EVE 161:
Microbial Phylogenomics

Era IV:
Shotgun Metagenomics

UC Davis, Winter 2016
Instructor: Jonathan Eisen

RESEARCH ARTICLE

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.

Methods

- Plasmid library
- Shotgun sequence
- Assembled
- No Major Binning
- Potential “nearly” complete genomes
- Annotation, population analysis, phylogenetic analysis

Sargasso Sea

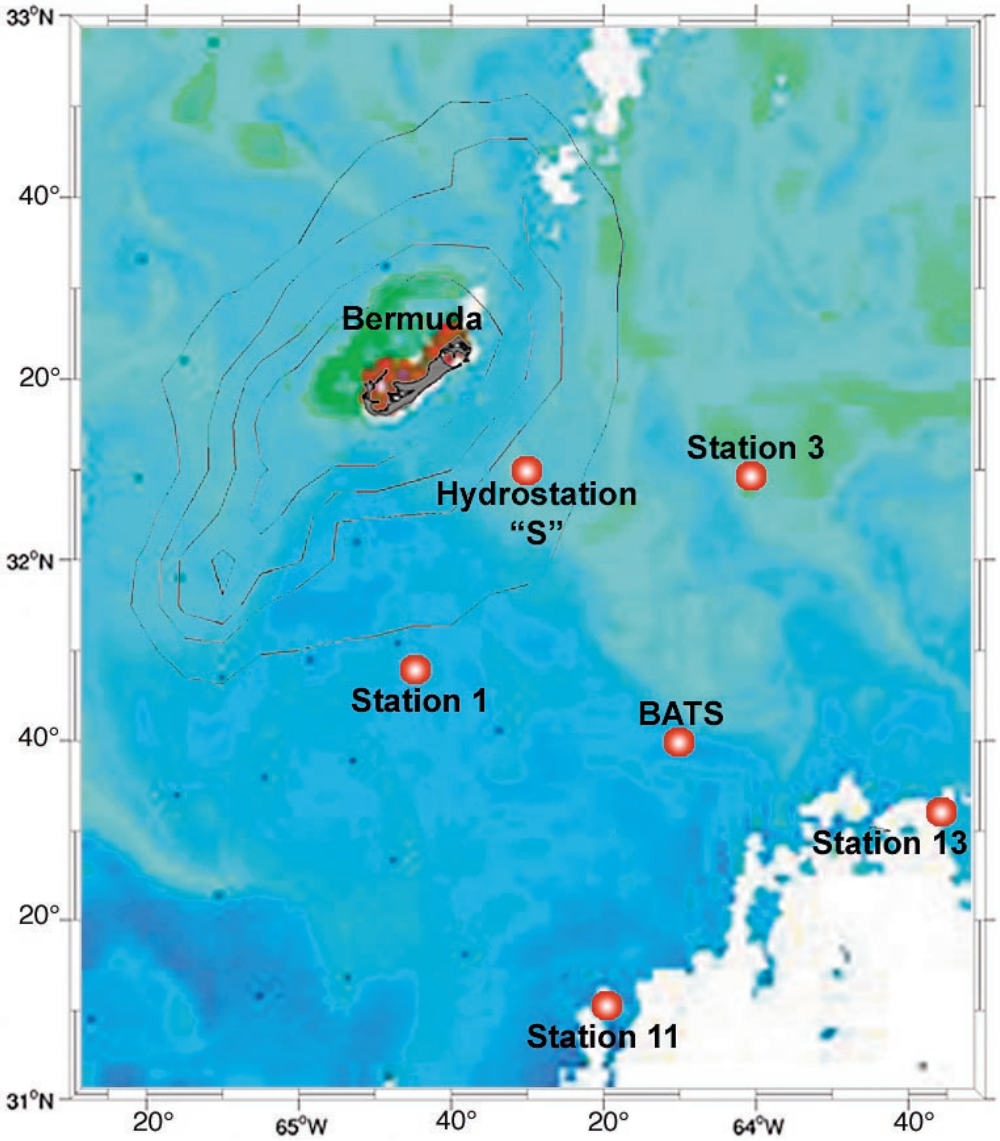


Fig. 1. MODIS-Aqua satellite image of ocean chlorophyll in the Sargasso Sea grid about the BATS site from 22 February 2003. The station locations are overlain with their respective identifications. Note the elevated levels of chlorophyll (green color shades) around station 3, which are not present around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

http://www.sciencemag.org/content/304/5667/66

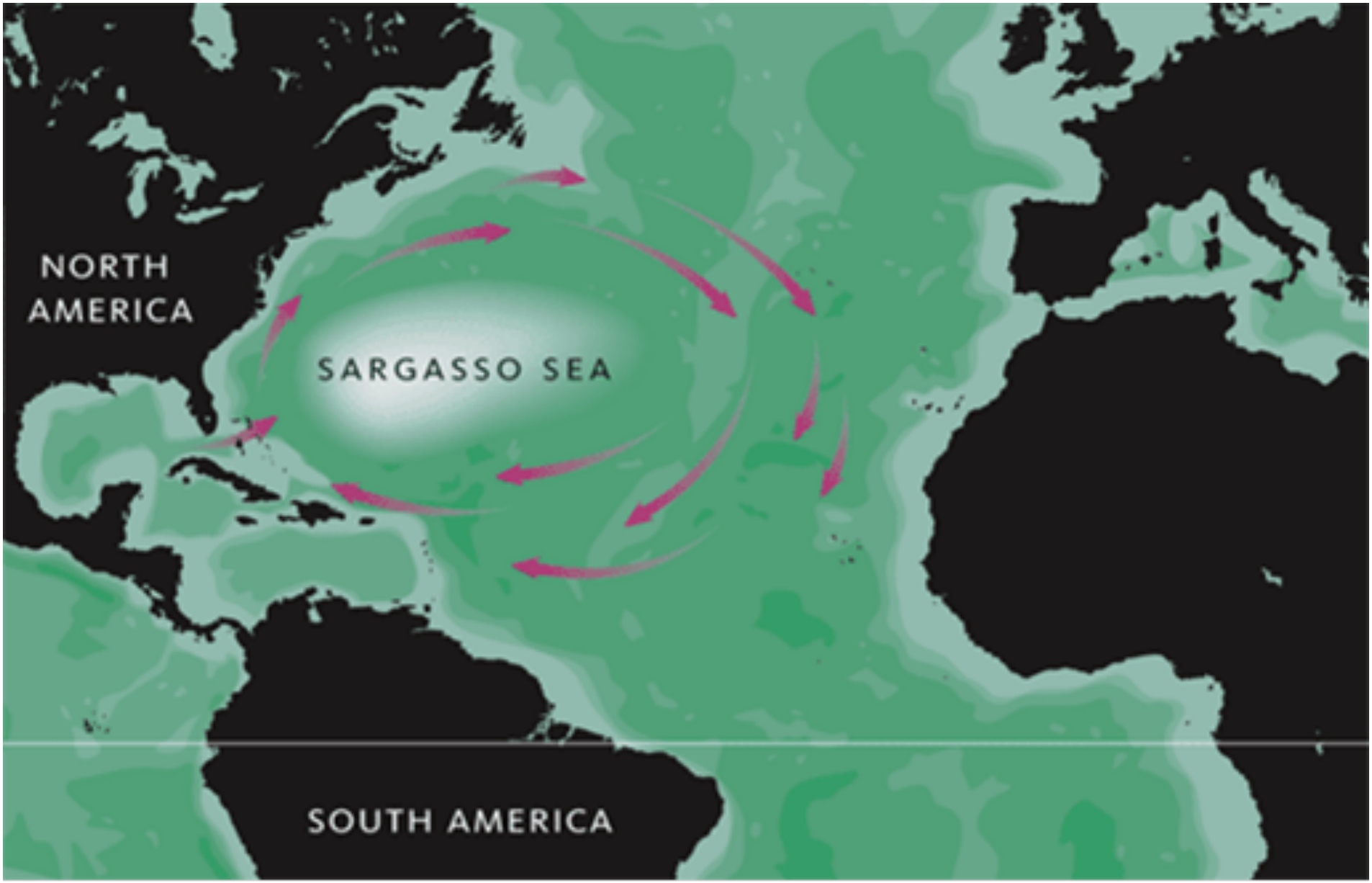
- Why did the authors focus on the Sargasso Sea?

