Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1]* Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

We have applied "whole-genome shotgun sequencing" to microbial populations collected en masse on tangential flow and impact filters from seawater samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. We have identified over 1.2 million previously unknown genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.

Microorganisms are responsible for most of the biogeochemical cycles that shape the environment of Earth and its oceans. Yet, these organisms are the least well understood on Earth, as the ability to study and understand the metabolic potential of microorganisms has been hampered by the inability to generate pure cultures. Recent studies have begun to explore environmental bacteria in a culture-independent manner by isolating DNA from environmental samples and transforming it into large insert clones. For example, a previously unknown light-driven proton pump, proteorhodopsin, was discovered within a bacterial artificial chromosome (BAC) from the genome of a SAR86 ribotype (*1*), and soil microbial DNA libraries have been constructed and screened for specific activities (*2*).

Here we have applied whole-genome shotgun sequencing to environmental-pooled DNA samples to test whether new genomic approaches can be effectively applied to gene and species discovery and to overall environmental characterization. To help ensure a tractable pilot study, we sampled in the Sargasso Sea, a nutrient-limited, open ocean environment. Further, we concentrated on the genetic material captured on filters sized to isolate primarily microbial inhabitants of the environment, leaving detailed analysis of dissolved DNA and viral particles on one end of the size spectrum and eukaryotic inhabitants on the other, for subsequent studies.

The Sargasso Sea. The northwest Sargasso Sea, at the Bermuda Atlantic Time-series Study site (BATS), is one of the best-studied and arguably most well-characterized regions of the global ocean. The Gulf Stream represents the western and northern boundaries of this region and provides a strong physical boundary, separating the low nutrient, oligotrophic open ocean from the more nutrient-rich waters of the U.S. continental shelf. The Sargasso Sea has been intensively studied as part of the 50-year time series of ocean physics and biogeochemistry (*3*, *4*) and provides an opportunity for interpretation of environmental genomic data in an oceanographic context. In this region, formation of subtropical mode water occurs each winter as the passage of cold fronts across the region erodes the seasonal thermocline and causes convective mixing, resulting in mixed layers of 150 to 300 m depth. The introduction of nutrient-rich deep water, following the breakdown of seasonal thermoclines into the brightly lit surface waters, leads to the blooming of single cell phytoplankton, including two cyanobacteria species, *Synechococcus* and *Prochlorococcus*, that numerically dominate the photosynthetic biomass in the Sargasso Sea.

Surface water samples (170 to 200 liters) were collected aboard the RV Weatherbird II from three sites off the coast of Bermuda in February 2003. Additional samples were collected aboard the SV Sorcerer II from "Hydrostation S" in May 2003. Sample site locations are indicated on Fig. 1 and described in table S1; sampling protocols were fine-tuned from one expedition to the next (*5*). Genomic DNA was extracted from filters of 0.1 to 3.0 μm, and genomic libraries with insert sizes ranging from 2 to 6 kb were made as described (*5*). The prepared plasmid clones were sequenced from both ends to provide paired-end reads at the J. Craig Venter Science Foundation Joint Technology Center on ABI 3730XL DNA sequencers (Applied Biosystems, Foster City, CA). Whole-genome random shotgun sequencing of the Weatherbird II samples (table S1, samples 1 to 4) produced 1.66 million reads averaging 818 bp in length, for a total of approximately 1.36 Gbp of microbial DNA sequence. An additional 325,561 sequences were generated from the Sorcerer II samples (table S1, samples 5 to 7), yielding approximately 265 Mbp of DNA sequence.

Environmental genome shotgun assembly. Whole-genome shotgun sequencing projects have traditionally been applied to identify the genome sequence(s) from one particular organism, whereas the approach taken here is intended to capture representative sequence from many diverse organisms simultaneously. Variation in genome size and relative abundance determines the depth of coverage of any particular organism in the sample at a given level of sequencing and has strong implications for both the application of assembly algorithms and for the metrics used in evaluating the resulting assembly. Although we would expect abundant species to be deeply covered and well assembled, species of lower abundance may be represented by only a few sequences. For a single genome analysis, assembly coverage depth in unique regions should approximate a Poisson distribution. The mean of this distribution can be estimated from the observed data, looking at the depth of coverage of contigs generated before any scaffolding. The assembler used in this study, the Celera Assembler (*6*), uses this value to heuristically identify clearly unique regions to form the backbone of the final assembly within the scaffolding phase. However, when the starting material consists of a mixture of genomes of varying abundance, a threshold estimated in this way would classify samples from the most abundant organism(s) as repetitive, due to their greater-than-average depth of coverage, paradoxically leaving the most abundant organisms poorly assembled. We therefore used manual curation of an initial

[1]The Institute for Biological Energy Alternatives, [2]The Center for the Advancement of Genomics, 1901 Research Boulevard, Rockville, MD 20850, USA. [3]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA. [4]The J. Craig Venter Science Foundation Joint Technology Center, 5 Research Place, Rockville, MD 20850, USA. [5]University of Southern California, 223 Science Hall, Los Angeles, CA 90089–0740, USA. [6]Bermuda Biological Station for Research, Inc., 17 Biological Lane, St George GE 01, Bermuda.

*To whom correspondence should be addressed. E-mail: jcventer@tcag.org

assembly to identify a set of large, deeply assembling nonrepetitive contigs. This was used to set the expected coverage in unique regions (to 23×) for a final run of the assembler. This allowed the deep contigs to be treated as unique sequence when they would otherwise be labeled as repetitive. We evaluated our final assembly results in a tiered fashion, looking at well-sampled genomic regions separately from those barely sampled at our current level of sequencing.

The 1.66 million sequences from the Weatherbird II samples (table S1; samples 1 to 4; stations 3, 11, and 13), were pooled and assembled to provide a single master assembly for comparative purposes. The assembly generated 64,398 scaffolds ranging in size from 826 bp to 2.1 Mbp, containing 256 Mbp of unique sequence and spanning 400 Mbp. After assembly, there remained 217,015 paired-end reads, or "mini-scaffolds," spanning 820.7 Mbp as well as an additional 215,038 unassembled singleton reads covering 169.9 Mbp (table S2, column 1). The Sorcerer II samples provided almost no assembly, so we consider for these samples only the 153,458 mini-scaffolds, spanning 518.4 Mbp, and the remaining 18,692 singleton reads (table S2, column 2). In total, 1.045 Gbp of nonredundant sequence was generated. The lack of overlapping reads within the unassembled set indicates that lack of additional assembly was not due to algorithmic limitations but to the relatively limited depth of sequencing coverage given the level of diversity within the sample.

The whole-genome shotgun (WGS) assembly has been deposited at DDBJ/EMBL/GenBank under the project accession AACY00000000, and all traces have been deposited in a corresponding TraceDB trace archive. The version described in this paper is the first version, AACY01000000. Unlike a conventional WGS entry, we have deposited not just contigs and scaffolds but the unassembled paired singletons and individual singletons in order to accurately reflect the diversity in the sample and allow searches across the entire sample within a single database.

Genomes and large assemblies. Our analysis first focused on the well-sampled genomes by characterizing scaffolds with at least 3× coverage depth. There were 333 scaffolds comprising 2226 contigs and spanning 30.9 Mbp that met this criterion (table S3), accounting for roughly 410,000 reads, or 25% of the pooled assembly data set. From this set of well-sampled material, we were able to cluster and classify assemblies by organism; from the rare species in our sample, we used sequence similarity based methods together with computational gene finding to obtain both qualitative and quantitative estimates of genomic and functional diversity within this particular marine environment.

