicomponent Na+:H+ antiporter
TIGR01836	174	GS33	3.3	Fatty	Biosynthesis	Poly(R)-hydroxyalkanoic acid synthase, class III, PhaC subunit
TIGR02110	135	GS33	3.2	Biosynthesis	Other	Coenzyme PQQ biosynthesis protein PqqF
TIGR01106	902	GS33	3.1	Energy	ATP-proton	Na,H/K antiporter P-type ATPase, alpha subunit
TIGR01497	726	GS33	3.1	Transport	Cations	K+-transporting ATPase, B subunit
TIGR01524	603	GS33	3.1	Transport	Cations	Magnesium-translocating P-type ATPase
TIGR02140	166	GS33	3.1	Transport	Anions	Sulfate ABC transporter, permease protein CysW
TIGR01409	122	GS33	3			Tat (twin-arginine translocation) pathway signal sequence
TIGR02195	282	GS33	3	Cell	Biosynthesis	Lipopolysaccharide heptosyltransferase II
TIGR01522	696	GS33	2.9			Calcium-transporting P-type ATPase, PMR1-type
TIGR01523	766	GS33	2.9			Potassium/sodium efflux P-type ATPase, fungal-type
TIGR00202	228	GS33	2.8	Regulatory	RNA	Carbon storage regulator
TIGR01116	798	GS33	2.8	Transport	Cations	Calcium-translocating P-type ATPase, SERCA-type
TIGR02094	204	GS33	2.8			Alpha-glucan phosphorylases
TIGR02051	296	GS33	2.7	Regulatory	DNA	Hg(II)-responsive transcriptional regulator
TIGR01005	259	GS33	2.6	Transport	Carbohydrates,	Exopolysaccharide transport protein family
TIGR01222	129	GS33	2.6	Cellular	Cell	Septum site-determining protein MinC
TIGR01334	157	GS33	2.6	Unknown	General	modD protein
TIGR01554	152	GS33	2.6	Mobile	Prophage	Phage major capsid protein, HK97 family
TIGR00640	1,059	GS33	2.5			Methylmalonyl-CoA mutase C-terminal domain
TIGR01708	801	GS33	2.5	Cellular	Pathogenesis	General secretion pathway protein H
TIGR02018	373	GS33	2.5	Regulatory	DNA	Histidine utilization repressor
TIGR00554	182	GS33	2.4	Biosynthesis	Pantothenate	Pantothenate kinase
TIGR01202	152	GS33	2.4	Biosynthesis	Chlorophyll	Chlorophyll synthesis pathway, bchC
TIGR01583	102	GS33	2.4	Energy	Electron	Formate dehydrogenase, gamma subunit
TIGR02092	241	GS33	2.4	Energy	Biosynthesis	Glucose-1-phosphate adenylyltransferase, GlgD subunit
TIGR00052	263	GS33	2.3	Hypothetical	Conserved	Conserved hypothetical protein TIGR00052
TIGR00824	166	GS33	2.3	Signal	PTS	PTS system, mannose/fructose/sorbose family, IIA component
TIGR01003	321	GS33	2.3	Transport	Carbohydrates,	Phosphocarrier, HPr family
TIGR01439	191	GS33	2.3	Regulatory	DNA	Transcriptional regulator, AbrB family
TIGR01457	220	GS33	2.3			HAD-superfamily subfamily IIA hydrolase, TIGR01457
TIGR01517	684	GS33	2.3			Calcium-translocating P-type ATPase, PMCA-type
TIGR02028	632	GS33	2.3	Biosynthesis	Chlorophyll	Geranylgeranyl reductase
TIGR00452	217	GS33	2.2	Unknown	Enzymes	Methyltransferase, putative
TIGR00609	2,555	GS33	2.2	DNA	DNA	Exodeoxyribonuclease V, beta subunit
TIGR00876	265	GS33	2.2	Energy	Pentose	Transaldolase

Table 9. Continued.

TIGRFAM	Number of Peptides	Sample[a]	Relative Abundance[b]	Major Category	Minor Category	Description
TIGR00996	302	GS33	2.2	Cellular	Pathogenesis	Virulence factor Mce family protein
TIGR01254	419	GS33	2.2	Transport	Other	ABC transporter periplasmic binding protein, thiB subfamily
TIGR01278	508	GS33	2.2	Biosynthesis	Chlorophyll	Light-independent protochlorophyllide reductase, B subunit
TIGR01512	1,820	GS33	2.2	Transport	Cations	Cadmium-translocating P-type ATPase
TIGR01525	2,037	GS33	2.2			Heavy metal translocating P-type ATPase
TIGR01543	418	GS33	2.2	Protein	Other	Phage prohead protease, HK97 family
TIGR01857	1,495	GS33	2.2	Purines	Purine	Phosphoribosylformylglycinamidine synthase
TIGR02015	183	GS33	2.2	Energy	Photosynthesis	Chlorophyllide reductase subunit Y
TIGR02072	1,052	GS33	2.2	Biosynthesis	Biotin	Biotin biosynthesis protein BioC
TIGR02099	273	GS33	2.2	Hypothetical	Conserved	Conserved hypothetical protein TIGR02099
TIGR00203	168	GS33	2.1	Energy	Electron	Cytochrome d ubiquinol oxidase, subunit II
TIGR00218	141	GS33	2.1	Energy	Sugars	Mannose-6-phosphate isomerase, class I
TIGR00915	4,834	GS33	2.1	Transport	Other	Transporter, hydrophobe/amphiphile efflux-1 (HAE1) family
TIGR01315	319	GS33	2.1			FGGY-family pentulose kinase
TIGR01330	253	GS33	2.1			3′(2′),5′-bisphosphate nucleotidase
TIGR01508	848	GS33	2.1			Diaminohydroxyphosphoribosylaminopyrimidine reductase
TIGR01511	1,928	GS33	2.1	Transport	Cations	Copper-translocating P-type ATPase
TIGR01764	387	GS33	2.1	Unknown	General	DNA binding domain, excisionase family
TIGR02014	298	GS33	2.1	Energy	Photosynthesis	Chlorophyllide reductase subunit Z
TIGR02047	222	GS33	2.1			Cd(II)/Pb(II)-responsive transcriptional regulator
TIGR00586	990	GS33	2	DNA	DNA	Mutator mutT protein
TIGR00853	114	GS33	2	Signal	PTS	PTS system, lactose/cellobiose family IIB component
TIGR00937	364	GS33	2	Transport	Anions	Chromate transporter, chromate ion transporter (CHR) family
TIGR01030	221	GS33	2	Protein	Ribosomal	Ribosomal protein L34
TIGR01214	2,834	GS33	2	Cell	Biosynthesis	dTDP-4-dehydrorhamnose reductase
TIGR01698	608	GS33	2			Purine nucleotide phosphorylase
TIGR02190	465	GS33	2			Glutaredoxin-family domain

[a]Reads associated with *Shewanella* and *Burkholderia* have been excluded.
[b]TIGRFAM is this many times more abundant than in the next most abundant sample.
doi:10.1371/journal.pbio.0050077.t009

sample is reanalyzed to exclude reads identified as belonging to these two groups, sample GS00a groups loosely with GS00b, GS00c, and GS00d (unpublished data). Finally, three subsamples from a single Sargasso sample (GS01a, GS01b, GS01c) group together, despite representing three distinct size fractions (3.0–20, 0.8–3.0, and 0.1–0.8 μm, respectively; Table 1).

The complete set of sample similarities is more complex than described above, and indeed is more complex than can be captured by a hierarchical clustering. For instance, the southern temperate samples are appreciably more similar to the tropical cluster than are the northern temperate samples. GS22 appears to constitute a mix of tropical types, showing strong similarity not only to the GS47–GS14 subcluster to which it was assigned, but also to the other tropical samples.

These results may be compared to the more traditional view of community structure afforded by 16S sequences (Figure 9). Some of the same groupings of samples are visible using both analyses. Several ribotypes recapitulated the temperate/tropical clustering described above. Others were restricted to the single instances of nonmarine habitats. Several of the most abundant organisms from the coastal mangrove, hypersaline lagoon, and freshwater lake were found exclusively in these respective samples. However, while several ribotypes recapitulated the temperate/tropical distinction revealed by the genomic sequence, others crosscut it. A few dominant 16S ribotypes, related to SAR11, SAR86, and SAR116, were found in every marine sample. The brackish waters from two mid-Atlantic estuaries (GS11 and GS12) contained a mixture of otherwise exclusively marine and freshwater ribotypes; similarity of these sites to the freshwater sample (GS20) was minimal at the metagenomic level, while the greater similarity of GS11 to coastal samples visible at the metagenomic level was not readily visible here. A fuller comparison of metagenome-based measurements of diversity based on a large dataset of PCR-derived 16S sequences will be presented in another paper (in preparation).

Variation in Gene Abundance

Differences in gene content between samples can identify functions that reflect the lifestyles of the community in the context of its local environment [20,32]. We examined the relative abundance of genes belonging to specific functional categories in the distinct GOS samples. Genes were binned into functional categories using TIGRFAM hidden Markov models [18], which are well annotated and manually curated [33].

The results can be filtered in various ways to highlight genes associated with specific environments. One catalog of possible interest is genes that were predominantly found in a single sample. We identified 95 TIGRFAMs that annotated large sets of genes (100 or more) that were significantly more frequent (greater than 2-fold) in one sample than in any other sample (Table 9). Not surprisingly, this approach disproportionately singles out genes from the samples collected on larger filters (GS01a, GS01b, and GS25) and from the nonmarine environments, particularly the hypersaline pond (sample GS33). Another contrast might be between the