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.

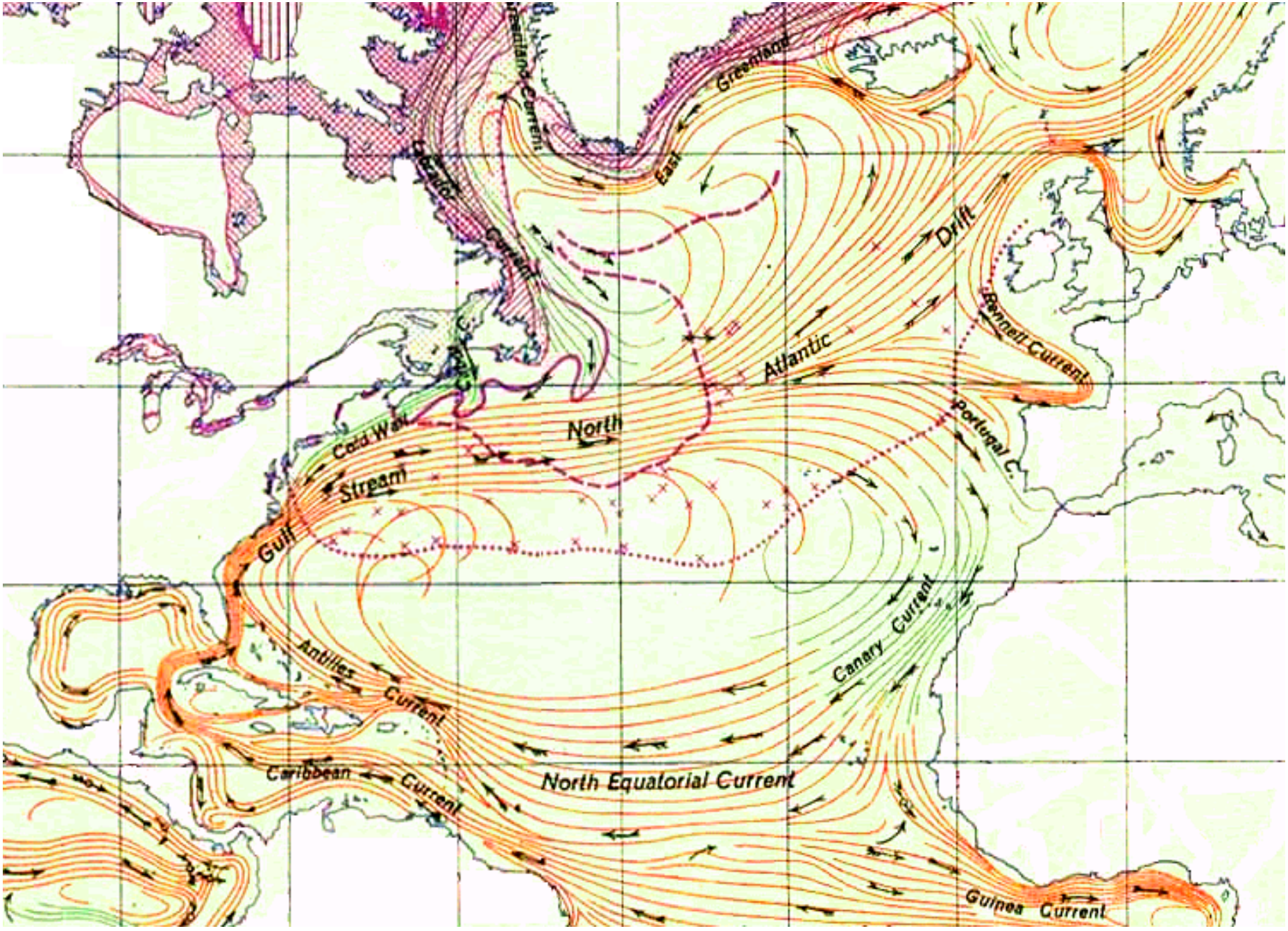
What is the Sargasso Sea?

The Sargasso Sea, located entirely within the Atlantic Ocean, is the **only sea without a land boundary**.

Sargasso Sea (from Wikipedia)



Sargasso Sea (from Wikipedia)



Mats of free-floating sargassum, a common seaweed found in the Sargasso Sea, provide shelter and habitat to many animals. Image credit: University of Southern Mississippi Gulf Coast Research Laboratory.

