Supporting Online Material for

Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota

Vaughn Iverson, Robert M. Morris, Christian D. Frazar, Chris T. Berthiaume, Rhonda L. Morales, E. Virginia Armbrust

*To whom correspondence should be addressed. E-mail: armbrust@uw.edu

Published 3 February 2012, *Science* **335**, 587 (2012)
DOI: 10.1126/science.1212665

This PDF file includes:

Supporting Online Text

Metagenome community analysis

A complication of using short sequence reads for metagenomic analysis is that many reads align equally well with multiple sequences in a reference database due to sequence conservation and reference redundancy (e.g. uneven taxonomic coverage). To address this, the short-read alignments were analyzed in an information theoretic (bit-scored) framework *(31)*, as different reads carry varying amounts of information (Fig. S1).

The sums of bit-scores calculated in this manner are shown for read alignments with each database (Table S2). Building on this information-based approach, algorithms were developed to select the most parsimonious reference sequences from each database and to estimate the relative abundance and depth of read coverage for these selected sequences.

The October and May microbial communities were characterized (Fig. S2) using information-based analysis of metagenomic reads aligned to the RDP 16S rDNA database *(32)* augmented with full-length 16S rDNA sequences cloned from the October sample (Fig. S3). Database 16S rDNA sequences selected by these analyses were taxonomically classified with the Bayesian method of Wang *et al*. *(33)* using a custom training set augmented with marine environmental clades, including a previously undescribed group of alpha-*Proteobacteria* dubbed "PS1" (Figs. S4-5).

The 16S rDNA analysis of metagenomic reads detected nearly twice as many family level taxa compared to the 16S rDNA clone sequences (48 versus 28 for the October sample, Figs. S3, S6 and S7). Abundance estimates calculated for the identified taxa were validated using simulations that quantified the sensitivity and accuracy of our information-based approach. Simulated populations of 16S rDNA taxa with relative abundance as low as 0.3% were consistently identified and those above 1.0% were typically quantified to within ±10% of their true proportion in the simulated sample (Fig. S8).

Overall community composition differed between the October and May samples, with major shifts in population abundances among alpha-*Proteobacteria*, *Flavobacteria* and both phyla of marine Archaea (Figs. S2, S6 and S9). Notably, the Marine Group-II *Euryarchaeota* increased in abundance by 2.5-fold between October and May (reaching ~7.5%), a seasonal pattern observed previously *(18, 19)*.

Analysis of MG-II protein homologs in RefSeq genomes

We performed an initial analysis of gene homologs shared among seven selected archaeal genomes (Table S9), and selected *A. boonei* (a marine thermoacidophile and the genome with the largest number of shared homologs, n=506), and *Nitrosopumilus maritimus (34)* (the only other free-living marine archaeon with a finished genome) for detailed comparison. Homologous genes shared among these three genomes (Fig. S19B),

reveals a small shared core (n=291) primarily composed of characteristically archaeal proteins required for replication, transcription, translation and core metabolic functions.

Curiously, over 60% of predicted MG-II proteins have no homologs in these archaeal genomes, with many having no significant homology to any RefSeq genome (n= 578). To explore potential relationships of the "non-shared" proteins to homologs that were found in RefSeq genomes, taxonomic assignments were calculated for the non-homologous proteins among the three comparison genomes (Figs. S19C, S20 and Table S10). About 50% of the "non-shared" MG-II proteins had homologs in the RefSeq database, with 60% of those (n=332) assigned to Bacterial homologs, which is striking in comparison to *A. boonei*, where only 19% were similarly assigned (n=119, Table S10). The opposite was also observed, with 59% (n=366) of *A. boonei* "non-shared" proteins with RefSeq homologs assigned to Archaea, whereas only 11% (n=64) of equivalent MG-II proteins were so assigned (Table S10).

Materials and Methods

Sample Collection

Samples for metagenomic analysis were collected on October 10, 2008 (October) and on May 4, 2009 (May). Approximately ninety liters of surface water (1m) were collected from the Puget Sound main basin in Seattle, WA (47º 41.24' N, 122º 24.14' W). The less than 0.8µm fraction was concentrated by tangential flow filtration (TFF) equipped with a 30kD Biomax Pellican 2 Cassette (Millipore Corporation, Billerica, MA). Cells were concentrated to a final volume of approximately 150mL. Cells were subsequently pelleted by centrifugation at 4ºC for 60 min at 17,000xg and resuspended in 20mM Tris (pH7.4). DNA was extracted from the concentrated cells using a DNeasy Blood and Tissue kit (QIAGEN, Germantown, MD) following the manufacturer's instructions.

Library Construction and SOLiD Sequencing

Metagenomic libraries were constructed according to the SOLiD™ v3.0 mate-pair library preparation protocol (Life Technologies, Foster City, CA). Briefly, 22.2 µg (October) and 20.6 µg (May) of input DNA were sheared and sized selected to produce DNA fragments between 2000 and 3000 base pairs. The DNA was end repaired, ligated with EcoP15i CAP adaptors and circularized. Subsequently, the circularized DNA was nick translated and digested, leaving 50 bases of genomic DNA on either side of the internal adaptor. Finally P1 and P2 adaptors were ligated to the ends of the genomic DNA and served as sites for PCR amplification. Lambda DNA was also prepared for mate pair sequencing, was added to the metagenomic library at a ratio of 1:1000 (Lambda: Genomic DNA) and served as an internal control. Libraries were deposited on full slides and sequenced using the SOLiD v3.0 system. The October library was

sequenced by Life Technologies (Foster City, CA) and the May library was sequenced at the University of Washington, Center for Environmental Genomics (Seattle, WA).

Cell Counts

Cell counts and TFF recovery efficiency were determined by staining cells from whole water and concentrated seawater samples with the nucleic acid stain 4',6-diamidino-2-phenylindole (DAPI) and by fluorescence in situ hybridization (FISH) as previously described *(35)*. Cells were imaged using a Nikon 80i microscope equipped with a CoolSNAP HQ2 camera (Photometrics, Tucson, AZ) and NIS Elements Basic Research software (Nikon Instruments, Melville, NY). More than 500 cells were counted from each slide (15 frames). Count data is summarized in Table S1.

Clone library construction

For the October sample bacterial and archaeal 16S rRNA gene clone libraries were constructed. Briefly, RNA genes were amplified from community genomic DNA by PCR using Taq polymerase (Genechoice, Frederick, MD) and variations of commonly used bacterial primers, 8F and 1492R, or archaeal primers, 21F and 1492R. Amplifications were performed in a C1000 thermal cycler (Bio-Rad Laboratories, Hercules, CA) using the following conditions: 35 cycles, annealing at 55ºC for 1 min, elongation at 72ºC for 2 min, and denaturation at 94ºC for 30 s. A single band of the predicted length was observed by agarose gel electrophoresis, excised and purified using a Minelute Gel Extraction kit (QIAGEN, Germantown, MD) according to the manufacturer's instructions.

Clone libraries were constructed and sequence identities were obtained as described by Morris *et al*. *(36)*. Briefly, amplicons were cloned into the pGEM T-easy vector (Promega, Madison, WI) and sequenced at the University of Washington High-Throughput Genomics Unit (Seattle, WA). Cloned 16S rRNA gene sequences were aligned to a custom ARB database *(37)* that contained 151,952 sequences from cultured organisms and environmental gene clone libraries. Taxonomic assignments were determined by phylogenetic inference.

Metagenome read preparation for alignment and assembly

Raw color-space SOLiD reads were converted to the FASTQ format for compatability with standard tools, de-duplicated to remove redundant PCR duplicated mate-pairs, and trimmed/filtered for quality, all using custom software. PCR duplicate mate-paired reads were removed using a method that indexes and compares mate-paired reads directly in an error tolerant manner, removing the lowest quality pairs when duplication is detected. For trimming, quality scores were used to determine the cumulative probability of zero uncorrected nucleotide-space errors occurring up to each position in a read. Reads used for alignment from each sequencing run were

independently trimmed using an error probability threshold tuned to maximize the sequence aligned with a known reference genome. For the October sample, the genome sequence of *Pelagibacter ubique* str. HTCC1062 (NC_007205) was used, yielding a trimming threshold of 0.4. For the May sample, the genome sequence of Enterobacteria phage lambda (NC_001416, an added standard) was used, yielding a threshold of 0.5. These values correspond with the predicted probability that a trimmed read is free of uncorrectable nucleotide errors, based on instrument generated quality scores. Reads trimmed below 34nt in length were discarded, as were low complexity reads with measured mean entropy of less than 3 bits per di-nucleotide after trimming. Reads containing low complexity sequence were discarded because this is a common error mode for SOLiD generated sequence, and such reads consume memory and complicate *de novo* assembly.

For *de novo* assembly, a further reduced set of higher quality reads was produced using the above methods, but tuning thresholds were determined to produce the approximate number of reads that would fit in 144 Gbytes of available memory during the assembly process (error thresholds 0.76 and 0.93 for the October and May samples, respectively.) Quality trimmed reads used for assembly were then error-corrected using the Applied Biosystems SOLiD™ Accuracy Enhancer Tool (SAET) tool version 2.2 (Foster City, CA) with an expected reference length of 200 megabases and parameters "-trustprefix 25 -localrounds 3 -globalrounds 2". These settings were selected as a reasonable tradeoff between error correction potential, and memory use and computation time required by the tool.

Alignment to reference databases

Quality trimmed reads from each sample were aligned to reference databases using BWA *(38)* version 0.5.7. 16S rDNA alignment was performed against the "good quality, ≥ 1200bp" subset of the Ribosomal Database Project *(32)* (RDP) database release 10 update 25, with the omission of sequences with significant homo-polymer runs, and the addition of clone library sequences from the October sample. Whole genome alignment was performed against the NCBI RefSeq *(17)* database release 46, with the omission of Eukaryotic nuclear genomes. Metagenome alignment was performed against assembled contigs from the Global Ocean Sampling *(10)* "GOS: Assembled Sequences (N) -- Scaffolds and Unassembled Sequences" database, downloaded from CAMERA as file "node10154399155649252 80.fasta.gz" on October 10, 2009 *(39)*. All BWA alignments were produced using the following parameters "bwa aln -c -n 0.001 -l 18 -k 2" and "bwa samse -n 500000". These parameters were selected to allow a tradeoff between allowing a liberal number of mismatches between reads and reference sequences (varying depending on trimmed read length) and maintaining reasonable computational complexity given the multi-gigabase reference databases used. Significantly, these

settings also explicitly return multiple candidate alignments (up to 500000) for any read that aligns equivalently well with more than one position in a reference database.

Alignment post-processing for reference selection and abundance/coverage estimation

Read alignments in the SAM format *(40)* output by BWA were post-processed using custom software called SEAStAR (Select and Estimate Abundance from ShorT Aligned Reads). Briefly, SEAStAR works by calculating an information bit-score *(31)* for each aligned read by treating it as an encoded symbol and using the number of aligned positions in the reference sequence database to estimate its probability. The information content I (as a bit-score) of each read aligned to one or more reference sequences in a given database was quantified as: $I(p) = -\log_2(p)$ where p is the measured probability of the read aligning with a randomly selected reference sequence in that database *(31)*. The most parsimonious database reference sequences (those that best explain the information contained in the read-reference alignments) are selected using a greedy algorithm that iteratively chooses sequences with the maximum length-normalized residual total aligned read bit-score, down to a floor threshold. The residual bit-score for a reference is the sum of the bit-scores of aligned reads that do not also align with previously selected reference sequences. This algorithm produces the set of "selected reference sequences" used by the following analyses.

Coverage is estimated by processing reads in order of bit-score (high to low), allocating coverage to selected reference sequences fractionally among each read's alignments, weighted by the relative mean coverage already accumulated among those reference sequences. That is, coverage is accumulated successively, with priority given to the most informative reads; those that align uniquely and those that align to the fewest positions among selected reference sequences. In this way, the coverage contributed by reads that map to multiple selected reference sequences is shared fractionally among those references, with the split proportion determined by more informative reads. Fractional abundance is estimated for each selected reference sequence as its fraction of the sum of the mean coverage of all selected sequences (see Fig. S1 for an illustrated example of this process).

Note, this method allows reference sequences that are selected but not fully covered by metagenome reads to be identified; indicating that novel taxa (not found in the database) are present in the sample. If desired, the covered (conserved) portions of the selected reference sequence can be used to design primers facilitating targeted reconstruction of the novel taxa using Sanger sequencing. We did not use this approach because our October 16S rDNA clone libraries recovered all taxa with estimated abundance greater than 1%.

Bayesian classification of environmental 16S rDNA sequences

To assign taxonomic information to full-length RDP and October clone library 16S rDNA reference sequences selected by SEAStAR (Figs. S6 and S9), these sequences were automatically classified using the RDP Bayesian Classifier *(33)* version 2.2 using a custom training set comprised of the standard RDP training set (version 6_032010) with the addition of taxonomy and Genbank training sequences for major environmental clades detected in our October clone libraries (resulting from ARB alignments previously described) that the standard RDP training set did not encompass (Fig. S4). A previously unnamed novel clade of marine alpha-*Proteobacteria* that was detected in the October clone libraries and both of our metagenomic samples has been dubbed the PS1 group. Figure S5 shows a bootstrapped neighbor-joining phylogeny including this group, computed from the standard RDP *(41)* alignment using the software package Geneious Pro *(42)* version 5.3. All October 16S rDNA clone library sequences and the reference sequences selected by SEAStAR from the October and May metagenome alignments with the modified RDP database were classified based on the custom training set described above (results in Figs. S3, S6, and S9, respectively.)

SEAStAR simulations and 16S database comparison

The specificity and accuracy of SEAStAR reference sequence selection and abundance estimation were evaluated through simulations using synthetic metagenomic datasets generated from randomized populations of 16S rDNA sequences. 10 random trial populations of 28 family-level taxa were drawn randomly from the October 16S rDNA clone library sequence classifications (excluding sequences that did not classify with $p \geq 0.85$, and taxonomic orphans, as shown in bold italic type on Fig. S3). For each trial population, a clone sequence was randomly selected to represent each selected family-level taxon, and each clone sequence was randomly and exclusively assigned one of four relative abundance classes: 9.8668% (~10.0), 3.1202% (~3.0), 0.9867% (~1.0) and ~0.3120% (~0.3), with each class containing 7 clones, yielding a total abundance of 100%. The four exact relative abundance class values used were selected to be equally spaced on a log scale while summing to 100%.

Each trial population was used to generate two large simulated read sets ($n=10^6$ reads) with errors introduced based on models for the October and May SOLiD sequencing runs respectively, using the “simutrain” and “simulate” tools in the Maq *(43)* software package version 0.7.1, with custom modifications to correctly handle generation of color-space reads. The resulting read sets contained reads generated from each clone sequence in the correct proportion for the trial population's family-abundance class it was selected to represent.

A subset of reads (n=80000, approximately the number of reads from our October and May metagenomes aligning with the RDP database) was randomly drawn from each simulated read set for use as the simulated population sample. Each of the 20 read

population samples was separately aligned against three 16S rDNA databases: October clones only, October clones added to the RDP database, and the RDP database only.

Alignments, post-processing using SEAStAR, and taxonomic classifications were performed as described above for the actual metagenome reads, with the exception that the 28 clone sequences used to generate each simulated read set were excluded from all alignment databases used for that read set. The resulting classifications and estimated abundances for each synthetic read set were compared to the corresponding simulated starting population and binned at the family-level as successful detections, false positives and false negatives. Detection and population abundance estimation results are summarized in Figure S8.

The effect of the sequence composition of 16S rDNA databases was further evaluated by aligning the October metagenome reads to the same three databases used for the simulations described above, run through the same analysis pipeline, and compared with the 16S abundances implied by counting classified clone library sequences (as shown in Fig. S3). The results of this comparison are shown in Figure S7.

De novo contig assembly of metagenomic reads

Color-space contigs were assembled from the SAET error-corrected read sets from each of the two sample metagenomes using the program Velvet *(44)* version 0.7.63 with a k-mer size of 21 and parameters "-scaffolding no -read_trkg yes -ins_length X -ins_length_sd 210 -exp_cov 1000 -cov_cutoff 4 -min_contig_lgth 75" where the insert length X was estimated from reference alignments to be 1100 and 2100 for the October and May samples, respectively. The effect of k-mer size was evaluated for all odd values between 16 and 28, with 21 selected because it maximized the N50 contig length metric. All other parameters were selected based on recommendations in the Velvet documentation. The resulting contigs are summarized in Table S2.

Re-alignment of reads to contigs, mate-pair processing, visualization and scaffolding

Quality trimmed reads from the each sample metagenome were aligned back to the corresponding color-space assembly contigs using BWA *(38)* as described for the reference databases. Color-space contigs were naively converted to nucleotide-space using a constant nucleotide prefix for the purposes of input to BWA, which requires nucleotide-space reference sequences but immediately converts them back to color-space internally when the "-c" parameter is used. The resulting SAM alignment files were post-processed with the custom software SEAStAR to produce nucleotide-space consensus contigs with coverage estimates and bit-scored mate-pair connections between contigs.

Briefly, SEAStAR uses a custom dynamic programming algorithm to generate consensus nucleotide-space contigs from color-space read alignments, and assigns bit-scores to mate-pair connections within and between contig sequences based on the sum of the minimum alignment bit-score (calculated as described above) assigned to each pair

of mated reads. SEAStAR emits a mate-pair connection graph which represents contigs as nodes, mate-pair connections as edges, and encodes for each element of the graph, statistics such as total bit-score, coverage, %GC content, sequence length, and mean aligned positions of mate-paired reads. These output graphs were post-processed (filtered by bit-score and colored by %GC) and visualized using the neato tool from GraphViz *(45)* version 2.27 (Figs. S10-11).

Sequence scaffolds were produced using custom software for each sample through a series of operations on the mate-pair connection graphs described above. First, all low-quality connections were filtered out of each graph by removing edges with a mate-pair bit-score below a minimum threshold of 25 bits. Each filtered graph was transformed into a set of maximum spanning trees with edge weights determined using a heuristic taking into account mate-pair bit-scores and the estimated coverage and %GC content of the connected contigs. In this way, the remaining connections maximized parsimony by eliminating the weakest connections and those between contigs with divergent %GC and coverage estimates.

Branches containing contigs totaling more than 5000bp of sequence (the minimum length anticipated to have usable tri-nucleotide usage statistics, see below) were cleaved off the tree by selectively removing the weakest mate-pair edges, determined using the same weighting heuristic used to calculate the maximum spanning tree. The remaining connected components were laid out into scaffold sequences (with gaps) using custom software that automatically determined the order and orientation of contigs from mate-pair alignment statistics. Potential chimeric (misassembled) contigs were detected automatically using custom software that searched for anomalous drops in mate-pair physical coverage within contigs (i.e. few mate-pairs spanning a particular region.) These cases were individually curated by hand, with contigs split at such positions when necessary.

Binning scaffolds by genome

For each sample metagenome, assembled scaffolds were binned, using custom software “tetracalc”, into candidate genomes based on aligned read coverage and a statistical analysis of nucleotide usage patterns. Briefly, tri- and tetra-nucleotide usage anomaly Z-statistics *(46)* were calculated for each scaffold and linear regressions of the resulting 256 (or 64, for tri-nucleotides) values were calculated for each pair of scaffolds (e.g. Fig. S13).

Scaffolds were clustered by building a graph of all scaffolds (nodes) connected by edges representing tetra-nucleotide Z-statistic correlation coefficients exceeding an empirically determined threshold ($R > 0.9$). Additional edges were then added for lower tetra-nucleotide correlations exceeding a lower threshold ($R > 0.7$) when at least one of the scaffolds in a pair contained less than 30Kbp of sequence and the tri-nucleotide correlation exceeded a third threshold ($R > 0.85$). Connected components of each

resulting graph containing more than 950Kbp of scaffold sequence became the candidate genome bins for each metagenome sample.

Candidate genome bins were then phylogenetically screened using aligned mate-pair connections between informative regions of 16S rDNA sequences selected from RDP database alignments (as described above), and positions within the candidate genomes. The set of reads mapping uniquely to selected 16S rDNA sequences classified to the same genus by the analysis of RDP database metagenome alignments were identified as "16S taxonomic anchors" for each sample. Contig alignments for mate-pairs of these 16S anchor reads were identified, and taxonomic assignments were made for scaffold bins containing contigs connected (via mate-pairs in anchors) to 16S rDNA sequences classifying to a single genus.

Alignment of binned scaffolds to the alpha-*Proteobacterium* str. HTCC2255 genome

The candidate genome bins from the October sample included a set of scaffolds with contigs connected to a 16S sequence (RDP ID: S000380128) classified as genus *Thalassobacter*. This sequence corresponds identically to the 16S sequence from the genome sequence of alpha-*Proteobacterium* str. HTCC2255 (NZ_DS022282), which was also identified as the most highly covered genome from the analysis of the October sample alignment with the RefSeq database (Table S3). This candidate genome bin contained 41 scaffolds (grey highlighted sequences, Fig. S10) which were then aligned in nucleotide-space with the alpha-*Proteobacterium* str. HTCC2255 reference genome using the nucmer tool from the MUMMER *(47)* software package version 3.22, using alignment parameters "-b 1500 -g 500 -c 500" and plotted using the mummerplot command with option "--layout" (Fig. S12).

Assembly of binned scaffolds into the Marine Group-II *Euryarchaeote* genome

The screened candidate genome bins from the May sample included a bin containing a set of 16 scaffolds (grey highlighted sequences, Figs. 1 and S11) with contigs connected to the informative regions of 16S sequences classified as Marine Group-II *Euryarchaeota* (Fig. S7). This candidate genome was selected for full assembly.

All quality filtered metagenome reads from the May sample were aligned to contigs in the Marine Group-II *Euryarchaeote* (MG-II) scaffolds, and the results analyzed with SEAStAR, yielding a mate-pair graph forming a single connected component. The set of quality filtered reads recruited in the alignment and their mates were then reassembled using Velvet *(44)* version 1.0.13 with a k-mer size of 21 and parameters "-scaffolding no -read_trkg no -ins_length auto -ins_length_sd auto -exp_cov 1000 -cov_cutoff 13 -min_contig_lgth 75". This reassembly process was repeated nine times, until the N50 of the assembled contigs stabilized. The final assembly of 743 contigs (N50 = 6819) was processed by aligning all quality filtered metagenome reads from the May sample back to the contigs, analyzing the SAM output with SEAStAR and processing the resulting

output graph as described above for the production of metagenome scaffolds. The resulting 11 scaffolds were hand checked for chimeric contigs and layout problems. In this process a highly repetitive region was identified, and a best effort was made to lay out its contigs by hand into the "Repeat region" (RPR). The curated assembly of 11 scaffolds plus the repeat region formed the "draft assembly" of the MG-II genome.