Table 10. Relative Abundance of TIGRFAM Matches in Temperate and Tropical Waters

TIGRFAM	Number of Peptides	Sample(s)	Relative Abundance[a]	Major Category	Minor Category	Description
TIGR01153	729	GS15–GS19	32.7	Energy	Photosynthesis	Photosystem II 44 kDa subunit reaction center protein
TIGR02093	673	GS15–GS19	29.6	Energy	Biosynthesis	Glycogen/starch/alpha-glucan phosphorylases
TIGR01335	813	GS15–GS19	26.6	Energy	Photosynthesis	Photosystem I core protein PsaA
TIGR01336	806	GS15–GS19	26.5	Energy	Photosynthesis	Photosystem I core protein PsaB
TIGR00975	648	GS15–GS19	11	Transport	Anions	Phosphate ABC transporter, phosphate-binding protein
TIGR00297	261	GS15–GS19	8.6	Hypothetical	Conserved	Conserved hypothetical protein TIGR00297
TIGR00992	302	GS15–GS19	8.5	Transport	Amino	Chloroplast envelope protein translocase, IAP75 family
TIGR02030	560	GS15–GS19	7.5	And	Chlorophyll	Magnesium chelatase ATPase subunit I
TIGR02041	359	GS15–GS19	6	Central	Sulfur	Sulfite reductase (NADPH) hemoprotein, beta-component
TIGR01151	2,095	GS15–GS19	4.7	Energy	Photosynthesis	Photosystem q(b) protein
TIGR01152	1,865	GS15–GS19	4.7	Energy	Photosynthesis	Photosystem II D2 protein (photosystem q(a) protein)
TIGR02031	800	GS15–GS19	4.2	Biosynthesis	Chlorophyll	Magnesium chelatase ATPase subunit D
TIGR01790	629	GS15–GS19	4			Lycopene cyclase family protein
TIGR02100	512	GS15–GS19	4	Energy	Biosynthesis	Glycogen debranching enzyme GlgX
TIGR00073	284	GS15–GS19	3.4	Protein	Protein	Hydrogenase accessory protein HypB
TIGR00159	497	GS15–GS19	3	Hypothetical	Conserved	Conserved hypothetical protein TIGR00159
TIGR01515	594	GS15–GS19	3	Energy	Biosynthesis	1,4-alpha-glucan branching enzyme
TIGR00217	601	GS15–GS19	2.7	Energy	Biosynthesis	4-alpha-glucanotransferase
TIGR01486	505	GS15–GS19	2.7			Mannosyl-3-phosphoglycerate phosphatase family
TIGR01098	720	GS15–GS19	2.6	Transport	Carbohydrates	Phosphonate ABC transporter, periplasmic phosphonate-binding protein
TIGR00101	495	GS15–GS19	2.5	Central	Nitrogen	Urease accessory protein UreG
TIGR01273	567	GS15–GS19	2.4	Central	Polyamine	Arginine decarboxylase
TIGR01470	179	GS5–GS10	25.7	Biosynthesis	Heme	Siroheme synthase, N-terminal domain
TIGR00361	374	GS5–GS10	12.1	Cellular	DNA	DNA internalization-related competence protein ComEC/Rec2
TIGR01537	333	GS5–GS10	6.2	Mobile	Prophage	Phage portal protein, HK97 family
TIGR00201	291	GS5–GS10	6	Cellular	DNA	comF family protein
TIGR00879	420	GS5–GS10	5	Transport	Carbohydrates	Sugar transporter
TIGR02018	373	GS5–GS10	4.1	Regulatory	DNA	Histidine utilization repressor
TIGR02183	294	GS5–GS10	4.1			Glutaredoxin, GrxA family
TIGR00427	602	GS5–GS10	4	Hypothetical	Conserved	Conserved hypothetical protein TIGR00427
TIGR01109	219	GS5–GS10	3.6	Energy	Other	Sodium ion-translocating decarboxylase, beta subunit
TIGR01262	840	GS5–GS10	2.8	Energy	Amino	Maleylacetoacetate isomerase

[a]Average abundance of TIGRFAM is that many times more abundant the average abundance in the given samples than in the other set of samples (in this case, GS15–GS19 were compared with GS5–GS10).

doi:10.1371/journal.pbio.0050077.t010

temperate and tropical clusters (Figures 10 and 11). We identified 32 proteins that were more than 2-fold more frequent in one or the other group (Table 10). The presence of various *Prochlorococcus*-associated genes in this list highlights some of the potential challenges with this sort of approach. Overrepresentation may reflect: a direct response to particular environmental pressures (as the excess of salt transporters plausibly do in the hypersaline pond); a lineage-restricted difference in functional repertoire (as exemplified by the excess of photosynthesis genes in samples containing *Prochlorococcus*); or a more incidental "hitchhiking" of a protein found in a single organism that happens to be present.

We explored whether clearer and more informative differences could be discovered between communities by focusing on groups of samples that are highly similar in overall taxonomic/genetic content. Two pairs of samples provide a particularly nice illustration of this approach. Samples GS17 and GS18 from the western Caribbean Sea and samples GS23 and GS26 from the eastern Pacific Ocean were all very similar based on the presence of abundant ribotypes and overall similarity in genetic content (Figures 9–11). Despite these similarities, several genes are found to be up to seven times more common in the pair of Caribbean samples than the Pacific pair (Table 11). No genes are more than 2-fold higher in the Pacific than the Caribbean pair of samples. Several of the most differentially abundant genes are related to phosphate transport and utilization. It is very plausible that this is a reflection of a functional adaptation: these differences correlate well with measured differences in phosphate abundance between the Atlantic and eastern Pacific samples [34,35], and phosphate abundance plays a critical role in microbial growth [36,37]. Indeed, the ability to acquire phosphate, especially under conditions where it is limited, is thought to determine the relative fitness of *Prochlorococcus* strains [38].

The single greatest difference between GS17 and GS18 on the one hand and GS23 and GS26 on the other was attributed to a set of genes annotated by the hidden Markov model TIGR02136 as a phosphate-binding protein *(PstS)*. This TIGRFAM identified a single gene in both *P. marinus* MIT9312 and *P. ubique* HTCC1062. In *P. marinus* MIT9312, this gene is located at 672 kb lying roughly in the middle of a 15-kb segment of the genome that recruits almost no GOS sequences from the Pacific sampling sites (Poster S1H). In *P. ubique* HTCC1062, the *PstS* gene is found at 1,133 kb in a 5-kb segment that also recruited far fewer GOS sequences from all the Pacific samples except for GS51 (Poster S1E). These genomic segments differ structurally among isolates but they are no more variable than the flanking regions, and thus are

Table 11. Relative Abundance of TIGRFAM Matches in Atlantic and Pacific Open Ocean Waters

TIGRFAM	Number of Peptides	Sample(s)	Relative Abundance[a]	Major Category	Minor Category	Description
TIGR02136	1,130	GS17, GS18	7.2	Transport	Anions	Phosphate-binding protein
TIGR00974	2,122	GS17, GS18	3.5	Transport	Anions	Phosphate ABC transporter, permease protein PstA
TIGR00975	648	GS17, GS18	3.5	Transport	Anions	Phosphate ABC transporter, phosphate-binding protein
TIGR02138	2,139	GS17, GS18	3.4	Transport	Anions	Phosphate ABC transporter, permease protein PstC
TIGR00206	459	GS17, GS18	2.8	Cellular	Chemotaxis	Flagellar M-ring protein FliF
TIGR01782	1,297	GS17, GS18	2.4	Transport	Unknown	TonB-dependent receptor
TIGR00642	862	GS17, GS18	2.3	Central	Other	Methylmalonyl-CoA mutase, small subunit
TIGR02135	899	GS17, GS18	2.3	Transport	Anions	Phosphate transport system regulatory protein PhoU

[a]Relative Abundance: average abundance of TIGRFAM is at least that many times more abundant the average abundance in the given samples than in the other set of samples (in this case, GS17–GS18 were compared to GS23 and GS26).
doi:10.1371/journal.pbio.0050077.t011

not hypervariable in the sense used previously (unpublished data). Nor are they particularly conserved when present, indicating that they are not the result of a recent lateral transfer. Phylogenetic analyses outside these segments did not produce any evidence of a Pacific versus Caribbean clade of either *Prochlorococcus* or SAR11 (Figure 3A–3B). The presence or absence of phosphate transporters is not limited to these two types of organisms. The number of phosphate transporters that were found in the Caribbean far exceeds the number that can be attributed to HTCC1062- and MIT9312-like organisms. However, these results indicate that within individual strains or subtypes the ability to acquire phosphate (in one or more of its forms) can vary without detectable differences in the surrounding genomic sequences.

Biogeographic Distribution of Proteorhodopsin Variants

Variation in gene content is only one aspect of the tremendous diversity in the GOS data. The functional significance of all the polymorphic differences between homologous proteins remains largely unknown. To look for functional differences, we analyzed members of proteorhodopsin gene family. Proteorhodopsins are fast, light-driven proton pumps for which considerable functional information is available though their biological role remains unknown. Proteorhodopsins were highly abundant in the Sargasso Sea samples [19] and continue to be highly abundant and evenly distributed (relative to *recA* abundance) in all the GOS samples. A total of 2,674 putative proteorhodopsin genes were identified in the GOS dataset. Although many of the sequences are fragmentary, 1,874 of these genes contain the residue that is primarily responsible for tuning the light-absorbing properties of the protein [39–41], and these properties have been shown to be selected for under different environmental conditions [42]. Variation at this residue is strongly correlated with sample of origin (Figure 12). The leucine (L) or green-tuned variant was highly abundant in the North Atlantic samples and in the nonmarine environments like the fresh water sample from Lake Gatun (GS20). The glutamine (Q) or blue-tuned variant dominated in the remaining mostly open ocean samples.

Given our limited understanding of the biological role for proteorhopsin, the reason for this differential distribution is not immediately clear. In coastal waters where nutrients are more abundant, phytoplankton is dominant. Phytoplankton absorbs primarily in the blue and red spectra; consequently, the water appears green [43]. Conversely, in the open ocean nutrients are rare and phytoplanktonic biomass is low, so waters appear blue because in the absence of impurities the red wavelengths are absorbed preferentially [44]. It may be that proteorhodopsin-carrying microbes have simply adapted to take advantage of the most abundant wavelengths of light in these systems.

Proteorhodopsins encoded on reads that were recruited to *P. ubique* HTCC1062 account for a fraction (~25%) of all the proteorhodopsin-associated reads, suggesting that the remainder must be associated with a variety of marine microbial taxa (see also [45–47]). Phylogenetic analysis of the SAR11-associated proteins revealed that each variant has arisen independently at least two times in the SAR11 lineage (Figure 3C). Consistent with other findings that proteorhodopsins are widely distributed throughout the microbial world [48], we conclude that multiple microbial lineages are responsible for proteorhodopsin spectral variation and that the abundance of a given variant reflects selective pressures rather than taxonomic effects. Similar mechanisms seem to be involved in the evolution and diversification of opsins that mediate color vision in vertebrates [49].

Discussion

Our results highlight the astounding diversity contained within microbial communities, as revealed through whole-genome shotgun sequencing carried out on a global scale. Much of this microbial diversity is organized around phylogenetically related, geographically dispersed populations we refer to as subtypes. In addition, there is tremendous variation within subtypes, both in the form of sequence variation and in hypervariable genomic islands. Our ability to make these observations derived from not only the large volumes of data but also from the development of new tools and techniques to filter and organize the information in manageable ways.