Genetic diversity in Sargasso Sea bacterioplankton

Stephen J. Giovannoni, Theresa B. Britschgi, Craig L. Moyer & Katharine G. Field

http://www.nature.com/nature/journal/v345/n6270/abs/345060a0.html

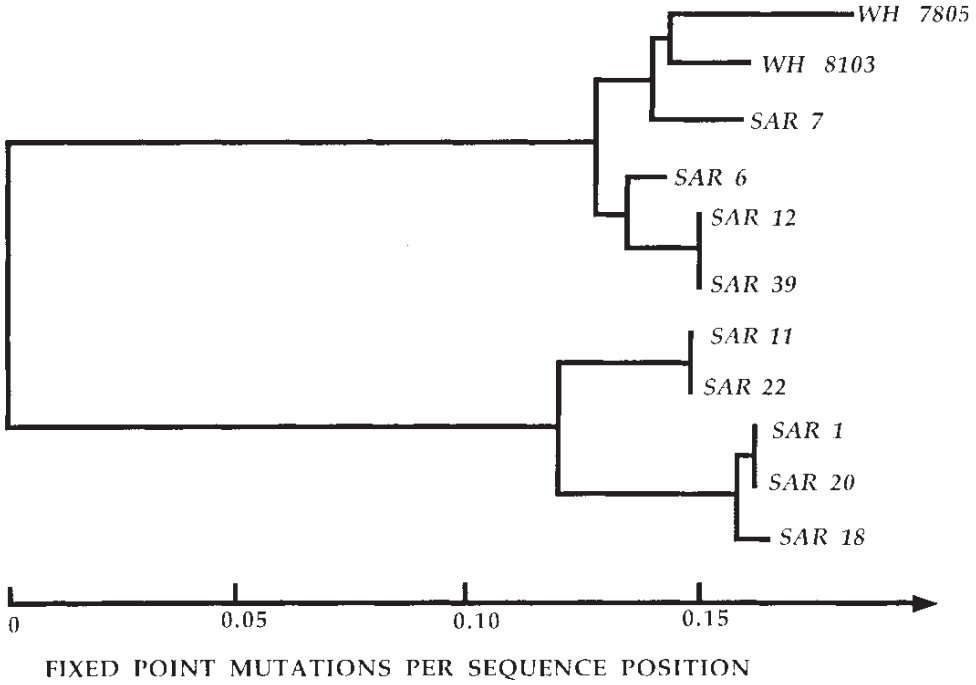


FIG. 1 Unrooted phylogenetic tree depicting relationships among Sargasso Sea bacterioplankton 16S rRNAs. The analysis includes the 16S rRNA sequences of two cultivated strains of marine cyanobacteria, WH8103 and WH7805. The sequences of clones SAR67 and SAR78 were identical to those of SAR1 and SAR20; SAR77 differed at one position (not shown). A total of 230 positions located at the 5′ region of the gene, including both conserved and hypervariable domains, were included within the analysis (*E. coli* positions 101–344). Phylogenetic trees were constructed by a distance matrix method[23,24,25].

METHODS. The picoplankton samples were collected in April by tangential flow filtration on Durapore 0.1 μm fluorocarbon membranes from a depth of 1–2 m at hydrostation S (32°4′ N 64°23′ W). The small subunit ribosomal RNA genes were amplified from bulk picoplankton DNA by a modification of the polymerase chain reaction[26]. The reaction conditions were: 1 μg template DNA; 2′ at 94 °C, 2′ at 37 °C, 7′ at 72 °C; 30 cycles. The amplified genes were cloned as *Bam*HI/*Pst*I fragments into M13 phage mp18 and sequenced twice, once with inosine substituting for guanosine, using the dideoxy chain termination method[22]. The amplification primers (OX1 and OX2) were designed with a bias towards the cyanobacterial phylum of the eubacteria. Primers OX1 and OX2 are, respectively, 90% and 82% similar to eubacterial consensus sequences. The sequences of the amplification primers are: OX1, GTGCTGCAG**AGAGTTYGATCCTGGCTAGG**; OX2, CACGGATCC**AAGGAGGTGATCCANCCNCACC**, where the domain complementary to the coding region is indicated in bold.

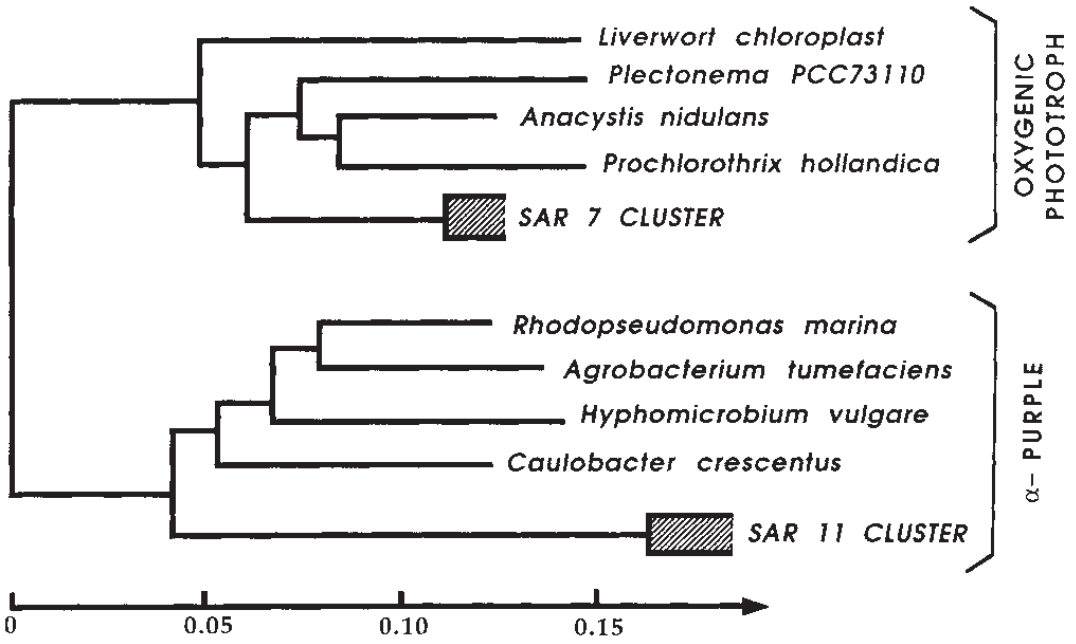


FIG. 3 Phylogenetic relationships of SAR7 and SAR11 16S rDNA sequence clusters to a collection of 16S rRNA sequences representing the oxygenic phototroph[8,28] and α-purple eubacterial phyla[29]. Four clones were sequenced completely (SAR6, SAR7, SAR1 and SAR11) and used for the inference of distant relationships. The tree was rooted using the sequences of *Bacillus subtilis* and *Heliobacterium chlorum*[8,30]. The analysis was restricted to 900 sequence positions. Regions of uncertain homology between phyla, including hypervariable domains, were excluded from this analysis. Hence, the variability within the clusters (indicated by the hatched boxes) is about 0.01 similarity units less in this figure than in Fig. 1. The 3′-terminal domain of the 16S rRNAs was excluded from the analysis because of an internal *Bam*HI restriction site in clones SAR1 and SAR11 at position 1,190.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, June 1991, p. 1707–1713 Vol. 57, No. 6
0099-2240/91/061707-07$02.00/0

Phylogenetic Analysis of a Natural Marine Bacterioplankton Population by rRNA Gene Cloning and Sequencing

THERESA B. BRITSCHGI AND STEPHEN J. GIOVANNONI*
Department of Microbiology, Oregon State University, Corvallis, Oregon 97331