We employed several criteria to sort the major assembly pieces into tentative organism "bins"; these include depth of coverage, oligonucleotide frequencies (*7*), and similarity to previously sequenced genomes (*5*). With these techniques, the majority of sequence assigned to the most abundant species (16.5 Mbp of the 30.9 Mb in the main scaffolds) could be separated based on several corroborating indicators. In particular, we identified a distinct group of scaffolds representing an abundant population clearly related to *Burkholderia* (fig. S2) and two groups of scaffolds representing two distinct strains closely related to the published *Shewanella oneidensis* genome (*8*) (fig. S3). There is a group of scaffolds assembling at over 6× coverage that appears to represent the genome of a SAR86 (table S3). Scaffold sets representing a conglomerate of *Prochlorococcus* strains (Fig. 2), as well as an uncultured marine archaeon, were also identified (table S3; Fig. 3). Additionally, 10 putative mega plasmids were found in the main scaffold set, covered at depths ranging from 4× to 36× (indicated with shading in table S3 with nine depicted in

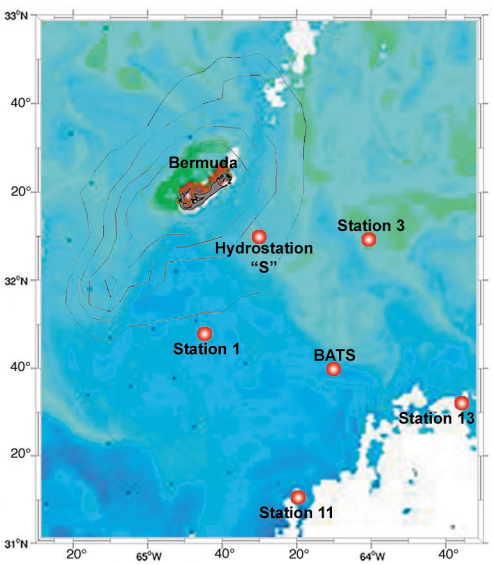

Fig. 1. MODIS-Aqua satellite image of ocean chlorophyll in the Sargasso Sea grid about the BATS site from 22 February 2003. The station locations are overlain with their respective identifications. Note the elevated levels of chlorophyll (green color shades) around station 3, which are not present around stations 11 and 13.

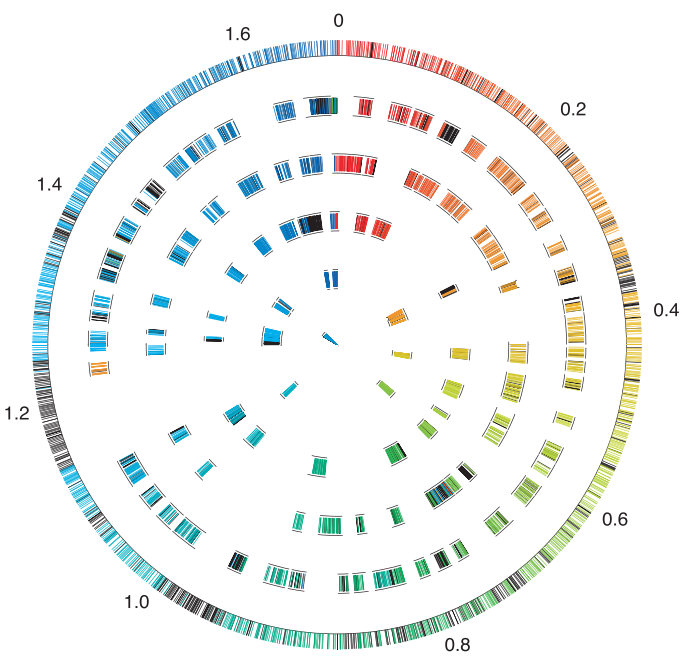

Fig. 2. Gene conservation among closely related *Prochlorococcus*. The outermost concentric circle of the diagram depicts the competed genomic sequence of *Prochlorococcus marinus* MED4 (*11*). Fragments from environmental sequencing were compared to this completed *Prochlorococcus* genome and are shown in the inner concentric circles and were given boxed outlines. Genes for the outermost circle have been assigned psuedospectrum colors based on the position of those genes along the chromosome, where genes nearer to the start of the genome are colored in red, and genes nearer to the end of the genome are colored in blue. Fragments from environmental sequencing were subjected to an analysis that identifies conserved gene order between those fragments and the completed *Prochlorococcus* MED4 genome. Genes on the environmental genome segments that exhibited conserved gene order are colored with the same color assignments as the *Prochlorococcus* MED4 chromosome. Colored regions on the environmental segments exhibiting color differences from the adjacent outermost concentric circle are the result of conserved gene order with other MED4 regions and probably represent chromosomal rearrangements. Genes that did not exhibit conserved gene order are colored in black.

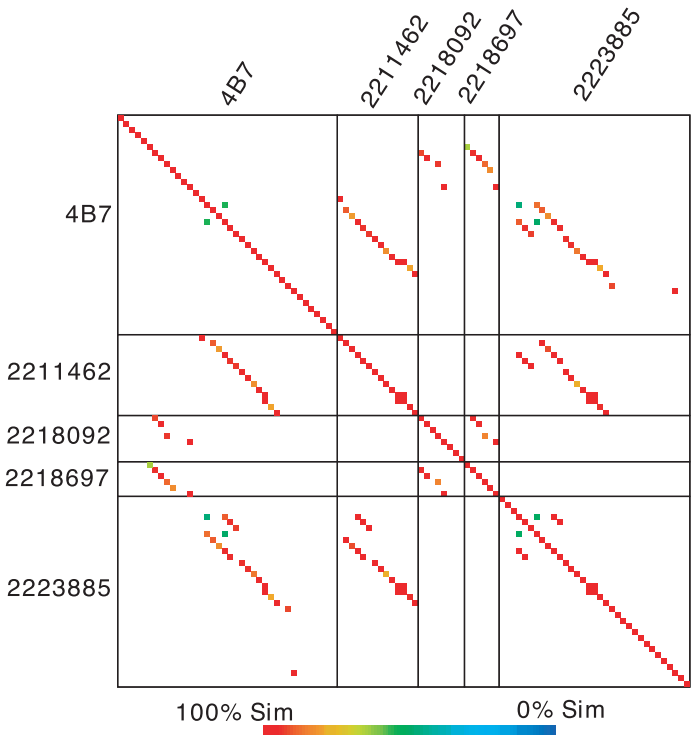

Fig. 3. Comparison of Sargasso Sea scaffolds to Crenarchaeal clone 4B7. Predicted proteins from 4B7 and the scaffolds showing significant homology to 4B7 by tBLASTx are arrayed in positional order along the *x* and *y* axes. Colored boxes represent BLASTp matches scoring at least 25% similarity and with an e value of better than 1e-5. Black vertical and horizontal lines delineate scaffold borders.

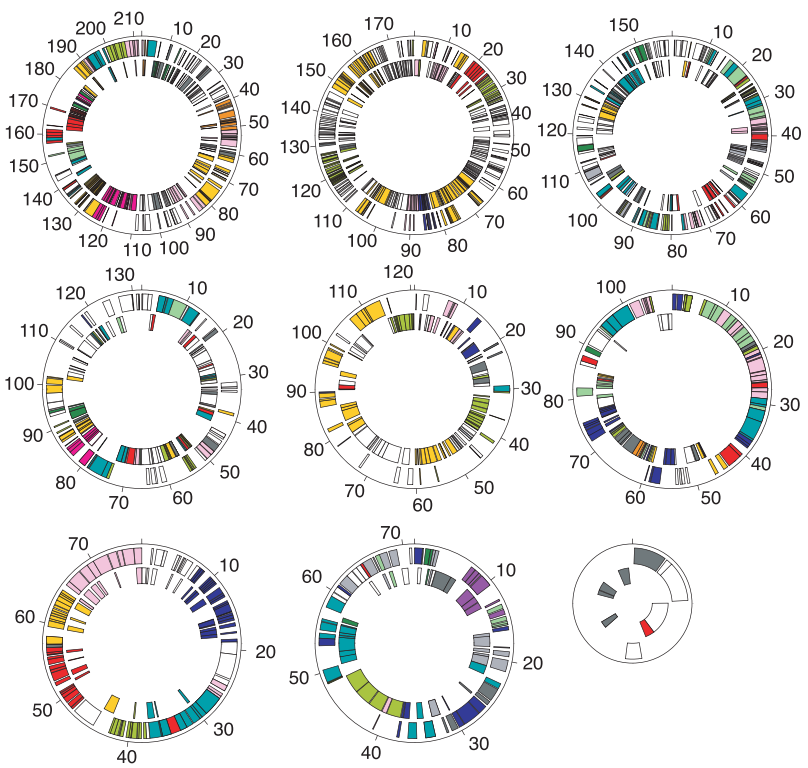

Fig. 4. Circular diagrams of nine complete megaplasmids. Genes encoded in the forward direction are shown in the outer concentric circle; reverse coding genes are shown in the inner concentric circle. The genes have been given role category assignment and colored accordingly: amino acid biosynthesis, violet; biosynthesis of cofactors, prosthetic groups, and carriers, light blue; cell envelope, light green; cellular processes, red; central intermediary metabolism, brown; DNA metabolism, gold; energy metabolism, light gray; fatty acid and phospholipid metabolism, magenta; protein fate and protein synthesis, pink; purines, pyrimidines, nucleosides, and nucleotides, orange; regulatory functions and signal transduction, olive; transcription, dark green; transport and binding proteins, blue-green; genes with no known homology to other proteins and genes with homology to genes with no known function, white; genes of unknown function, gray; Tick marks are placed on 10-kb intervals.