Variation and Diversity

Our data demonstrate to an unprecedented degree the nature and evolution of genetic variation below the species level. Variation can be analyzed in several ways, including observed differences in sequence, genomic structure, and

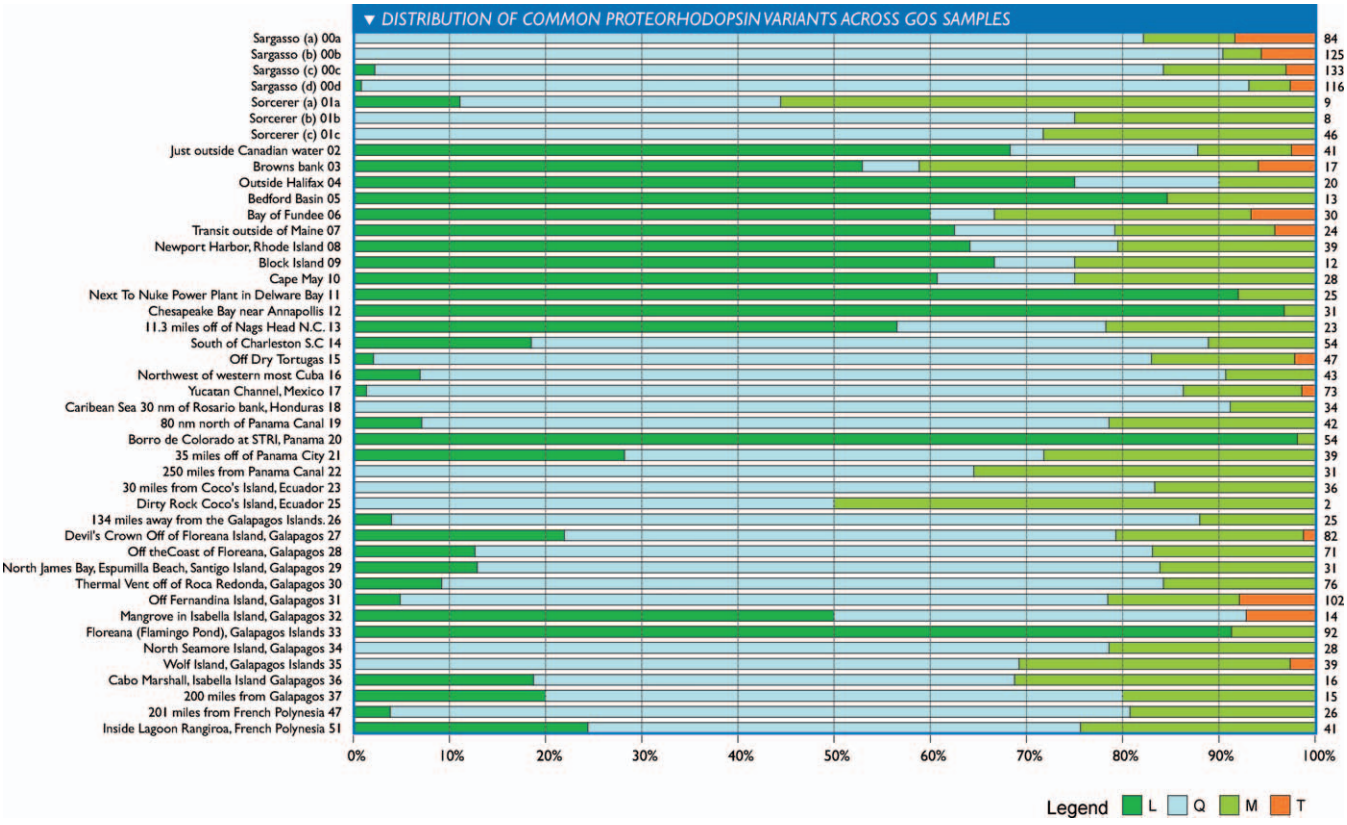


Figure 12. Distribution of Common Proteorhodopsin Variants across GOS Samples

The leucine (L) and methionine (M) variants absorb maximally in the green spectrum (Oded Beja, personal communication) while the glutamine (Q) variant absorbs maximally in the blue spectrum. The relative abundance of each variant is shown as a percentage (*x*-axis) per sample (*y*-axis). Total abundance for all variants in read equivalents normalized by the abundance of recA protein are shown on the right side of the *y*-axis. The L and Q variants show a nonrandom distribution. The L variant is abundant in temperate Atlantic waters close to the U.S. and Canadian coast. The Q variant is abundant in warmer waters further from land. The M variant is moderately abundant in a wide range of samples with no obvious geographic/environmental association.

doi:10.1371/journal.pbio.0050077.g012

gene complement. The observed patterns of variation shed light on the mechanisms by which marine prokaryotes evolve. Gene synteny seems to be more highly conserved than the nucleotide and protein sequences. This variation is seen over essentially the entire genome in every abundant group of organisms sufficiently related for us to recognize a population by fragment recruitment. (These include, but are not limited to, the organisms shown in Figure 2 and Poster S1.) Notably, we found no evidence of widespread low-diversity organisms such as *B. anthracis* [50].

Phylogenetic trees and fragment recruitment plots (Figures 7 and 8) indicate that the variation within a species is not an unstructured swarm or cloud of variants all equally diverged from one another. Instead, there are clearly distinct subtypes, in terms of sequence similarity, gene content, and sample distribution. Similar findings have been shown for specific organisms, based on evaluation of one or a few loci [2,51–53]. These results rule out certain trivial models of population history and evolution for what is commonly considered a bacterial "species." For instance, it argues against a recent explosive population growth from a single successful individual (selective sweep) [54]. Equally, it argues against a perfectly mixed population, suggesting instead some barriers to competition and exchange of genetic material.

In principle, this variation could reflect some combination of physical barriers (true biogeography), short-term stochastic effects, and/or functional differentiation. Given the confounding variables of geography, time, and environmental conditions in the current collection of samples, it is difficult to definitively separate these effects, but various observations argue for functional differentiation between subtypes (i.e., they constitute distinct ecotypes). First, individual subtypes may be found in a wide range of locations; *P. ubique* HTCC1062 was isolated in the Pacific Ocean off the coast of Oregon [55], but closely related sequences are relatively abundant in our samples taken in the Atlantic Ocean. Second, geography per se cannot fully explain differences in subtype distributions, as multiple subtypes are found simultaneously in a single sample. Third, the collection of samples in which a given subtype was found generally exhibits similar environmental conditions. A strong independent illustration of this comes from the correlation of temperature with the distribution of *Prochlorococcus* subtypes [56]. Fourth, the extensive variation within each subtype (i.e., the fact that subtypes are not clonal populations) indicates that it cannot be chance alone that makes genetically similar organisms have similar observed distributions.

Taken together, these results argue that subtype classification is more informative for categorizing microbial popula-

tions than classification using 16S-based ribotypes, or fingerprinting techniques based on length polymorphism, such as T-RFLPs [57] or ARISA [58]. For example, the grouping of such disparate microbial populations under the umbrella *P. marinus* dilutes the significance of the term "species." Indeed, numerous papers have been devoted to comparing and contrasting the differences and variability in *P. marinus* isolates to better understand how this particularly abundant group of organisms has evolved and adapted within the dynamic marine environment [28,52,56,59–66]. Prior to the widespread use of marker-based phylogenetic approaches, microbial systematics relied on a wide range of variables to distinguish microbial populations [67]. Subtypes bring us back to these more comprehensive approaches since they reflect the influences of a wide range of factors in the context of an entire genome.

Although subtypes are a salient feature of our data, variation within a ribotype does not stop at the level of subtypes. Variation within subtypes is so extensive that few GOS reads can be aligned at 100% identity to any other GOS read, despite the deep coverage of several taxonomic groups. Related findings have been shown for the ITS region in various organisms [2,51,52], and in a limited number of organisms for individual protein coding and intergenic regions [2,53,68]. High levels of diversity within the ribotype can be convincingly demonstrated in the 16S gene itself [69]. The applicability of these results over the entire genome were recently shown for *P. marinus* [28] using data from the Sargasso Sea samples taken as a pilot project for the expedition reported here [19]. We have definitively demonstrated the generality of these findings, greatly increased our understanding of the minimum number of variants of a given organism, and shown that these observations apply to the entire genome for a wide range of abundant taxonomic groups and across a wide range of geographic locations.

Average pairwise differences of several percent between overlapping *P. marinus* or SAR11 reads imply that this variation did not arise recently. If one uses substitution rates estimated for *E. coli* [70], one could conclude that on average any two *P. marinus* cells must have diverged millions of years ago. Mutational rates are notoriously variable and hard to estimate, and assumptions of molecular clocks are equally chancy, but clearly within-subtype variants have persisted side by side for quite some time. This raises a question related to the classic "Paradox of the Plankton": how can so many similar organisms have coexisted for so long [71,72]? One explanation, which we favor, is that not only subtypes but also individual variants are sufficiently different phenotypically to prevent any one strain from completely replacing all others (discussed further below; see [71] for a recent theoretical treatment). An alternative is that recombination might prevent selective sweeps within ecotypes, as proposed by Cohan (reviewed in [73]).

The Significance of Within-Subtype Variation

Given the apparent generality of subtypes and intra-subtype variation, it is important to understand if and how these subpopulations are functionally distinct. At the level of DNA sequence, a substantial fraction of substitutions are silent in terms of amino acid sequence, and others may be nonsynonymous but functionally neutral. However, two organisms that differ by 5% in their genetic sequence (e.g., 100,000 substitutions in 2 Mbp of shared sequence) will inevitably have at least minor functional differences such as in the optimal temperature or pH for the activity of some enzyme. At the level of gene content, the observation of hypervariable segments ([28] and here) implies that there is an additional dimension to functional variability. Hypervariable genomic islands with preferential insertion sites could potentially be associated with a wide range of functions, though to date they have been most closely examined for their role in pathogenicity (for a review, see [74]). However, given their apparent variability within even a single sampling site, it seems unlikely that these elements reflect a specific adaptive advantage to the local population. Identifying the source(s), diversity, and range of functionality associated with these islands by fully sequencing a large number of these segments and understanding how their individual abundances fluctuate should be quite informative.

Some might still argue that these differences must be moot for the purpose of understanding the role these organisms play in an ecosystem. Yet even small differences in optimal conditions may have profound effects. They may prevent any single genotype from being universally fittest, allowing and/or necessitating the coexistence of multiple variants [2,51,69]. Moreover, variation within subtype might afford a form of functional "buffering," such that the population as a whole may be more stable in its ecosystem role than any one clone could be (see also [51]). That is, while any one strain of *Prochlorococcus* might thrive and provide energy input to the rest of the community at a limited range of temperatures, light conditions, etc., the ensemble might provide such inputs over a wider range of environmental conditions. In this way, microdiversity might provide system stability or robustness through functional redundancy and the "insurance effect" (reviewed in [75]). Thus, while the extent of microdiversity suggests that knowing the behavior of any one isolate in exquisite detail might not be as useful to reductionist modeling as one might hope, this buffering could afford a more stable ensemble behavior, facilitating the development and maintenance of an ecosystem and allowing for system-level modeling.