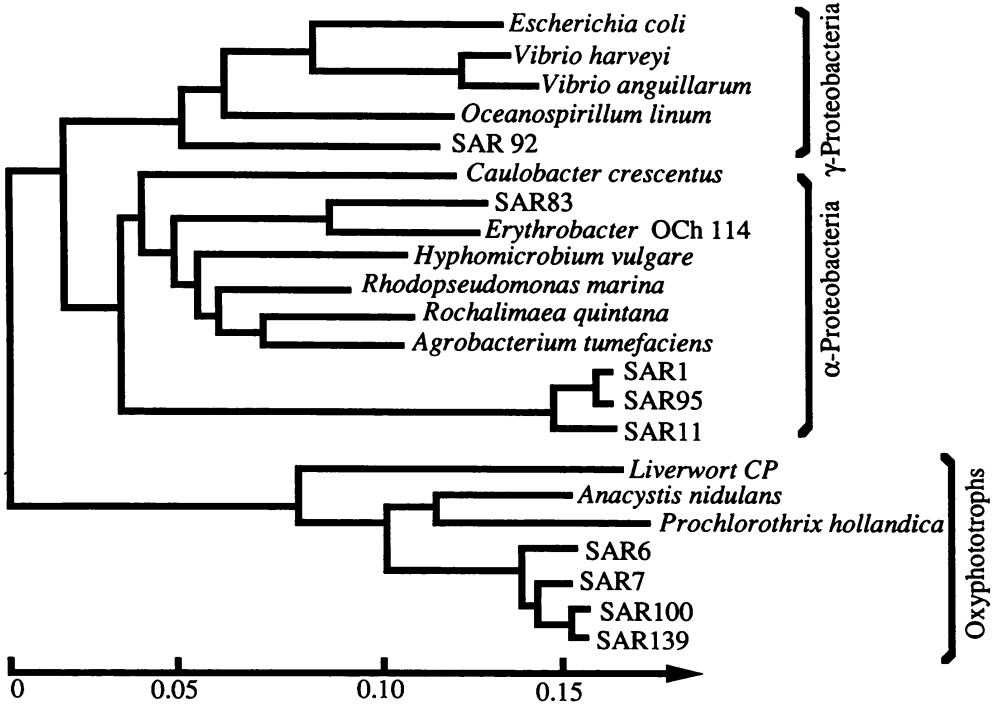


FIG. 4. Phylogenetic tree showing relationships of the rDNA clones from the Sargasso Sea to representative, cultivated species (2, 6, 18, 30, 31, 32, 35, 40). Positions of uncertain homology in regions containing insertions and deletions were omitted from the analysis. Evolutionary distances were calculated by the method of Jukes and Cantor (15), which corrects for the effects of superimposed mutations. The phylogenetic tree was determined by a distance matrix method (20). The tree was rooted with the sequence of *Bacillus subtilis* (38). Sequence data not referenced were provided by C. R. Woese and R. Rossen.

- What did they sample and how did they sample?

Sargasso Sea

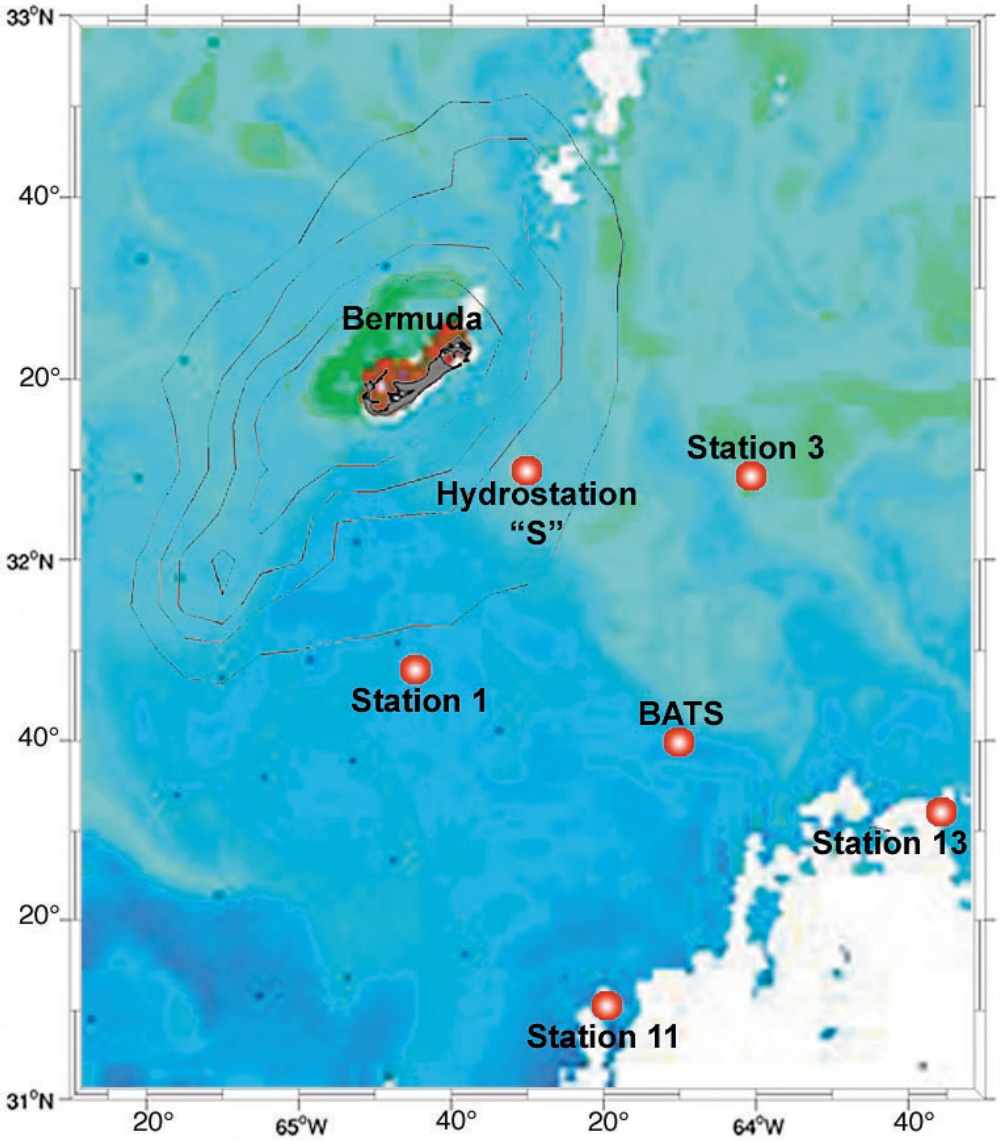
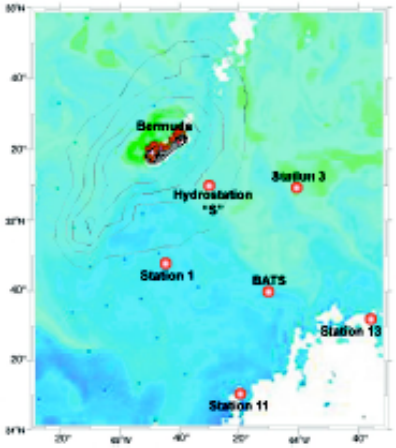
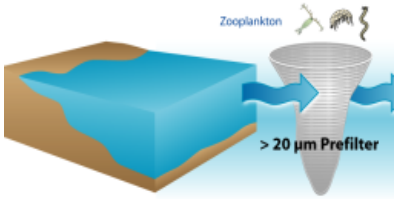