Fig. 4). Other organisms were not so readily separated, presumably reflecting some combination of shorter assemblies with less "taxonomic signal," less distinctive sequence, and greater divergence from previously sequenced genomes (*9*).

Discrete species versus a population continuum. The most deeply covered of the scaffolds (21 scaffolds with over 14× coverage and 9.35 Mb of sequence), contain just over 1 single nucleotide polymorphism (SNP) per 10,000 base pairs, strongly supporting the presence of discrete species within the sample. In the remaining main scaffolds (table S3), the SNP rate ranges from 0 to 26 per 1000 bp, with a length-weighted average of 3.6 per 1000 bp. We closely examined the multiple sequence alignments of the contigs with high SNP rates and were able to classify these into two fairly distinct classes: regions where several closely related haplotypes have been collapsed, increasing the depth of coverage accordingly (*10*), and regions that appear to be a relatively homogenous blend of discrepancies from the consensus without any apparent separation into haplotypes, such as the *Prochlorococcus* scaffold region (Fig. 5). Indeed, the *Prochlorococcus* scaffolds display considerable heterogeneity not only at the nucleotide sequence level (Fig. 5) but also at the genomic level, where multiple scaffolds align with the same region of the MED4 (*11*) genome but differ due to gene or genomic island insertion, deletion, rearrangement events. This observation is consistent with previous findings (*12*). For instance, scaffolds 2221918 and 2223700 share gene synteny with each other and MED4 but differ by the insertion of 15 genes of probable phage origin, likely representing an integrated bacteriophage. These genomic differences are displayed graphically in Fig. 2, where it is evident that up to four conflicting scaffolds can align with the same region of the MED4 genome. More than 85% of the *Prochlorococcus* MED4 genome can be aligned with Sargasso Sea scaffolds greater than 10 kb; however, there appear to be a couple of regions of MED4 that are not represented in the 10-kb scaffolds (Fig. 2). The larger of these two regions (PMM1187 to PMM1277) consists primarily of a gene cluster coding for surface polysaccharide biosynthesis, which may represent a MED4-specific polysaccharide absent or highly diverged in our Sargasso Sea *Prochlorococcus* bacteria. The heterogeneity of the *Prochlorococcus* scaffolds suggest that the scaffolds are not derived from a single discrete strain, but instead probably represent a conglomerate assembled from a population of closely related *Prochlorococcus* biotypes.

The gene complement of the Sargasso. The heterogeneity of the Sargasso sequences complicates the identification of microbial genes. The typical approach for microbial annotation, model-based gene finding, relies entirely on training with a subset of manually

identified and curated genes. With the vast majority of the Sargasso sequence in short (less than 10 kb), unassociated scaffolds and singletons from hundreds of different organisms, it is impractical to apply this approach. Instead, we developed an evidence-based gene finder (*5*). Briefly, evidence in the form of protein alignments to sequences in the bacterial portion of the nonredundant amino acid (nraa) data set (*13*) was used to determine the most likely coding frame. Likewise, approximate start and stop positions were determined from the bounding coordinates of the alignments and refined to identify specific start and stop codons. This approach identified 1,214,207 genes covering over 700 MB of the total data set. This represents approximately an order of magnitude more sequences than currently archived in the curated SwissProt database (*14*), which contains 137,885 sequence entries at the time of writing; roughly the same number of sequences as have been deposited into the uncurated REM-TrEMBL database (*14*) since its inception in 1996. After excluding all intervals covered by previously identified genes, additional hypothetical genes were identified on the basis of the presence of conserved open reading frames (*5*). A total of 69,901 novel genes belonging to 15,601 single link clusters were identified. The predicted genes were categorized

Table 1. Gene count breakdown by TIGR role category. Gene set includes those found on assemblies from samples 1 to 4 and fragment reads from samples 5 to 7. A more detailed table, separating Weatherbird II samples from the Sorcerer II samples is presented in the SOM (table S4). Note that there are 28,023 genes which were classified in more than one role category.

TIGR role category	Total genes
Amino acid biosynthesis	37,118
Biosynthesis of cofactors, prosthetic groups, and carriers	25,905
Cell envelope	27,883
Cellular processes	17,260
Central intermediary metabolism	13,639
DNA metabolism	25,346
Energy metabolism	69,718
Fatty acid and phospholipid metabolism	18,558
Mobile and extrachromosomal element functions	1,061
Protein fate	28,768
Protein synthesis	48,012
Purines, pyrimidines, nucleosides, and nucleotides	19,912
Regulatory functions	8,392
Signal transduction	4,817
Transcription	12,756
Transport and binding proteins	49,185
Unknown function	38,067
Miscellaneous	1,864
Conserved hypothetical	794,061
Total number of roles assigned	1,242,230
Total number of genes	1,214,207