A direct equation of subtypes with ecotypes is tempting, but not entirely clear-cut. The correlation of *PstS* distribution with phosphate abundance suggests a functional adaptation, but within *Prochlorococcus* and SAR11 the presence or absence of *PstS* subdivides subtypes without apparent respect for phylogenetic structure. This contrasts markedly with the distribution of proteorhodopsin-tuning variants within SAR11, which, despite a few convergent substitutions, are strongly congruent with phylogeny. It is interesting to ask what distinguishes pressures or adaptations that respect (or that lead to) lineage splits from those that show little or no phylogenetic structuring. These two specific examples plausibly reflect two different mechanisms (i.e., convergent but independent mutation in proteorhodopsin genes and the acquisition by horizontal transfer of genes involved in phosphate uptake). Yet, we must wonder: given the evidence that proteorhodopsin has been transferred laterally [48], and that only a small number of mutations, in some circumstances even a single base-pair change, are required to switch between the blue-absorbing and green-absorbing forms [39,40], why should proteorhodopsin variants show any lineage restriction? Perhaps this relates to the modularity of the system in question: proteorhodopsin tuning may be part of a larger collection of synergistic adaptations that are

collectively not easily evolved, acquired, or lost, while the *PstS* and surrounding genes may represent a functional unit that can be readily added and removed over relatively short evolutionary time scales. If so, perhaps subtypes are indeed ecotypes, but rapidly evolving characters can lead to phenotypes that crosscut or subdivide ecotypes.

Phage provide one possible mechanism for rapid evolution of microbial populations or strains, and have been found in abundance with this and other marine metagenomic datasets [18,20]. It has been proposed that hypervariable islands are phage mediated [28]. However, there are reasons to be cautious about invoking phage as an explanation for rapidly evolving characteristics. While we see variability of *PstS* and neighboring genes in both SAR11 and *P. marinus* populations, this variation does not seem to be linked to recent phage activity. Initially, the distribution of *PstS* seems similar to the variation associated with the hypervariable islands, which may be phage mediated [28]. Indeed, phosphate-regulating genes including *PstS* have been identified in phage genomes [64], presumably because enhanced phosphate acquisition is required during the replication portion of their life cycle. However, the regions containing the *PstS* genes in both SAR11 and *Prochlorococcus* do not behave in the same fashion as clearly hypervariable regions, being effectively bimorphic (modulo the level of sequence variation observed elsewhere in the genome), whereas clearly hypervariable regions are so diverse that nearly every sampled clone falling in such a region appears completely unrelated to every other. Nor do the other genes in *PstS*-containing regions appear to be phage associated. These observations suggest that differences in *PstS* presence or absence arose in the distant past, or that different mechanisms are at work. It seems likely that phage may mediate lateral transfer of *PstS* and other phosphate acquisition genes, but it is unclear whether these genes then can become fixed within the population. Phage require enhanced phosphate acquisition as part of their life cycle [64], so regulatory or functional differences in these genes may limit their suitability for being acquired by the host cell for its own purposes. The rate of phage-mediated horizontal transfer of genes may reflect a combination of the gene's value to the host and to the agent mediating the transfer (e.g., phage), suggesting that *PstS* may have much greater immediate value than do proteorhodopsin genes and their variants.

In practical terms, these results highlight the limitations associated with marker-based analysis and the use of these approaches to infer the physiology of a particular microbial population. At the resolution used here, marker-based approaches are not always informative regarding differences in gene content (e.g., the *PstS* gene as well as neighboring genes), especially those associated with hypervariable segments. Though phosphate acquisition is known to vary within different strains of *P. marinus* [64,76], our results clearly show that this variability can happen within a single subtype (as represented by MIT9312), effectively identifying distinct ecotypes. Given the correct samples from the appropriate environments, other core genes might also show similar variation and allow us to more fully assess the reliability of reference genomes as indicators of physiological potential.

Tools and Techniques

Analysis of the GOS dataset has benefited from the development of new tools and techniques. Many of these approaches rely on fairly well-known techniques but have been modified to take greater advantage of the metadata.

The technique of fragment recruit and the corresponding fragment recruitment plots have proven highly useful for examining the biogeography and genomic variation of abundant marine microbes when a close reference genome exists. Ultimately, this approach derives from the percent identity plots of PipMaker [27]. Similar approaches have been used to examine variation with respect to metagenomic datasets. For example, hypervariable segments and sequence variation have been visualized in *P. marinus* MIT9312 using the Sargasso Sea data [28] and in human gut microbes *Bifidobacterium longum* and *Methanobrevibacter smithii* [77].

Our primary advance associated with fragment recruitment plots is the incorporation of metadata associated with the isolation or production of the sequencing data. While simple in nature, the resulting plots can be extremely informative due to the volume of data being presented. Being able to present the sequence similarity and metadata visually allows a researcher to quickly identify interesting portions of the data for further examination. This is one of the first tools to make extensive use of the metadata collected during a metagenomic sequencing project. The use of sample and recruitment metadata is just the beginning. It is not difficult to imagine displaying other variations such as water temperature, salinity, phosphate abundance, and time of year with this approach. Even sample independent metadata such as phylogenetic information may produce informative views of the data. The usefulness of this and related approaches will only grow as the robust collection of metadata becomes routine and the variables that are most relevant to microbial communities are further elucidated.

The greatest limitation of fragment recruitment is the lack of appropriate reference sequences, particularly finished genomes. Using a series of modifications to the Celera Assembler referred to as "extreme assembly," we have produced large assemblies for cultivated and uncultivated marine microbes. On its own, the extreme assembly approach would be excessively prone to producing chimeric sequences. However, when extreme assemblies are used as references for fragment recruitment, the metadata provides additional criteria to validate the sampling consistency along the length of the scaffold. Chimeric joins can be rapidly detected and avoided. This argues that future metagenomic assemblers could be specifically designed to make use of the metadata to produce more accurate assemblies, and that metagenomic assemblies will be improved by using data from multiple sources. Finding ways to represent the full diversity in these assemblies remains a pressing issue.

Extreme assembly can produce much larger assemblies but it is still limited by overall coverage. While many ribotypes are presumably present in sufficient quantities that reasonable assemblies of these genomes might be expected, this did not occur even for the most abundant organisms, including SAR11 and *P. marinus*. Many of the problems can be attributed to the diversity associated with the hypervariable segments where the effective coverage drops precipitously. If these are indeed commonplace in the microbial world, it is unlikely that complete genomes will be produced using the small insert libraries presented here. However, the ability to bin the larger sequences based on their coverage profiles across multiple samples, oligonucleotide frequency profiles,

and phylogenetic markers suggests that large portions of a microbial genome can be reconstructed from the environmental data. This in turn should provide critical insights into the physiology and biochemistry of these microbial lineages that will inform culture techniques to allow cultivation of these recalcitrant organisms under laboratory conditions.

Not every technique described herein relies on metadata. The marker-less, overlap-based metagenomic comparison provides a quantitative approach to comparing the overall genetic similarity of two samples (Figures 10 and 11). In essence, genomic similarity acts as a proxy for community similarity. Marker-based approaches such as ARISA including the use of 16S sequences described herein can also be used to infer community similarity, though these approaches more aptly generate a census of the community members [51,69,78,79]. This census is biased to the extent that 16S genes can vary in copy number and relies on linkage of the marker gene to infer genome composition. While our metagenome comparison does not directly provide a census, the sensitivity can be tuned by restricting the identity of matches. This means that even subtype-level differences can be detected across samples. It would also identify the substantial gene content differences between the K12 and O157:H7 *E. coli* strains [12]. Such large-scale gene content differences have yet to be seen between closely related marine microbes, but may be a factor in other environments. Although the requisite amount of data will vary with the complexity of the environment or the degree of resolution required, we have found that 10,000 sequencing reads is sufficient to reliably measure the similarity of two surface water samples (unpublished data). This analysis may become a general tool for allocating sequencing resources by allowing a shallow survey of many samples followed by deep sequencing of a select number of "interesting" ones.

The application of this technique for comparing samples along with detailed analysis of fragments recruiting to a given reference sequence can also help explicate differences among communities in gene content or sequence variation. For example, recent metagenomic studies have reported differences in abundance of various gene families or differing functional roles between samples. Some of these differences correspond to plausible differences in physiology and biochemistry, such as the relative overabundance of photosynthetic or light-responsive genes in surface water samples [20,32]. Other differences however are less obvious, such as the abundance of ribosomal proteins at 130 m or the abundance of tranposase at 4,000 m [20]; some of these may reflect "taxonomic hitchhiking," such that a sample rich in Archaea or *Firmicutes* or *Cyanobacteria,* etc., has an over-representation of genes more reflective of their recent evolutionary history than of a response to environmental conditions. Being able to control or account for these taxonomic effects is crucial to understanding how microbial populations have adapted to environmental conditions and how they may behave under changing conditions. The metagenomic comparison method described here provides a new tool to more accurately measure the impact of taxonomic effects.

In conclusion, this study reveals the wealth of biological information that is contained within large multi-sample environmental datasets. We have begun to quantify the amount and structure of the variation in natural microbial populations, while providing some information about how these factors are structured along phylogenetic and environmental factors. At the same time, many questions remain unanswered. For example, although microbial populations are structured and therefore genetically isolated, we do not understand the mechanisms that lead to this isolation. Their isolation seems contradictory given overwhelming evidence that horizontal gene transfer associated with hypervariable islands is a common phenomenon in marine microbial populations. Whatever the mechanism, the role and rate at which gene exchange occurs between populations will be crucial to understanding population structure within microbial communities and whether these communities are chance associations or necessary collections. The hypervariable islands could be a source for tremendous genetic innovation and novelty as evidenced by the rate of discovery of novel protein families in the GOS dataset [18]. However, it is not clear whether these entities are the main source of this novelty or whether this novelty resides in the vast numbers of rare microbes [4] that cannot be practically accessed using current metagenomic approaches. Altogether, this research reaffirms our growing wealth and complexity of data and paucity of understanding regarding the biological systems of the oceans.