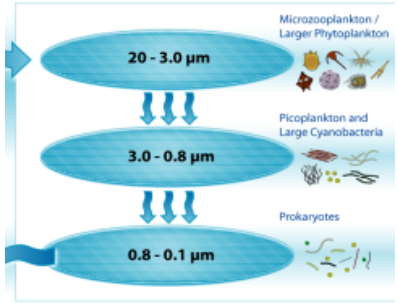
Fig. 1. MODIS-Aqua satellite image of ocean chlorophyll in the Sargasso Sea grid about the BATS site from 22 February 2003. The station locations are overlain with their respective identifications. Note the elevated levels of chlorophyll (green color shades) around station 3, which are not present around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

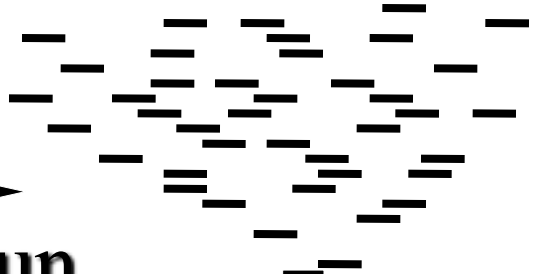
http://www.sciencemag.org/content/304/5667/66

Shotgun metagenomics

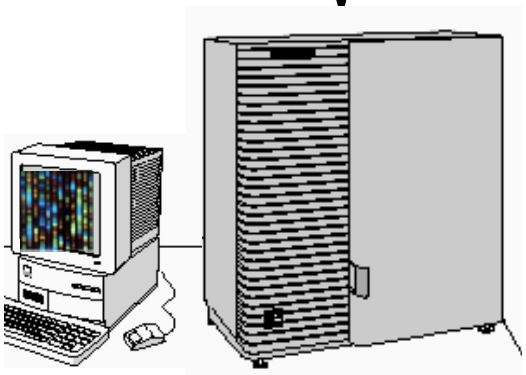








shotgun sequence



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.

Sampling Protocols. Sampling on the RV Weatherbird II was done as follows: Seawater (170 liters) from stations 11 and 13 was directly filtered through a 0.8µm Supor membrane disc filter (Pall Life Sciences) followed in series by a 0.22µm Supor membrane disc filter (Pall Life Sciences). The sample from station 3 was pumped into a 250 L carboy prior to being filtered through the impact filters. The length of time from collection of the sample until the end of the filtration step was approximately one hour. Filters were placed in 5ml of sucrose lysis buffer (20mM EDTA, 400mM NaCl, 0.75 M Sucrose, 50mM Tris-HCl, pH 9.0) and stored in liquid nitrogen on the Weatherbird then placed at -80°C until DNA extractions were done. Alternatively seawater (340 liters) was collected from 5 meters below the surface into a carboy then filtered through a 0.8µm Supor membrane disc filter (Pall Life Sciences), followed by concentration to 1 liter using a Pellicon tangential flow filtration system (Millipore) with a 0.1µm Durapore VVPP cartridge (Millipore); again the total time for the filtration and concentration was approximately one hour. Cells were pelleted at 10,000 rpm, 4°C for 30 minutes.). The impact filters and the retentate from the TFF were then handled as described above. The carboys, tubing and filter systems were cleaned with a 10% hydrochloric acid wash prior to each leg of the sampling. Any of the sampling equipment (tubing, etc.) that could reasonably be soaked was soaked in an acid bath is for at least 24 hours. Sampling carboys were filled with the acid wash and "soaked" for at least 24 hours as well. All acid washed items were subsequently rinsed very liberally with Milli-Q water. A liberal Milli-Q water rinse was also conducted between samples on the same leg. All spigots from the carboys were covered with a ziploc bag until needed. Tubing was stored in clean ziploc bags until needed.

Sample preparation. The impact filters were cut into quarters and placed in individual 50 ml conical tubes. TE buffer (5 ml, pH 8) containing 150 ug/ml lysozyme was added to each tube. The tubes were incubated at 37oC for 2 hours. SDS was added to 0.1% and the samples were then put through three freeze/thaw cycles. The lysate was then treated with Proteinase K (100 ug/mL) for one hour at 55oC followed by three aqueous phenol extractions and one extraction with phenol/chloroform. The supernatant was then precipitated with two volumes of 100% ethanol and the DNA pellet washed with 70% ethanol.

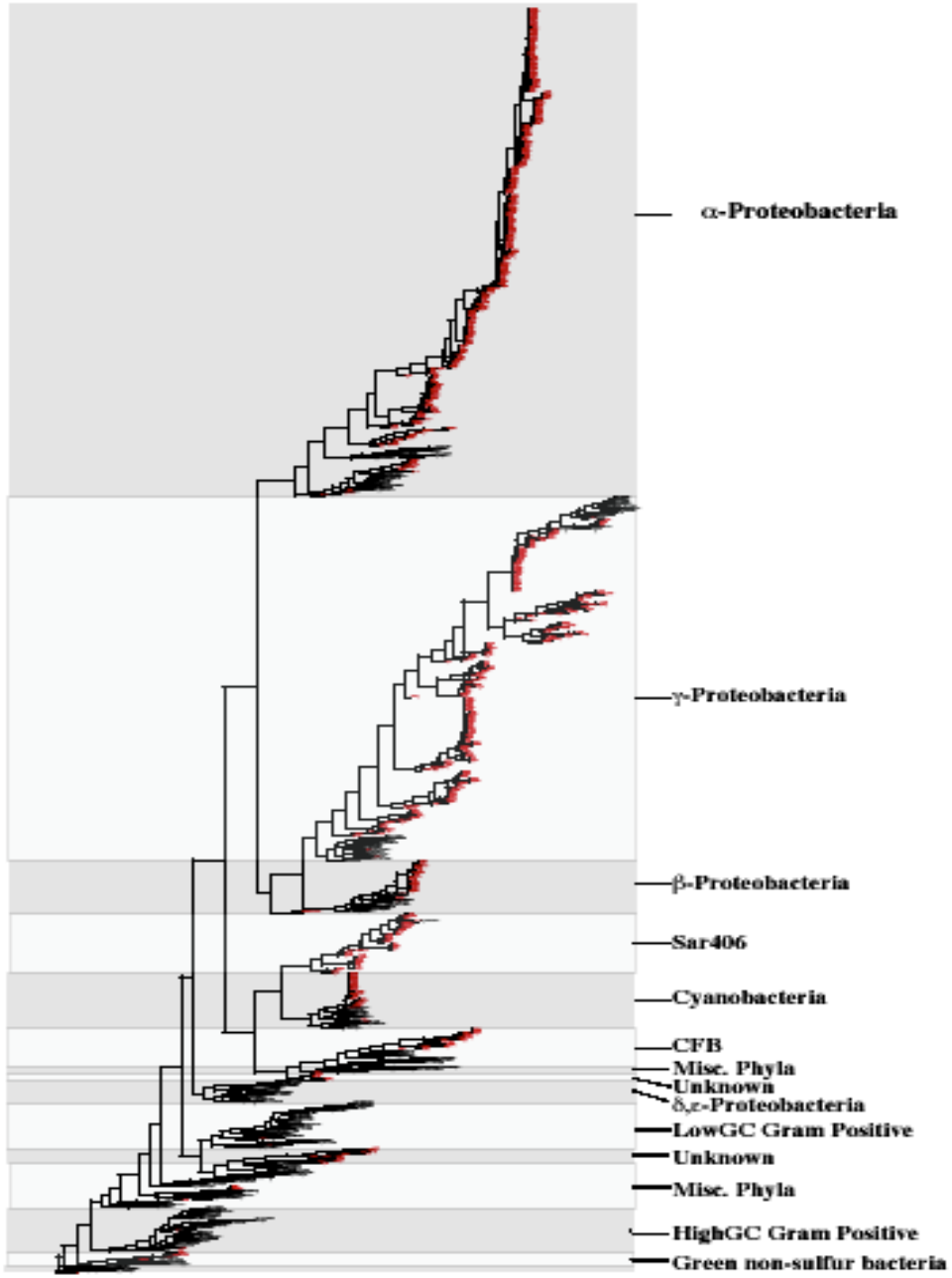
DNA preparation. DNA was randomly sheared, end-polished with consecutive BAL31 nuclease and T4 DNA polymerase treatments, and size-selected by electrophoresis on 1% low-melting-point agarose. After ligation to Bst XI adapters (Invitrogen, catalog no. N408-18), DNA was purified by three rounds of gel electrophoresis to remove excess adapters, and the fragments, now with 3'-CACA overhangs, were inserted into Bst XI-linearized plasmid vector with 3'-TGTG overhangs. Fragments were cloned in a medium-copy pBR322 derivative.