The final assembled version of the MG-II genome was hand curated. Many contigs were merged by identifying unassembled overlaps between neighboring contigs within a scaffold using the Geneious Pro *(42)* assembler version 5.3. A candidate circular arrangement of the scaffolds was determined using the highest scoring mate-pair connections at the scaffold boundaries, and several ambiguous cases were resolved using PCR and Sanger sequencing (see next Method section). This process yielded a single linear scaffold layout with "dangling" mate-pairs connecting to relatively low coverage metagenome contigs at each end.

A process of iterative reassembly similar to that described for the whole genome above, but with relaxed minimum coverages: "-cov_cutoff 1" (first 20 iterations) and "-cov_cutoff auto" (subsequent iterations) was then employed. Proceeding from the last contig on each end of the scaffold layout, after 40 iterations a series of short mate-pair connected contigs forming two distinct low coverage paths across this final gap were observed; each estimated from the mean mate-pair insert size to be less than 20Kbp long.

One of these paths was selected, and a complete assembly across the gap along that set of connected contigs was produced using Sanger sequencing (see next Method section) and the Geneious Pro *(42)* assembler version 5.3, and subsequently verified by realigning with May metagenome reads, obtaining complete coverage across the gap. This assembled sequence forms the ~14 Kbp Hyper-Variable Region (HVR) of the Mar. Group-II final assembly. Quality filtered metagenomic reads from the May sample were aligned back to the final MG-II assembly, the results analyzed with SEAStAR, and position specific genome statistics were generated (Fig S14).

<u>PCR amplification and Sanger sequencing across MG-II assembly scaffold gaps.</u>

Gaps between SOLiD Sequencing scaffolds were closed by designing custom primers to span the gap. Amplification was performed in a C1000 thermal cycler (Bio-Rad Laboratories, Hercules, CA) using varied temperatures and extension times based on the melting temperatures of the primer sets and the anticipated length of the product. Amplification products were visualized by gel electrophoresis and purified from the gel using the Qiagen Minelute Gel Extraction Kit (Qiagen, Gaithersburg, MD). The resulting purified DNA was cloned into a TOPO TA Cloning Kit for Sequencing vector (Invitrogen, Carlsbad, CA) according to the manufacturers instructions. Plasmids containing the insert of interest were purified using the Qiaprep Spin Miniprep Kit (Qiagen). The resulting plasmids were sequenced at the University of Washington Department of Biochemistry using a ABI 3730XL sequencer (Foster City, CA).

Genome annotation

A preliminary annotation of the draft MG-II genome was prepared using the RAST *(48)* service version 4.0. This annotation was used to hand adjust the length of a few small gaps (denoted with 15-17 Ns) in the assembly to avoid the introduction of artifactual frame-shifts within genes that span these gaps. Noncoding RNA genes were identified using RNAmmer *(49)* version 1.2 (Fig. S15). Final gene models for protein coding genes were identified using GeneMarkS *(50)* version 2.6r and Glimmer *(51)* version 3, with differences between the two methods resolved through hand curation.

The proteins coded by these final models were then annotated using InterProScan *(52)*, CDD batch search *(53)*, and by using BLASTP *(54)* version 2.2.25+ to search the following protein databases: COG *(55)*, arCOG *(56)* (2009 update), KEGG *(57)* (Jan 2011), NCBI ProtClustDB *(58)* (November 2010), MEROPS *(59)* release 9.4 and NCBI non-redundant proteins (downloaded April 27, 2011), all with an e-value cutoff of 10^{-5}. Final annotations for each protein were determined by hand, integrating the results of all of the above searches.

Euryarchaeal and Marine Group-II phylogenies

A maximum likelihood phylogeny of a concatenation of 31 conserved archaeal proteins from ten *Euryarchaeota* with available genomes was produced with PHYML *(60)* version 2.4.4 using the JTT substitution model, based on the concatenated amino acid sequences. The proteins were selected following Gao and Gupta *(61)*, and were concatenated in the order shown in Table S5, and aligned with the Geneious Pro *(42)* version 5.3 global alignment tool using the following settings: (Cost matrix: Blosum62, Gap open penalty: 12, Gap extension penalty: 3, Refinement iterations: 3). The tree is shown rooted with the same concatenation of proteins from *Nitrosopumilus maritimus* (NC_010085), a member of phylum *Thaumarchaeota*. Branch lengths are proportional to the number of substitutions per site. Bootstraps are shown for 100 replicate trees. The results are shown in Figure S19A.

A maximum likelihood phylogeny of full-length Marine Group-II 16S rRNA sequences was produced with PHYML *(60)* version 2.4.4 using the JC69 substitution model. 16S rDNA sequences were obtained from the Ribosomal Database Project *(41)* website except for clones generated from the October clone libraries (this study). The standard RDP alignment was used with October clone library sequences added to that alignment (preserving existing gaps) with the Geneious Pro *(42)* version 5.3 global alignment tool using the following settings: (Cost matrix: 70% similarity (IUB), Gap open penalty: 12, Gap extension penalty: 3, Refinement iterations: 2, Type: free end gaps). The tree is shown rooted with the 16S rDNA sequence of *Aciduliprofundum boonei.* Branch lengths are proportional to the number of substitutions per site. Bootstraps are shown for 100 replicate trees. The results are shown in Figure S16.

Alignment of environmental sequences to the MG-II genome

Candidate environmental clones were identified from the top Genbank hits of the MG-II protein BLASTP *(54)* search of the NCBI nr protein database (see genome annotation methods above). Candidate metagenomic assemblies were identified from the top hits of a MG-II protein BLASTP search of the NCBI env-nr protein database. Corresponding nucleotide sequences for both sets of environmental sequence hits were then retrieved from Genbank.

Each of these sets of environmental sequences were separately aligned with the MG-II genome final assembly using the PROMER tool from the MUMMER *(47)* software package version 3.22 using alignment parameters "--maxmatch -c 15 -b 100 -g 50". The show-tiling command was then used with parameters "-a -i 50 -V 0 -v 50 -g 20000" to produce the tiled placements of environmental clones and metagenomic assemblies shown in Figure 2A. Gene model alignments between selected environmental clone sequences and the MG-II genome were made using the Reciprocal Best Hits approach described below, and plotted using a customized version of the 'R' language package genoPlotR *(62)* version 0.6.1 (Figs. 2B and S17).

Proteorhodopsin phylogeny

Selected rhodopsin proteins from sequenced genomes and environmental clones were downloaded from GenBank. Full-length rhodopsin proteins from the NCBI env-nr database were identified using Marine Group-II rhodopsins identified through environmental sequence alignments (Fig. 2B) as BLASTP *(54)* version 2.2.25+ query sequences with an e-value cutoff of 10^{-30}, with sequence hits covering less than 75% of query genes rejected. The 80 rhodopsin amino acid sequences selected were aligned with the Geneious Pro *(42)* version 5.3 global alignment tool using the following settings: (Cost matrix: Blosum62, Gap open penalty: 12, Gap extension penalty: 3, Refinement iterations: 3).

A Bayesian phylogeny of rhodopsin proteins was produced from this alignment with MrBayes *(63)* version 3.1.2 using a poisson rate matrix, gamma rate variation with 4 categories and the sequence YP_325981, the halorhodopsin from *Natronomonas pharaonis*, as the outgroup. Additional MrBayes parameters used were: "Chain length: 1000000, Subsampling Freq: 200, Heated chains: 4, Burn-in length: 100000, Heated chain temp: 0.2, and Unconstrained branch lengths: 10". An unrooted cladogram of the resulting phylogeny is shown in Figure 3 and the full tree in Figure S18. Neighbor-joining and Maximum-likelihood trees made from this alignment (not shown) produced substantially the same topology at all major branches as the above Bayesian analysis.

Syntenous gene models flanking the rhodopsin protein consistent with the MG-II genome assembly were identified using Reciprocal Best Hit (RBH) analysis *(64)* (described below) of all protein models corresponding to the environmental assembles

selected for their rhodopsin proteins above. The positions of environmental models homologous to syntenous MG-II proteins were then hand curated. Results of this analysis are summarized in Figure 3 and detailed in Figure S18.

Reciprocal blast homologies

All protein homology inferences used for genome comparisons and alignment of gene models between environmental sequences and the MG-II genome were produced using the Reciprocal Best Hits (RBH) approach *(64)*. Analysis was performed with custom software utilizing BLASTP *(54)* version 2.2.25+ with an e-value cutoff of 10^{-5} for all amino acid searches. Pairwise RBH analyses were performed for seven archaeal genomes: MG-II (this study), *Aciduliprofundum boonei* (NC_013926), *Picrophilus torridus* (NC_005877), *Thermoplasma volcanium* (NC_002689), *Archaeoglobus fulgidus* (NC_000917), *Pyrococcus abyssi* (NC_000868), and *Nitrosopumilus maritimus* (NC_010085), see Table S9 for results.

A 3-way comparison of homologous proteins coded by the MG-II, *A. boonei*, and *N. maritimus* genomes was also performed, with transitivity of RBHs as a requirement for classification in the "core" set of 291 archaeal homologs detected (Figs. S15, S19B). The set of ribosomal proteins in the MG-II and *A. boonei* genomes was cataloged and found to be identical with all syntenous arrangements preserved between the two genomes (Table S6). "Non-homologous proteins", those without RBHs to either other comparison genome, were identified for the MG-II, *A. boonei*, and *N. maritimus* genomes.

Taxonomic classification of proteins

"Non-homologous" proteins coded by genes found among the MG-II, *A. boonei*, and *N. maritimus* geneomes were used to query the RefSeq subset of the NCBI nr protein database (downloaded April 27, 2011) using BLASTP *(54)* version 2.2.25+ with an e-value cutoff of 10^{-5} and a maximum of 500 hits. The software package MEGAN *(65)* version 4.40.5 was used to assign a taxonomic classification to each "non-homologous" protein based on its lowest common ancestor (LCA) algorithm.

An evaluation of the impact of taxonomic filtering of blast hits on the MEGAN classification results was made for all three genomes (Fig. S20E) and filtering was subsequently used only for the *A. boonei* and *N. maritimus* analyses. For MG-II and *A. boonei*, taxonomic filtering excluded hits to RefSeq proteins from five genomes: *A. boonei*, taxid: 439481; *Thermoplasma acidophilum*, taxid: 273075; *Thermoplasma volcanium*, taxid: 273116; *Picrophilus torridus*, taxid: 263820; and *Ferroplasma acidarmanus*, taxid: 333146. For *N. maritumus*, taxonomic filtering excluded hits to proteins from three RefSeq genomes: *N. maritumus*, taxid: 436308; *Candidatus Nitrosoarchaeum limnia*, taxid: 886738; and *Cenarchaeum symbosium*, taxid: 414004.

An evaluation of the effect of three settings (5, 10 and 20%) of the "Top Percent" parameter of MEGAN's LCA algorithm was conducted (Fig. S20F), and the value of 10%

selected for all subsequent use. Results of the MEGAN protein classifications are summarized in Figures S19C, S20A-S20D, and Table S10. Gene models from the "non-homologous" proteins in the MG-II genome that MEGAN classified within the domain Bacteria are plotted on Figure S15.

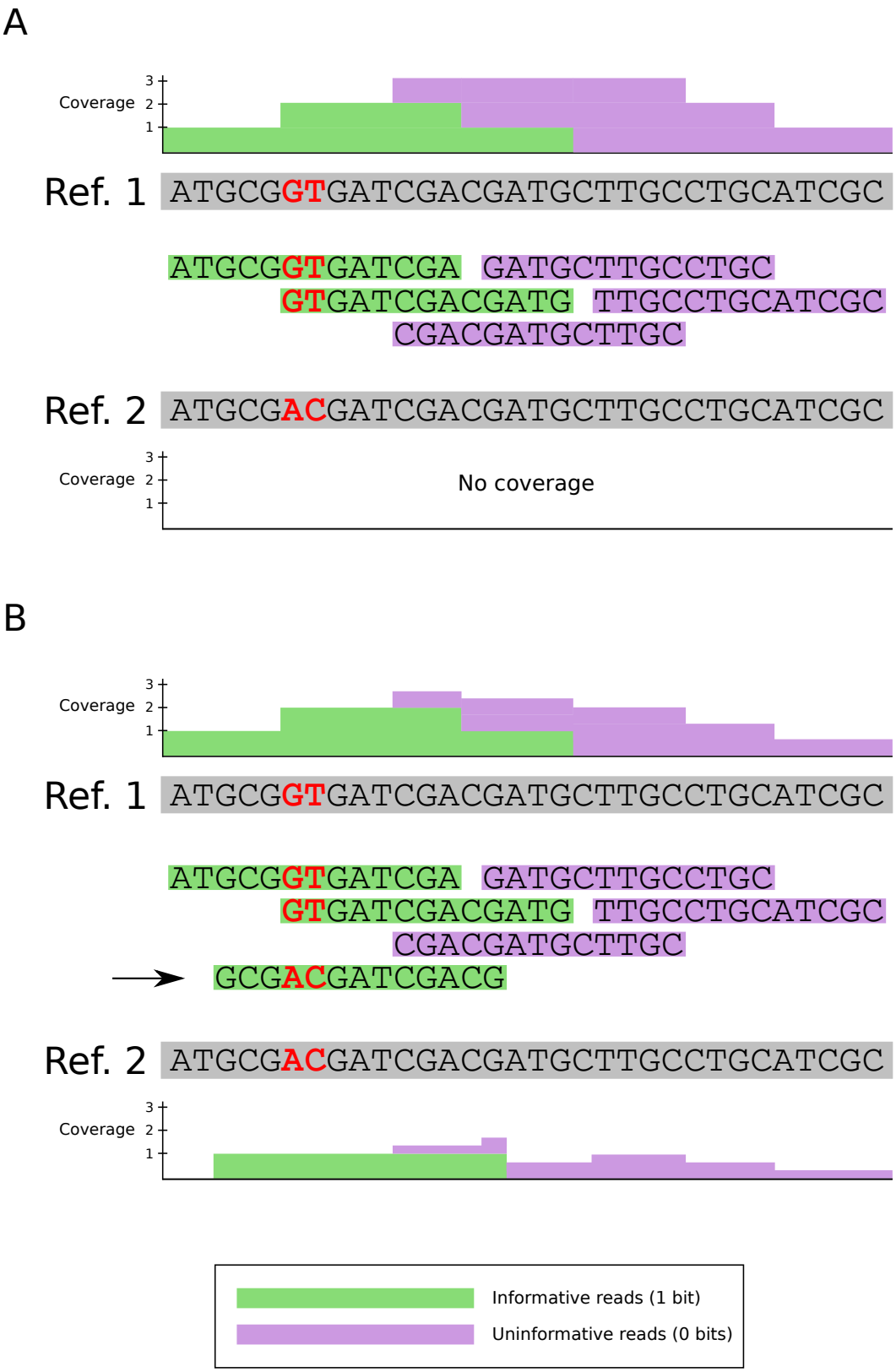

A
Coverage
3
2
1
Ref. 1
ATGCGGTGATCGACGATGCTTGCCTGCATCGC
ATGCGGTGATCGA
GATGCTTGCCTGC
GTGATCGACGATG
TTGCCTGCATCGC
CGACGATGCTTGC
Ref. 2
ATGCGACGATCGACGATGCTTGCCTGCATCGC
Coverage
3
2
1
No coverage
B
Coverage
3
2
1
Ref. 1
ATGCGGTGATCGACGATGCTTGCCTGCATCGC
ATGCGGTGATCGA
GATGCTTGCCTGC
GTGATCGACGATG
TTGCCTGCATCGC
CGACGATGCTTGC
GCGACGATCGACG
Ref. 2
ATGCGACGATCGACGATGCTTGCCTGCATCGC
Coverage
3
2
1
Informative reads (1 bit)
Uninformative reads (0 bits)

Fig. S1

Illustration of reference selection, fractional coverage allocation, and relative abundance estimation using sequence read alignments to a reference database. The examples in this figure use a simple database with only two reference sequences ("Ref. 1" and "Ref. 2"). These two sequences differ by only two nucleotides (red). Aligned reads are shown located between the two references, and are shaded by information content relative to the reference database. Informative reads (green) each provide 1 bit of information, because they align equally well with one of the two references ($-\log_2(1/2) = 1$ bit). Uninformative reads (purple) each provide 0 bits of information, because they align equally well with both of the references ($-\log_2(2/2) = 0$ bits). **(A)** In this example, Ref. 1 is selected from the database because reads scoring 2 bits have best alignments with it. Ref. 2 is not selected because no informative reads have a best alignment with it. All coverage from aligned reads (shown on graphs) goes to Ref. 1 because it is the only selected reference. Its relative abundance is 100% because its mean coverage equals the total coverage from all reads. **(B)** This example is identical to (A) above, except that an additional informative read (arrow) is added with a best alignment to Ref. 2. Ref. 1 is selected from the database first because 2 bits of reads best align with it, and that is the maximum for all reference sequences. Ref. 2 is selected next from the database because it has 1 bit of (residual) information after reads that have best alignments with previously selected references are removed from consideration (e.g. if Ref 1. and Ref 2. were identical, only one of them could be selected). Coverage for reads with unique best alignments is allocated first (shown on graphs in green), and coverage for reads aligning equivalently with the two references is allocated fractionally (i.e. shared) with weighting determined by the relative mean amount of previously accumulated coverage (in this case 2/3 for Ref. 1 and 1/3 for Ref. 2, shown on graphs in purple). Once the coverage for all reads is allocated, the relative abundance for the selected sequences is calculated from the mean coverage values relative to total coverage allocated: Ref. 1 ~66.7%, Ref. 2 ~33.3%.

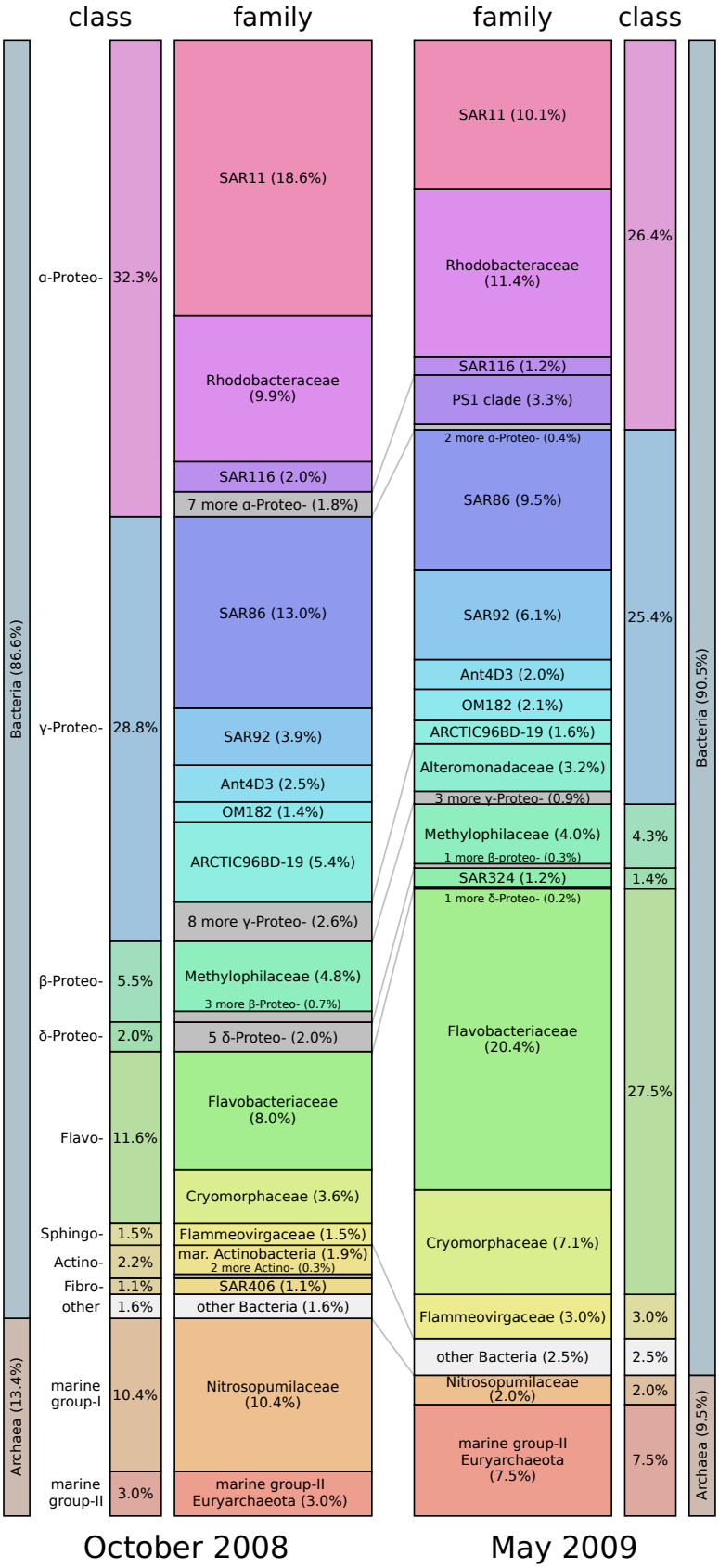

class
family
family
class
Bacteria (86.6%)
Archaea (13.4%)
Bacteria (90.5%)
Archaea (9.5%)
α-Proteo-
γ-Proteo-
β-Proteo-
δ-Proteo-
Flavo-
Sphingo-
Actino-
Fibro-
other
marine group-I
marine group-II
32.3%
28.8%
5.5%
2.0%
11.6%
1.5%
2.2%
1.1%
1.6%
10.4%
3.0%
SAR11 (18.6%)
Rhodobacteraceae (9.9%)
SAR116 (2.0%)
7 more α-Proteo- (1.8%)
SAR86 (13.0%)
SAR92 (3.9%)
Ant4D3 (2.5%)
OM182 (1.4%)
ARCTIC96BD-19 (5.4%)
8 more γ-Proteo- (2.6%)
Methylophilaceae (4.8%)
3 more β-Proteo- (0.7%)
5 δ-Proteo- (2.0%)
Flavobacteriaceae (8.0%)
Cryomorphaceae (3.6%)
Flammeovirgaceae (1.5%)
mar. Actinobacteria (1.9%)
2 more Actino- (0.3%)
SAR406 (1.1%)
other Bacteria (1.6%)
Nitrosopumilaceae (10.4%)
marine group-II Euryarchaeota (3.0%)
SAR11 (10.1%)
Rhodobacteraceae (11.4%)
SAR116 (1.2%)
PS1 clade (3.3%)
2 more α-Proteo- (0.4%)
SAR86 (9.5%)
SAR92 (6.1%)
Ant4D3 (2.0%)
OM182 (2.1%)
ARCTIC96BD-19 (1.6%)
Alteromonadaceae (3.2%)
3 more γ-Proteo- (0.9%)
Methylophilaceae (4.0%)
1 more β-proteo- (0.3%)
SAR324 (1.2%)
1 more δ-Proteo- (0.2%)
Flavobacteriaceae (20.4%)
Cryomorphaceae (7.1%)
Flammeovirgaceae (3.0%)
other Bacteria (2.5%)
Nitrosopumilaceae (2.0%)
marine group-II Euryarchaeota (7.5%)
26.4%
25.4%
4.3%
1.4%
27.5%
3.0%
2.5%
2.0%
7.5%
October 2008
May 2009

Fig. S2

Relative abundance of taxonomic groups of Bacteria and Archaea for the October and May samples, based on analysis of metagenomic reads recruited to sequences in the RDP 16S rDNA database. Stacked bars correspond to relative 16S abundance at the domain, class and family taxonomic levels (or approximate equivalent for environmental clades). Colors correspond to the same groups in each sample. White bars represent classes (and their families) that are < 1% relative abundance. Grey bars represent families within a class that are < 1%. Grey lines in the center encompass families collapsed into white or grey bars in one sample or the other.