Materials and Methods

Sampling sites. A more detailed description of the sampling sites provides additional context in which to understand the individual samples. The northernmost site (GS05) was at Compass Buoy in the highly eutrophic Bedford Basin, a marine embayment encircled by Halifax, Nova Scotia, that has a 15-y weekly record of biological, physical, and chemical monitoring (http://www.mar.dfo-mpo.gc.ca/science/ocean/BedfordBasin/index.htm). Other temperate sites included a coastal station sample near Nova Scotia (GS4), a station in the Bay of Fundy estuary at outgoing tide (GS06), and three Gulf of Maine stations (GS02, GS03, and GS07). These were followed by sampling coastal stations from the New England shelf region of the Middle Atlantic Bight (Newport Harbor through Delaware Bay; GS08–GS11). The Delaware Bay (GS11) was one of several estuary samples along the Global Expedition path. Estuaries are complex hydrodynamic environments that exhibit strong gradients in oxygen, nutrients, organic matter, and salinity and are heavily impacted by anthropogenic nutrients. The Chesapeake Bay (GS12) is the largest estuary in the United States and has microbial assemblages that are diverse mixtures of freshwater and marine-specific organisms [80]. GS13 was collected near Cape Hatteras, North Carolina, inside and north of the Gulf Stream, and GS14 was taken along the western boundary frontal waters of the Gulf Stream off the coast of Charleston, South Carolina. The vessel stopped at five additional stations as it transited through the Caribbean Sea (GS15–GS19) to the Panama Canal. In Panama, we sampled the freshwater Lake Gatun, which drains into the Panama Canal (GS20). The first of the eastern Pacific coastal stations GS21, GS22, and GS23 were sampled on the way to Cocos Island (~500 km southwest of Costa Rica), followed by a coastal Cocos Island sample (GS25). Near the island, ocean currents diverge and nutrient rich upwellings mix with warm surface waters to support a highly productive ecosystem. Cocos Island is distinctive in the eastern Pacific because it belongs to one of the first shallow undersea ridges in the region encountered by the easterly flowing North Equatorial Counter/Cross Current in the Far Eastern Pacific [81,82]. After departing Cocos Island, the vessel continued southwest to the Galapagos Islands, stopping for an open ocean station (GS26). An intensive sampling program was then conducted in the Galapagos. The Galapagos Archipelago straddles the equator 960 km west of mainland Ecuador in the eastern Pacific. These islands are in a hydrographically complex region due to their proximity to the Equatorial Front and other major oceanic currents and regional front systems [83]. The coastal and marine parts of the Galapagos Islands ecosystem harbor an array of distinctive habitats, processes, and endemic species. Several distinct zones were targeted including a shallow-water, warm seep (GS30), below the thermocline in an

upwelling zone (GS31), a coastal mangrove (GS32), and a hypersaline lagoon (GS33). The last stations were collected from open ocean sites (GS37 and GS47) and a coral reef atoll lagoon (GS51) in the immense South Pacific Gyre. The open ocean samples come from a region of lower nutrient concentrations where picoplankton are thought to represent the single most abundant and important factor for biogeochemical structuring and nutrient cycling [84–87]. In the atoll systems, ambient nutrients are higher, and bacteria are thought to constitute a large biomass that is one to three times as large as that of the phytoplankton [88–90].

Sample collection. A YSI (model 6600) multiparameter instrument (http://www.ysi.com) was deployed to determine physical characteristics of the water column, including salinity, temperature, pH, dissolved oxygen, and depth. Using sterilized equipment [91], 40–200 l of seawater, depending on the turbidity of the water, was pumped through a 20-μm nytex prefilter into a 250-l carboy. From this sample, two 20-ml subsamples were collected in acid-washed polyethylene bottles and frozen (–20 °C) for nutrient and particle analysis. At each station the biological material was size fractionated into individual "samples" by serial filtration through 20-μm, 3-μm, 0.8-μm, and 0.1-μm filters that were then sealed and stored at –20 °C until transport back to the laboratory. Between 44,160 and 418,176 clones per station were picked and end sequenced from short-insert (1.0–2.2 kb) sequencing libraries made from DNA extracted from filters [19]. Data from these six *Sorcerer II* expedition legs (37 stations) were combined with the results from samples in the Sargasso Sea pilot study (four stations; GS00a–GS00d and GS01a–GS01c; [19]. The majority of the sequence data presented came from the 0.8- to 0.1-μm size fraction sample that concentrated mostly bacterial and archaeal microbial populations. Two samples (GS01a, GS01b) from the Sargasso Sea pilot study dataset and one GOS sample (GS25) came from other filter size fractions (Table 1).

Filtration and storage. Microbes were size fractionated by serial filtration through 3.0-μm, 0.8-μm, and 0.1-μm membrane filters (Supor membrane disc filter; Pall Life Sciences, http://www.pall.com), and finally through a Pellicon tangential flow filtration (Millipore, http://www.millipore.com) fitted with a Biomax-50 (polyethersulfone) cassette filter (50 kDa pore size) to concentrate a viral fraction to 100 ml. Filters were vacuum sealed with 5 ml sucrose lysis buffer (20 mM EDTA, 400 mM NaCl, 0.75 M sucrose, 50 mM Tris-HCl [pH 8.0]) and frozen to –20 °C on the vessel until shipment back to the Venter Institute, where they were transferred to a –80 °C freezer until DNA extraction. Glycerol was added (10% final concentration) as a cryoprotectant for the viral/phage sample.

DNA isolation. In the laboratory, the impact filters were aseptically cut into quarters for DNA extraction. Unused quarters of the filter were refrozen at –80 °C for storage. Quarters used for extraction were aseptically cut into small pieces and placed in individual 50-ml conical tubes. TE buffer (pH 8) containing 50 mM EGTA and 50 mM EDTA was added until filter pieces were barely covered. Lysozyme was added to a final concentration of 2.5 mg/ml^{-1}, and the tubes were incubated at 37 °C for 1 h in a shaking water bath. Proteinase K was added to a final concentration of 200 μg/ml^{-1}, and the samples were frozen in dry ice/ethanol followed by thawing at 55 °C. This freeze–thaw cycle was repeated once. SDS (final concentration of 1%) and an additional 200 μg/ml^{-1} of proteinase K were added to the sample, and samples were incubated at 55 °C for 2 h with gentle agitation followed by three aqueous phenol extractions and one phenol/chloroform extraction. The supernatant was then precipitated with two volumes of 100% ethanol, and the DNA pellet was washed with 70% ethanol. Finally, the DNA was treated with CTAB to remove enzyme inhibitors. Size fraction samples not utilized in this study were archived for future analysis.

Library construction. DNA was randomly sheared via nebulization, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected using gel electrophoresis on 1% low-melting-point agarose. After ligation to BstXI adapters, DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments were inserted into BstXI-linearized medium-copy pBR322 plasmid vectors. The resulting library was electroporated into *E. coli*. To ensure construction of high-quality random plasmid libraries with few to no clones with no inserts, and no clones with chimeric inserts, we used a series of vectors (pHOS) containing BstXI cloning sites that include several features: (1) the sequencing primer sites immediately flank the BstXI cloning site to avoid excessive resequencing of vector DNA; (2) elimination of strong promoters oriented toward the cloning site; and (3) the use of BstXI sites for cloning facilitates the preparation of libraries with a low incidence of no-insert clones and a high frequency of single inserts. Clones were sequenced from both ends to produce pairs of linked sequences representing ~820 bp at the end of each insert.

Template preparation. Libraries were transformed, and cells were plated onto large format (16 × 16cm) diffusion plates prepared by layering 150 ml of fresh molten, antibiotic-free agar onto a previously set 50-ml layer of agar containing antibiotic. Colonies were picked for template preparation using the Qbot or QPix colony-picking robots (Genetix, http://www.genetix.com), inoculated into 384-well blocks containing liquid media, and incubated overnight with shaking. High-purity plasmid DNA was prepared using the DNA purification robotic workstation custom-built by Thermo CRS (http://www.thermo.com) and based on the alkaline lysis miniprep [92]. Bacterial cells were lysed, cell debris was removed by centrifugation, and plasmid DNA was recovered from the cleared lysate by isopropanol precipitation. DNA precipitate was washed with 70% ethanol, dried, and resuspended in 10 mM Tris HCl buffer containing a trace of blue dextran. The typical yield of plasmid DNA from this method is approximately 600–800 ng per clone, providing sufficient DNA for at least four sequencing reactions per template.

Automated cycle sequencing. Sequencing protocols were based on the di-deoxy sequencing method [93]. Two 384-well cycle-sequencing reaction plates were prepared from each plate of plasmid template DNA for opposite-end, paired-sequence reads. Sequencing reactions were completed using the Big Dye Terminator chemistry and standard M13 forward and reverse primers. Reaction mixtures, thermal cycling profiles, and electrophoresis conditions were optimized to reduce the volume of the Big Dye Terminator mix (Applied Biosystems, http://www.appliedbiosystems.com) and to extend read lengths on the AB3730xl sequencers (Applied Biosystems). Sequencing reactions were set up by the Biomek FX (Beckman Coulter, http://www.beckmancoulter.com) pipetting workstations. Robots were used to aliquot and combine templates with reaction mixes consisting of deoxy- and fluorescently labeled dideoxynucleotides, DNA polymerase, sequencing primers, and reaction buffer in a 5 μl volume. Bar-coding and tracking promoted error-free template and reaction mix transfer. After 30–40 consecutive cycles of amplification, reaction products were precipitated by isopropanol, dried at room temperature, and resuspended in water and transferred to one of the AB3730xl DNA analyzers. Set-up times were less than 1 h, and 12 runs per day were completed with average trimmed sequence read length of 822 bp.

Fosmid end sequencing. Fosmid libraries [24] were constructed using approximately 1 μg DNA that was sheared using bead beating to generate cuts in the DNA. The staggered ends or nicks were repaired by filling with dNTPs. A size selection process followed on a pulse field electrophoresis system with lambda ladder to select for 39–40 Kb fragments. The DNA was then recovered from a gel, ligated to the blunt-ended pCC1FOS vector, packaged into lambda packaging extracts, incubated with the host cells, and plated to select for the clones containing an insert. Sequencing was performed as described for plasmid ends.

Metagenomic assembly. Assembly was conducted with the Celera Assembler [21], with modifications as follows. The "genome length" was artificially set at the length of the dataset divided by 50 to allow unitigs of abundant organisms to be treated as unique, as previously described [19]. Several distinct assemblies were computed. In the primary assembly, all pairs of mated reads were tested to see whether the paired reads overlapped one another; if so, they were merged into a single pseudo-read that replaced the two original reads; further, only overlaps of 98% identity or higher were used to construct unitigs. A second assembly was conducted in the same fashion with the exception of using a 94% identity cutoff to construct unitigs. Finally, series of assemblies at various stringencies were computed for subsets of the GOS data; in these assemblies, overlapping mates were not preassembled and the Celera Assembler code was modified slightly to allow for overlapping and multiple sequence alignment at lower stringency.

Construction of a low-identity overlap database. An all-against-all comparison of unassembled (but merged and duplicate-stripped) sequences from the combined dataset was performed using a modified version of the overlapper component of the Celera Assembler [21]. The code was modified to find overlap alignments (global alignments allowing free end gaps) starting from pairs of reads that share an identical substring of at least 14 bp. An alignment extension was then performed with match/mismatch scores set to yield a positive outcome if an overlap alignment was found with ≥65% identity. Overlaps involving alignments of ≥40 bp were retained for various analyses. For the GOS dataset described here, this process resulted in a dataset of 1.2 billion overlaps. Due to the 14-

bp requirement and certain heuristics for early termination of apparently hopeless extensions, not all alignments at ≥65% were found. In addition, some of the lowest-identity overlaps are bound to be chance matches; however, this was a relatively uncommon event. Approximately one in 5×10^6 pairs of 800-bp random sequences (all sites independent, A = C = G = T = 25%) can be aligned to overlap ≥40 bp at ≥65% identity using the same procedure. At a 70% cutoff, the value is reduced to one in 4×10^7, and one in 5×10^8 at a 75% cut off.