Sequence assembly. With default parameter settings, the highly covered genome sequences would have been treated as repetitive DNA by the Celera Assembler. Since the Celera Assembler constructs scaffolds only from a backbone of sequence heuristically classified as unique, these organisms would not have been eligible for scaffolding and would have been absent from the final assembly. However, by tuning the threshold parameter for classifying unique sequence, we were able to compensate for the apparent repetitiveness of these genomic regions, and scaffold them appropriately. This was accomplished by identifying the most deeply assembling, obviously non-repetitive contigs in an initial run of the assembler (in this case, the strong assemblies at 21-36x coverage which were identified as gene-rich Burkholderia-like and plasmid scaffolds), and using a value slightly below the calculated "A-statistic" (an empirical uniqueness measure within the Assembler) of these contigs as the threshold parameter in a subsequent run. This allows the deep contigs to be treated as unique sequence, when they would otherwise be labeled as repetitive. At the other end of the spectrum, rare organisms in the sample have been sampled by sequencing only to a shallow depth of coverage. Routine assembly would not have considered the small fragment overlap based assemblies with shallow coverage as an eligible basis for scaffolding, due to a minimum length requirement of 1000bp, which is typically in place for efficiency. Therefore, in the present use case, the organisms represented by these sequences would not have been ordered and oriented with mate-pairs without adjusting the default minimum length to compensate for the low anticipated coverage depth and assembly length. With this selection of parameters, more suitable to the enivironmental project at hand, we were able to adequately assemble both the dominant and rare species simultaneously.

- What are some questions one might want to answer about the Sargasso Sea samples / sequences?

Who is out there?

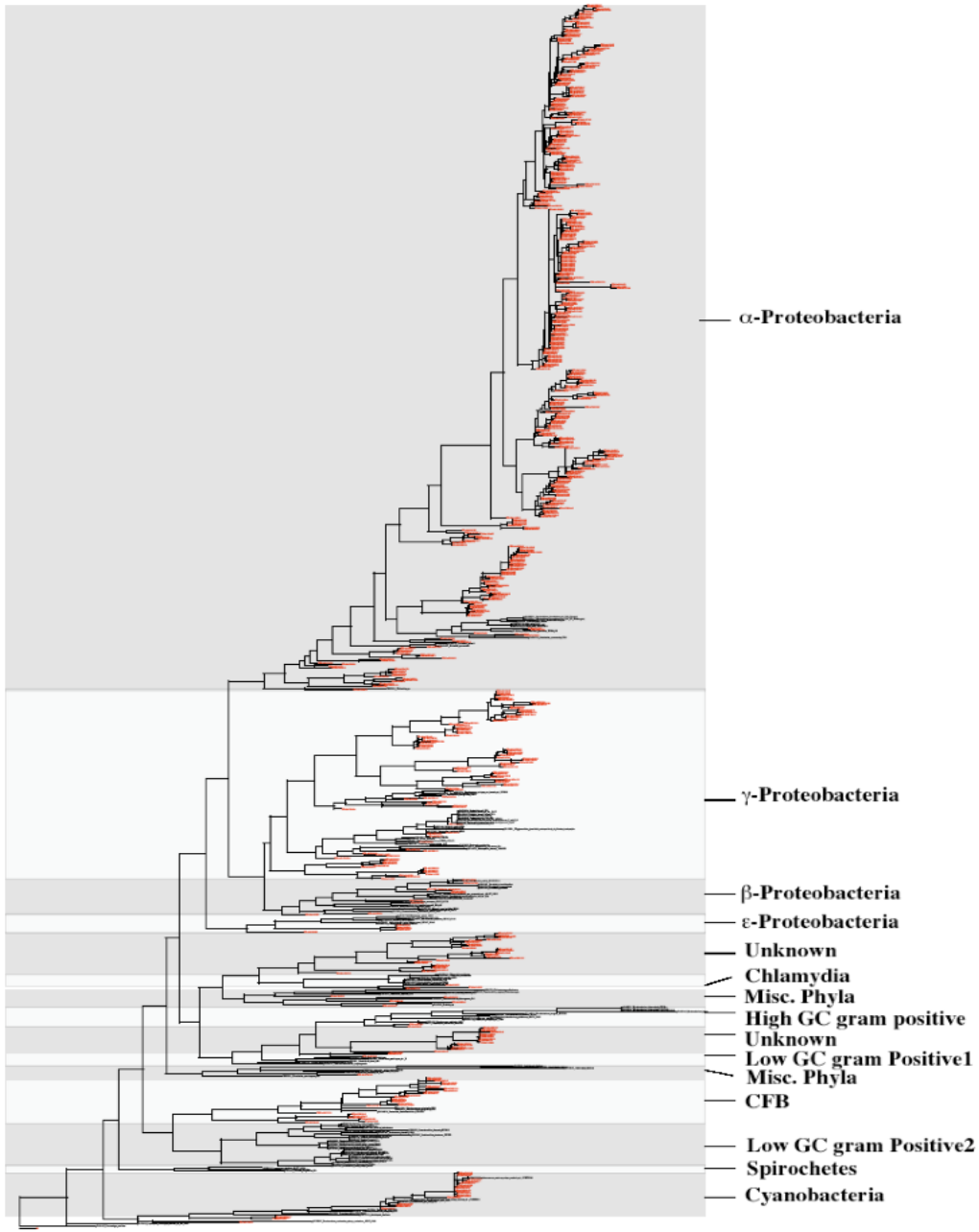
rRNA phylotyping from metagenomics



http://www.sciencemag.org/content/304/5667/66

- How can one do phylotyping with genes other than rRNA?