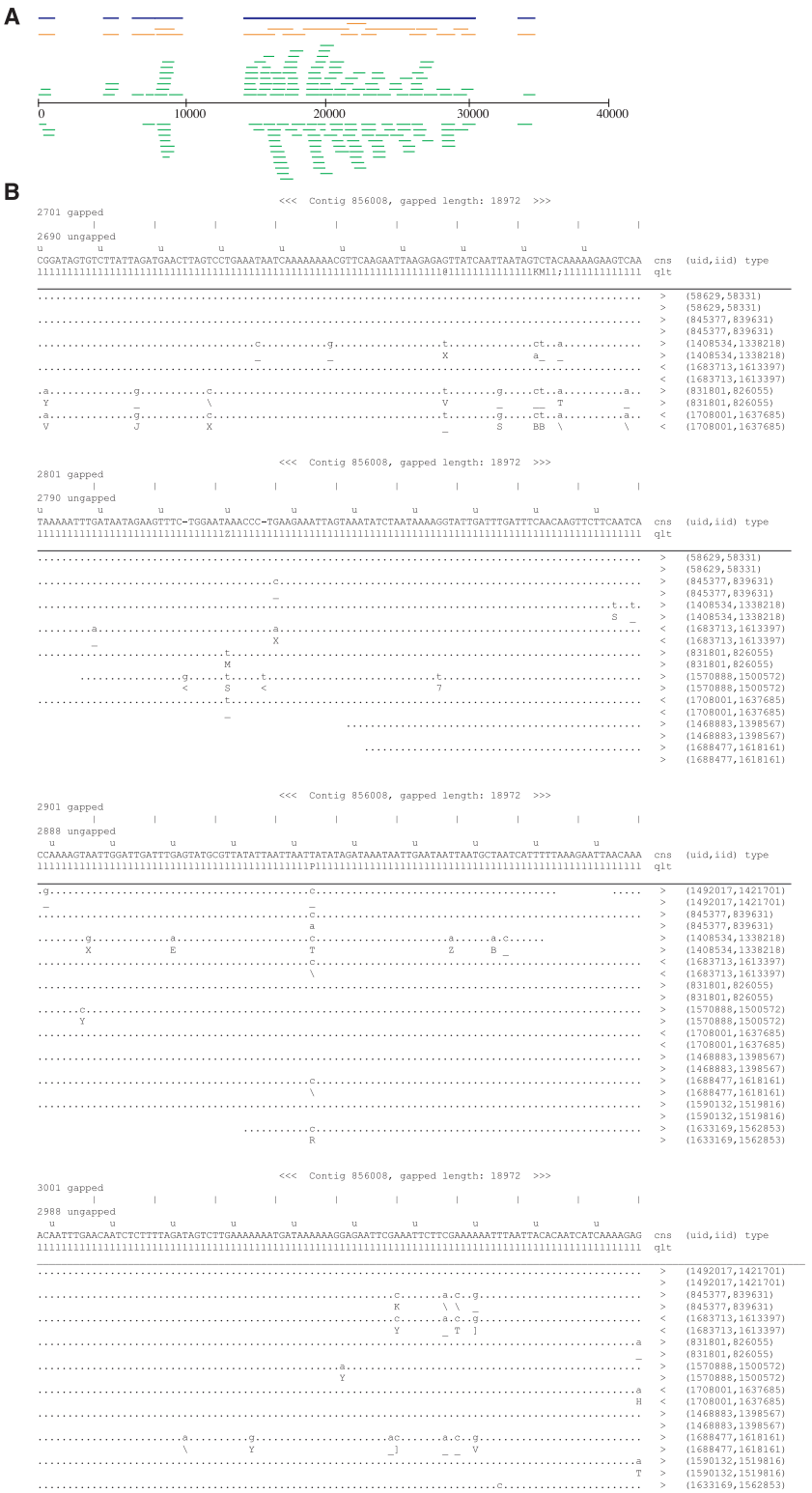

Fig. 5. *Prochlorococcus*-related scaffold 2223290 illustrates the assembly of a broad community of closely related organisms, distinctly nonpunctate in nature. The image represents (**A**) global structure of Scaffold 2223290 with respect to assembly and (**B**) a sample of the multiple sequence alignment. Blue segments, contigs; green segments, fragments; and yellow segments, stages of the assembly of fragments into the resulting contigs. The yellow bars indicate that fragments were initially assembled in several different pieces, which in places collapsed to form the final contig structure. The multiple sequence alignment for this region shows a homogenous blend of haplotypes, none with sufficient depth of coverage to provide a separate assembly.

using the curated TIGR role categories (*5*). A breakdown of predicted genes by category is given in Table 1.

The samples analyzed here represent only specific size fractions of the sampled environment, dictated by the pore size of the collection filters. By our selection of filter pore sizes, we deliberately focused this initial study on the identification and analysis of microbial organisms. However, we did examine the data for the presence of eukaryotic content as well. Although the bulk of known protists are 10 μm and larger, there are some known in the range of 1 to 1.5 μm in diameter [for example, *Ostreococcus tauri* (*15*) and the *Bolidomonas* species (*16*)], and such organisms could potentially work their way through a 0.80 μm prefilter. An initial screening for 18S ribosomal RNA (rRNA), a commonly used eukaryotic marker, identified 69 18S rRNA genes, with 63 of these on singletons and the remaining 6 on very small, lowcoverage assemblies. These 18S rRNAs are similar to uncultured marine eukaryotes and are indicative of a eukaryotic presence but inconclusive on their own. Because bacterial DNA contains a much greater density of genes than eukaryotic DNA, the relative proportion of gene content can be used as another indicator to distinguish eukaryotic material in our sample. An inverse relation was observed between the pore size of the pre-filters and collection filters and the fraction of sequence coding for genes (table S5). This relation, together with the presence of 18S rRNA genes in the samples, is strong evidence that eukaryotic material was indeed captured.

Diversity and species richness. Most phylogenetic surveys of uncultured organisms have been based on studies of rRNA genes using polymerase chain reaction (PCR) with primers for highly conserved positions in those genes. More than 60,000 small subunit rRNA sequences from a wide diversity of prokaryotic taxa have been reported (*17*). However, PCR-based studies are inherently biased, because not all rRNA genes amplify with the same "universal" primers. Within our shotgun sequence data and assemblies, we identified 1164 distinct small subunit rRNA genes or fragments of genes in the Weatherbird II assemblies and another 248 within the Sorcerer II reads (*5*). Using a 97% sequence similarity cutoff to distinguish unique phylotypes, we identified 148 previously unknown phylotypes in our sample when compared against the RDP II database (*17*). With a 99% similarity cutoff, this number increases to 643. Though sequence similarity is not necessarily an accurate predictor of functional conservation and sequence divergence does not universally correlate with the biological notion of "species," defining species (also known as phylotypes) by sequence similarity within the rRNA genes is the accepted standard in studies of uncultured microbes. All sampled rRNAs were then assigned to taxonomic groups using an automated rRNA classification program (*5*). Our samples are dominated by rRNA genes from Proteobacteria (primarily members of the α, β, and γ subgroups) with moderate contributions from Firmicutes (low-GC Gram positive), Cyanobacteria, and species in the CFB phyla (Cytophaga, Flavobacterium, and Bacteroides) (fig. S4A; Fig. 6). The patterns we see are similar in broad outline to those observed by rRNA PCR studies from the Sargasso Sea (*18*), but with some quantitative differences that reflect either biases in PCR studies or differences in the species found in our sample versus those in other studies.

An additional disadvantage associated with relying on rRNA for estimates of species diversity and abundance is the varying number of copies of rRNA genes between taxa (more than an order of magnitude among prokaryotes) (*19*). Therefore, we constructed phylogenetic trees (fig. S4, B to E) using other represented phylogenetic markers found in our data set, [RecA/RadA, heat shock protein 70 (HSP70), elongation factor Tu (EF-Tu), and elongation factor G (EF-G)]. Each marker gene interval in our data set (with a minimum length of 75 amino acids) was assigned to a putative taxonomic group using the phylogenetic analysis described for rRNA. For example, our data set contains over 600 *recA* homologs from throughout the bacterial phylogeny, including representatives of Proteobacteria, low- and high-GC Gram positives, Cyanobacteria, green sulfur and green nonsulfur bacteria, and other groups. Assignment to phylogenetic groups shows a broad consensus among the different phylogenetic markers. For most taxa, the rRNA-based proportion is the highest or lowest in comparison to the other markers. We believe this is due to the large amount of variation in copy number of rRNA genes between species. For example, the rRNA-based estimate of the proportion of γProteobacteria is the highest, while the estimate for cyanobacteria is the lowest, which is consistent with the reports that members of the γ-Proteobacteria frequently have more than five rRNA operon copies, whereas cyanobacteria frequently have fewer than three (*19*).

Just as phylogenetic classification is strengthened by a more comprehensive marker set, so too is the estimation of species richness. In this analysis, we define "genomic" species as a clustering of assemblies or unassembled reads more than 94% identical on the nucleotide level. This cutoff, adjusted for the protein-coding marker genes, is roughly comparable to the 97% cutoff traditionally used for rRNA. Thus