Extreme assembly. Like many assembly algorithms, the extreme assembler proceeds in three phases: overlap, layout, and consensus. The overlap phase is provided by the all-against-all comparison described above. The consensus phase is performed by a version of the Celera Assembler, modified to accept higher rates of mismatch. The layout phase begins with a single sequencing read ("seed") that is chosen at random or specified by the user and is considered the "current" read. The following steps are performed off one or both ends of the seed. (1) Starting from the current fragment end, add the fragment with the best overlap off that end and mark the current fragment as "used," thus making the added fragment the new current fragment. (2) Mark as used any alternative overlap that would have resulted in a shorter extension. The simplest notion of "best overlap" is simply the one having the highest identity alignment, but more complicated criteria have certain advantages. A simple but useful refinement is to favor fragments whose other ends have overlaps over those which are dead ends. For an unsupervised extreme assembly, when the sequence extension terminates because there are no more overlaps, a new unused fragment is chosen as the next seed and the process is repeated until all fragments have been marked used.

Construction of multiple SAR11 variants. Sequencing reads mated to SAR11-like 16S sequences but themselves outside of the ribosomal operon (n = 348) were used as seeds in independent extreme assemblies. Since the assemblies were independent, the results were highly redundant, with a given chain of overlapping fragments typically being used in multiple assemblies. A subset of 24 assemblies that shared no fragments over their first 20 kb was identified as follows. (1) Connected components were determined in a graph defined by nodes corresponding to extreme assemblies. If the assemblies shared at least one fragment in the first 20 kb of each assembly, the two nodes were connected by an edge. (2) A single assembly was chosen at random from each of the connected components. The consensus sequence over the 20-kb segment of each such representative was used as the reference for fragment recruitment.

Phylogeny. Phylogenies of sequences homologous to a given portion of a reference sequence (typically 500 bp) were determined in the following manner. A set of homologous fragments was identified based on fragment recruitment to the reference as described above. Fragments that fully spanned the segment of interest and had almost full-length alignments to the reference sequence of a user-defined percent identity (typically, 70%) were used for further analysis. A preliminary master–slave multiple sequence alignment of the recruited reads (slaves) to the reference segment (master) was performed with a modified version of the consensus module of the Celera Assembler. Based on this alignment, reads were trimmed to the portion aligning to the reference segment of interest. A refined multiple sequence alignment was then computed with MUSCLE [94]. Distance based phylogenies were computed using the programs DNADIST and NEIGHBOR from the PHYLIP package [95] using default settings. Trees were visualized using HYPERTREE [96].

Measurement of library-to-library similarity. Based on the low-identity overlap database described above, the similarity of a library i to another library j at a given percent identity cutoff was computed as follows. For each sequence s of i, let $n_{s,i}$ = the number of overlaps to other fragments of i satisfying the cutoff; $n_{s,j}$ = the number of overlaps to fragments of j satisfying the cutoff; and $f_{s,i} = n_{s,i}/(n_{s,i} + n_{s,j})$ = fraction of reads overlapping s from i or j that are from i.

$$r_{i,i} = \sum_s f_{s,i} \quad (1)$$

$$r_{i,j} = \sum_s 1 - f_{s,i} \quad (2)$$

$$s_{i,j} = 0.5^*(r_{i,j} + r_{j,i})/\sqrt{r_{i,i}{}^*r_{j,j}} \quad (3)$$

$$S_{i,j} = S_{j,i} = 2s_{i,j}/(1 + s_{i,j}) \quad (4)$$

A read that can be overlapped to another at sufficiently high-sequence identity was taken to indicate that they were from similar organisms, and, relatedly, that similar genes were present in the samples. Only reads with such overlaps contributed to the calculation. Other reads reflect genes or segments of genomes that were so lightly sampled (i.e., at such low abundance) that they were not informative regarding the similarity of two samples. Consequently, the analysis automatically corrects for differences in the amount of sequencing, and can be computed over sets of samples that vary considerably in diversity. The resulting measure of similarity $S_{i,j}$ takes on a value between 0 and 1, where 0 implies no overlaps between i and j, and 1 implies that a fragment from i and a fragment from j are as likely to overlap one another as are two fragments from i or two fragments from j. As with the Bray-Curtis coefficient [97], abundance of categories affects the computation. In an idealized situation where two libraries can each be divided into some number k of "species" at equal abundance, and the libraries have l of the species in common, the similarity statistic will approach l/k for large samplings; in this sense, $S_{i,j} = x$ indicates that the two samples share approximately a fraction x of their genetic material. It is frequently useful to define $D_{i,j} = 1 - S_{i,j}$, the "dissimilarity" or distance between two samples.

Ribotype clustering and identification of representatives. An all-against-all comparison of predicted 16S sequences was performed to determine the alignment between pairs of overlapping sequences using a version of an extremely fast bit-vector algorithm [98]. A hierarchical clustering was determined using percent-mismatch in the resulting alignments as the distance between pairs of sequences. Order of clustering and cluster identity scores were based on the average-linkage criterion, with distances between nonoverlapping partial sequences treated as missing data. Ribotypes were the maximal clusters with an identity score above the cutoff (typically 97%). Representative sequences were chosen for each cluster based on both length and highest average identity to other sequences in the cluster.

Taxonomic classification. Taxonomic classification of 16S sequences was conducted using phylogenetic techniques based on clade membership of similar sequences with 16S sequences with defined taxonomic membership. Representative sequences from clustered sequences were analyzed as described previously [19,99] and by addition into an ARB database of small subunit rDNAs [100,101]. Results were spot-checked against the Ribosomal Database Project II Classifier server [102] and the taxonomic labels of the best BLASTN hits against the nonredundant database at NCBI.

Fragment recruitment. Global ocean sequences were aligned to genomic sequences of different bacteria and phage using NCBI BLASTN [26]. The following blast parameters were designed to identify alignments as low as 55% identity that could contain large gaps: -F "m L" -U T -p blastn -e 1e-4 -r 8 -q -9 -z 3000000000 -X 150. Reads were filtered in several steps to identify the reads that were aligned over more or less their entire length. Reads had to be aligned for more than 300 bp at >30% identity with less than 25 bp of unaligned bases on either end, or reads had to be aligned over more than 100 bp at >30% identity with less than 20 bp of overhang off either end. Identity was calculated ignoring gaps. In some instances a read might be placed, but the mate would not be placed under these criteria. In such cases, if 80% or more of the mate were successfully aligned, then the mate would be rescued and considered successfully aligned.

Generation of shredded artificial reads from finished genomes. Random pieces of DNA from the genome in question with a length between 1,800 to 2,500 bp were selected. For each piece a read length N1 was selected from the distribution of lengths using the GOS dataset. If that GOS sequence had a mate pair, then a second length N2 was again randomly selected. The length N1 was used to generate a read from the 5′ end of the DNA. The piece of DNA was then reverse complemented and if appropriate, a second length N2 was used to generate a second read. The relationship between these two reads was then recorded and used to produce a fasta file. This approach successfully mimics the types of reads found in the GOS data with similar rates of missing mates.

Abundance of proteorhodopsin variants. A total of 2,644 proteorhodopsin genes were identified from the clustered open reading frames derived from the GOS assembly [18]. These genes could be linked back to 3,608 GOS clones. Open reading frames were predicted from these clones as described in [18]. The peptide sequences were aligned with NCBI blastpgp with the following parameters: -j 5 -U T -e 10 -W 2 -v 5 -b 5000 -F "m L" -m 3. The search was performed with a previously described blue-absorbing proteorhodopsin protein BPR (gi|32699602) as the query. The amino acid associated with light absorption is found within a short

conserved motif RYVDWLLTVPL*IVEF, where the asterisk indicates tuning amino acid [39–41]. In total, 1,938 clones were found to contain this motif. Clones and the sample metadata were then associated with the tuning amino acid to determine the relative abundance of the different amino acids at these positions. Clones could be associated with SAR11 if both mated sequencing reads (when available) were recruited to *P. ubique* HTCC1062.

Site abundance estimates and comparisons. Given a set of genes identified on the GOS sequences, we can identify the scaffolds on which these genes were annotated. A vector indicating the number of sequences contributed by every sample is determined for every gene. This vector reflects the number of sequences from every sample that assembled into the scaffold on which the gene was identified after normalizing for the proportion of scaffold covered by the gene. For example, if a 10-kb scaffold contains a 1-kb gene, then each sample will contribute one GOS sequence for every ten GOS sequences it contributed to the entire scaffold. The vectors are then summed and normalized to account for either the total number of GOS sequences obtained from each sample or based on the number of typically single copy *recA* genes (identified as in [18]). Unless stated otherwise, *recA* was used to normalize abundance across samples. When comparisons using groups of samples were performed, the average value for the samples was compared.

Oligonucleotide composition profile. A 1-D profile representing oligonucleotide frequencies was computed as follows. A sequence was converted into a series of overlapping 10,000-bp segments, each segment offset by 1,000 bp from the previous one, using perl and shell scripting. Dinucleotide frequencies are computed on each segment using a C program written for this purpose. Higher-order oligonucleotides were examined and gave similar results for the genomes of interest. Remaining calculations were performed using the R package [103]. Principle component analysis (function *princomp* with default settings) was applied to the matrix of frequencies per window position. The value of the first component for each position was normalized by the standard deviation of these values, and truncated to the range [−5, 5]. For visualizations, the resulting values were plotted at the center of each window.

Estimating frequency of large-scale translocations and inversions. The unrecruited mated sequencing reads of reads recruited to *P. marinus* MIT9312 at or above 80% identity were examined. An unrecruited mate indicated a potential translocation or inversion if it aligned to the MIT9312 genome in two and only two distinct alignments separated by at least 50 kb, if each aligned portion was at least 250 bp long, if there was less than 100 bp of unaligned sequence and no more than 100 bp of overlapping sequence between the two aligned portions in read coordinates, and if each aligned portion was anchored to one end of the sequencing read with less than 25 bp of unaligned sequence from each end. In total, 18 rearrangements were identified, six of which appear to be unique events.

The rate of discovery was estimated by determining the number of rearrangements in a given volume of sequence. We estimated the volume of sequence that was potentially examined by identifying recruited mated sequencing reads that fit the "good" category (i.e., which were recruited in the correct orientation at the expected distance from each other). For a given read, if the mate was recruited at greater than or equal to 80% identity, then the expected amount of sequence examined should be the current (as opposed to mate) read length minus 500 bp. This produces an estimate of the search space to be ~47 Mbp. Given 18 rearrangements, this leads to an estimate of one rearrangement per 2.6 Mbp.