Shotgun Sequencing Allows Alternative Anchors (e.g., RecA)



Who is out there?

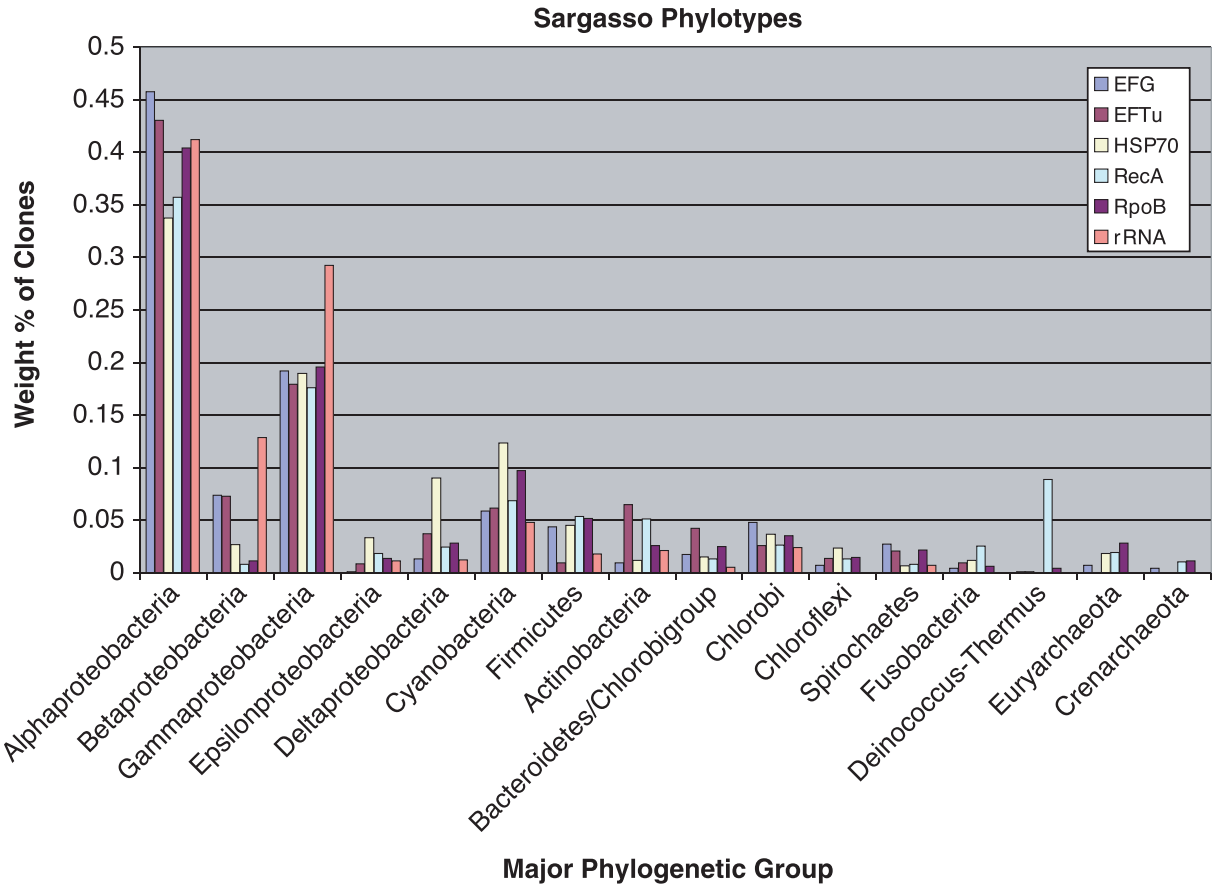


Fig. 6. Phylogenetic diversity of Sargasso Sea sequences using multiple phylogenetic markers. The relative contribution of organisms from different major phylogenetic groups (phylotypes) was measured using multiple phylogenetic markers that have been used previously in phylogenetic studies of prokaryotes: 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.

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

http://www.sciencemag.org/content/304/5667/66

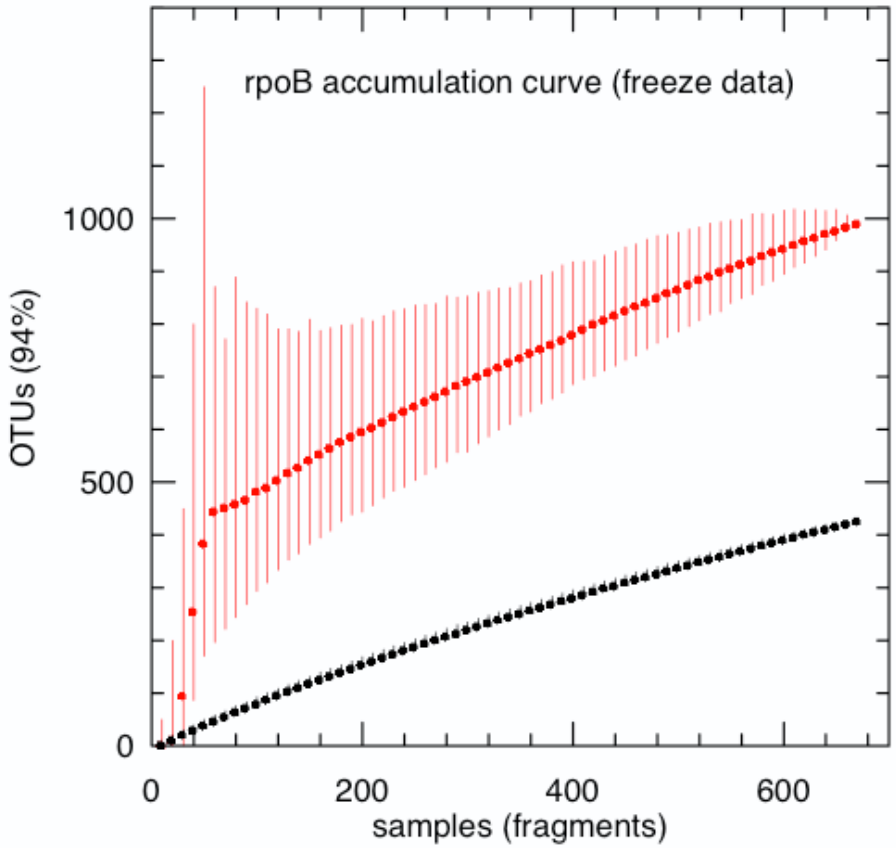


Figure S6. Accumulation curve for rpoB. Observed (black) OTU counts for rpoB (based on the fragment grouping summarized in Table 2), as well as the Chao1-corrected estimate of total species (red; see (*3*)). Points are mean values of 1000 shufflings of the observed data, while bars show 90% confidence intervals.

http://www.sciencemag.org/content/304/5667/66

What are they doing?

Community Function?