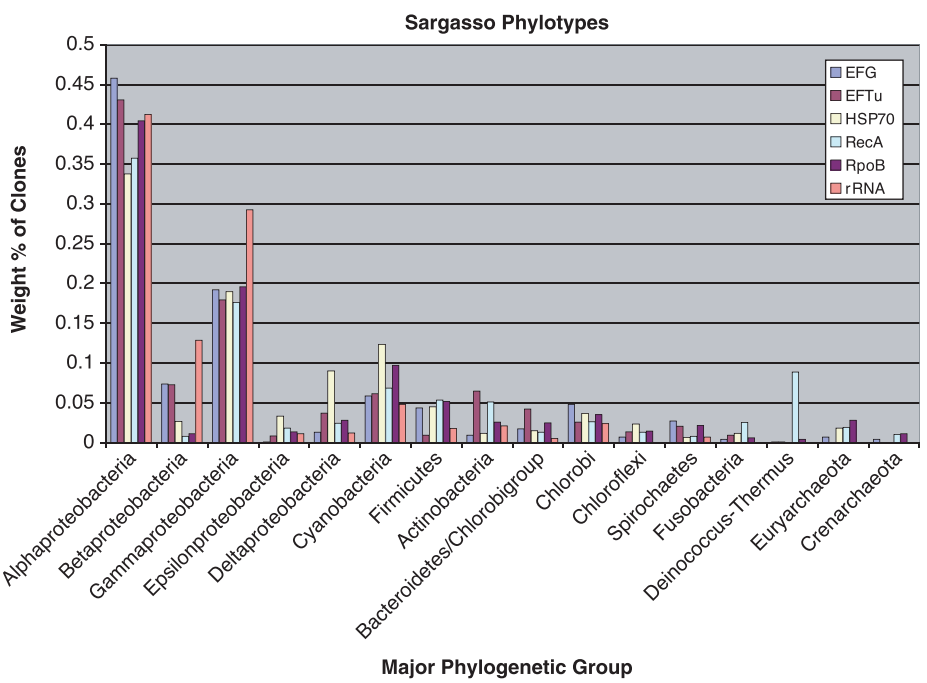

Fig. 6. Phylogenetic diversity of Sargasso Sea sequences using multiple phylogenetic markers. The relative contribution of organisms from different major phylogenetic groups (phylotypes) was measured using multiple phylogenetic markers that have been used previously in phylogenetic studies of prokaryotes: 16S rRNA, RecA, EF-Tu, EF-G, HSP70, and RNA polymerase B (RpoB). The relative proportion of different phylotypes for each sequence (weighted by the depth of coverage of the contigs from which those sequences came) is shown. The phylotype distribution was determined as follows: (i) Sequences in the Sargasso data set corresponding to each of these genes were identified using HMM and BLAST searches. (ii) Phylogenetic analysis was performed for each phylogenetic marker identified in the Sargasso data separately compared with all members of that gene family in all complete genome sequences (only complete genomes were used to control for the differential sampling of these markers in GenBank). (iii) The phylogenetic affinity of each sequence was assigned based on the classification of the nearest neighbor in the phylogenetic tree.

defined, the mean number of species at the point of deepest coverage was 451 [averaged over the six genes analyzed; range 341 to 569 (Table 2)]; this serves as the most conservative estimate of species richness.

Although counts of observed species in a sample are directly obtainable, the true number of distinct species within a sample is almost certainly greater than that which can be observed by finite sequence sampling. That is, given a diverse sample at any given level of sequencing, random sampling is likely to entirely miss some subset of the species at lowest abundance. We considered three approaches to estimating the true diversity: nonparametric methods for small sample corrections (*20*), parametric methods assuming a log-normal distribution of species abundance (*21*), and a novel method based on fitting the observed depth of coverage to a theoretical model of assembly progress for a sample corresponding to a mixture of organisms at different abundances. All three methods agree on a minimum of at least 300 species per sample, with more than 1000 species in the combined sample (*5*). We describe in detail a model based on assembly depth of coverage (*5*). Assuming standard random models for shotgun sequencing, sequencing of an environmental sample should result in depths of sequence coverage reflecting a mixture of Poisson distributions. We computed the empirical distribution of coverage depth at every position in the full set of assemblies (including single fragment contigs, but not counting gaps between contigs), and compared it with hand-constructed mixtures of Poisson distributions. The three depth of coverage–based models shown in Table 3 indicate that there are at least 1800 species in the combined sample, and that a minimum of 12-fold deeper sampling would be required to obtain 95% of the unique sequence. However, these are only lower bounds. The depth of coverage modeling is consistent with as much as 80% of the assembled sequence being contributed by organisms at very low individual abundance and thus would be compatible with total diversity orders of magnitude greater than the lower bound. The assembly coverage data also implies that more than 100 Mbp of genome (i.e., probably more than 50 species) is present at coverage high enough to permit assembly of a complete or nearly complete genome were we to sequence to 5- to 10-fold greater sampling depth.

Taking the well-known marine rRNA clade SAR11 as an example, one can readily get a more tangible view of the diversity within our sample. The SAR11 rRNA group accounts for 26% of all RNA clones that have been identified by culture-independent PCR amplification of seawater (*22*) and has been found in nearly every pelagic marine bacterioplankton community. However, there are very few cultured representatives of this clade (*23*), and little is known about its metabolic diversity. In total, 89 scaffolds and 291 singletons from our data set contain a SAR11 rRNA sequence. However, even with these nearly 400 representatives of SAR11 within our sample, assembly depth of coverage ranges from only from 0.94- to 2.2-fold, and the largest scaffold is quite small at 21,000 bp. This indicates a much more diverse population of organisms than previously attributed to SAR11 based on rRNA PCR methods.

Variability in species abundance. A key advantage to the random sampling approach described here is that it allows assessment of the stoichiometry (that is, estimation of the relative abundance of the dominant organisms). For example, we found that more than half of all assemblies with more than 50 fragments were from organisms that were not at equal relative abundance in all of the samples collected, suggesting widespread "patchiness"

Table 2. Diversity of ubiquitous single copy protein coding phylogenetic markers. Protein column uses symbols that identify six proteins encoded by exactly one gene in virtually all known bacteria. Sequence ID specifies the GenBank identifier for corresponding *E. coli* sequence. Ortholog cutoff identifies BLASTx e-value chosen to identify orthologs when querying the *E. coli* sequence against the complete Sargasso Sea data set. Maximum fragment depth shows the number of reads satisfying the ortholog cutoff at the point along the query for which this value is maximal. Observed "species" shows the number of distinct clusters of reads from the maximum fragment depth column, after grouping reads whose containing assemblies had an overlap of at least 40 bp with > 94% nucleotide identity (single-link clustering). Singleton "species" shows the number of distinct clusters from the observed "species" column that consist of a single read. Most abundant column shows the fraction of the maximum fragment depth that consists of single largest cluster.

Protein	Sequence ID	Ortholog cutoff	Max. fragment depth	Observed "species"	Singleton "species"	Most abundant (%)
AtpD	NTL01EC03653	1e-32	836	456	317	6
GyrB	NTL01EC03620	1e-11	924	569	429	4
Hsp70	NT01EC0015	1e-31	812	515	394	4
RecA	NTL01EC02639	1e-21	592	341	244	8
RpoB	NTL01EC03885	1e-41	669	428	331	7
TufA	NTL01EC03262	1e-41	597	397	307	3

Table 3. Diversity models based on depth of coverage. Each row corresponds to an abundance class of organisms. The first column in each model "fr(asm)" gives the fraction of the assembly consensus modeled due to organisms at an abundance giving "Depth" coverage depth (second column) in the sample. The third column ["E(s)"] gives the fraction of such a genome expected to be sampled. The fourth column ("Genomes") gives the resulting estimated number of genomes in the abundance class.

Model 1				Model 2				Model 3			
fr(asm)	Depth	E(s)	Genomes	fr(asm)	Depth	E(s)	Genomes	fr(asm)	Depth	E(s)	Genomes
0.0055	25	1.00E+00	2.5	0.0055	25	1.00E+00	2.5	0.0055	25	1.00E+00	2.5
0.005	21	1.00E+00	2.3	0.005	21	1.00E+00	2.3	0.005	21	1.00E+00	2.3
0.0035	13	1.00E+00	1.6	0.0035	13	1.00E+00	1.6	0.0035	13	1.00E+00	1.6
0.004	9	1.00E+00	1.8	0.004	9	1.00E+00	1.8	0.004	9	1.00E+00	1.8
0.008	7	9.99E-01	3.6	0.0088	7	9.99E-01	4	0.0088	7	9.99E-01	4
0.0047	6	9.98E-01	2.1	0.0047	6	9.98E-01	2.1	0.0047	6	9.98E-01	2.1
0.01	4	9.82E-01	4.6	0.029	2.4	9.09E-01	14.4	0.029	2.4	9.09E-01	14.4
0.0258	2.4	9.09E-01	12.8	0.096	2	8.65E-01	50	0.097	2	8.65E-01	50.5
0.07	2	8.65E-01	36.4	0.0235	1	6.32E-01	16.7	0.0225	1	6.32E-01	16
0.8635	0.25	2.21E-01	1,756.7	0.06	0.5	3.93E-01	68.6	0.06	0.5	3.93E-01	68.6
				0.76	0.09	8.61E-02	3,973.6	0.66	0.124	1.17E-01	2,546.7
								0.1	0.001	1.00E-03	45,022.5
Total			1824.4				4137.6				47,733

in the bacterioplankton distributions (table S3). A patchy distribution in the abundance of larger eukaryotic planktonic organisms (i.e., phytoplankton and zooplankton) within marine ecosystems has been well documented (*24*, *25*). However, equivalent patchiness in the bacterioplankton (including autotrophic cyanobacteria) has not been well studied, largely due to observations that bacterial numbers appear not to exhibit large spatial or temporal changes within specific oceanographic regimes (*26*, *27*). Our approach provides a means to better elucidate bacterioplankton distributions.