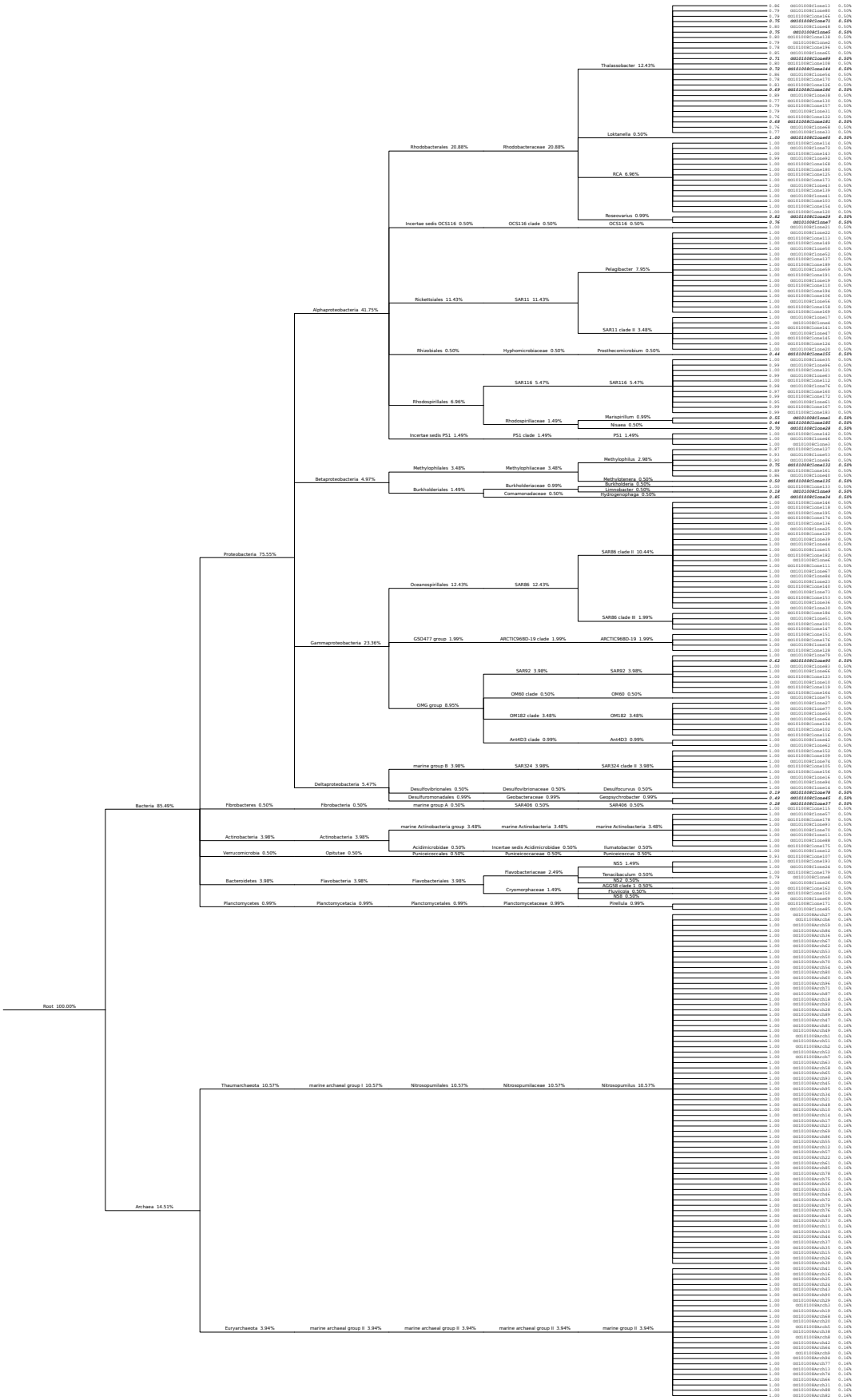

Root 100.00%
Bacteria 85.49%
Archaea 14.51%
Proteobacteria 75.55%
Alphaproteobacteria 41.75%
Betaproteobacteria 4.97%
Gammaproteobacteria 23.36%
Deltaproteobacteria 5.47%
Fibrobacteres 0.50%
Actinobacteria 3.98%
Verrucomicrobia 0.50%
Bacteroidetes 3.98%
Planctomycetes 0.99%
Thaumarchaeota 10.57%
Euryarchaeota 3.94%
Fibrobacteria 0.50%
Actinobacteria 3.98%
Opitutae 0.50%
Flavobacteria 3.98%
Planctomycetacia 0.99%
marine archaeal group I 10.57%
marine archaeal group II 3.94%
marine group A 0.50%
marine Actinobacteria group 3.48%
Acidimicrobidae 0.50%
Puniceicoccales 0.50%
Flavobacteriales 3.98%
Planctomycetales 0.99%
Nitrosospumilales 10.57%
marine archaeal group II 3.94%
SAR406 0.50%
marine Actinobacteria 3.48%
Incertae sedis Acidimicrobidae 0.50%
Puniceicoccaceae 0.50%
Flavobacteriaceae 2.49%
Cryomorphaceae 1.49%
Planctomycetaceae 0.99%
Nitrosospumilaceae 10.57%
marine archaeal group II 3.94%
SAR406 0.50%
marine Actinobacteria 3.48%
Ilumatobacter 0.50%
Puniceicoccus 0.50%
NS5 1.49%
Tenacibaculum 0.50%
NS7 0.50%
AGG58 clade 1 0.50%
Flavobacteria 0.50%
NS9 0.50%
Pirellula 0.99%
Nitrosospumilus 10.57%
marine group II 3.94%
Rhodobacterales 20.88%
Incertae sedis OCS116 0.50%
Rickettsiales 11.43%
Rhizobiales 0.50%
Rhodospirillales 6.96%
Incertae sedis PS1 1.49%
Methylophilales 3.48%
Burkholderiales 1.49%
Oceanospirillales 12.43%
GSO477 group 1.99%
OMG group 8.95%
marine group B 3.98%
Desulfovibrionales 0.50%
Desulfuromonadales 0.99%
Rhodobacteraceae 20.88%
OCS116 clade 0.50%
SAR11 11.43%
Hyphomicrobiaceae 0.50%
SAR116 5.47%
Rhodospirillaceae 1.49%
PS1 clade 1.49%
Methylophilaceae 3.48%
Burkholderiaceae 0.99%
Comamonadaceae 0.50%
SAR86 12.43%
ARCTIC96BD-19 clade 1.99%
SAR92 3.98%
OM60 clade 0.50%
OM182 clade 3.48%
Arn4D3 clade 0.99%
SAR324 3.98%
Desulfovibrionaceae 0.50%
Geobacteraceae 0.99%
Thalassobacter 12.43%
Loktanella 0.50%
RCA 6.96%
Roseovarius 0.99%
OCS116 0.50%
Pelagibacter 7.95%
SAR11 clade II 3.48%
Prosthecomicrobium 0.50%
SAR116 5.47%
Marispirillum 0.99%
Nisaea 0.50%
PS1 1.49%
Methylophilus 2.98%
Methylotenera 0.50%
Burkholderia 0.50%
Limnobacter 0.50%
Hydrogenophaga 0.50%
SAR86 clade II 10.44%
SAR86 clade III 1.99%
ARCTIC96BD-19 1.99%
SAR92 3.98%
OM60 0.50%
OM182 3.48%
Arn4D3 0.99%
SAR324 clade II 3.98%
Desulfocurvus 0.50%
Geopsychrobacter 0.99%

Fig. S3

Taxonomic cladogram showing the Bayesian classification of all 16S rDNA clones generated from the October 2008 sample. Taxonomic groups, labelled on branches, are from the RDP taxonomy extended with marine environmental clades (see Methods). Clone identifiers are noted at the leaves. Bayesian classification p-values for the clone sequences are shown to the left of the clone identifiers. Percentages noted on each branch and leaf represent the abundance of that taxonomic unit as a fraction of total clones generated from the October 2008 sample, with the fraction of bacterial and archaeal clones fixed as 85.49% and 14.51% of the population respectively, as determined from aligning metagenomic reads to the October 16S clone sequences (Fig. S9). Clones identified in bold-italic type were excluded from abundance analysis simulations.

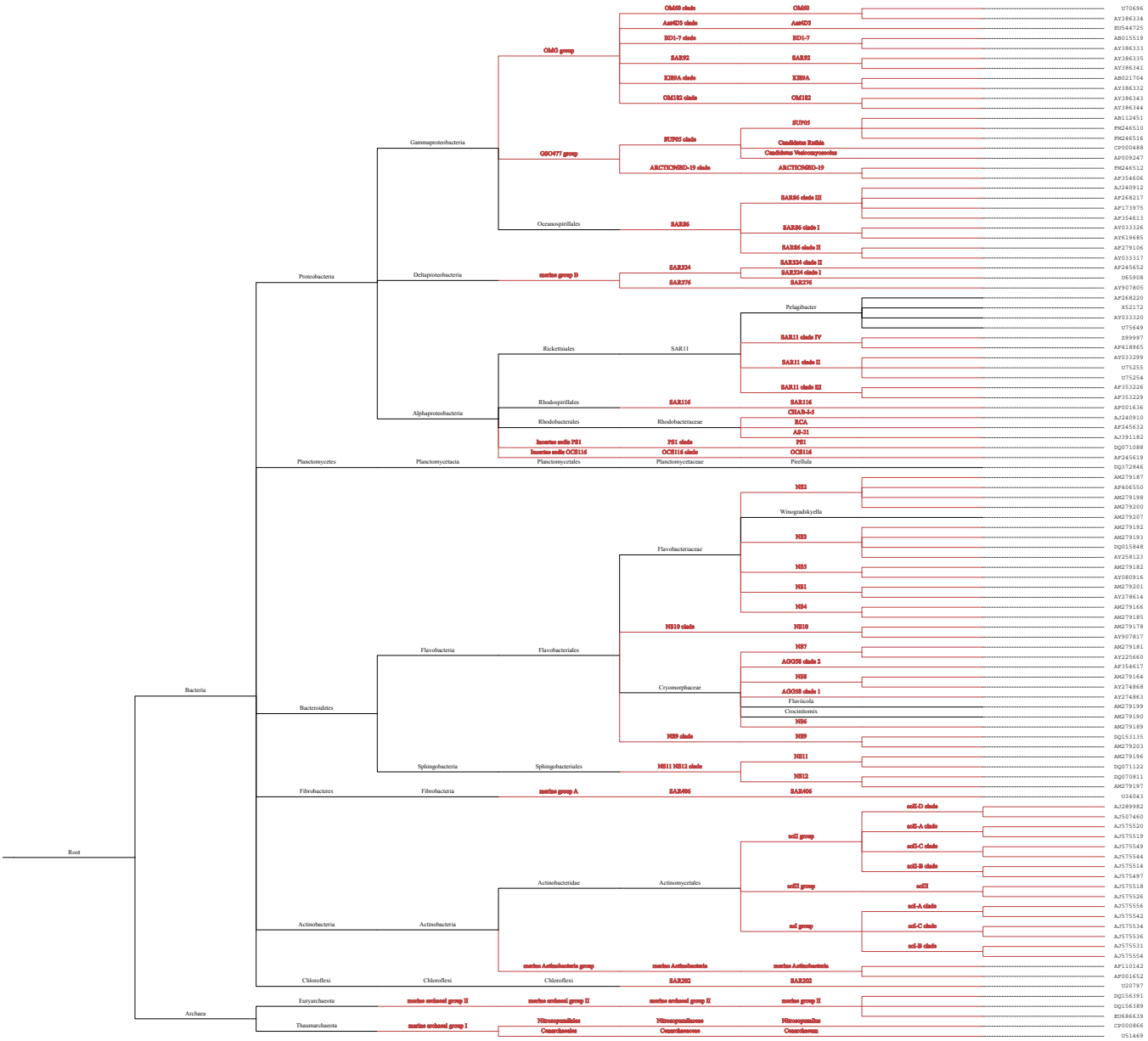


Fig. S4

Taxonomic cladogram of environmental 16S rDNA sequences added to the training set of the RDP Bayesian classifier. Branches in black denote pre-existing RDP taxonomy. Red branches denote environmental clades added to the RDP taxonomy for this study. Leaves of the tree show accession numbers of the sequences added to the training set.

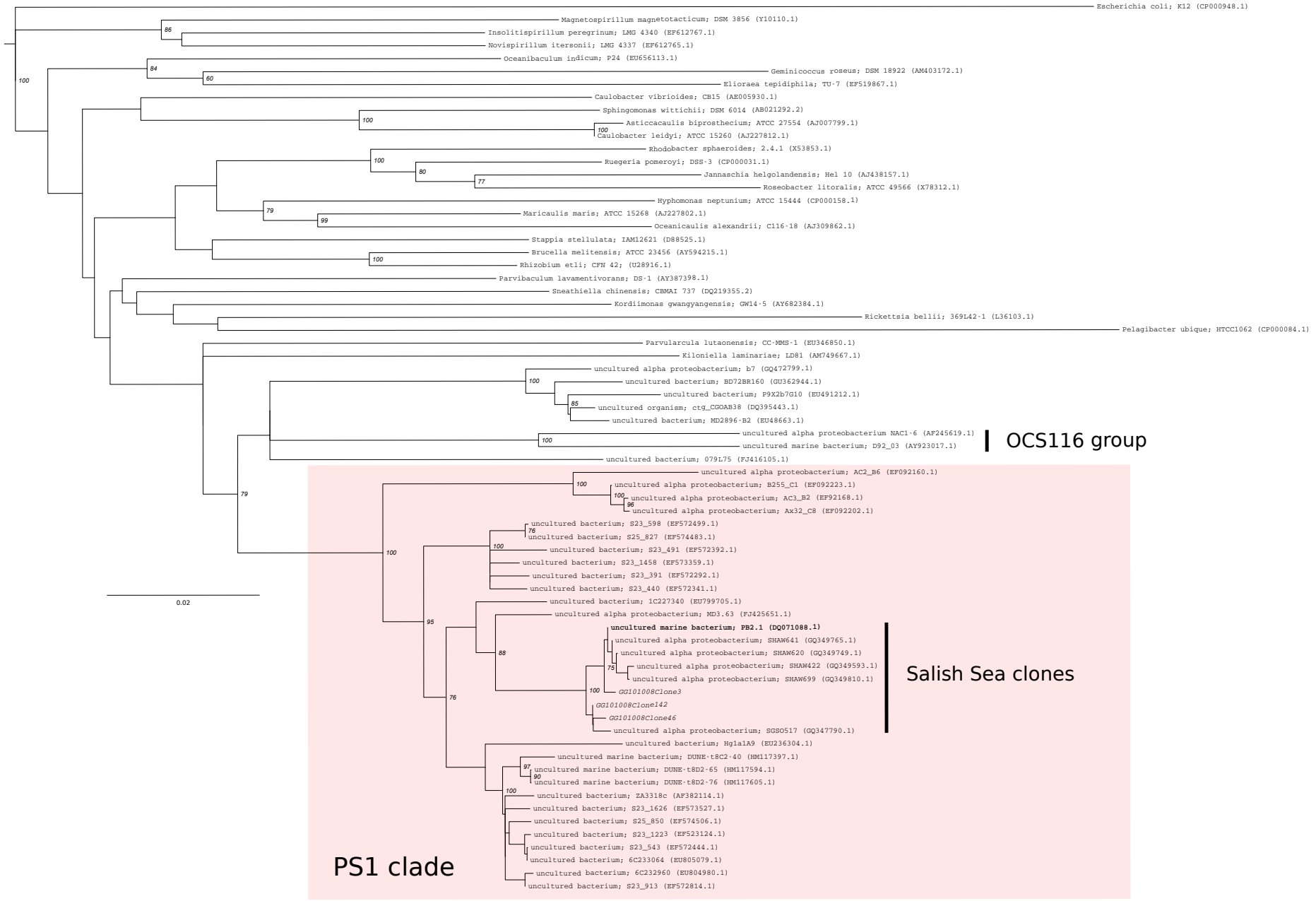

Escherichia coli; K12 (CP000948.1)
Magnetospirillum magnetotacticum; DSM 3856 (Y10110.1)
Insolitispirillum peregrinum; LMG 4340 (EF612767.1)
Novispirillum itersonii; LMG 4337 (EF612765.1)
Oceanibaculum indicum; P24 (EU656113.1)
Geminicoccus roseus; DSM 18922 (AM403172.1)
Elioraea tepidiphila; TU-7 (EF519867.1)
Caulobacter vibrioides; CB15 (AE005930.1)
Sphingomonas wittichii; DSM 6014 (AB021292.2)
Asticcacaulis biprosthecium; ATCC 27554 (AJ007799.1)
Caulobacter leidyi; ATCC 15260 (AJ227812.1)
Rhodobacter sphaeroides; 2.4.1 (X53853.1)
Ruegeria pomeroyi; DSS-3 (CP000031.1)
Jannaschia helgolandensis; Hel 10 (AJ438157.1)
Roseobacter litoralis; ATCC 49566 (X78312.1)
Hyphomonas neptunium; ATCC 15444 (CP000158.1)
Maricaulis maris; ATCC 15268 (AJ227802.1)
Oceanicaulis alexandrii; C116-18 (AJ309862.1)
Stappia stellulata; IAM12621 (D88525.1)
Brucella melitensis; ATCC 23456 (AY594215.1)
Rhizobium etli; CFN 42; (U28916.1)
Parvibaculum lavamentivorans; DS-1 (AY387398.1)
Sneathiella chinensis; CBMAI 737 (DQ219355.2)
Kordiimonas gwangyangensis; GW14-5 (AY682384.1)
Rickettsia bellii; 369L42-1 (L36103.1)
Pelagibacter ubique; HTCC1062 (CP000084.1)
Parvularcula lutaonensis; CC-MMS-1 (EU346850.1)
Kiloniella laminariae; LD81 (AM749667.1)
uncultured alpha proteobacterium; b7 (GQ472799.1)
uncultured bacterium; BD72BR160 (GU362944.1)
uncultured bacterium; P9X2b7G10 (EU491212.1)
uncultured organism; ctg_CGOAB38 (DQ395443.1)
uncultured bacterium; MD2896-B2 (EU48663.1)
uncultured alpha proteobacterium NAC1-6 (AF245619.1)
uncultured marine bacterium; D92_03 (AY923017.1)
uncultured bacterium; 079L75 (FJ416105.1)
OCS116 group
uncultured alpha proteobacterium; AC2_B6 (EF092160.1)
uncultured alpha proteobacterium; B255_C1 (EF092223.1)
uncultured alpha proteobacterium; AC3_B2 (EF92168.1)
uncultured alpha proteobacterium; Ax32_C8 (EF092202.1)
uncultured bacterium; S23_598 (EF572499.1)
uncultured bacterium; S25_827 (EF574483.1)
uncultured bacterium; S23_491 (EF572392.1)
uncultured bacterium; S23_1458 (EF573359.1)
uncultured bacterium; S23_391 (EF572292.1)
uncultured bacterium; S23_440 (EF572341.1)
uncultured bacterium; 1C227340 (EU799705.1)
uncultured alpha proteobacterium; MD3.63 (FJ425651.1)
uncultured marine bacterium; PB2.1 (DQ071088.1)
uncultured alpha proteobacterium; SHAW641 (GQ349765.1)
uncultured alpha proteobacterium; SHAW620 (GQ349749.1)
uncultured alpha proteobacterium; SHAW422 (GQ349593.1)
uncultured alpha proteobacterium; SHAW699 (GQ349810.1)
GG101008Clone3
GG101008Clone142
GG101008Clone46
uncultured alpha proteobacterium; SGSO517 (GQ347790.1)
Salish Sea clones
uncultured bacterium; Hg1a1A9 (EU236304.1)
uncultured marine bacterium; DUNE-t8C2-40 (HM117397.1)
uncultured marine bacterium; DUNE-t8D2-65 (HM117594.1)
uncultured marine bacterium; DUNE-t8D2-76 (HM117605.1)
uncultured bacterium; ZA3318c (AF382114.1)
uncultured bacterium; S23_1626 (EF573527.1)
uncultured bacterium; S25_850 (EF574506.1)
uncultured bacterium; S23_1223 (EF523124.1)
uncultured bacterium; S23_543 (EF572444.1)
uncultured bacterium; 6C233064 (EU805079.1)
uncultured bacterium; 6C232960 (EU804980.1)
uncultured bacterium; S23_913 (EF572814.1)
PS1 clade
0.02

Fig. S5

Neighbor-joining phylogeny of alpha-*Proteobacterial* 16S rDNA sequences, showing a novel clade of marine alpha-*Proteobacteria* designated as “PS1”. The tree is rooted with *Escherichia coli*, a member of the gamma-*Proteobacteria*. Branch lengths are proportional to the number of substitutions per site. Bootstraps are shown for 100 replicate trees. The PS1 clade is boxed in pink. A sub-clade of sequences isolated from the vicinity of Puget Sound, WA, USA and the Straights of Georgia, BC, Canada are noted as “Salish Sea clones”. The taxon in bold type was sequenced from Puget Sound in a previous study. Taxa in italic type are clones sequenced in this study.