Quantification and assessment of sequences associated with gaps. GOS reads assigned to the "missing mate" category that were recruited at greater than 80% identity outside the gap in question were identified. The mates of these reads were then identified and clustering was attempted with Phrap (http://www.phrap.org). Reads that were incorporated end to end into the Phrap assemblies were identified. For most small gaps a single assembly included all the missing mate reads and identified the precise difference between the reference and the environmental sequences. For the hypervariable segments, most of the reads failed to assemble at all, and those that did show greater sequence divergence than typically seen. In the case of SAR11-recruited reads, to increase the number of reads associated with the hypervariable gaps we identified reads that did not recruit to the *P. ubique* HTCC1062 but aligned in a single HSP (high-scoring pair) over at least 500 bp with one end unaligned because it extended into the hypervariable gap.

Data and tool release. To facilitate continued analysis of this and other metagenomic datasets, the tools presented here along with their source code will be available via the Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) website (http://camera.calit2.net). The dataset and associated metadata will be accessible via CAMERA (using the dataset tag CAM__PUB__Rusch07a). Given the exceptional abundance of *Burkholderia* and *Shewanella* sequences in the first Sargasso Sea sample and the feeling that these may be contaminants, we are also providing a list of the scaffold IDs and sequencing read IDs associated with these organisms to facilitate analyses with or without the sequences. In addition to CAMERA, the GOS scaffolds and annotations will be available via the public sequence repositories such as NCBI (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj&cmd=Retrieve&dopt=Overview&list__uids=13694), and the reads will be available via the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?).

Supporting Information

Poster S1. Fragment Recruitment of GOS Data to Finished Microbial Genomes

Found at doi:10.1371/journal.pbio.0050077.sd001 (21 MB PDF).

Accession Numbers

The GenBank (http://www.ncbi.nlm.nih.gov/Genbank) accession number for proteorhodopsin protein BPR is gi|32699602.

Acknowledgments

We acknowledge the Department of Energy (DOE), Office of Science, and Office of Biological and Environmental Research (DE-FG02-02ER63453), the Gordon and Betty Moore Foundation, the Discovery Channel, and the J. Craig Venter Science Foundation for funding to undertake this study. We are also indebted to a large group of individuals and groups for facilitating our sampling and analysis. We thank the governments of Canada, Mexico, Honduras, Costa Rica, Panama, Ecuador, French Polynesia, and France for facilitating sampling activities. All sequencing data collected from waters of the above-named countries remain part of the genetic patrimony of the country from which they were obtained. Canada's Bedford Institute of Oceanography provided a vessel and logistical support for sampling in Bedford Basin. The Universidad Nacional Autónoma de México (UNAM) facilitated permitting and logistical arrangements and identified a team of scientists for collaboration. The scientists and staff of the Smithsonian Tropical Research Institute (STRI) hosted our visit in Panama. Representatives from Costa Rica's Organization for Tropical Studies (Jorge Arturo Jimenez and Francisco Campos Rivera), the University of Costa Rica (Jorge Cortés), and the National Biodiversity Institute (INBio) provided assistance with planning, logistical arrangements, and scientific analysis. Our visit to the Galapagos Islands was facilitated by assistance from the Galapagos National Park Service Director, Washington Tapia, and the Charles Darwin Research Institute, especially Howard Snell and Eva Danulat. We especially thank Greg Estes (guide), Héctor Cháuz Campo (Institute of Oceanography of the Ecuador Navy), and a National Park Representative, Simon Ricardo Villemar Tigrero, for field assistance while in the Galapagos Islands. Martin Wilkalski (Princeton University) and Rod Mackie (University of Illinois) provided planning advice for the Galapagos sampling plan. We thank Matthew Charette (Woods Hole Oceanographic Institute) for nutrient data analysis. We also acknowledge the help of Michael Ferrari and Jennifer Clark for remote sensing data. The U.S. Department of State facilitated Governmental communications on multiple occasions. John Glass (J. Craig Venter Institute [JCVI]) provided valuable assistance in methods development. The dedicated efforts of the quality systems, library construction, template, and sequencing teams at the JCVI Joint Technology Center produced the high quality sequence data that was the basis of this paper. We thank Matthew LaPointe, Creative Director of JCVI, for assistance with figure design, and the JCVI information technology support team who facilitated many of the vessel related technical needs. Special thanks are due for Charles H. Howard, captain of the *Sorcerer II,* and fellow crew members Cyrus Foote and Brooke A. Dill for their time and effort in support of this research. We gratefully acknowledge Dr. Michael Sauri, who oversaw medical related issues for the crew of the *Sorcerer II*.

Author contributions. DBR, ALH, KB, HS, CAP, JFH, MF, and JCV conceived and designed the experiments. DBR, ALH, JMH, KB, BT, HBT, CS, JT, JF, CAP, and JCV performed the experiments. DBR, ALH, GS, KBH, SW, DW, JAE, KR, JEV, TU, YHR, MRF, KN, and RF analyzed the data. DBR, ALH, GS, KBH, SY, JMH, KR, KB, BT, HS,

HBT, CS, JT, JF, CAP, KL, SK, JFH, TU, YHR, LIF, VS, GBR, LEE, DMK, SS, TP, EB, VG, GTC, MRF, RLS, MF, and JCV contributed reagents/materials/analysis tools. DBR, ALH, GS, KBH, SW, SY, JAE, RLS, KN, RF, MF, and JCV wrote the paper.

Funding. Funding for this study was received from the US Department of Energy, Office of Science, and Office of Biological and Environmental Research (DE-FG02-02ER63453), the Gordon and Betty Moore Foundation, the Discovery Channel, and the J. Craig Venter Science Foundation.

Competing interests. The authors have declared that no competing interests exist.

References

1. Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. Proc Natl Acad Sci U S A 95: 6578–6583.
2. Beja O, Koonin EV, Aravind L, Taylor LT, Seitz H, et al. (2002) Comparative genomic analysis of archaeal genotypic variants in a single population and in two different oceanic provinces. Appl Environ Microbiol 68: 335–345.
3. DeLong EF, Pace NR (2001) Environmental diversity of bacteria and archaea. Systematic Biol 50: 1–9.
4. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, et al. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc Natl Acad Sci U S A 103: 12115–12120.
5. Garrity GM (2001) Bergey's manual of systematic bacteriology. New York; Springer-Verlag.
6. Madigan M, Martinko JM, Parker J (2000) Brock biology of microorganisms. Upper Saddle River (NJ); Prentice Hall. 991 p.
7. Fuhrman JA, McCallum K, Davis AA (1992) Novel major archaebacterial group from marine plankton. Nature (London) 356: 148–149.
8. Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345: 60–63.
9. Pace NR (1997) A molecular view of microbial diversity and the biosphere. Science 276: 734–740.
10. Rappe MS, Giovannoni SJ (2003) The uncultured microbial majority. Ann Rev Microbiol 57: 369–394.
11. Stackebrandt E, Goebel BM (1994) Taxonomic note: A place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int J Syst Bacteriol 44: 846–849.
12. Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, et al. (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A 99: 17020–17024.
13. Linklater E (1972) The voyage of the Challenger. Garden City (NJ); Doubleday. 280 p.
14. Mosley HN (1879) Notes by a naturalist on the "Challenger," being an account of various observations made during the voyage of H.M.S. "Challenger" round the world, in the years 1872–1876. London: Macmillian and Company. 540 p.
15. Thompson SCW, Murray SJ, Nares GS, Thompson FT (1895) Report on the scientific results of the voyage of H.M.S. Challenger during the years 1873–76 under the command of Captain George S. Nares, R.N., F.R.S. and the late Captain Frank Tourle Thomson, R.N. Prepared under the superintendence of the late Sir C. Wyville Thomson, 1885–1895: Edinburgh: printed for H.M. Stationery off. (by order of Her Majesty's Government).
16. Fuhrman JA, Mccallum K, Davis AA (1993) Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. Appl Environ Microbiol 59: 1294–1302.
17. Hewson I, Steele JA, Capone DG, Fuhrman JA (2006) Temporal and spatial scales of variation in bacterioplankton assemblages of oligotrophic surface waters. Mar Ecol Prog Ser 311: 67–77.
18. Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, et al. (2007) The *Sorcerer II* Global Ocean Sampling expedition: Expanding the universe of protein families. PLoS Biol 5: e16. doi:10.1371/journal.pbio.0050016
19. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304: 66–74.
20. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, et al. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496–503.
21. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, et al. (2000) A whole-genome assembly of *Drosophila*. Science 287: 2196–2204.
22. Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics 2: 231–239.
23. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, et al. (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298: 129–149.
24. Kim UJ, Shizuya H, Dejong PJ, Birren B, Simon MI (1992) Stable propagation of cosmid sized human DNA inserts in an F-factor based vector. Nucleic Acids Res 20: 1083–1085.
25. Stein JL, Marsh TL, Wu KY, Shizuya H, DeLong EF (1996) Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J Bacteriol 178: 591–599.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
27. Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, et al. (2000) PipMaker: A Web server for aligning two genomic DNA sequences. Genome Res 10: 577–586.
28. Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, et al. (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. Science 311: 1768–1770.
29. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309: 1242–1245.
30. Hagstrom A, Pommier T, Rohwer F, Simu K, Stolte W, et al. (2002) Use of 16S ribosomal DNA for delineation of marine bacterioplankton species. Appl Environ Microbiol 68: 3628–3633.
31. Giovannoni S, Rappe M (2002) Evolution, diversity and molecular ecology of marine Prokaryotes. In: Kirchman DL, editor. Microbial ecology of the oceans. New York: Wiley-Liss. pp. 47–84.
32. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, et al. (2005) Comparative metagenomics of microbial communities. Science 308: 554–557.
33. Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Res 31: 371–373.
34. Conkright M, Levitus S, Boyer T. (1994) World ocean atlas 1994. Volume 1: Nutrients. Washington, D.C. U.S. Department of Commerce.
35. Levitus S, Burgett R, Boyer T (1994) World ocean atlas 1994. Volume 3: Nutrients. Washington, D.C.: U.S. Department of Commerce.
36. Parekh P, Follows MJ, Boyle E (2004) Modeling the global ocean iron cycle. Global Biogeochem Cycles 18: GB1002.
37. Scanlan DJ, Wilson WH (1999) Application of molecular techniques to addressing the role of P as a key effector in marine ecosystems. Hydrobiologia 401: 149–175.
38. Moore L, Ostrowski M, Scanlan D, Feren K, Sweetsir T (2005) Ecotypic variation in phosphorus acquisition mechanisms within marine picocyanobacteria. Aquat Microb Ecol 39: 257–269.
39. Kelemen BR, Du M, Jensen RB (2003) Proteorhodopsin in living color: Diversity of spectral properties within living bacterial cells. Biochim Biophys Acta 1618: 25–32.
40. Man D, Wang W, Sabehi G, Aravind L, Post AF, et al. (2003) Diversification and spectral tuning in marine proteorhodopsins. EMBO J 22: 1725–1731.
41. Man-Aharonovich D, Sabehi G, Sineshchekov OA, Spudich EN, Spudich JL, et al. (2004) Characterization of RS29, a blue-green proteorhodopsin variant from the Red Sea. Photochem Photobiol Sci 3: 459–462.
42. Bielawski JP, Dunn KA, Sabehi G, Beja O (2004) Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. Proc Natl Acad Sci U S A 101: 14824–14829.
43. Johnsen S, Sosik H. (2005) Shedding light on light in the ocean. Oceanus Mag 43: 24–28.
44. Braun C, Smirnov S (1993) Why is water blue. J Chem Educ 70: 612–615.
45. Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, et al. (2000) Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. Science 289: 1902–1906.
46. de la Torre JR, Christianson LM, Beja O, Suzuki MT, Karl DM, et al. (2003) Proteorhodopsin genes are distributed among divergent marine bacterial taxa. Proc Natl Acad Sci U S A 100: 12830–12835.
47. Sabehi G, Beja O, Suzuki MT, Preston CM, DeLong EF (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. Environ Microbiol 6: 903–910.
48. Frigaard NU, Martinez A, Mincer TJ, DeLong EF (2006) Proteorhodopsin lateral gene transfer between marine planktonic bacteria and archaea. Nature 439: 847–850.
49. Yokoyama S (2000) Phylogenetic analysis and experimental approaches to study color vision in vertebrates. Methods Enzymol 315: 312–325.
50. Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, et al. (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science 296: 2028–2033.
51. Brown MV, Fuhrman JA (2005) Marine bacterial microdiversity as revealed by internal transcribed spacer analysis. Aquat Microb Ecol 41: 15–23.
52. Rocap G, Distel DL, Waterbury JB, Chisholm SW (2002) Resolution of *Prochlorococcus* and *Synechococcus* ecotypes by using 16S-23S ribosomal DNA internal transcribed spacer sequences. Appl Environ Microbiol 68: 1180–1191.
53. Schleper C, DeLong EF, Preston CM, Feldman RA, Wu KY, et al. (1998) Genomic analysis reveals chromosomal variation in natural populations of the uncultured psychrophilic archaeon *Cenarchaeum symbiosum*. J Bacteriol 180: 5003–5009.
54. Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9: 552–569.
55. Rappe MS, Connon SA, Vergin KL, Giovannoni SJ (2002) Cultivation of