Microheterogeneity of environments and the existence of microbes on easily disrupted particulate matter (such as the tiny bits of debris known as "marine snow" or copepod fecal pellets) is an increasingly wellrecognized component of biology and biogeochemistry of the ocean (*28*–*30*) and is supported by our findings. Within water samples as large as the ones used in this study (~200 to 340 l), there are likely to be numerous particles providing microhabitats for specific bacterioplankton, and disruption during the collection process would allow attached organisms to be captured within our size fraction. For example, unique to sample 1 were scaffolds with 6- and 7-fold coverage that seem to represent two distinct strains of *Shewanella*, comprising roughly 14.4% of the sample 1 data, and a group of 21-fold coverage scaffolds with similarity to a *Burkholderia* spp. (table S3), comprising roughly 38.5% of the sample 1 data. These findings were initially quite surprising; *Shewanella* is an abundant genus in aquatic, usually nutrient-rich environments (*31*), and *Burkholderia* is typically found only in terrestrial environmental samples. Our rigorous protocols (*5*) lead us to believe these sequences do not represent contamination. The finding of *Burkholderia* in samples taken from marine mammals (*32*) provides a plausible mechanism for the transfer of this typically terrestrial or coastal organism to the open ocean, and microhabitats such as marine snow could certainly provide point sources of protein material for growth of these species.

Some organisms expected to be more common at lower depths were also represented in our surface samples in significant numbers. For example, scaffolds from an uncultured marine archaeon were identified (based on a 16S rDNA sequence and two 5S ribosomal sequences). A more inclusive set of 18 scaffolds covering 396 kb were identified and putatively designated as being derived from Archaea (*33*) and examined further. Several of these scaffolds (Fig. 3) were similar to (>90% identity) and have synteny with a genomic fragment that was derived from planktonic marine Archaea clone 4B7 (*34*), collected at a depth of 200 m in the eastern North Pacific. Further, the collection of scaffolds related to clone 4B7 can be separated from another group of archaeal-like assemblies, with sequence depth approximately fourfold, with between 40 and 60% sequence identity to sequences derived from various completely sequenced archaea including *Pyrococcus horikoshii* and *Sulfolobus solfataricus*. Among the activities encoded in these archaeal assemblies are pathways for chorismate biosynthesis, oxoacid-ferrodoxin oxidoreductases, and glutamate and phosphoglycerate dehydrogenases. The separation of the archaeal assemblies into two groups of ~2× and ~4× coverage, respectively, the presence of at least two distinct groups of ribosomal proteins, and the association of the assemblies with different species within this phylum suggest that these assemblies represent genomic sequences from at least two distinct archaeal lineages.

Plasmids. Interesting as vehicles of lateral gene transfer and as reservoirs of genes for important environmental adaptations, plasmids are also abundant in the assembly. In the scaffold set, we find evidence for six putative plasmids larger than 100 kbp in length, two plasmids ~70 to 80 kbp, and two plasmids under 10 kbp (table S3; Fig. 4). In addition to genes for plasmid replication, putative genes for arsenate, mercury, copper, and cadmium resistance were also found. Additionally, putative conjugal transfer systems (type IV secretion) and putative plasmid transfer genes (i.e., *tra* genes) were identified, suggesting that at least some of the plasmids from this sample may be readily mobilized. One of the plasmids encodes homologs of the UmuCD DNA damage-induced DNA polymerase of *Escherchia coli*. In the Sargasso Sea, UmuCD genes could play a role in ultraviolet (UV) resistance (the UmuC DNA polymerase of *E. coli* is able to replicate DNA even when it has been damaged by UV irradiation). Alternatively, these genes could convey a mutator phenotype on their hosts, as other plasmid-encoded UmuCD homologs have been shown to do in various bacterial species. Further studies on the specificity of the putative trace metal resistance genes will further our understanding of the role of microbes in trace metal cycling in the oligotrophic ocean.

Bacteriophage. Due to the experimental design and approach in this sampling, only double-stranded DNA bacteriophages were observable. Searching the assembly against known phage sequence databases (*5*), we identified 71 scaffolds greater than 10 kb in length containing identifiable clusters of phage genes, with *Burkholderia*- and *Shewanella*-associated scaffolds accounting for roughly one-third of these. Because each functional phage contains one copy of the major capsid, portal protein, and large terminase gene (among others), we used matches to these to determine that there are at least 50 major capsid, portal, and/or large terminase groupings genes in the scaffolds and three times as many of these in the singletons.

The primary phage sequences identified were matches to lytic phages and the archaeal phage, *Sulfolobus islandicus* filamentous virus (SIFV), with the marine *Vibrio* T-4–like KVP40 phage that appears most frequently. The distribution of best hits across viral families is similar to observations in a recent study of samples from Mission Bay and Scripps Pier (*35*). Among the matches to lytic phages are proteins classically involved in recycling or scavenging phosphate (e.g., ribonucleoside or -tide reductases, phoH, thymidine synthetases, and endo- or exonucleases). Historically, 1 to 4% of planktonic free-living bacteria are visibly infected (*36*). On the basis of gene order, no complete replicative phage genomes could be identified to corroborate this number. Given the high diversity of phages in the population, it is reasonable to hypothesize that our current data set was not sampled deeply enough to completely assemble a clonal phage.

Photobiology of the Sargasso Sea. Microbial photosynthesis in the Sargasso Sea is thought to be dominated by the cyanobacteria *Prochlorococcus* and *Synechococcus.* However, more than 90% of the cyanobacterial-like scaffolds across our samples appear to be derived from *Prochlorococcus* rather than *Synechococcus* species. This probably reflects, in part, the relative distribution of these two genera in different gradients of open ocean waters (*37*) as well as the larger average size of *Synechococcus* cells that would have excluded many from 0.2- to 0.8-μm size fraction. The diversity of homologs of ribulose bis-phosphate carboxylase (RubisCO) was low in our samples (e.g., we only found 37 BLASTp matches to the large subunit of RubisCO in the predicted gene set). This was surprising, given that RubisCO is the key enzyme of the Calvin Cycle, the main process for carbon fixation in plants, cyanobacteria, and many types of photosynthetic bacteria, but might simply reflect the dominance of a single group of autotrophs (i.e., *Prochlorococcus*) in the size fractions studied here.