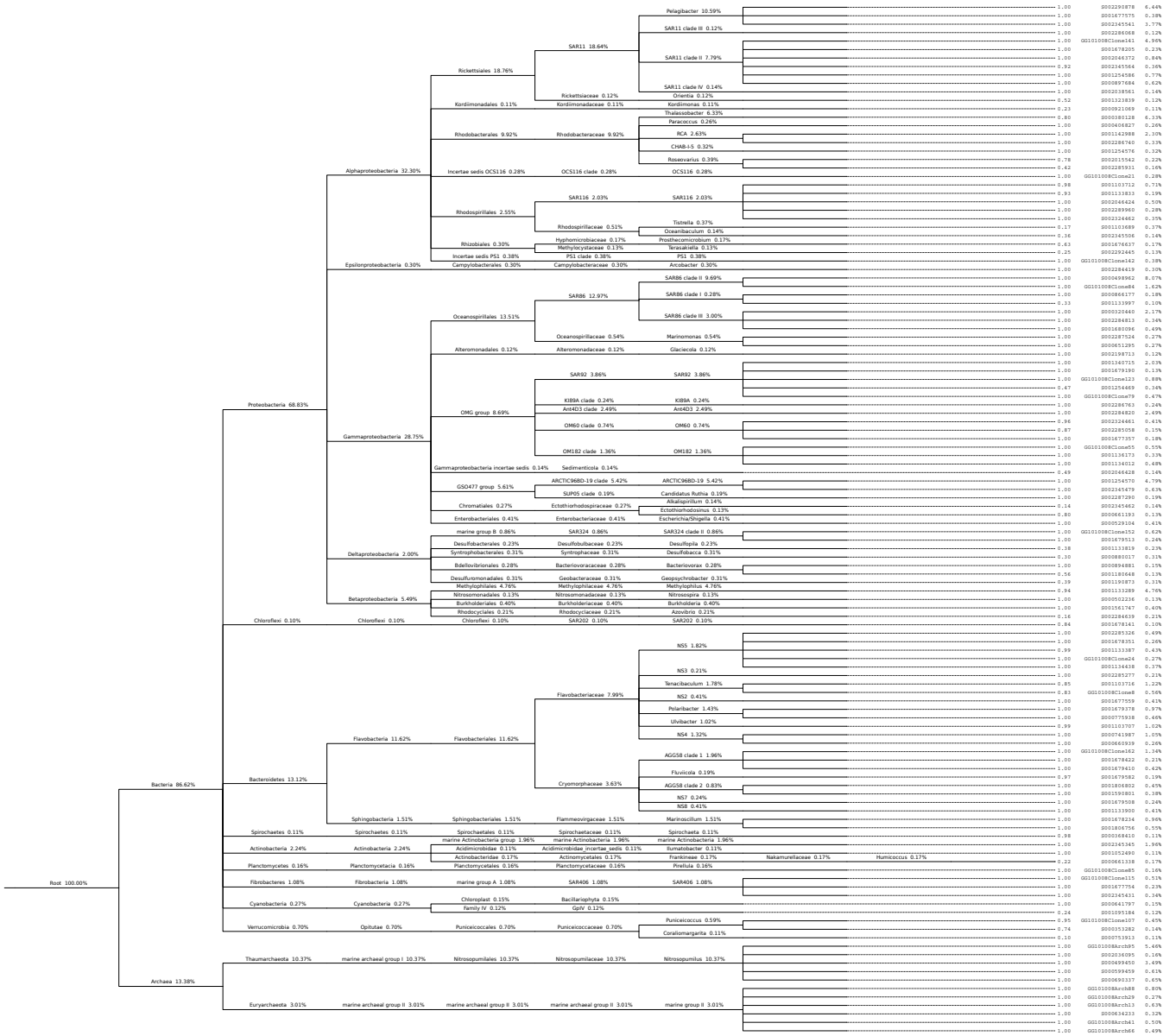


Fig. S6

Taxonomic cladogram detailing all 16S rDNA sequences selected for the October 2008 sample. Taxonomic groups, labelled on branches, are from the extended RDP taxonomy. RDP identifiers for selected sequences are noted at the leaves, except for those prefaced with "GG" which denote clones generated in this study. Bayesian classification p-values for the selected sequences are shown to the left of the clone identifiers. Percentages noted on each branch and leaf represent the estimated abundance of that taxonomic unit as a fraction of total 16S rDNA in the sample.

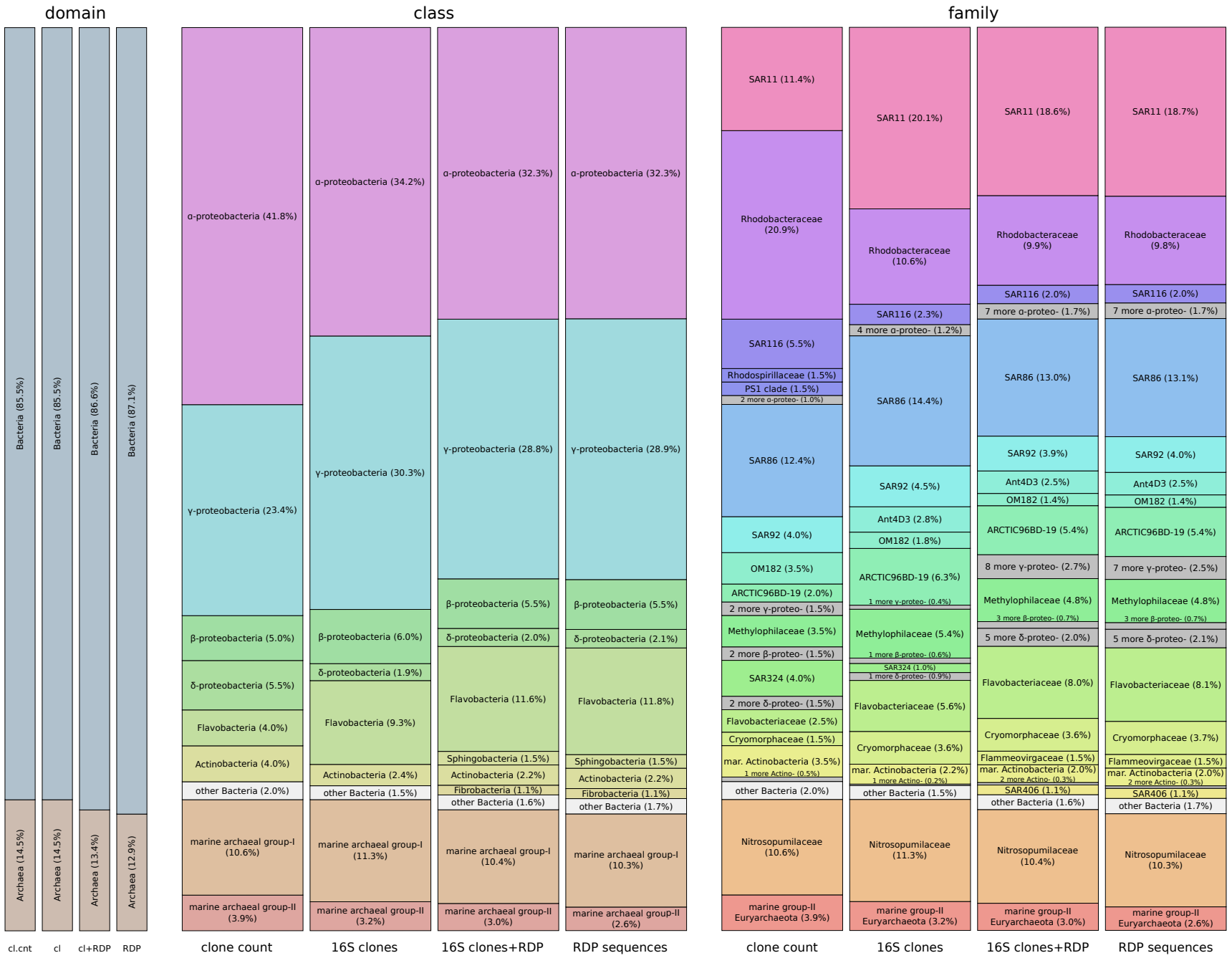

domain
class
family
Bacteria (85.5%)
Bacteria (85.5%)
Bacteria (86.6%)
Bacteria (87.1%)
Archaea (14.5%)
Archaea (14.5%)
Archaea (13.4%)
Archaea (12.9%)
α-proteobacteria (41.8%)
γ-proteobacteria (23.4%)
β-proteobacteria (5.0%)
δ-proteobacteria (5.5%)
Flavobacteria (4.0%)
Actinobacteria (4.0%)
other Bacteria (2.0%)
marine archaeal group-I (10.6%)
marine archaeal group-II (3.9%)
α-proteobacteria (34.2%)
γ-proteobacteria (30.3%)
β-proteobacteria (6.0%)
δ-proteobacteria (1.9%)
Flavobacteria (9.3%)
Actinobacteria (2.4%)
other Bacteria (1.5%)
marine archaeal group-I (11.3%)
marine archaeal group-II (3.2%)
α-proteobacteria (32.3%)
γ-proteobacteria (28.8%)
β-proteobacteria (5.5%)
δ-proteobacteria (2.0%)
Flavobacteria (11.6%)
Sphingobacteria (1.5%)
Actinobacteria (2.2%)
Fibrobacteria (1.1%)
other Bacteria (1.6%)
marine archaeal group-I (10.4%)
marine archaeal group-II (3.0%)
α-proteobacteria (32.3%)
γ-proteobacteria (28.9%)
β-proteobacteria (5.5%)
δ-proteobacteria (2.1%)
Flavobacteria (11.8%)
Sphingobacteria (1.5%)
Actinobacteria (2.2%)
Fibrobacteria (1.1%)
other Bacteria (1.7%)
marine archaeal group-I (10.3%)
marine archaeal group-II (2.6%)
SAR11 (11.4%)
Rhodobacteraceae (20.9%)
SAR116 (5.5%)
Rhodospirillaceae (1.5%)
PS1 clade (1.5%)
2 more α-proteo- (1.0%)
SAR86 (12.4%)
SAR92 (4.0%)
OM182 (3.5%)
ARCTIC96BD-19 (2.0%)
2 more γ-proteo- (1.5%)
Methylophilaceae (3.5%)
2 more β-proteo- (1.5%)
SAR324 (4.0%)
2 more δ-proteo- (1.5%)
Flavobacteriaceae (2.5%)
Cryomorphaceae (1.5%)
mar. Actinobacteria (3.5%)
1 more Actino- (0.5%)
other Bacteria (2.0%)
Nitrosopumilaceae (10.6%)
marine group-II Euryarchaeota (3.9%)
SAR11 (20.1%)
Rhodobacteraceae (10.6%)
SAR116 (2.3%)
4 more α-proteo- (1.2%)
SAR86 (14.4%)
SAR92 (4.5%)
Ant4D3 (2.8%)
OM182 (1.8%)
ARCTIC96BD-19 (6.3%)
1 more γ-proteo- (0.4%)
Methylophilaceae (5.4%)
1 more β-proteo- (0.6%)
SAR324 (1.0%)
1 more δ-proteo- (0.9%)
Flavobacteriaceae (5.6%)
Cryomorphaceae (3.6%)
mar. Actinobacteria (2.2%)
1 more Actino- (0.2%)
other Bacteria (1.5%)
Nitrosopumilaceae (11.3%)
marine group-II Euryarchaeota (3.2%)
SAR11 (18.6%)
Rhodobacteraceae (9.9%)
SAR116 (2.0%)
7 more α-proteo- (1.7%)
SAR86 (13.0%)
SAR92 (3.9%)
Ant4D3 (2.5%)
OM182 (1.4%)
ARCTIC96BD-19 (5.4%)
8 more γ-proteo- (2.7%)
Methylophilaceae (4.8%)
3 more β-proteo- (0.7%)
5 more δ-proteo- (2.0%)
Flavobacteriaceae (8.0%)
Cryomorphaceae (3.6%)
Flammeovirgaceae (1.5%)
mar. Actinobacteria (2.0%)
2 more Actino- (0.3%)
SAR406 (1.1%)
other Bacteria (1.6%)
Nitrosopumilaceae (10.4%)
marine group-II Euryarchaeota (3.0%)
SAR11 (18.7%)
Rhodobacteraceae (9.8%)
SAR116 (2.0%)
7 more α-proteo- (1.7%)
SAR86 (13.1%)
SAR92 (4.0%)
Ant4D3 (2.5%)
OM182 (1.4%)
ARCTIC96BD-19 (5.4%)
7 more γ-proteo- (2.5%)
Methylophilaceae (4.8%)
3 more β-proteo- (0.7%)
5 more δ-proteo- (2.1%)
Flavobacteriaceae (8.1%)
Cryomorphaceae (3.7%)
Flammeovirgaceae (1.5%)
mar. Actinobacteria (2.0%)
2 more Actino- (0.3%)
SAR406 (1.1%)
other Bacteria (1.7%)
Nitrosopumilaceae (10.3%)
marine group-II Euryarchaeota (2.6%)
cl.cnt
cl
cl+RDP
RDP
clone count
16S clones
16S clones+RDP
RDP sequences
clone count
16S clones
16S clones+RDP
RDP sequences

Fig. S7

Relative abundance of Bacterial and Archaeal taxonomic groups for the October sample, a comparison of four different estimation methods. Stacked bars correspond to the relative prokaryotic 16S abundance for hierarchical groups at the domain, class and family taxonomic levels (or the approximate equivalent for environmental clades.) The abundance estimation methods are: October 2008 16S rDNA clone counts (cl.cnt) as shown in Figure S3; and metagenomic read alignments to October 16S clones only (cl), October 16S clones added to the RDP 16S database (cl+RDP) and the RDP 16S database only (RDP). Colors correspond to the same groups for each estimation method. White bars contain classes (and their families) that are < 1% relative abundance. Grey bars contain families within a class that are < 1%.

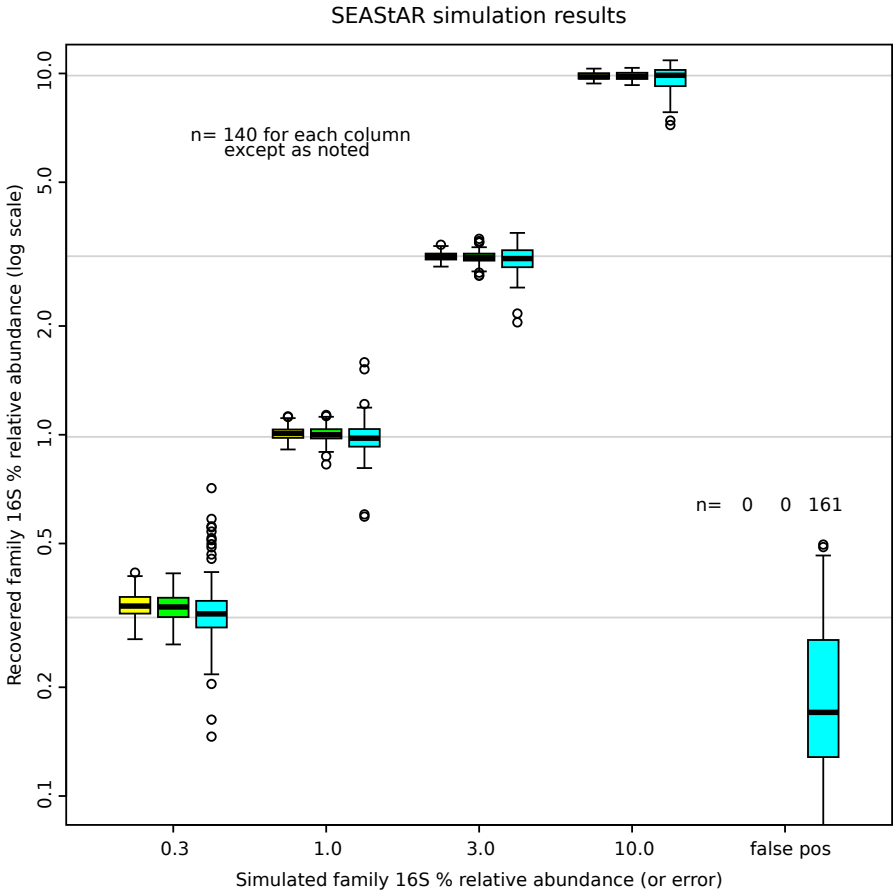


Fig. S8

Results of abundance estimation for synthetic metagenomic datasets generated from simulated populations of 16S rDNA sequences. This figure depicts simulation results from the analysis of twenty synthetic metagenomic datasets, each constructed by randomly drawing a fixed number of reads from significantly larger pools generated from randomly selected 16S sequences in fixed proportions, shown on the y-axis. The x-axis shows the four classes of simulated taxa abundances that were successfully recovered and classified to the family taxonomic level. A fifth category (false pos) on the x-axis contains taxa that were detected, but which did not correspond (at the family level) to one of the taxa used to generate that simulated population. There were no false negatives detected (i.e. all taxa in all trial populations were recovered in the simulations.) The three columns within each x-axis category are for analysis using different reference databases (left to right): October 16S clones only (yellow), October 16S clones added to the RDP 16S database (green), and the RDP 16S database only (blue). The y-axis shows the estimated abundance of recovered family-level taxa on a log scale. The box-and-whiskers plotted show the median value (black bar), the interquartile (filled box), range of values (whiskers), and outliers beyond 1.5-times the range of the interquartile (open circles). Grey horizontal lines show the simulated initial populations of the four abundance classes.

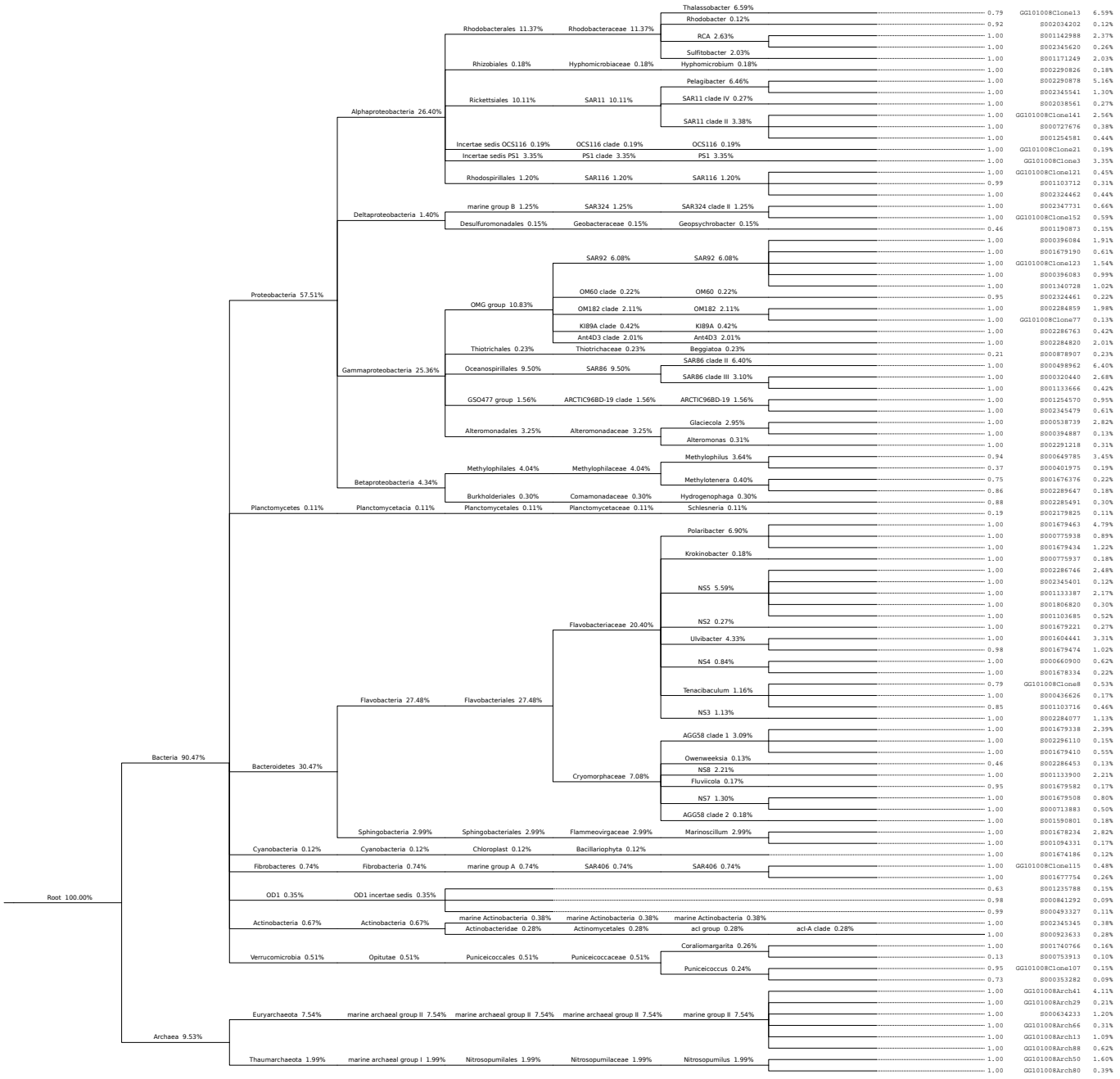


Fig. S9

Taxonomic cladogram detailing all 16S rDNA sequences selected for the May 2009 sample. Taxonomic groups, labelled on branches, are from the extended RDP taxonomy. RDP identifiers for selected sequences are noted at the leaves, except for those prefaced with “GG” which denote clones generated in this study. Bayesian classification p-values for the selected sequences are shown to the left of the clone identifiers. Percentages noted on each branch and leaf represent the estimated abundance of that taxonomic unit as a fraction of total 16S rDNA in the sample.