- Many attempts to treat community as a bag of genes
- The run pathway prediction tools on entire data set to try and predict “community metabolism”
- Does not work very well

Functional Inference from Metagenomics

- Can work well for individual genes

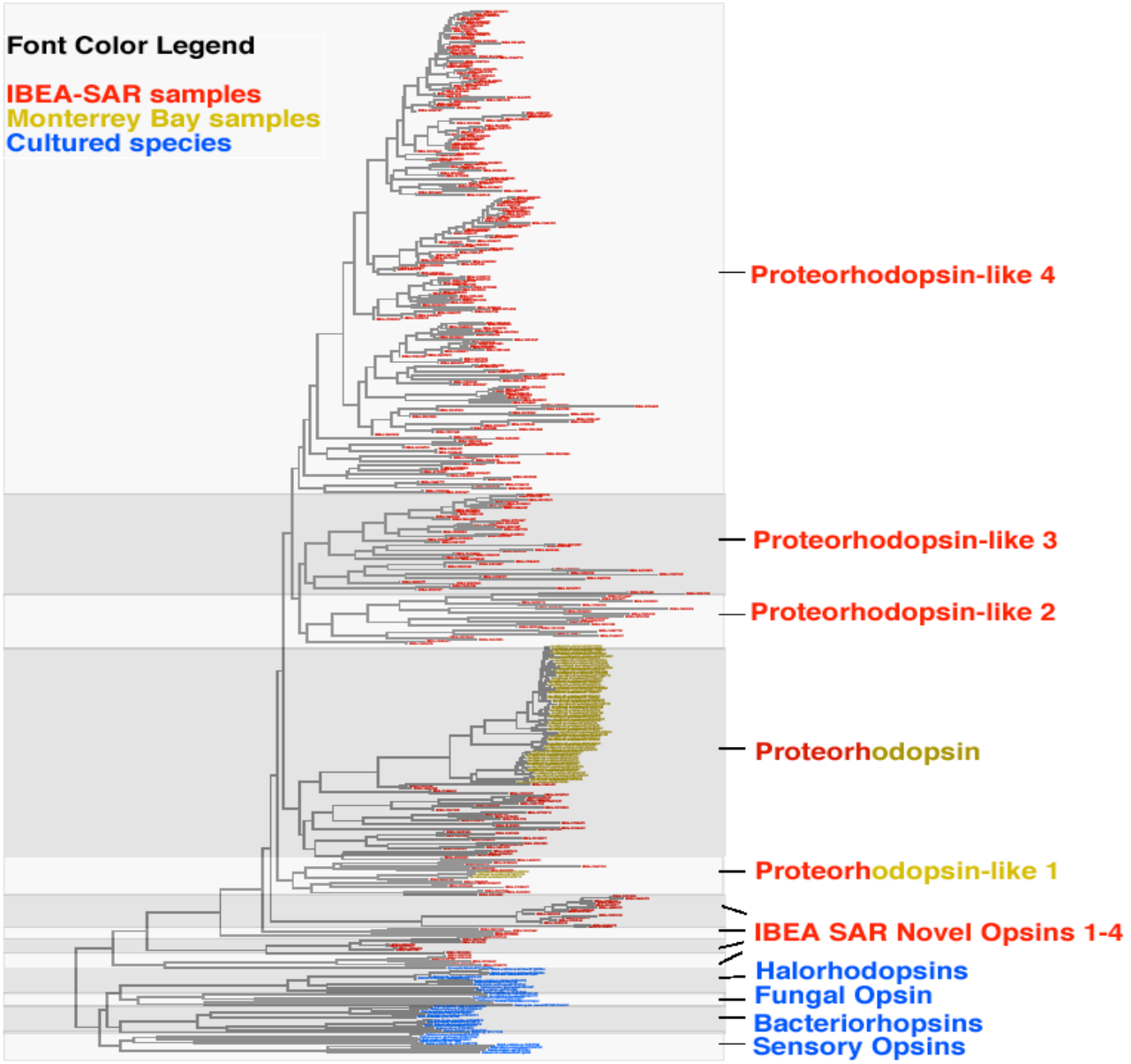
What are they doing?

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

http://www.sciencemag.org/content/304/5667/66

Functional Diversity of Proteorhodopsins?



http://www.sciencemag.org/content/304/5667/66

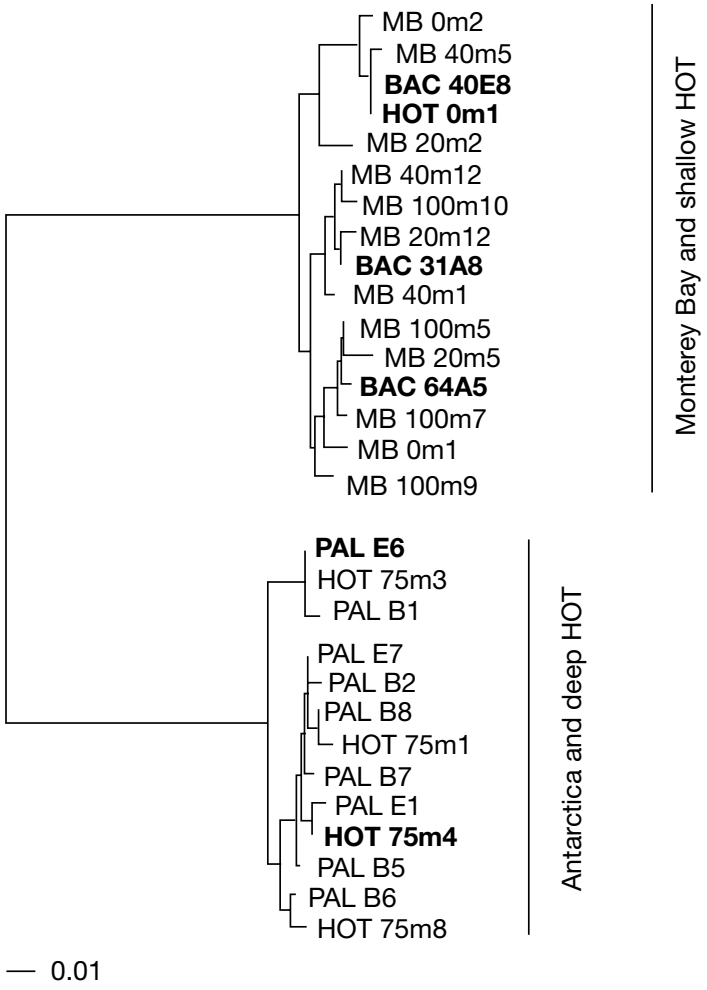


Figure 3 Phylogenetic analysis of the inferred amino-acid sequence of cloned proteorhodopsin genes. Distance analysis of 220 positions was used to calculate the tree by neighbour-joining using the PaupSearch program of the Wisconsin Package version 10.0 (Genetics Computer Group; Madison, Wisconsin). *H. salinarum* bacteriorhodopsin was used as an outgroup, and is not shown. Scale bar represents number of substitutions per site. Bold names indicate the proteorhodopsins that were spectrally characterized in this study.