The recent discovery of a homolog of bacteriorhodopsin in an uncultured γ-Proteobacteria from the Monterey Bay revealed the basis of a form of phototrophy in marine systems (*38*) that was observed previously by oceanographers (*39*, *40*). Bacteriorhodopsin allows for the coupling of light energy harvesting and carbon cycling in the ocean through a non–chlorophyll-based pathway. Environmental culture-independent gene surveys with PCR have since shown that proteorhodopsin is not limited to a single oceanographic location and revealed some 67 additional closely related proteorhodopsin homologs (*41*). More than 650 rhodopsin homologs were identified within the Weatherbird II samples (samples 1 to 4), and an additional 132 were identified in the Sorcerer II samples (samples 5 to 7), increasing the total number of proteorhodopsin examples by almost an order of magnitude. Phylogenetic analysis of all available rhodopsin-like genes reveals that the additional sequences we identified in the Sargasso

Sea populations also represent a great increase in the sequence diversity of proteorhodopsin-like genes, in comparison to what had been seen previously (Fig. 7). In total, we identified 13 distinct subfamilies of rhodopsin-like genes. These include four families of proteins known from cultured organisms (halorhodpsin, bacterorhodopsin, sensory opsins, and fungal opsin), and nine families from uncultured species of which seven are only known from the Sargasso Sea populations. Of these subfamilies, many are quite distant from either proteorhodopsin or from rhodopsins in cultured species. In situ protein profiles for this area are not available, so expression levels of these genes remain to be determined. The capability of our approach to identify genes of particular biological function in context with phylogenetically informative markers allowed us to look beyond the diversity of the rhodopsins themselves to the phylogenetic diversity of the organisms containing the identified rhodopsins. Analysis of the scaffolds containing both a rhodopsin and a phylogenetic marker demonstrate that the rhodopsins are found well outside the groups of proteobacteria where they had previously been discovered (*39*). For example, on one scaffold we observed these as well as a DNA-directed RNA polymerase σ subunit originating from the CFB group (fig. S10). Such findings are consistent with the hypothesis that rhodopsins are widespread and abundant in marine planktonic bacteria; however, this study demonstrates that the distribution of rhodopsins is even greater than previously established.

Conclusion. We demonstrate here that shotgun sequencing provides a wealth of phylogenetic markers that can be used to assess the phylogenetic diversity of a sample with more power than conventional PCR-based rRNA studies allow. We find that although the qualitative picture that emerges is similar to that based on analysis of rRNA genes alone, the quantitative picture is significantly different for certain taxonomic groups. Further, just as shotgun sequencing provides a relatively unbiased way to examine species diversity, it can also allow a relatively unbiased identification of the diversity of genes in particular gene families. Using these results to drive gene expression surveys and physiological studies provides a more comprehensive approach to environmental biology than previously available.

We are a long way from a full understanding of the biology of the organisms sampled here, but this relatively small genomics study demonstrates areas where important insights may be gained. For example, because it has been believed that only members of the bacterial domain were capable of oceanic nitrification, it is interesting to note that an ammonium monooxygenase gene was found on an archaeal-associated scaffold within our data set. Though ultraviolet (UV) light in the surface ocean is likely to inhibit ammonia oxidation by chemoautotrophs (*42*, *43*), nitrite concentrations are nearly as great as nitrate concentrations during the period of winter mixing when our samples were collected. Together with our finding, this suggests that the high nitrite concentrations may result from archaeal ammonium oxidation, which would not be inhibited by UV light. Our study also sheds light on the mechanisms by which the Sargasso Sea inhabitants may efficiently utilize the available phosphorus in this extremely phosphorus-limited environment. Transport mechanisms for phosphonates that were recently been identified in the *Prochlorococcus* and *Synechococcus* genomes (*11*, *44*) appear in our data set, in addition to a large number of published genes responsible for utilization of polyphosphates and pyrophosphates, which are also available in this environment. Further, the high-affinity inorganic phosphorus transporter PstS (*45*, *46*), one of the Pho Regulon group of genes associated with *Synechococcus*, was also identified within our samples. Indeed, it would appear that a great variety of the microbial population present at the time of this sampling possessed mechanisms to utilize the dominant form of phosphorus in the Sargasso Sea. This opens a door to further studies of expression and physiology.

There are obviously challenges to genome assembly in the environmental context. Particularly well-conserved regions (e.g., 16S rRNA) may assemble across species, similar to an over-collapsed repeat in a standard eukaryotic assembly project. Such over-collapsing, which results in a pattern of inconsistent mate-pair relationships, causes scaffolds to break, resulting in suboptimal assembly. Similarly, two closely related organisms may contain stretches of high similarity that co-assemble, whereas other, more dissimilar regions assemble appropriately by organism. A third challenge is that posed by the low sequence coverage of the less abundant organisms, which may call for more aggressive use of mate links to achieve assembly. With deeper coverage, multiple mate links between contigs provide greater statistical support for scaffolds. At lower coverage, there generally will be fewer confirming mate links between contigs, and therefore less statistical support for the scaffolds. Mismated mate reads, while rare, occur with some probability and, in a low coverage scenario, could contribute to misassembly. None of these issues is unique to environmental assembly, and advances in these areas will find application in single organism assembly projects as well. Our assembly results demonstrate that despite these challenges, one can apply whole-genome assembly algorithms successfully in an environmental context, with the only real limitation being the sequencing cost.

Although it may be expensive to sequence at great depths at this particular point in time, the economics of sequencing is rapidly changing. Were sequencing costs not to drop dramatically (as we fully expect they will), normalization procedures to remove over-sampled organisms from the DNA sample to better study the

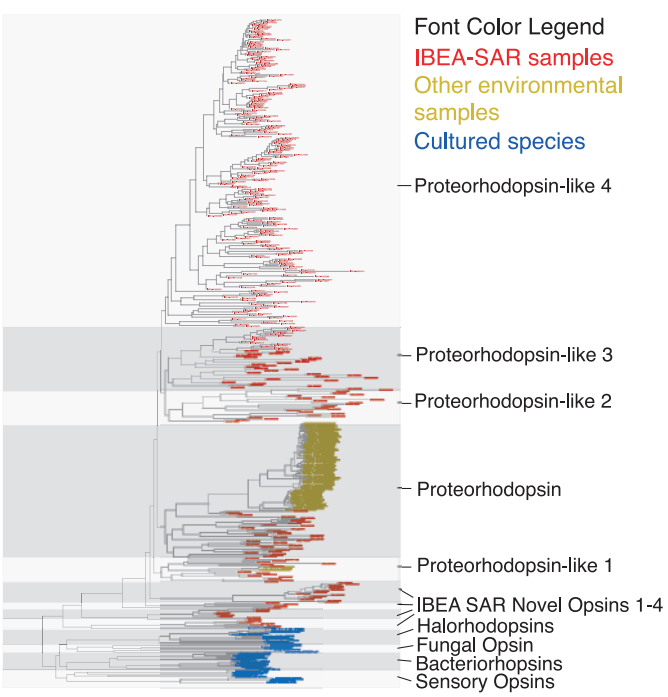

Fig. 7. Phylogenetic tree of rhodopsinlike genes in the Sargasso Sea data along with all homologs of these genes in GenBank. The sequences are colored according to the type of sample in which they were found: blue, cultured species; yellow, sequences from uncultured organisms in other environmental samples; and red, sequences from uncultured species in the Sargasso Sea. The tree was divided into what we propose are distinct subfamilies of sequences, which are labeled on the right. The tree was constructed as follows: (i) All homologs of halorhodopsin were identified in the predicted proteins from the Sargasso Sea assemblies using BLASTp searches with representatives of previously identified halorhodpsinlike protein families as query sequences. (ii) All sequences greater than 75 amino acids in length were aligned to each other using CLUSTALw, and a neighbor-joining phylogenetic tree was inferred using the protdist and neighbor programs of Phylip.

remainder or to isolate DNA of particular interest based on the initial survey make this a very practical method for examining the genomic content of environmental communities.