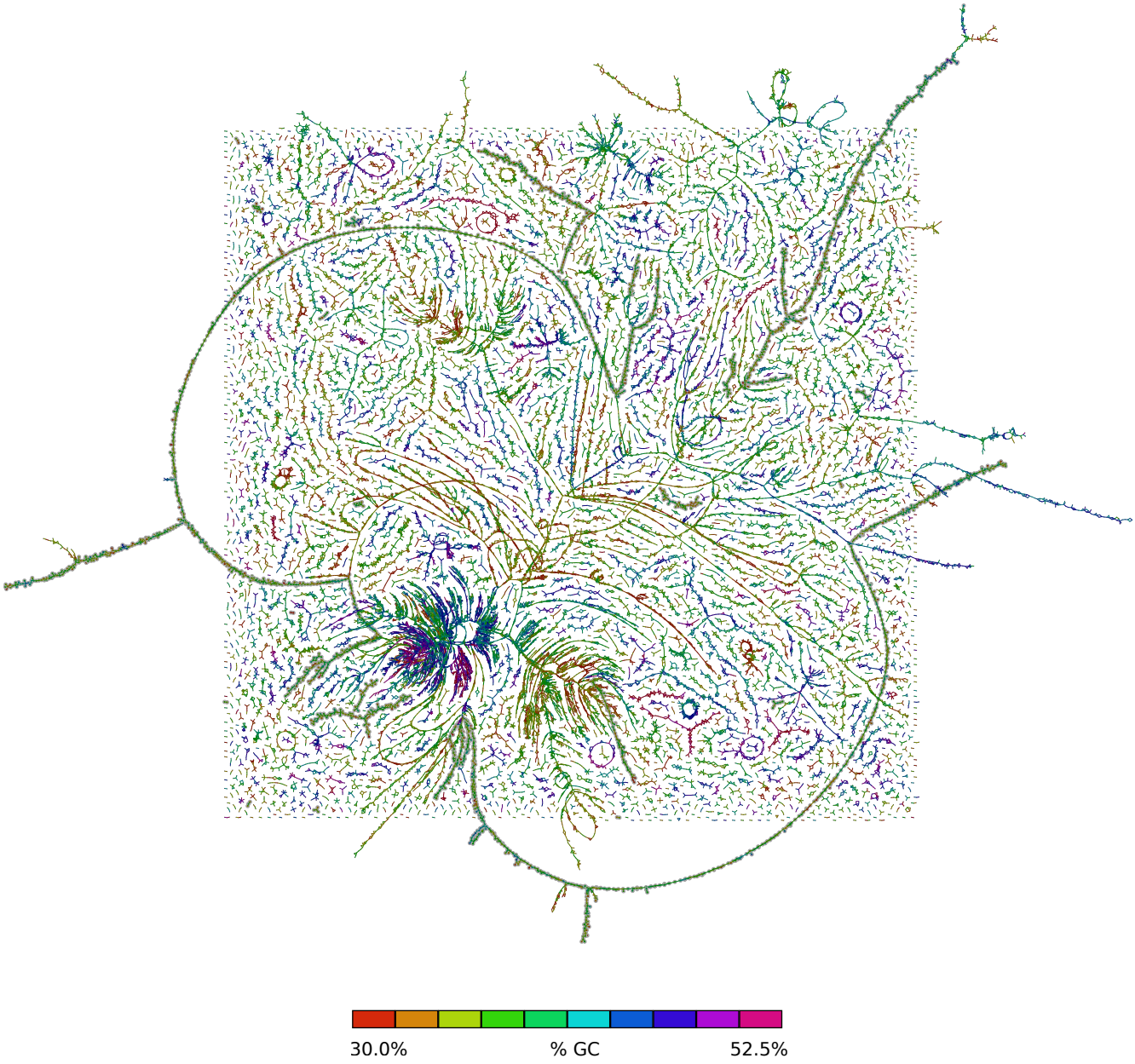


Fig. S10
Mate-pair connection graph illustrating the October 2008 metagenome *de novo* assembly. Lines represent contigs with mate-pair connections scoring greater than 400 bits (~75% of the assembly). Long strands represent prokaryote genome sequences and small circular strands show likely virus/plasmid sequences. Contigs in the candidate genome assembly related to alpha-*Proteobacterium* str. HTCC2255 are indicated (shaded in gray).

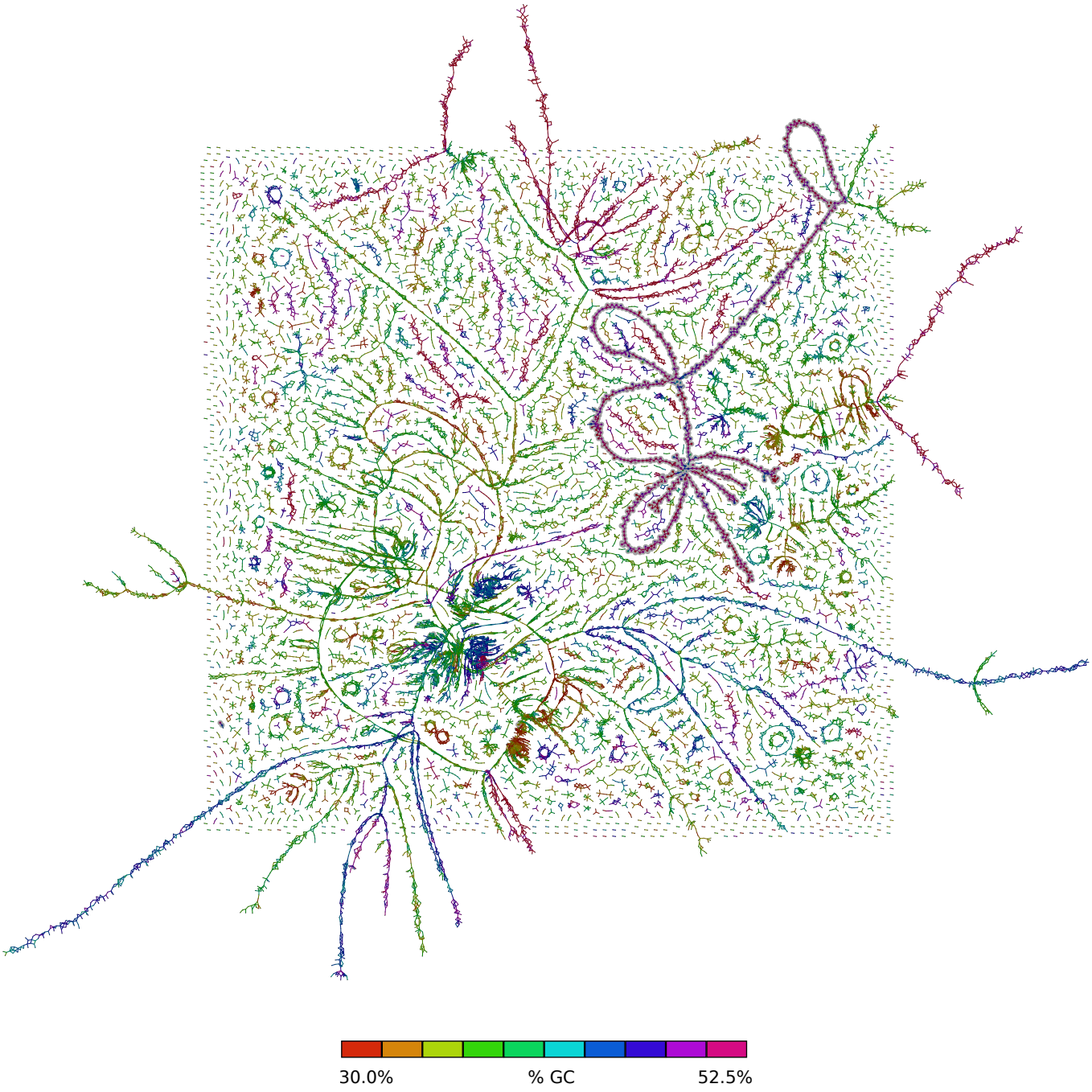


Fig. S11
Mate-pair connection graph illustrating the May 2009 metagenome *de novo* assembly. Lines represent contigs with mate-pair connections scoring greater than 750 bits (~60% of the assembly). Long strands represent prokaryote genome sequences and small circular strands show likely virus/plasmid sequences. Contigs aligning with the MG-II genome assembly are indicated (gray shading.) This figure is a full-resolution representation of Figure 1.

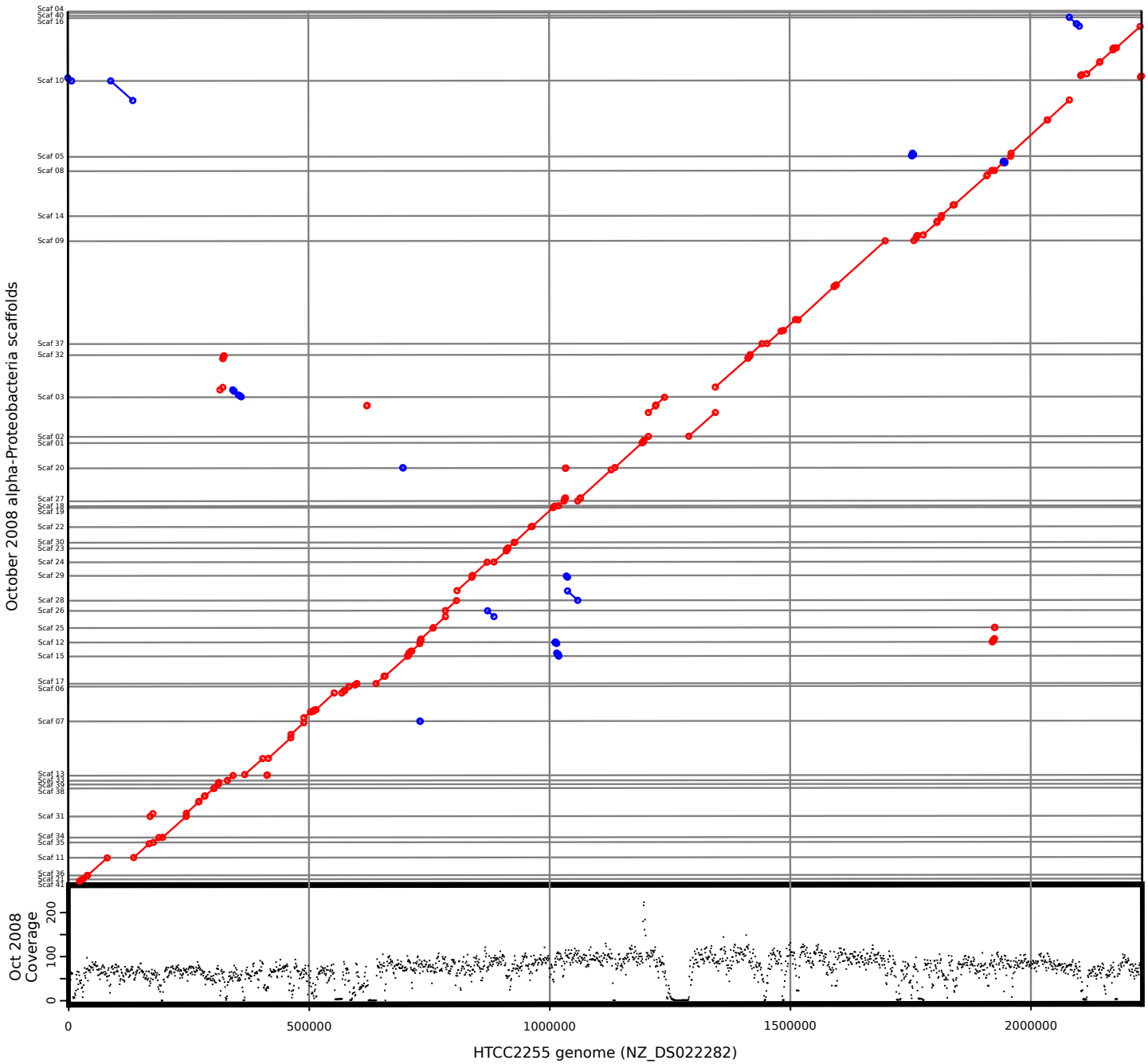

October 2008 alpha-Proteobacteria scaffolds
Scaf 04
Scaf 40
Scaf 16
Scaf 10
Scaf 05
Scaf 08
Scaf 14
Scaf 09
Scaf 37
Scaf 32
Scaf 03
Scaf 02
Scaf 01
Scaf 20
Scaf 27
Scaf 18
Scaf 19
Scaf 22
Scaf 30
Scaf 23
Scaf 24
Scaf 29
Scaf 28
Scaf 26
Scaf 25
Scaf 12
Scaf 15
Scaf 17
Scaf 06
Scaf 07
Scaf 13
Scaf 33
Scaf 39
Scaf 38
Scaf 31
Scaf 34
Scaf 35
Scaf 11
Scaf 36
Scaf 21
Scaf 41
Oct 2008 Coverage
0
100
200
0
500000
1000000
1500000
2000000
HTCC2255 genome (NZ_DS022282)

Fig. S12
Comparison of an October 2008 statistical cluster of assembled scaffolds to the genome of alpha-Proteobacterium str. HTCC2255. The x-axis represents coordinates of the HTCC2255 reference genome (NZ_DS022282). The y-axis of the upper plot shows coordinates of the 41 scaffolds automatically clustered together using tetra-nucleotide statistics (Fig. S10, shaded in gray), sorted and reversed as necessary to produce the most parsimonious alignment. Grey horizontal lines show scaffold boundaries. Diagonal lines in the upper plot show the extent of NUCMER nucleotide alignments between scaffolds and the reference genome, with red lines showing direct alignments and blue lines showing alignments of scaffold regions that were reversed relative to the primary scaffold alignment. The y-axis of the lower plot shows mean coverage (averaged over 1000bp segments) of the HTCC2255 reference genome by aligned October 2008 metagenomic reads.

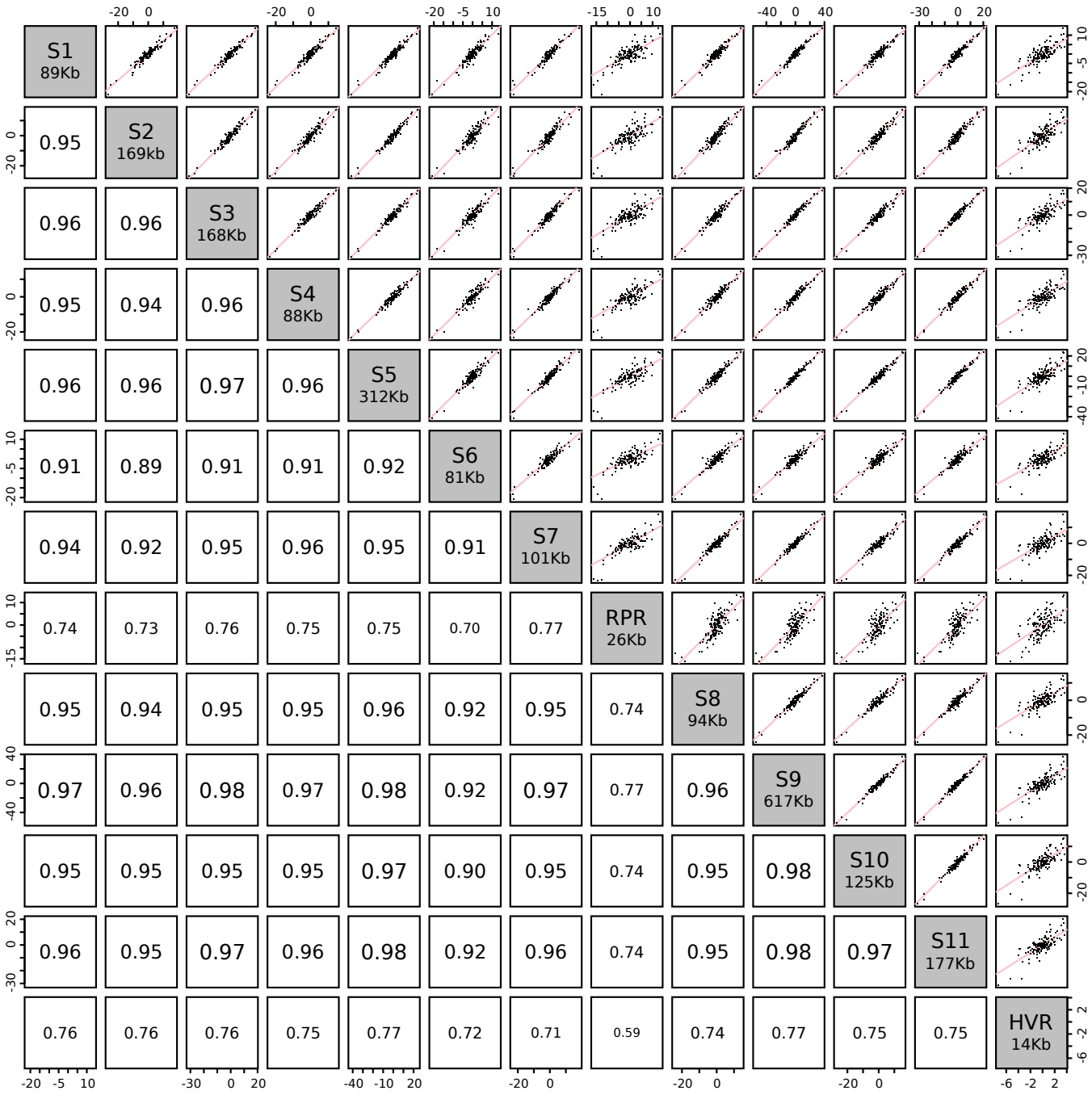


Fig. S13

Pairwise tetra-nucleotide usage anomaly correlations between scaffolds of the Marine Group-II Euryarchaeote genome assembly. The eleven assembly scaffolds (S1 – S11) and the repeat region (RPR) and hyper-variable region (HVR) are shown on the matrix diagonal, with nucleotide sequence lengths noted. The upper half of the matrix shows pair-wise linear-regression plots of the tetra-nucleotide usage anomaly Z-statistics (x and y-axes, black dots, n=256), with the pink-lines indicating the best linear model fits. The lower half of the matrix shows the correlation coefficients of the corresponding pair-wise regressions shown in the upper half.

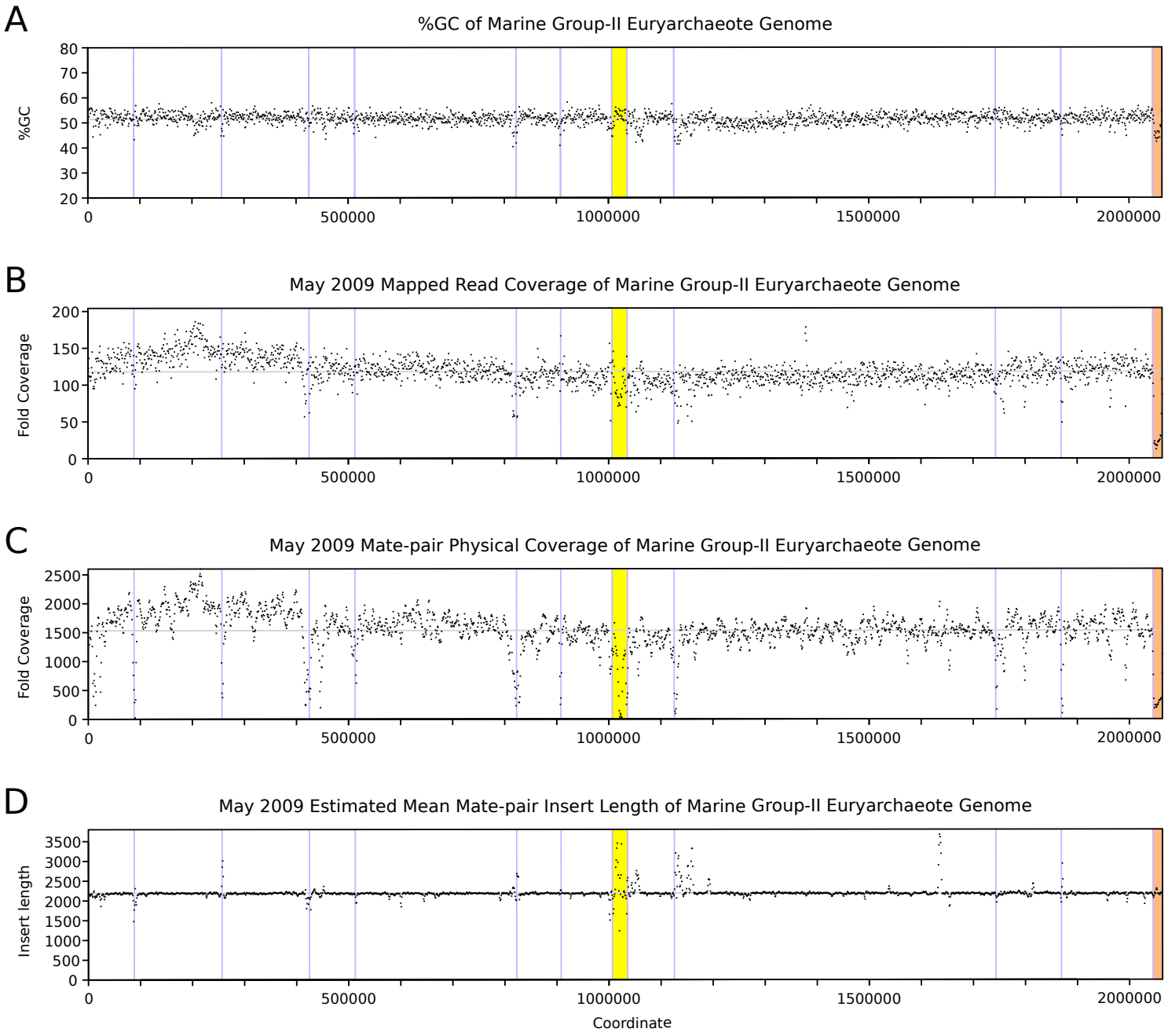

A
%GC of Marine Group-II Euryarchaeote Genome
%GC
80
70
60
50
40
30
20
0
500000
1000000
1500000
2000000
B
May 2009 Mapped Read Coverage of Marine Group-II Euryarchaeote Genome
Fold Coverage
200
150
100
50
0
0
500000
1000000
1500000
2000000
C
May 2009 Mate-pair Physical Coverage of Marine Group-II Euryarchaeote Genome
Fold Coverage
2500
2000
1500
1000
500
0
0
500000
1000000
1500000
2000000
D
May 2009 Estimated Mean Mate-pair Insert Length of Marine Group-II Euryarchaeote Genome
Insert length
3500
3000
2500
2000
1500
1000
500
0
0
500000
1000000
1500000
2000000
Coordinate

Fig. S14

Sequence statistics for the Marine Group-II *Euryarchaeote* genome (MG-II). For all plots the x-axis corresponds to coordinates of MG-II, and values shown are averaged over 1000bp segments. Purple vertical lines denote scaffold boundaries, and the yellow and orange regions show the repeat region (RPR) and hyper-variable region (HVR), respectively. Horizontal grey lines show the genome-wide mean value for the plotted statistic. **(A)** Mean %GC content of MG-II. **(B)** Mean coverage of MG-II by aligned May 2009 metagenomic reads. **(C)** Mean physical coverage of MG-II by aligned mate-paired reads from the May 2009 metagenome. **(D)** Estimated mean mate-pair insert length for mate-pair alignments spanning positions along MG-II.