the ubiquitous SAR11 marine bacterioplankton clade. Nature 418: 630–633.
56. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, et al. (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science 311: 1737–1740.
57. Liu WT, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. Appl Environ Microbiol 63: 4516–4522.
58. Fisher MM, Triplett EW (1999) Automated approach for ribosomal intergenic spacer analysis of microbial diversity and its application to freshwater bacterial communities. Appl Environ Microbiol 65: 4630–4636.
59. Hess WR, Rocap G, Ting CS, Larimer F, Stilwagen S, et al. (2001) The photosynthetic apparatus of *Prochlorococcus:* Insights through comparative genomics. Photosynth Res 70: 53–71.
60. Martiny AC, Coleman ML, Chisholm SW (2006) Phosphate acquisition genes in *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. Proc Natl Acad Sci U S A 103: 12552–12557.
61. Moore LR, Chisholm SW (1999) Photophysiology of the marine cyanobacterium *Prochlorococcus:* Ecotypic differences among cultured isolates. Limnol Oceanogr 44: 628–638.
62. Moore LR, Rocap G, Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. Nature 393: 464–467.
63. Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424: 1042–1047.
64. Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW (2005) Three *Prochlorococcus* cyanophage genomes: Signature features and ecological interpretations. PLoS Biol 3: e144.
65. Ting CS, Rocap G, King J, Chisholm SW (2002) Cyanobacterial photosynthesis in the oceans: The origins and significance of divergent light-harvesting strategies. Trends Microbiol 10: 134–142.
66. Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, 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.
67. Wayne LG, Brenner DJ, Colwell RR, Grimont PAD, Kandler O, et al. (1987) Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol 37: 463–464.
68. Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, et al. (2005) Genotypic diversity within a natural coastal bacterioplankton population. Science 307: 1311–1313.
69. Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, et al. (2004) Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430: 551–554.
70. Lawrence JG, Ochman H (1998) Molecular archaeology of the *Escherichia coli* genome. Proc Natl Acad Sci U S A 95: 9413–9417.
71. Scheffer M, Rinaldi S, Huisman J, Weissing FJ (2003) Why plankton communities have no equilibrium: Solutions to the paradox. Hydrobiologia 491: 9–18.
72. Hutchinson GE (1961) The paradox of the plankton. Am Nat 95: 137–145.
73. Cohan F (2002) Concepts of bacterial biodiversity for the age of genomics. In: Fraser CM, Read TD, Nelson KE, editors. Microbial genomes. Totowa (New Jersey): Humana Press. pp. 175–194.
74. Hacker J, Blum-Oehler G, Hochhut B, Dobrindt U (2003) The molecular basis of infectious diseases: Pathogenicity islands and other mobile genetic elements. A review. Acta Microbiol Immunol Hung 50: 321–330.
75. McCann KS (2000) The diversity-stability debate. Nature 405: 228–233.
76. Fuller NJ, West NJ, Marie D, Yallop M, Rivlin T, et al. (2005) Dynamics of community structure and phosphate status of picocyanobacterial populations in the Gulf of Aqaba, Red Sea. Limnol Oceanogr 50: 363–375.
77. Gill SR, Pop M, Deboy RT, Eckburg PB, Turnbaugh PJ, et al. (2006) Metagenomic analysis of the human distal gut microbiome. Science 312: 1355–1359.
78. Brown MV, Schwalbach MS, Hewson I, Fuhrman JA (2005) Coupling 16S-ITS rDNA clone libraries and automated ribosomal intergenic spacer analysis to show marine microbial diversity: Development and application to a time series. Environ Microbiol 7: 1466–1479.
79. Garcia-Martinez J, Rodriguez-Valera F (2000) Microdiversity of uncultured marine prokaryotes: The SAR11 cluster and the marine Archaea of Group I. Mol Ecol 9: 935–948.
80. Jenkins BD, Steward GF, Short SM, Ward BB, Zehr JP (2004) Fingerprinting Diazotroph communities in the Chesapeake Bay by using a DNA macroarray. Appl Environ Microbiol 70: 1767–1776.
81. Legeckis R (1988) Upwelling off the Gulfs of Panama and Papagayo in the tropical Pacific during March 1985. J Geophys Res 93: 15485–15489.
82. McCreary JP, Lee HS, Enfield DB (1989) The response of the coastal ocean to strong offshore winds: With application to circulation in the gulfs of Tehuantepec and Papagayo. J Mar Res 47: 81–109.
83. Palacios DM (2003) Oceanographic conditions around the Galápagos Archipelago and their influence on Cetacean community structure. [PhD diss]. Corvallis: Oregon State University. 178 p.
84. Christian JR, Lewis MR, Karl DM (1997) Vertical fluxes of carbon, nitrogen, and phosphorus in the North Pacific Subtropical Gyre near Hawaii. J Geophys Res 102: 15667–15677.
85. Doney SC, Abbott MR, Cullen JJ, Karl DM, Rothstein L (2004) From genes to ecosystems: The ocean's new frontier. Front Ecol Environ 2: 457–466.
86. McGillicuddy DJ, Anderson LA, Doney SC, Maltrud ME (2003) Eddy-driven sources and sinks of nutrients in the upper ocean: Results from a 0.1° resolution model of the North Atlantic. Global Biogeochem Cycles 17: 1035.
87. van der Staay SYM, van der Staay GWM, Guillou L, Vaulot D, Claustre H, et al. (2000) Abundance and diversity of prymnesiophytes in the picoplankton community from the equatorial Pacific Ocean inferred from 18S rDNA sequences. Limnol Oceanogr 45: 98–109.
88. Blanchot J, Charpy L, Borgne RL (1989) Size composition of particulate organic matter in the lagoon of Tikehau Atoll (Tuiamotu Archipelago). Mar Biol 102: 329–339.
89. Torréton JP, Dufour P (1996) Bacterioplankton production determined by DNA synthesis, protein synthesis and frequency of dividing cells in Tuamotu atoll lagoons and surrounding ocean. Microb Ecol 32: 185–202.
90. Torréton JP, Dufour P (1996) Temporal and spatial stability of bacterioplankton biomass and productivity in an atoll lagoon. Aquat Microb Ecol 11: 251–261.
91. Rutala WA, Weber DJ (1997) Uses of inorganic hypochlorite (bleach) in health-care facilities. Clin Microbiol Rev 10: 597–610.
92. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning. A laboratory manual. Cold Spring Harbor (NY): Cold Spring Laboratory Press.
93. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74: 5463–5467.
94. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
95. Felsenstein J (1989) PHYLIP: Phylogeny Inference Package (Version 3.2). Cladistics 5: 164–166.
96. Bingham J, Sudarsanam S (2000) Visualizing large hierarchical clusters in hyperbolic space. Bioinformatics 16: 660–661.
97. Bray JR, Curtis JT. (1957) An ordination of upland forest communities of southern Wisconsin. Ecol Monogr 27: 325–349.
98. Myers G (1999) A fast bit-vector algorithm for approximate string matching based on dynamic programming. J ACM 46: 395–415.
99. Penn K, Wu D, Eisen JA, Ward N (2006) Characterization of bacterial communities associated with deep-sea corals on Gulf of Alaska Seamounts. Appl Environ Microbiol 72: 1680–1683.
100. Ludwig W, Strunk O, Westram R, Richter L, Meier H, et al. (2004) ARB: A software environment for sequence data. Nucleic Acids Res 32: 1363–1371.
101. Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. Genome Biol 3: REVIEWS0003.
102. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41.
103. R Development Core Team (2004) R: A language and environment for statistical computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.
104. Gomez-Consarnau L, Gonzalez JM, Coll-Llado M, Gourdon P, Pascher T, et al. (2007) Light stimulates growth of proteorhodopsin-containing marine Flavobacteria. Nature 445: 210–213.

Note Added in Proof

Recently, Gomez-Consarnau et al. provided credible evidence for the biological role of proteorhodopsins [104]. These results indicate that proteorhodopsins blur the line between heterotrophic and autotrophic microbes by allowing a wide range of organisms to harness light energy for respiration and growth. This reinforces the notion that the differential distribution of proteorhodopsin variants identified here reflects functional adaptation to the wavelengths of available light. Furthermore, these adaptations may be driven by the makeup of the microbial community. Thus, these distributional differences could reflect competition between microbes for light resources.