References and Notes

1. O. Beja *et al.*, *Science* **289**, 1902 (2000).
2. M. R. Rondon *et al.*, *Appl. Environ. Microbiol.* **66**, 2541 (2000).
3. M. H. Conte, *Oceanus* **40**, 15 (1998).
4. D. K. Steinberg *et al.*, *Deep-Sea Res. II* **48**, 1405 (2001).
5. Materials and Methods are available as supporting online material (SOM) on *Science* Online.
6. E. W. Myers *et al.*, *Science* **287**, 2196 (2000).
7. S. Karlin, C. Burge, *Trends Genet.* **11**, 283 (1995).
8. J. F. Heidelberg *et al.*, *Nature Biotechnol.* **20**, 1118 (2002).
9. For example, *Burkholderia* may be the only well-represented genus of β-Proteobacteria, with highly distinctive GC content, whereas there may be many different γ-Proteobacteria.
10. These putative strains can be easily separated post-hoc by reassembly with more stringent overlap criteria. See SOM for details.
11. G. Rocap *et al.*, *Nature* **424**, 1042 (2003).
12. B. Palenik, *Appl. Environ. Microbiol.* **60**, 3212 (1994).
13. Searches were performed on the sequences longer than 40 bp with BLASTx against the entire nraa data set archived at GenBank.
14. B. Boeckmann *et al.*, *Nucleic Acids Res.* **31**, 365 (2003).
15. M. J. Chretiennotdinet *et al.*, *Phycologia* **34**, 285 (1995).
16. L. Guillou *et al.*, *J. Phycol.* **35**, 368 (1999).
17. J. R. Cole *et al.*, *Nucleic Acids Res.* **31**, 442 (2003).
18. C. A. Carlson *et al.*, *Aquat. Microb. Ecol.* **30**, 19 (2002).
19. J. A. Klappenbach, J. M. Dunbar, T. M. Schmidt, *Appl. Environ. Microbiol.* **66**, 1328 (2000).
20. A. Chao, *Scand. J. Stat.* **11**, 265 (1984).
21. T. P. Curtis, W. T. Sloan, J. W. Scannell, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 10494 (2002).
22. S. Giovanonni, M. S. Rappe, in *Microbial Ecology of the Oceans*, D. L. Kirchman, Ed. (Wiley, New York, 2000), pp. 47–84.
23. M. S. Rappe, S. A. Connon, K. L. Vergin, S. J. Giovannoni, *Nature* **418**, 630 (2002).
24. J. Raymont, *Plankton and Productivity in the Oceans* (William Clowes Limited, London, ed. 2, 1980).
25. T. Parsons, M. Takahashi, B. Hargrave, *Biological Oceanographic Processes* (BPCC Wheatons Ltd., Exeter, UK, ed. 3, 1984).
26. J. E. Hobbie, R. J. Daley, S. Jasper, *Appl. Environ. Microbiol.* **33**, 1225 (1977).
27. P. del Giorgio, Y. Prairie, D. Bird, *Microb. Ecol.* **34**, 144 (1997).
28. A. Alldredge, Y. Cohen, *Science* **235**, 689 (1987).
29. E. F. DeLong, D. Franks, A. Alldredge, *Limnol. Oceanogr.* **38**, 924 (1993).
30. U. Passow, *Prog. Oceanogr.* **55**, 287 (2002).
31. K. H. Nealson, J. Scott, in *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community*, E. A. Dworkin, Ed. (Springer-Verlag, NY, 2003).
32. C. L. Hicks, R. Kinoshita, P. W. Ladds, *Aust. Vet. J.* **78**, 193 (2000).
33. The criteria for this filtering were that more than 50% of the assembly had a database hit and more than 30% of the sequence had the best hit to archaeal gene sequences.
34. J. L. Stein, T. L. Marsh, K. Y. Wu, H. Shizuya, E. F. DeLong, *J. Bacteriol.* **178**, 591 (1996).
35. M. Breitbart *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14250 (2002).
36. L. M. Proctor, J. A. Fuhrman, *Nature* **343**, 60 (1990).
37. K. K. Cavender-Bares, D. M. Karl, S. W. Chisholm, *Deep-Sea Res. I* **48**, 2373 (2001).
38. O. Beja *et al.*, *Science* **289**, 1902 (2000).
39. J. R. de la Torre *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12830 (2003).
40. O. Beja, E. N. Spudich, J. L. Spudich, M. Leclerc, E. F. DeLong, *Nature* **411**, 786 (2001).
41. G. Sabehi *et al.*, *Environ. Microbiol.* **5**, 842 (2003).
42. M. Guerero, R. Jones, *Mar. Ecol. Prog. Ser.* **141**, 193 (1996).
43. M. Guerrero, R. Jones, *Mar. Ecol. Prog. Ser.* **141**, 183 (1996).
44. B. Palenik *et al.*, *Nature* **424**, 1037 (2003).
45. D. Scanlan *et al.*, *Appl. Environ. Microbiol.* **63**, 2411 (1997).
46. D. Scanlan, W. H. Wilson, *Hydrobiologia* **401**, 149 (1999).

47. The authors would like to acknowledge P. Lethaby, M. Roadman, D. Clougherty, N. Buck, J. Selengut, A. Delcher, M. Pop, H. Koo, R. Doering, M. Wu, J. Badger, K. Moffat, S. Yooseph, E. Kirkness, D. Karl, K. Heidelberg, B. Friedman, H. Kowalski, and the staff of the J. Craig Venter Science Foundation Joint Technology Center. We also acknowledge the help of N. Nelson at UCSB-ICESS for assistance in acquiring the satellite image in Fig. 1. Further, we acknowledge the NSF Division of Ocean Sciences for its ongoing support of the BATS Program and the RV Weatherbird II, and the Department of Energy Genomes to Life program for its support of K. Nealson. This research was supported by the Office of Science (B.E.R.), U.S. Department of Energy grant no. DE-FG02-02ER63453, and the J. Craig Venter Science Foundation. This is contribution No. 1646 of the Bermuda Biological Station for Research, Inc.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1093857/DC1
Materials and Methods
SOM Text
Figs. S1 to S10
Tables S1 to S5
References

20 November 2003; accepted 20 February 2004
Published online 4 March 2004;
10.1126/science.1093857
Include this information when citing this paper.

REPORTS

Approaching the Quantum Limit of a Nanomechanical Resonator

M. D. LaHaye,[1,2] O. Buu,[1,2] B. Camarota,[1,2] K. C. Schwab[1]*

By coupling a single-electron transistor to a high–quality factor, 19.7-megahertz nanomechanical resonator, we demonstrate position detection approaching that set by the Heisenberg uncertainty principle limit. At millikelvin temperatures, position resolution a factor of 4.3 above the quantum limit is achieved and demonstrates the near-ideal performance of the single-electron transistor as a linear amplifier. We have observed the resonator's thermal motion at temperatures as low as 56 millikelvin, with quantum occupation factors of N_{TH} = 58. The implications of this experiment reach from the ultimate limits of force microscopy to qubit readout for quantum information devices.

Since the development of quantum mechanics, it has been appreciated that there is a fundamental limit to the precision of repeated position measurements (*1*). This is a consequence of the Heisenberg uncertainty principle (*2*), which places a limit on the simultaneous knowledge of position x and momentum p: $\Delta x \cdot \Delta p \geq \frac{\hbar}{2}$, where $2\pi \cdot \hbar$ is Planck's constant. When applied to a simple harmonic oscillator of mass m and angular resonant frequency ω_0, this relationship places a limit on the precision of two instantaneous, strong position measurements, what is called the "standard quantum limit," $\Delta x_{SQL} = \sqrt{\frac{\hbar}{2m\omega_0}}$ (*3*).

Although the standard quantum limit captures the physics of the uncertainty principle, it is far from the situation found when one continuously measures the position with a linear detector. Linear amplifiers not only detect and amplify the incoming desired signal but also impose back-action onto the object under study (*4*); the current noise emanating from the input of a voltage preamplifier or the momentum noise imparted to a mirror in an optical interferometer are manifestations of this back-action. The uncertainty principle again appears and places a quantum limit on the minimum possible back-action for a linear amplifier. Previous work (*5*) has concluded that the minimum possible amplifier noise temperature is $T_{QL} = \frac{\hbar\omega_0}{\ln 3 \cdot k_B}$. Applying this result to the continuous readout of a simple harmonic oscillator yields the ultimate position resolution (*6*): $\Delta x_{QL} = \sqrt{\frac{\hbar}{\ln 3 \cdot m\omega_0}} \approx 1.35 \cdot \Delta x_{SQL}$, which

[1]Laboratory for Physical Sciences, 8050 Greenmead Drive, College Park, MD 20740, USA. [2]Department of Physics, University of Maryland, College Park, MD 20740, USA.

*To whom correspondence should be addressed. E-mail: schwab@lps.umd.edu