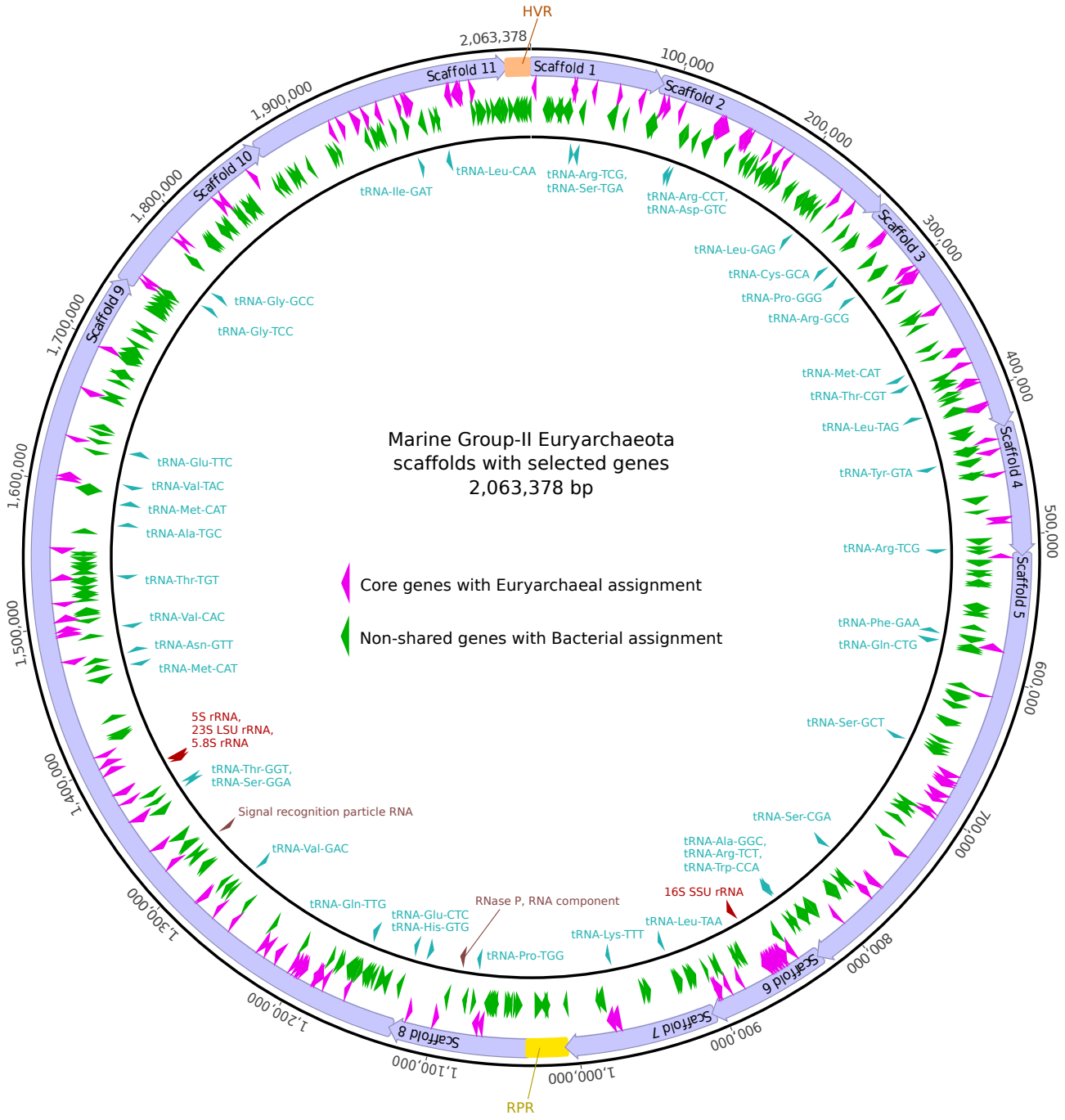


Fig. S15

Schematic representation of the assembled circular Marine Group-II *Euryarchaeote* genome. From the outside inwards the rings show: Genome coordinates, scaffold regions, "Core" gene models, "Bacterial" gene models, and ncRNA genes. RPR (yellow) and HVR (orange) denote the repeat and hyper-variable regions respectively. "Core" gene models (magenta) code for proteins homologous with proteins in both *Aciduliprofundum boonei* (NC_013926) and *Nitrosopumilus maritimus* (NC_010085) – based on reciprocal best blast hits – that are classified taxonomically, by MEGAN, within the phylum Euryarchaeota. "Bacterial" gene models (green) code for proteins without homologs in either *A. boonei* or *N. maritimus* that classify taxonomically within the domain Bacteria. Non-coding RNA genes include tRNAs (blue), rRNA (red), other ncRNAs (brown).

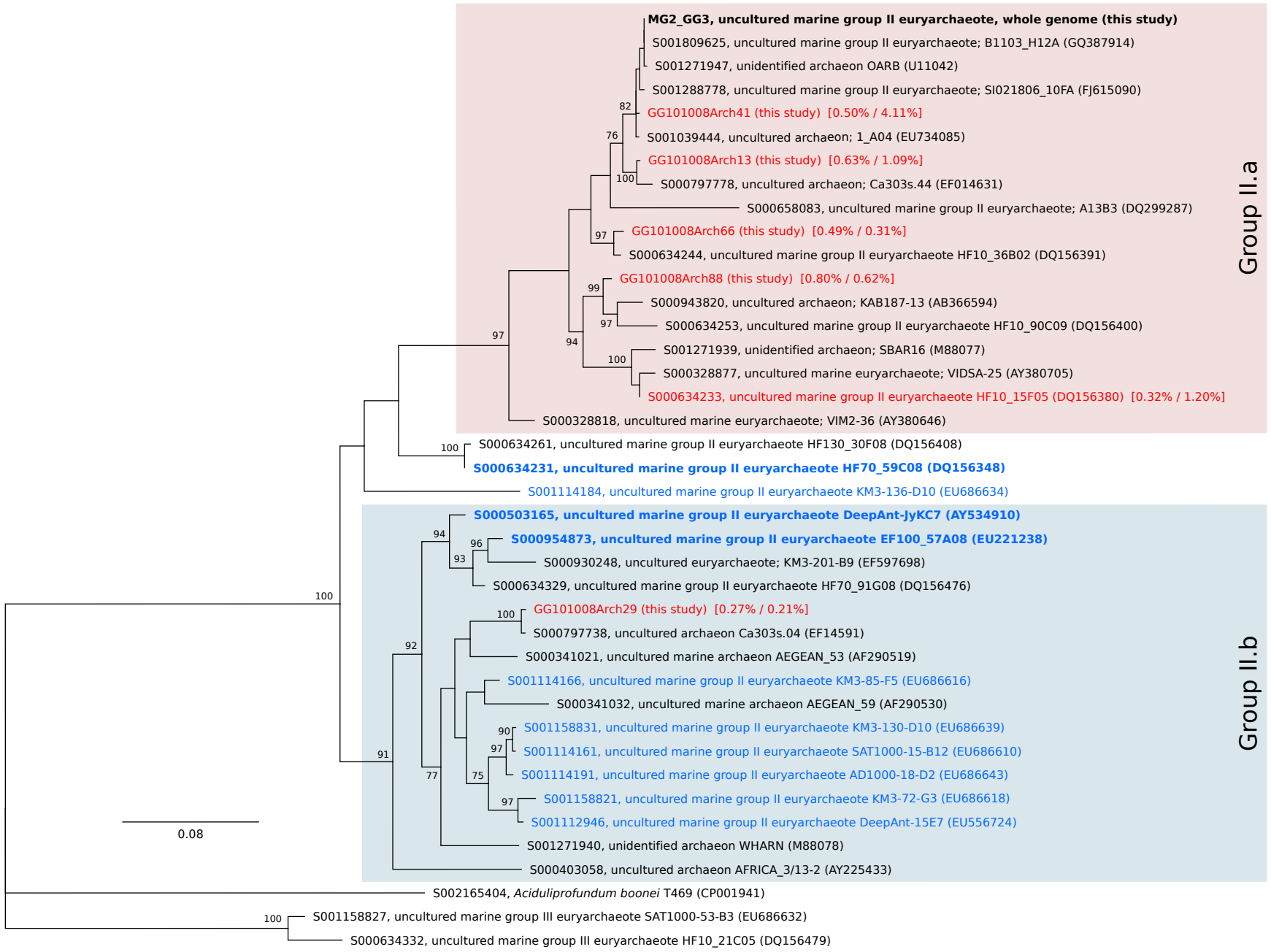

MG2_GG3, uncultured marine group II euryarchaeote, whole genome (this study)
S001809625, uncultured marine group II euryarchaeote; B1103_H12A (GQ387914)
S001271947, unidentified archaeon OARB (U11042)
S001288778, uncultured marine group II euryarchaeote; SI021806_10FA (FJ615090)
GG101008Arch41 (this study) [0.50% / 4.11%]
S001039444, uncultured archaeon; 1_A04 (EU734085)
GG101008Arch13 (this study) [0.63% / 1.09%]
S000797778, uncultured archaeon; Ca303s.44 (EF014631)
S000658083, uncultured marine group II euryarchaeote; A13B3 (DQ299287)
GG101008Arch66 (this study) [0.49% / 0.31%]
S000634244, uncultured marine group II euryarchaeote HF10_36B02 (DQ156391)
GG101008Arch88 (this study) [0.80% / 0.62%]
S000943820, uncultured archaeon; KAB187-13 (AB366594)
S000634253, uncultured marine group II euryarchaeote HF10_90C09 (DQ156400)
S001271939, unidentified archaeon; SBAR16 (M88077)
S000328877, uncultured marine euryarchaeote; VIDSA-25 (AY380705)
S000634233, uncultured marine group II euryarchaeote HF10_15F05 (DQ156380) [0.32% / 1.20%]
S000328818, uncultured marine euryarchaeote; VIM2-36 (AY380646)
S000634261, uncultured marine group II euryarchaeote HF130_30F08 (DQ156408)
S000634231, uncultured marine group II euryarchaeote HF70_59C08 (DQ156348)
S001114184, uncultured marine group II euryarchaeote KM3-136-D10 (EU686634)
S000503165, uncultured marine group II euryarchaeote DeepAnt-JyKC7 (AY534910)
S000954873, uncultured marine group II euryarchaeote EF100_57A08 (EU221238)
S000930248, uncultured euryarchaeote; KM3-201-B9 (EF597698)
S000634329, uncultured marine group II euryarchaeote HF70_91G08 (DQ156476)
GG101008Arch29 (this study) [0.27% / 0.21%]
S000797738, uncultured archaeon Ca303s.04 (EF14591)
S000341021, uncultured marine archaeon AEGEAN_53 (AF290519)
S001114166, uncultured marine group II euryarchaeote KM3-85-F5 (EU686616)
S000341032, uncultured marine archaeon AEGEAN_59 (AF290530)
S001158831, uncultured marine group II euryarchaeote KM3-130-D10 (EU686639)
S001114161, uncultured marine group II euryarchaeote SAT1000-15-B12 (EU686610)
S001114191, uncultured marine group II euryarchaeote AD1000-18-D2 (EU686643)
S001158821, uncultured marine group II euryarchaeote KM3-72-G3 (EU686618)
S001112946, uncultured marine group II euryarchaeote DeepAnt-15E7 (EU556724)
S001271940, unidentified archaeon WHARN (M88078)
S000403058, uncultured archaeon AFRICA_3/13-2 (AY225433)
S002165404, Aciduliprofundum boonei T469 (CP001941)
S001158827, uncultured marine group III euryarchaeote SAT1000-53-B3 (EU686632)
S000634332, uncultured marine group III euryarchaeote HF10_21C05 (DQ156479)
Group II.a
Group II.b
0.08

Fig. S16

Maximum likelihood phylogeny of full-length Marine Group-II *Euryarchaeote* 16S rRNA sequences. The tree is rooted with *Aciduliprofundum boonei.* Branch lengths are proportional to the number of substitutions per site. Bootstraps are shown for 100 replicate trees. All taxa are identified by RDP ID, except those prefaced with "GG", which are October 2008 clones from this study, and the sequence for the Marine Group-II *Euryarchaeote* genome (MG2_GG3) which is shown in black bold type. Taxa in blue type are from fully sequenced large-insert environmental clones, and those in bold blue type are shown in Figure 2B. Taxa in red type were selected from the October 16S clones +RDP database by the method used to estimate abundance from aligned metagenomic reads, and the estimated relative abundances attributed to each taxa for the October and May samples, respectively, are shown in square brackets. The red and blue regions show taxa that group within the two previously identified major 16S environmental clades within the Marine Group-II *Euryarchaeota* (Groups II.a and II.b).

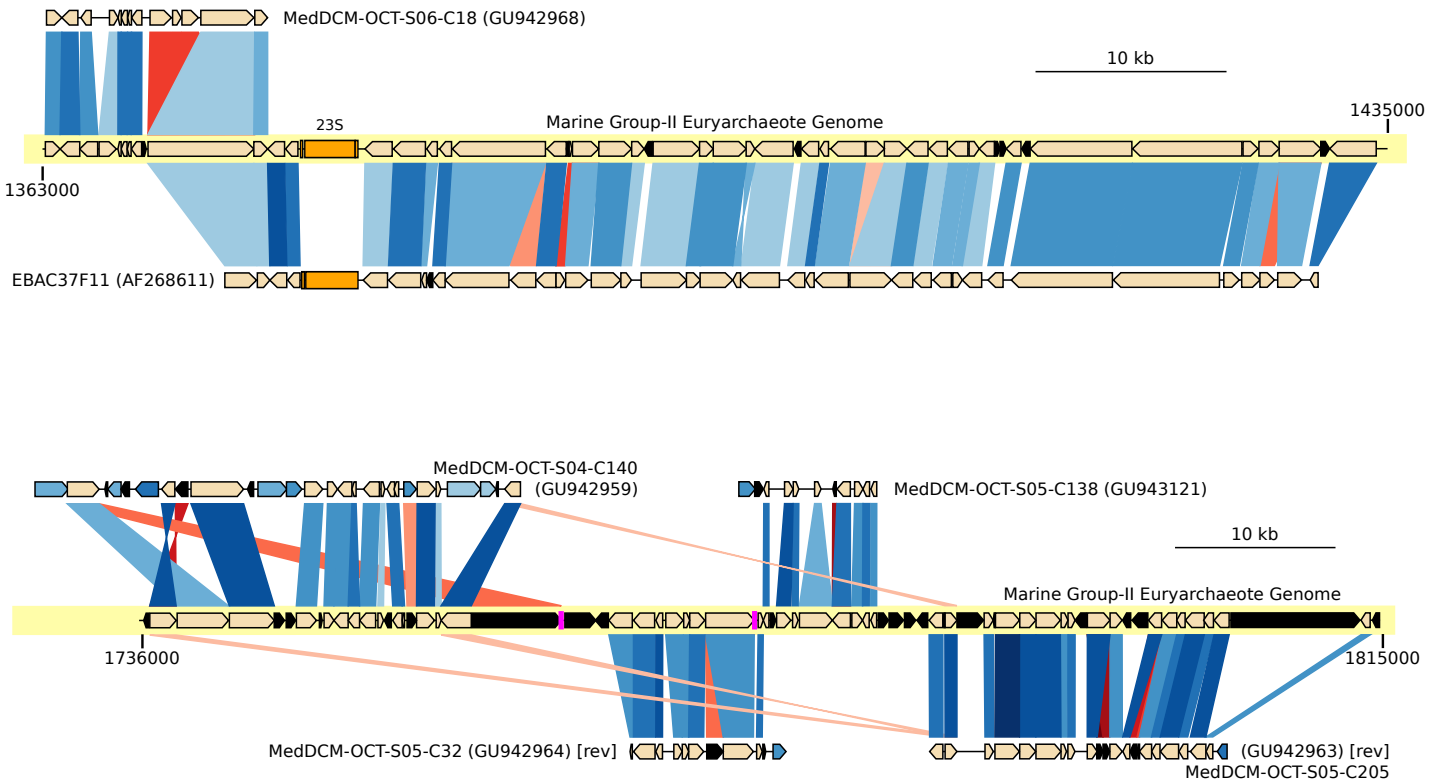

MedDCM-OCT-S06-C18 (GU942968)
10 kb
23S
Marine Group-II Euryarchaeote Genome
1435000
1363000
EBAC37F11 (AF268611)
MedDCM-OCT-S04-C140
(GU942959)
MedDCM-OCT-S05-C138 (GU943121)
10 kb
Marine Group-II Euryarchaeote Genome
1736000
1815000
MedDCM-OCT-S05-C32 (GU942964) [rev]
(GU942963) [rev]
MedDCM-OCT-S05-C205
reciprocal best hits
12.5
25
37.5
50
62.5
75
87.5
100
% AA identity
unidirectional best hits

Fig. S17

Comparison of homologous proteins between environmental clones and two additional regions of the Marine Group-II *Euryarchaeote* genome (MG-II). MG-II regions correspond to grey shaded regions shown on Figure 2A. Homologous genes are connected by shaded regions; blue and red indicate reciprocal and unidirectional best hits, and the depth of shading indicates percent amino acid identity on the scale shown. Gene models, depicted as pointed boxes, are shaded black for genes with no blast hits, blue for environmental genes with a reciprocal best hit to an undepicted location in MG-II, and beige otherwise. rRNA and tRNA genes are shown in orange and magenta, respectively. Clone identifiers and accession numbers are shown for environmental sequences, and those marked "[rev]" are depicted reversed relative to the deposited sequence.

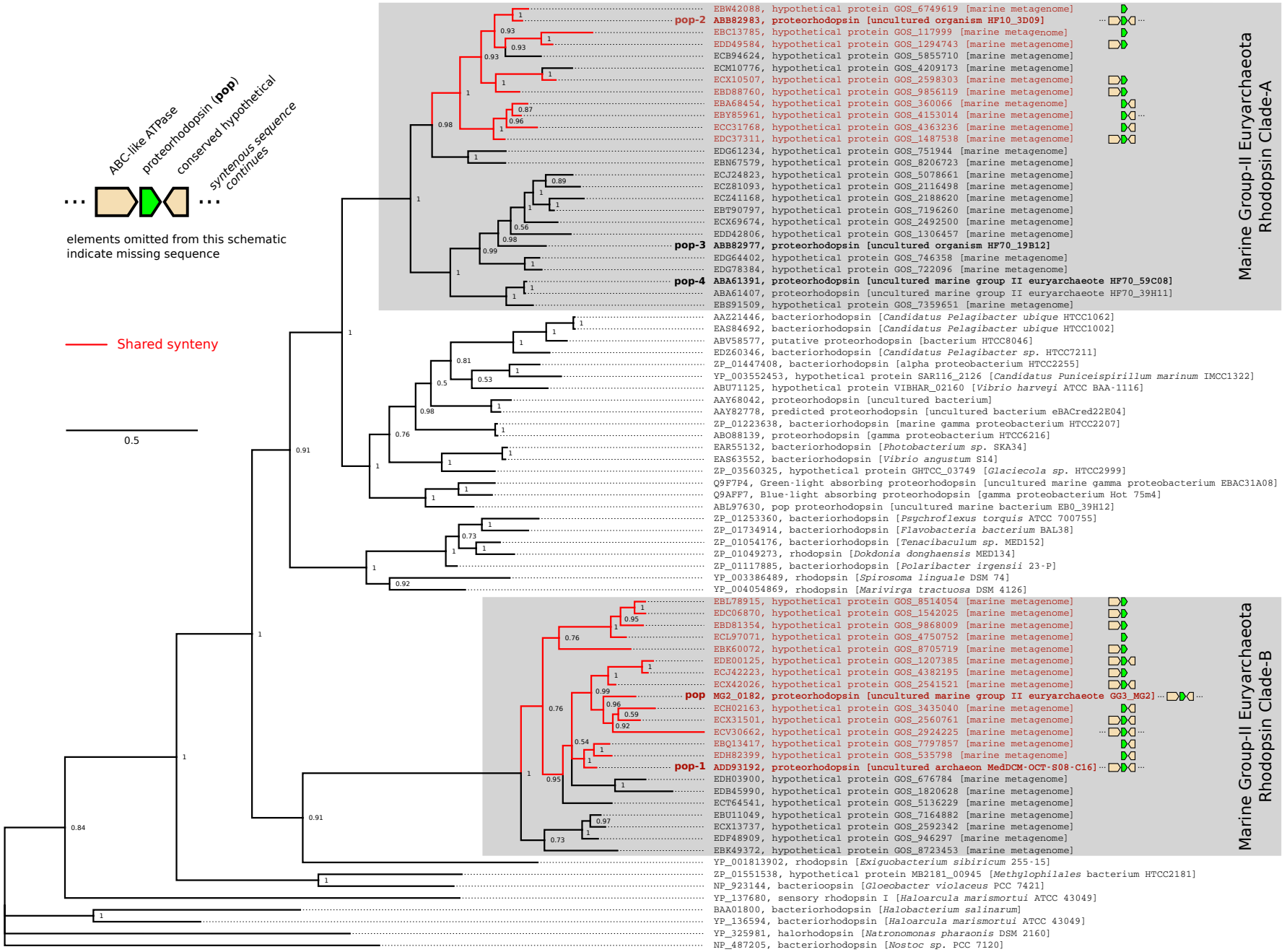
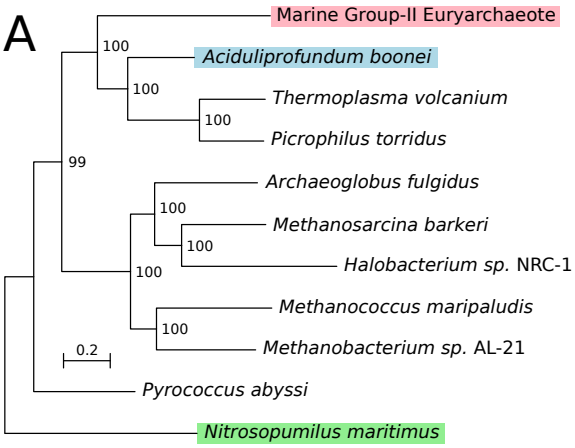
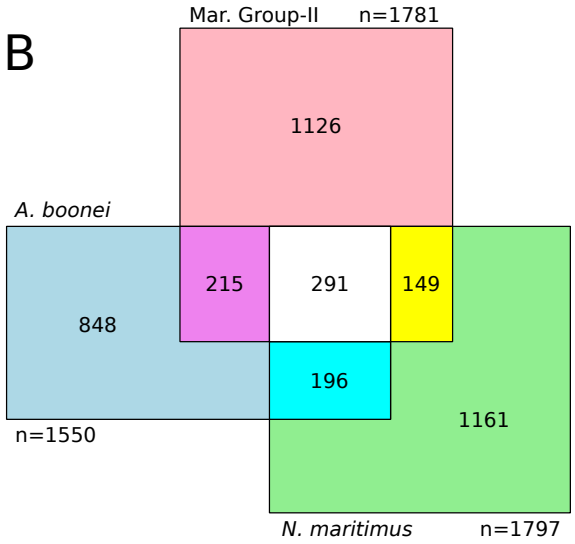

ABC-like ATPase
proteorhodopsin (pop)
conserved hypothetical
syntenous sequence continues
elements omitted from this schematic indicate missing sequence
Shared synteny
0.5
Marine Group-II Euryarchaeota Rhodopsin Clade-A
Marine Group-II Euryarchaeota Rhodopsin Clade-B
pop-2
pop-3
pop-4
pop
pop-1
EBW42088, hypothetical protein GOS_6749619 [marine metagenome]
ABB82983, proteorhodopsin [uncultured organism HF10_3D09]
EBC13785, hypothetical protein GOS_117999 [marine metagenome]
EDD49584, hypothetical protein GOS_1294743 [marine metagenome]
ECB94624, hypothetical protein GOS_5855710 [marine metagenome]
ECM10776, hypothetical protein GOS_4209173 [marine metagenome]
ECX10507, hypothetical protein GOS_2598303 [marine metagenome]
EBD88760, hypothetical protein GOS_9856119 [marine metagenome]
EBA68454, hypothetical protein GOS_360066 [marine metagenome]
EBY85961, hypothetical protein GOS_4153014 [marine metagenome]
ECC31768, hypothetical protein GOS_4363236 [marine metagenome]
EDC37311, hypothetical protein GOS_1487538 [marine metagenome]
EDG61234, hypothetical protein GOS_751944 [marine metagenome]
EBN67579, hypothetical protein GOS_8206723 [marine metagenome]
ECJ24823, hypothetical protein GOS_5078661 [marine metagenome]
ECZ81093, hypothetical protein GOS_2116498 [marine metagenome]
ECZ41168, hypothetical protein GOS_2188620 [marine metagenome]
EBT90797, hypothetical protein GOS_7196260 [marine metagenome]
ECX69674, hypothetical protein GOS_2492500 [marine metagenome]
EDD42806, hypothetical protein GOS_1306457 [marine metagenome]
ABB82977, proteorhodopsin [uncultured organism HF70_19B12]
EDG64402, hypothetical protein GOS_746358 [marine metagenome]
EDG78384, hypothetical protein GOS_722096 [marine metagenome]
ABA61391, proteorhodopsin [uncultured marine group II euryarchaeote HF70_59C08]
ABA61407, proteorhodopsin [uncultured marine group II euryarchaeote HF70_39H11]
EBS91509, hypothetical protein GOS_7359651 [marine metagenome]
AAZ21446, bacteriorhodopsin [Candidatus Pelagibacter ubique HTCC1062]
EAS84692, bacteriorhodopsin [Candidatus Pelagibacter ubique HTCC1002]
ABV58577, putative proteorhodopsin [bacterium HTCC8046]
EDZ60346, bacteriorhodopsin [Candidatus Pelagibacter sp. HTCC7211]
ZP_01447408, bacteriorhodopsin [alpha proteobacterium HTCC2255]
YP_003552453, hypothetical protein SAR116_2126 [Candidatus Puniceispirillum marinum IMCC1322]
ABU71125, hypothetical protein VIBHAR_02160 [Vibrio harveyi ATCC BAA-1116]
AAY68042, proteorhodopsin [uncultured bacterium]
AAY82778, predicted proteorhodopsin [uncultured bacterium eBACred22E04]
ZP_01223638, bacteriorhodopsin [marine gamma proteobacterium HTCC2207]
ABO88139, proteorhodopsin [gamma proteobacterium HTCC6216]
EAR55132, bacteriorhodopsin [Photobacterium sp. SKA34]
EAS63552, bacteriorhodopsin [Vibrio angustum S14]
ZP_03560325, hypothetical protein GHTCC_03749 [Glaciecola sp. HTCC2999]
Q9F7P4, Green-light absorbing proteorhodopsin [uncultured marine gamma proteobacterium EBAC31A08]
Q9AFF7, Blue-light absorbing proteorhodopsin [gamma proteobacterium Hot 75m4]
ABL97630, pop proteorhodopsin [uncultured marine bacterium EB0_39H12]
ZP_01253360, bacteriorhodopsin [Psychroflexus torquis ATCC 700755]
ZP_01734914, bacteriorhodopsin [Flavobacteria bacterium BAL38]
ZP_01054176, bacteriorhodopsin [Tenacibaculum sp. MED152]
ZP_01049273, rhodopsin [Dokdonia donghaensis MED134]
ZP_01117885, bacteriorhodopsin [Polaribacter irgensii 23-P]
YP_003386489, rhodopsin [Spirosoma linguale DSM 74]
YP_004054869, rhodopsin [Marivirga tractuosa DSM 4126]
EBL78915, hypothetical protein GOS_8514054 [marine metagenome]
EDC06870, hypothetical protein GOS_1542025 [marine metagenome]
EBD81354, hypothetical protein GOS_9868009 [marine metagenome]
ECL97071, hypothetical protein GOS_4750752 [marine metagenome]
EBK60072, hypothetical protein GOS_8705719 [marine metagenome]
EDE00125, hypothetical protein GOS_1207385 [marine metagenome]
ECJ42223, hypothetical protein GOS_4382195 [marine metagenome]
ECX42026, hypothetical protein GOS_2541521 [marine metagenome]
MG2_0182, proteorhodopsin [uncultured marine group II euryarchaeote GG3_MG2]
ECH02163, hypothetical protein GOS_3435040 [marine metagenome]
ECX31501, hypothetical protein GOS_2560761 [marine metagenome]
ECV30662, hypothetical protein GOS_2924225 [marine metagenome]
EBQ13417, hypothetical protein GOS_7797857 [marine metagenome]
EDH82399, hypothetical protein GOS_535798 [marine metagenome]
ADD93192, proteorhodopsin [uncultured archaeon MedDCM-OCT-S08-C16]
EDH03900, hypothetical protein GOS_676784 [marine metagenome]
EDB45990, hypothetical protein GOS_1820628 [marine metagenome]
ECT64541, hypothetical protein GOS_5136229 [marine metagenome]
EBU11049, hypothetical protein GOS_7164882 [marine metagenome]
ECX13737, hypothetical protein GOS_2592342 [marine metagenome]
EDF48909, hypothetical protein GOS_946297 [marine metagenome]
EBK49372, hypothetical protein GOS_8723453 [marine metagenome]
YP_001813902, rhodopsin [Exiguobacterium sibiricum 255-15]
ZP_01551538, hypothetical protein MB2181_00945 [Methylophilales bacterium HTCC2181]
NP_923144, bacterioopsin [Gloeobacter violaceus PCC 7421]
YP_137680, sensory rhodopsin I [Haloarcula marismortui ATCC 43049]
BAA01800, bacteriorhodopsin [Halobacterium salinarum]
YP_136594, bacteriorhodopsin [Haloarcula marismortui ATCC 43049]
YP_325981, halorhodopsin [Natronomonas pharaonis DSM 2160]
NP_487205, bacteriorhodopsin [Nostoc sp. PCC 7120]

Fig. S18

Bayesian phylogeny of rhodopsin protein sequences. The tree is rooted with sequence YP_325981, the halorhodopsin from *Natronomonas pharaonis.* Branch lengths are proportional to the number of substitutions per site. Bayesian p-values are shown for each branch. All taxa are identified by NCBI accession number, except "MG2_0182", which is the locus id for the rhodopsin found in the Marine Group-II *Euryarchaeote* genome (MG-II). Taxa in bold type and prefaced by a "pop" identifier correspond with those labelled in Figure 3. The two clades of Marine Group-II *Euryarchaeote* rhodopsin proteins noted in Figure 3 are shown with grey shading. Red branches and leaves on the tree show Marine Group-II *Euryarchaeote* rhodopsins of both types that appear in a syntenic context consistent with that seen in MG-II. The gene model schematic for each red taxa shows which shared neighboring gene models are present in the corresponding nucleotide sequence. Gene models omitted from a schematic indicate missing nucleotide sequence.


A
Marine Group-II Euryarchaeote
Aciduliprofundum boonei
Thermoplasma volcanium
Picrophilus torridus
Archaeoglobus fulgidus
Methanosarcina barkeri
Halobacterium sp. NRC-1
Methanococcus maripaludis
Methanobacterium sp. AL-21
Pyrococcus abyssi
Nitrosopumilus maritimus
100
100
100
99
100
100
100
100
0.2


B
Mar. Group-II n=1781
1126
A. boonei
848
215
291
149
196
1161
n=1550
N. maritimus n=1797

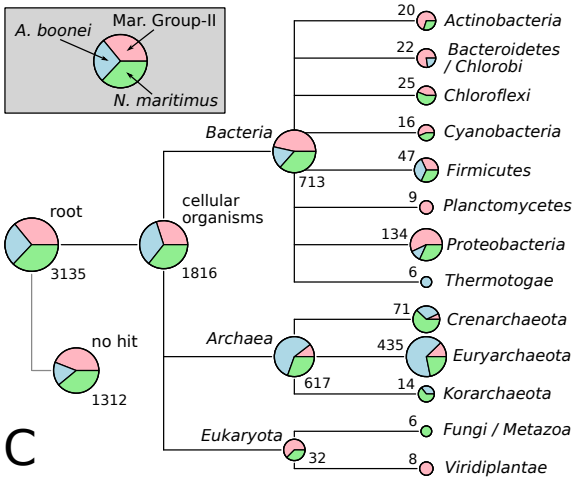

A. boonei
Mar. Group-II
N. maritimus
root
3135
cellular organisms
1816
no hit
1312
Bacteria
713
Archaea
617
Eukaryota
32
20 Actinobacteria
22 Bacteroidetes / Chlorobi
25 Chloroflexi
16 Cyanobacteria
47 Firmicutes
9 Planctomycetes
134 Proteobacteria
6 Thermotogae
71 Crenarchaeota
435 Euryarchaeota
14 Korarchaeota
6 Fungi / Metazoa
8 Viridiplantae
C

Fig. S19

Comparison of the Marine Group-II Euryarchaeote genome (MG-II) to other sequenced Archaea. **(A)** Maximum likelihood phylogeny of ten selected *Euryarchaeota* with available genomes, based on the concatenated amino acid sequences of 31 conserved Archaeal proteins. The tree is rooted with *Nitrosopumilus maritimus* (NC_010085), a member of the *Thaumarchaeota*. Branch lengths are proportional to the number of substitutions per site. Bootstraps are shown for 100 replicate trees. The genomes of *N. maritimus* and *Aciduliprofundum boonei* (NC_013926) are selected for comparison with MG-II. **(B)** Venn diagram depicting protein homologs shared among *A. boonei*, *N. maritimus* and MG-II. Homologous proteins were defined by reciprocal blast hits (RBHs) with an e-value cutoff of 10^{-5}. Areas are proportional to the number of proteins indicated. **(C)** Taxonomic assignment of proteins without shared homologs from *A. boonei*, *N. maritimus* and MG-II, based on classification of blast hits to a RefSeq protein database. The tree represents NCBI taxonomic hierarchy, to the level of phyla. The pie chart at each node represents the number of proteins from each genome assigned to that level, with total pie area scaling as the log of the (noted) total number of proteins in each pie. The "root" node represents all non-homologous proteins in the comparison genomes. The "no hits" node represents proteins without blast hits to the database. All other nodes represent the number of proteins classifiable to that taxonomic level or below (Table S10).

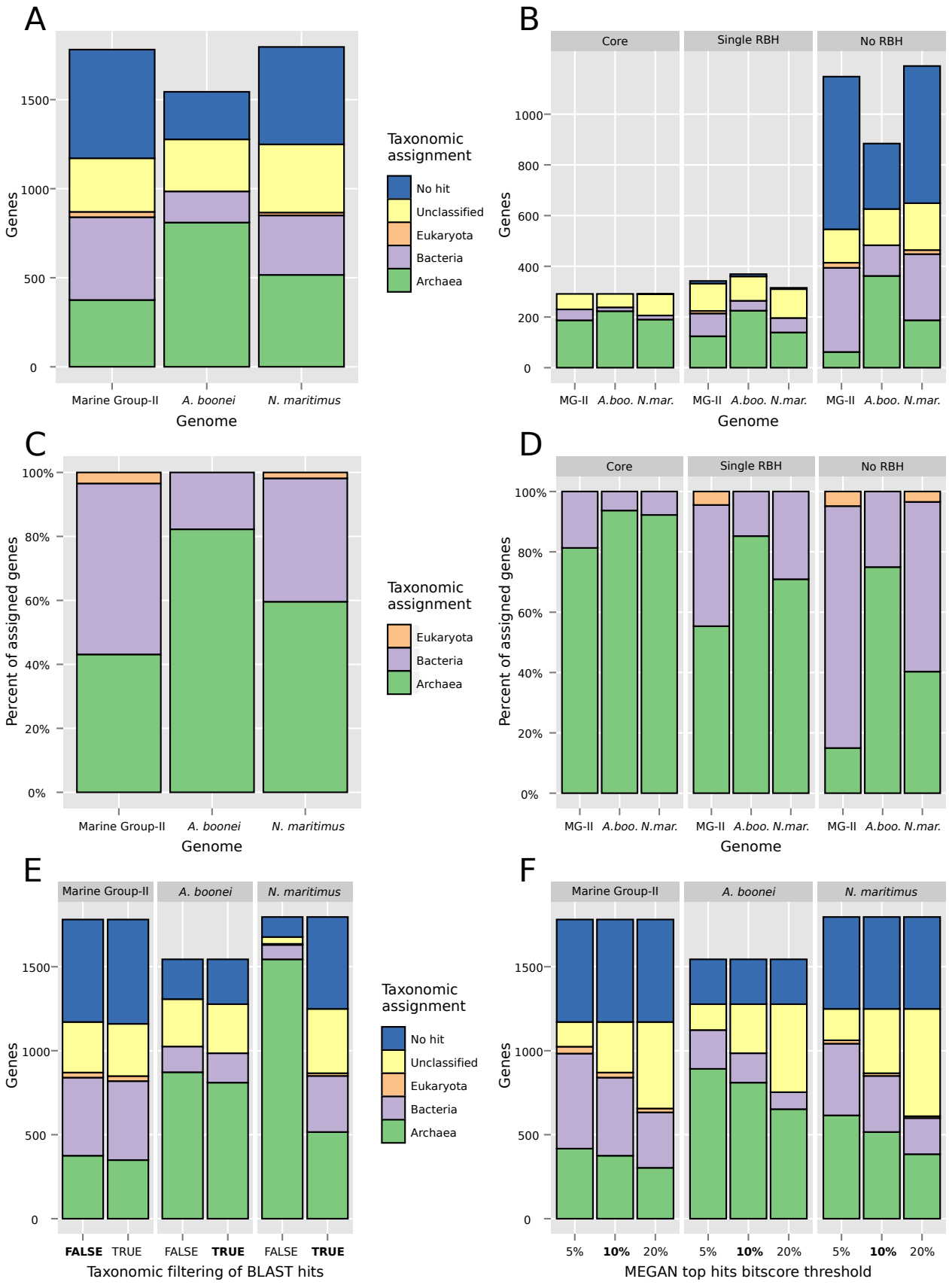

A
Genes
1500
1000
500
0
Marine Group-II
A. boonei
N. maritimus
Genome
Taxonomic assignment
No hit
Unclassified
Eukaryota
Bacteria
Archaea
B
Core
Single RBH
No RBH
1000
800
600
400
200
0
MG-II
A.boo.
N.mar.
Genome
C
Percent of assigned genes
100%
80%
60%
40%
20%
0%
Marine Group-II
A. boonei
N. maritimus
Genome
Taxonomic assignment
Eukaryota
Bacteria
Archaea
D
Core
Single RBH
No RBH
100%
80%
60%
40%
20%
0%
MG-II
A.boo.
N.mar.
Genome
E
Marine Group-II
A. boonei
N. maritimus
Genes
1500
1000
500
0
FALSE
TRUE
Taxonomic filtering of BLAST hits
Taxonomic assignment
No hit
Unclassified
Eukaryota
Bacteria
Archaea
F
Marine Group-II
A. boonei
N. maritimus
Genes
1500
1000
500
0
5%
10%
20%
MEGAN top hits bitscore threshold

Fig. S20

Details of the taxonomic assignment of *Archaeal* proteins, based on analysis of blast hits to the RefSeq protein database using MEGAN. (**A)** Comparison of the taxonomic assignments of proteins from the Marine Group-II genome, *Aciduliprofundum boonei* (NC_013926) and *Nitrosopumilus maritimus* (NC_010085). The y-axis shows a count of genes. Colored bars correspond to taxonomic classification of proteins by MEGAN. "No hit" indicates proteins with no blast homologs in the RefSeq database, and "Unclassified" indicates proteins that MEGAN could not place within a specific taxonomic domain. **(B)** Comparison of the taxonomic assignments of proteins from the three reference genomes, further categorized into one of three groups via reciprocal best hit (RBH) blast analysis as shown in Figure S19B. Core proteins are those with symmetrical RBHs among all three genomes. Single RBH are those proteins that have RBH homologs within only two of the genomes, and No RBH are those without shared homologs. **(C)** Relative comparison of the taxonomically assigned proteins from the three comparison genomes. The y-axis shows the percentage of all proteins coded by a genome that MEGAN could place within a specific taxonomic domain, as indicated by the colored bars. **(D)** Relative comparison of the taxonomically assigned proteins from the three comparison genomes, further categorized into one of three groups as described for Figure S20B. **(E)** Comparison of the effect of taxonomic filtering of the RefSeq database on taxonomic assignments. The x-axis categories FALSE and TRUE, indicate whether blast hits to selected RefSeq taxa (other than the reference genome itself, which is always removed) were excluded from the MEGAN analysis. Bold type indicates the setting selected for use for all other MEGAN analyses in this study. See Methods for a list of RefSeq taxa excluded for each reference genome. **(F)** Comparison of the effect of varying the "Top Percent" parameter of MEGAN's lowest common ancestor (LCA) algorithm. The x-axis categories 5%, 10% and 20% show the three parameter values evaluated. Bold type indicates the setting selected for use for all other MEGAN analyses in this study.

Table S1
Metagenome sample statistics and environmental measurements.

	October	May	
Sample collection			
Collection date	10-Oct-08	9-May-09	
Time of Day	09:00	09:30	
Depth (approximate)	1	1	meters
Location	(47.6906558, -122.404411)		degrees (lat, long)
Sample filtration			
Sample volume	91	91	liters
Prefilter	0.8	0.8	µmeters
Concentrated volume	160	150	milliliters
Nutrients and chlorophyll			
Nitrate	20.2 ± 0.2	27.4 ± 0.2	µmol / liter
Nitrite	0.76 ± 0.01	0.49 ± 0.01	µmol / liter
Ammonium	2.7 ± 0.09	1.5 ± 0.06	µmol / liter
Phosphate	2.3 ± 0.02	1.8 ± 0.01	µmol / liter
Silicate	53.5 ± 0.71	85.8 ± 1.61	µmol / liter
Chlorophyll-a	3.1	6.2	µgrams / milliliter
Clone libraries			
Bacterial 16S clones	172	-	
Archaeal 16S clones	92	-	
Cell counts - Whole Water			
Total cells (DAPI)	9.35	8.01	x 10^5 cells / milliliter
Bacteria	7.44 (80%)	6.26 (78%)	x 10^5 cells / milliliter (% of total)
SAR-11	1.76 (19%)	1.28 (16%)	x 10^5 cells / milliliter (% of total)
SAR-86	1.13 (12%)	0.81 (10%)	x 10^5 cells / milliliter (% of total)
Cell counts - Concentrate			
Total cells (DAPI)	12.8	4.66	x 10^7 cells / milliliter
Bacteria	10.2 (87%)	3.12 (67%)	x 10^7 cells / milliliter (% of total)
SAR-11	2.96 (23%)	0.67 (14%)	x 10^7 cells / milliliter (% of total)
SAR-86	1.4 (11%)	0.53 (11%)	x 10^7 cells / milliliter (% of total)

Table S2
Metagenome sequencing, analysis, alignment, and de novo assembly statistics.

	October	May
Metagenome sequence statistics		
Sample collection date	10-Oct-08	9-May-09
Raw reads (millions)	540	631
Raw sequence (gigabases)	27.0	31.5
Mate-pair insert length (bases ± stddev)	1078 ± 312	2084 ± 477
Reads filtered for alignments		
Filtered reads (millions, % of raw)	322 (59.6%)	301 (47.7%)
Filtered sequence (gigabases, % of raw)	15.2 (56.3%)	14.5 (46.0%)
Mean trimmed read length (bases)	47.2	48.1
RDP 16S rDNA alignment		
Reads recruited (thousands, % of filtered)	97 (0.031%)	76 (0.025%)
Information (megabits)	0.86	0.75
RefSeq genomes alignment		
Reads recruited (millions, % of filtered)	6.2 (1.9%)	6.4 (2.1%)
Information (megabits)	91.3	90.5
GOS assembly alignment		
Reads recruited (millions, % of filtered)	32.1 (10.0%)	24.6 (8.2%)
Information (megabits)	619	458
Contig assembly and alignment		
Contigs assembled (thousands)	589	517
Contig sequence (megabases)	153	155
Contig N50 (bases)	369	489
Reads recruited (millions, % of filtered)	97 (30.1%)	95 (31.6%)

Table S3

Summary of metagenomic read alignment with the RefSeq genome database. All genomes with mean read coverage >= 1-fold are listed. Enterobacteria phage lambda DNA (green) was spiked into the May sample DNA as a control. Read coverage as reported is competitive and shared coverage is allocated to the best alignment position(s) in the database, and coverage from reads aligning equally well to more than one position in the database is divided among those positions.

A) October 2008 reads aligned to RefSeq46 genomes. Results for genomes with >= 1x mean coverage

TaxID	Name	Sequence length	Mean Fold Coverage	% Covered	Genome 16S Classification (genus p-value when < 1.0)
367336	Rhodobacterales bacterium HTCC2255	2296803	70.5	96.0%	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Thalassobacter (0.64)
335992	Candidatus Pelagibacter ubique HTCC1062	1308759	35.6	95.3%	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; SAR11; Pelagibacter
314261	Candidatus Pelagibacter ubique HTCC1002	1324008	32.9	94.1%	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; SAR11; Pelagibacter
487796	Flavobacteria bacterium MS024-2A	1905484	5.5	90.3%	Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales; Flavobacteriaceae; NS5
136084	Roseobacter phage SIO1	39898	5.1	75.4%	N/A
487797	Flavobacteria bacterium MS024-3C	1515248	2.0	84.6%	Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales; Flavobacteriaceae; NS3
439493	Candidatus Pelagibacter sp. HTCC7211	1456888	1.9	23.2%	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; SAR11; Pelagibacter
436308	Nitrosopumilus maritimus SCM1	1645259	1.8	8.3%	Archaea; Thaumarchaeota; marine archaeal group I; Nitrosopumilales; Nitrosopumilaceae; Nitrosopumilus
247639	marine gamma proteobacterium HTCC2080	3576081	1.6	68.2%	Bacteria; Proteobacteria; Gammaproteobacteria; OMG group; OM60 clade; OM60
383631	Methylophilales bacterium HTCC2181	1304428	1.0	41.2%	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; Methylophilus (0.78)

B) May 2009 reads aligned to RefSeq46 genomes. Results for genomes with >= 1x mean coverage

TaxID	Name	Sequence length	Mean Fold Coverage	% Covered	Genome 16S Classification (genus p-value when < 1.0)
10710	*Enterobacteria phage lambda*	*48502*	*118.1*	*99.0%*	*N/A*
367336	Rhodobacterales bacterium HTCC2255	2296803	56.5	96.5%	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodobacterales; Rhodobacteraceae; Thalassobacter (0.64)
487796	Flavobacteria bacterium MS024-2A	1905484	23.0	92.9%	Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales; Flavobacteriaceae; NS5
335992	Candidatus Pelagibacter ubique HTCC1062	1308759	22.8	94.0%	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; SAR11; Pelagibacter
314261	Candidatus Pelagibacter ubique HTCC1002	1324008	22.5	92.8%	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; SAR11; Pelagibacter
314287	gamma proteobacterium HTCC2207	2633173	17.9	93.7%	Bacteria; Proteobacteria; Gammaproteobacteria; OMG group; SAR92; SAR92
487797	Flavobacteria bacterium MS024-3C	1515248	8.8	93.9%	Bacteria; Bacteroidetes; Flavobacteria; Flavobacteriales; Flavobacteriaceae; NS3
383631	Methylophilales bacterium HTCC2181	1304428	4.0	88.1%	Bacteria; Proteobacteria; Betaproteobacteria; Methylophilales; Methylophilaceae; Methylophilus (0.78)
136084	Roseobacter phage SIO1	39898	1.7	24.4%	N/A

Table S4
Marine Group-II *Euryarchaeote* genome statistics, in comparison with *Aciduliprofundum boonei* and *Nitrosopumilus maritimus*.

	Mar. Grp. II	*A. boonei*	*N. maritimus*
Length	2,063,378	1,486,778	1,645,259
%GC	51.5%	31.2%	34.2%
rRNAs	4	3	4
tRNAs	41	32	40
ncRNAs	2	2	1
Coding genes	1781	1550	1797
Mean length	367	304	277
% coding	95.0%	95.0%	90.6%
Unknown function	577 (32.4%)	444 (28.6%)	713 (39.7%)
Peptidases	52 (2.9%)	41 (2.6%)	20 (1.1%)
Mean length	2151	1665	1170
% of coding	5.7%	4.8%	1.6%
Non-peptidase homologs	14	19	4

Table S5
Conserved Archaeal proteins (n=31) used in the generation of the euryarchaeal phylogeny shown in figure S19A.

Protclust	KO	arCOG	COG	Annotation	MG-II	A. boonei	T. volcanum	P. torridus	M. barkeri	A. fulgidus	P. abyssi	M. sp. AL-21	H. sp. NRC-1	N. marinus	M. maripaludis
PRK09602	K06942	arCOG00357	COG0012	GTP-binding protein, probable translation factor	MG2_0309	Aboo_0538	TVN1009	PTO0852	Mbar_A1073	AF1364	PAB1357	Metbo_0024	VNG0504G	Nmar_1534	MMP1122
PRK04172	K01889	arCOG00410	COG0016	phenylalanyl-tRNA synthetase	MG2_1359	Aboo_0186	TVN1037	PTO0125	Mbar_A1375	AF1955	PAB2426	Metbo_1081	VNG2504G	Nmar_1489	MMP1496
CLSK2789673	K01887	arCOG00487	COG0018	Arginyl-tRNA synthetase	MG2_0811	Aboo_0562	TVN1320	PTO0603	Mbar_A1008	AF0894	PAB0469	Metbo_0089	VNG6312G	Nmar_0686	MMP1026
PRK04211	K02950	arCOG04255	COG0048	30S ribosomal protein S12	MG2_1069	Aboo_0070	TVN0163	PTO0853	Mbar_A3688	AF1892	PAB0427	Metbo_2392	VNG2658G	Nmar_0354	MMP1367
PRK04027	K02992	arCOG04254	COG0049	30S ribosomal protein S7	MG2_1068	Aboo_0071	TVN0162	PTO0854	Mbar_A3687	AF1893	PAB0428	Metbo_2391	VNG2657G	Nmar_0355	MMP1368
PRK04020	K02967	arCOG04245	COG0052	30S ribosomal protein S2	MG2_452	Aboo_1302	TVN0399	PTO0517	Mbar_A1423	AF1133	PAB0368	Metbo_0816	VNG1143G	Nmar_0316	MMP0667
PRK06039	K01870	arCOG00807	COG0060	Isoleucyl-tRNA synthetase	MG2_1459	Aboo_0154	TVN0829	PTO0411	Mbar_A3613	AF0633	PAB0616	Metbo_0154	VNG2190G	Nmar_1120	MMP1474
PRK01143	K02867	arCOG04372	COG0080	50S ribosomal protein L11	MG2_0382	Aboo_0863	TVN0423	PTO0437	Mbar_A0615	AF0538	PAB2353	Metbo_2250	VNG1108G	Nmar_0385	MMP1433
PRK04203	K02863	arCOG04289	COG0081	50S ribosomal protein L1	MG2_0383	Aboo_0862	TVN0422	PTO0438	Mbar_A0614	AF1490	PAB1166	Metbo_2251	VNG1105G	Nmar_0382	MMP0260
PRK08565	K13798 or K03044/K03045	arCOG01762	COG0085	DNA-directed RNA polymerase subunit beta	MG2_0882	Aboo_1091	TVN1182	PTO0259	Mbar_A3693 + Mbar_A3694	AF1887 + AF1886	PAB0423	Metbo_2397 + Metbo_2398	VNG2665G + VNG2666G	Nmar_0347	MMP1362 + MMP1361
PRK04231	K02906	arCOG04070	COG0087	50S ribosomal protein L3	MG2_0756	Aboo_1521	TVN0324	PTO0640	Mbar_A0110	AF1925	PAB2120	Metbo_0774	VNG1689G	Nmar_0809	MMP1543
PRK04223	K02890	arCOG04098	COG0091	50S ribosomal protein L22	MG2_0761	Aboo_1516	TVN0329	PTO0645	Mbar_A0105	AF1920	PAB2396	Metbo_0779	VNG1695G	Nmar_0805	MMP1403
PRK04191	K02982	arCOG04097	COG0092	30S ribosomal protein S3	MG2_0762	Aboo_1515	TVN0330	PTO0646	Mbar_A0104	AF1919	PAB2125	Metbo_0780	VNG1697G	Nmar_0804	MMP1404
PRK08571	K02874	arCOG04095	COG0093	50S ribosomal protein L14	MG2_0767	Aboo_1511	TVN0335	PTO0650	Mbar_A0100	AF1915	PAB2436	Metbo_0785	VNG1701G	Nmar_0800	MMP1409
PRK04219	K02931	arCOG04092	COG0094	50S ribosomal protein L5	MG2_0770	Aboo_1508	TVN0338	PTO0653	Mbar_A0097	AF1912	PAB2130	Metbo_0788	VNG1705G	Nmar_0797	MMP1412
PRK04034	K02994	arCOG04091	COG0096	30S ribosomal protein S8	MG2_0772	Aboo_1506	TVN0340	PTO0655	Mbar_A0095	AF1910	PAB2131	Metbo_0790	VNG1707G	Nmar_0795	MMP1414
PRK05518	K02933	arCOG04090	COG0097	50S ribosomal protein L6	MG2_0773	Aboo_1505	TVN0341	PTO0656	Mbar_A0094	AF1909	PAB2132	Metbo_0791	VNG1709G	Nmar_0794	MMP1415
PRK04044	K02988	arCOG04087	COG0098	30S ribosomal protein S5	MG2_0777	Aboo_1501	TVN0345	PTO0660	Mbar_A0090	AF1905	PAB2136	Metbo_0795	VNG1715G	Nmar_0399	MMP1419
PRK04053	K02952	arCOG01722	COG0099	30S ribosomal protein S13	MG2_0134	Aboo_1396	TVN0562	PTO1219	Mbar_A0078	AF2285	PAB0360	Metbo_0805	VNG1132G	Nmar_0325	MMP1319
PRK09607	K02948	arCOG04240	COG0100	30S ribosomal protein S11	MG2_0132	Aboo_1398	TVN0564	PTO1221	Mbar_A0076	AF2283	PAB0362	Metbo_0807	VNG1134G	Nmar_1450	MMP1321
PRK06394	K02871	arCOG04242	COG0102	50S ribosomal protein L13P	MG2_0262	Aboo_1475	TVN1135	PTO0323	Mbar_A1427	AF1128	PAB0365	Metbo_0810	VNG1138G	Nmar_0426	MMP1324
PRK00474	K02996	arCOG04243	COG0103	30S ribosomal protein S9	MG2_0263	Aboo_1476	TVN1136	PTO0324	Mbar_A1426	AF1129	PAB0366	Metbo_0811	VNG1139Gm	Nmar_0425	MMP1325
PRK08561	K02956	arCOG04185	COG0184	30S ribosomal protein S15	MG2_1260	Aboo_1376	TVN1208	PTO0244	Mbar_A1873	AF0801	PAB0033	Metbo_0103	VNG0790G	Nmar_1508	MMP1579
PRK08572	K02961	arCOG04096	COG0186	30S ribosomal protein S17	MG2_0766	Aboo_1512	TVN0334	PTO0649	Mbar_A0101	AF1916	PAB2127	Metbo_0784	VNG1700G	Nmar_0801	MMP1408
PRK04199	K02866	arCOG04113	COG0197	50S ribosomal protein L10/L16	MG2_0678	Aboo_0067	TVN0539	PTO0715	Mbar_A1371	AF1339	PAB1444	Metbo_0379	VNG0099G	Nmar_0449	MMP1289
PRK06419	K02876	arCOG00779	COG0200	50S ribosomal protein L15	MG2_0779	Aboo_1499	TVN0347	PTO0662	Mbar_A0088	AF1903	PAB2138	Metbo_0797	VNG1718G	Nmar_0401	MMP1421
CLSK2560501	K03076	arCOG04169	COG0201	Preprotein translocase subunit SecY	MG2_0780	Aboo_1498	TVN0348	PTO0663	Mbar_A0087	AF1902	PAB2139	Metbo_0798	VNG1719G	Nmar_0402	MMP1422
PRK00783	K03047	arCOG04241	COG0202	DNA-directed RNA polymerase subunit alpha	MG2_0131	Aboo_1399	TVN0565	PTO1222	Mbar_A0075	AF2282	PAB2410	Metbo_0808	VNG1136G	Nmar_0428	MMP1322
PRK08569	K02881	arCOG04088	COG0256	50S ribosomal protein L18	MG2_0776	Aboo_1502	TVN0344	PTO0659	Mbar_A0091	AF1906	PAB2135	Metbo_0794	VNG1714G	Nmar_0398	MMP1418
PRK04051	K02986	arCOG04239	COG0522	30S ribosomal protein S4	MG2_0133	Aboo_1397	TVN0563	PTO1220	Mbar_A0077	AF2284	PAB0361	Metbo_0806	VNG1133G	Nmar_0324	MMP1320
PRK09605	K07174	arCOG01183	COG0533	O-sialoglycoprotein endopeptidase	MG2_1145	Aboo_0655	TVN1276	PTO0374	Mbar_A0279	AF0665	PAB1047	Metbo_0101	VNG2045G	Nmar_1531	MMP0415

Table S6

Ribosomal proteins found in the Marine Group-II Euryarchaeote and Aciduliprofundum boonei (NC_013926) genomes. Blue and green shading alternates and denotes groups of syntenic genes. Locus Ids and accessions are noted for MG-II and A. boonei proteins, respectively.

MG2_Loc	Abo_Acc	Name
MG2_0132	YP_003483767	ribosomal protein S11P
MG2_0133	YP_003483766	ribosomal protein S4
MG2_0134	YP_003483765	ribosomal protein S13P
MG2_0175	YP_003483627	ribosomal protein LX
MG2_0261	YP_003483842	ribosomal protein L15
MG2_0262	YP_003483843	ribosomal protein L13
MG2_0263	YP_003483844	ribosomal protein S9P
MG2_0362	YP_003483136	ribosomal protein L21e
MG2_0382	YP_003483234	ribosomal protein L11
MG2_0383	YP_003483233	ribosomal protein L1
MG2_0384	YP_003483232	ribosomal protein L10
MG2_0385	YP_003483231	ribosomal protein L12
MG2_0423	YP_003483497	ribosomal protein L15e
MG2_0452	YP_003483672	ribosomal protein S2
MG2_0625	YP_003483173	ribosomal protein S6e
MG2_0678	YP_003482441	ribosomal protein L10e
MG2_0687	YP_003483524	ribosomal protein S27E
MG2_0688	YP_003483525	ribosomal protein L44E
MG2_0691	YP_003483759	ribosomal protein L37e
MG2_0756	YP_003483889	ribosomal protein L3
MG2_0757	YP_003483888	ribosomal protein L4P
MG2_0758	YP_003483887	ribosomal protein L23
MG2_0759	YP_003483886	ribosomal protein L2
MG2_0760	YP_003483885	ribosomal protein S19
MG2_0761	YP_003483884	ribosomal protein L22
MG2_0762	YP_003483883	ribosomal protein S3
MG2_0763	YP_003483882	ribosomal protein L29
MG2_0766	YP_003483880	ribosomal protein S17P
MG2_0767	YP_003483879	ribosomal protein L14P
MG2_0768	YP_003483878	ribosomal protein L24
MG2_0769	YP_003483877	ribosomal protein S4e
MG2_0770	YP_003483876	ribosomal protein L5
MG2_0771	YP_003483875	ribosomal protein S14
MG2_0772	YP_003483874	ribosomal protein S8
MG2_0773	YP_003483873	ribosomal protein L6P
MG2_0774	YP_003483872	ribosomal protein L32e
MG2_0775	YP_003483871	ribosomal protein L19e
MG2_0776	YP_003483870	ribosomal protein L18P
MG2_0777	YP_003483869	ribosomal protein S5
MG2_0778	YP_003483868	ribosomal protein L30P
MG2_0779	YP_003483867	ribosomal protein L15
MG2_0885	YP_003483459	ribosomal protein L30e
MG2_0923	YP_003483006	ribosomal protein S3Ae
MG2_1068	YP_003482445	ribosomal protein S7
MG2_1069	YP_003482444	ribosomal protein S12
MG2_1079	YP_003482554	ribosomal protein L37a
MG2_1161	YP_003483616	ribosomal protein S17e
MG2_1162	YP_003482401	ribosomal protein L31e
MG2_1163	YP_003482400	ribosomal protein L39e
MG2_1165	YP_003482398	ribosomal protein S19e
MG2_1232	YP_003482937	ribosomal protein S27a
MG2_1233	YP_003482938	ribosomal protein S24e
MG2_1260	YP_003483746	ribosomal protein S15P
MG2_1387	YP_003483682	ribosomal protein L40e
MG2_1704	YP_003483606	ribosomal protein L7Ae
MG2_1705	YP_003483607	ribosomal protein S28e
MG2_1706	YP_003483608	ribosomal protein L24E
MG2_1729	YP_003482448	ribosomal protein S10
MG2_1736	YP_003483781	ribosomal protein S8e

Table S7

Putative peptidase genes and non-peptidase homologs identified in the Marine Group-II Euryarchaeote genome. MEGAN classifications indicate the most specific taxonomic group with which the protein identified. Archaeal classifications are shaded for visibility

A) Putative peptidases

locus	length (nt)	MEROPS id	e-value	MEGAN classification	product
MG2_0002	1013	MER001728	6.60E-21	Euryarchaeota	methionyl aminopeptidase 2 (M24A subfamily)
MG2_0059	3407	MER000383	3.70E-50	Chroococcales	proprotein convertase, kexin type 2, (S8B subfamily)
MG2_0167	1835	MER161420	6.70E-104	Actinomycetales	X-Pro dipeptidyl-peptidase (S15 family)
MG2_0176	3128	MER085659	1.10E-25	Deltaproteobacteria	KP-43 subtilisin peptidase (S8A subfamily)
MG2_0201	1745	MER001284	2.30E-05	Bacteria	Zn-dependent peptidase (M28E subfamily)
MG2_0207	1586	MER002161	2.60E-10	Bacteroidetes	Zn-dependent peptidase (M28A subfamily)
MG2_0265	4724	MER133814	4.80E-06	none	putative subtilisin peptidase (S8A subfamily)
MG2_0275	1184	MER180647	1.20E-08	Halobacteriaceae	conserved cell surface protein similar to PKD/Chitinase/Collegenase domain proteins (putative S8A and M9 family peptidases)
MG2_0303	965	MER000431	4.00E-72	Cellular organisms	prolyl aminopeptidase (S33 family)
MG2_0348	1274	MER002038	4.00E-09	Bacteria	putative X-Pro metallopeptidase (M24B subfamily)
MG2_0389	3815	MER133814	1.80E-05	Bacteria	putative subtilisin peptidase (S8A subfamily)
MG2_0401	2123	MER001212	3.20E-16	Bacteria	metallocarboxypeptidase T (M14A subfamily)
MG2_0429	4538	MER016991	3.70E-74	cellular organism	KP-43 subtilisin peptidase (S8A subfamily)
MG2_0447	1805	MER133814	6.40E-06	cellular organisms	putative peptidase (S8A subfamily)
MG2_0492	1958	MER161420	2.40E-110	Bacteria	X-Pro dipeptidyl-peptidase (S15 family)
MG2_0512	1496	MER001186	1.10E-116	cellular organism	Zn-dependent carboxypeptidase Taq (M32 family)
MG2_0565	1367	MER058960	7.30E-05	none	putative subtilisin peptidase (S8A subfamily)
MG2_0567	2009	MER019280	6.10E-56	cellular organisms	tengconlysin-like subtilisin peptidase (S8A subfamily)
MG2_0636	1835	MER073084	3.00E-50	Bacteria	BcepAP-like glycyl aminopeptidase (M61 family)
MG2_0645	1229	MER100667	3.40E-11	Chloroflexi	predicted aminopeptidase (M28A subfamily)
MG2_0646	4208	MER180647	6.10E-09	Flavobacteriales	conserved putative peptidase (M30 and S8A family domains)
MG2_0652	1448	MER001205	9.60E-05	cellular organisms	putative metallocarboxypeptidase (M14 family), similar to Succinylglutamate desuccinylase (ASTE)/aspartoacylase (ASPA)
MG2_0659	1436	MER145055	3.20E-78	Euryarchaeota	peptidase U62, putative microcin-like modulator of DNA gyrase
MG2_0660	1403	MER051962	2.60E-26	Euryarchaeota	peptidase U62, putative microcin-like modulator of DNA gyrase
MG2_0747	2195	MER004123	5.50E-30	cellular organisms	subtilisin peptidase with homology to subA cytotoxin (S8A subfamily)
MG2_0818	1784	MER005239	1.50E-12	cellular organisms	Zn-dependent peptidase (M28B subfamily) with an SAP DNA binding domain
MG2_0905	5318	MER133814	4.80E-06	Bacteria	conserved putative beta-lytic metallopeptidase (M23B subfamily)
MG2_0906	6119	MER133814	9.20E-07	Bacteria	conserved putative peptidase (U69 and S8A family homology)
MG2_0909	3326	MER145093	4.10E-07	Bacteria	conserved putative peptidase (U69 and S8A family homology)
MG2_0910	3695	MER134526	2.30E-09	Bacteria	conserved putative peptidase (U69 and S8A family homology)
MG2_0955	626	MER000548	8.60E-47	Euryarchaeota	proteasome, beta component (T1A subfamily)
MG2_1084	722	MER017631	2.40E-59	Euryarchaeota	proteasome, alpha component (T1 family)
MG2_1145	569	MER006226	2.40E-38	Euryarchaeota	bifunctional UGMP fusion protein (Bud32 protein kinase / M22 family, Kae1 putative peptidase)
MG2_1183	2120	MER075049	5.90E-19	cellular organisms	Immune inhibitor A-like metalloendopeptidase (M6 family)
MG2_1200	1412	MER001244	1.20E-67	Bacteria	X-Pro aminopeptidase (M24B subfamily)
MG2_1217	2582	MER145048	3.80E-22	Bacteria	bacillopeptidase F-like subtilisin (S8A subfamily)
MG2_1226	1253	MER019280	1.10E-34	cellular organisms	subtilisin peptidase with homology to subA cytotoxin (S8A subfamily)
MG2_1245	4985	MER051661	2.40E-47	cellular organisms	KP-43 subtilisin peptidase (S8A subfamily)
MG2_1281	1220	MER019280	4.10E-20	cellular organisms	subtilisin peptidase with homology to subA cytotoxin (S8A subfamily)
MG2_1361	1277	MER161420	1.50E-118	Bacteria	putative X-Pro dipeptidyl-peptidase (S15 family)
MG2_1362	521	MER161420	3.70E-37	none	putative X-Pro dipeptidyl-peptidase (S15 family)
MG2_1374	746	MER173648	1.50E-10	Bacteria	putative prolyl oligopeptidase (S9 family)
MG2_1382	2078	MER059659	2.70E-71	Euryarchaeota	Lon-B peptidase, ATP-dependent (S16 family)
MG2_1463	1808	MER161420	9.20E-127	Actinomycetales	X-Pro dipeptidyl-peptidase (S15 family)
MG2_1486	2783	MER040490	5.60E-33	cellular organisms	subtilisin peptidase with homology to subA cytotoxin (S8A subfamily)
MG2_1487	989	MER017194	9.40E-08	Bacteria	Rhomboid family protein, possible intramembrane serine protease (S54 family)
MG2_1493	635	MER180577	5.10E-10	Bacteria	putative lysostaphin peptidase (M23B subfamily)
MG2_1525	5642	MER051661	8.20E-36	cellular organisms	KP-43 subtilisin peptidase (S8A subfamily) with thrombospondin type 3 repeats
MG2_1528	1490	MER001236	2.00E-69	cellular organisms	multifunctional PepA aminopeptidase (M17 family)
MG2_1529	1346	MER161667	3.00E-15	cellular organisms	putative carboxypeptidase T (M14A subfamily)
MG2_1636	1472	MER001288	1.10E-06	Bacteria	putative Zn-dependent exopeptidase (M28E subfamily)
MG2_1667	1553	MER058052	2.00E-06	Euryarchaeota	putative zinc metalloprotease (M50B subfamily)

B) Putative Non-peptidase homologs

locus	length (nt)	MEROPS id	e-value		product
MG2_0003	608	MER151562	3.60E-20	cellular organisms	putative translation factor SUA5, (M22 family non-peptidase homolog)
MG2_0123	524	MER090044	1.60E-24	Archaea	GMP synthase / gamma-glutamyl hydrolase, (C26 family unassigned peptidase)
MG2_0184	851	MER042827	2.80E-08	Bacteria	phosphoribosylformylglycinamidine synthase (C56 family non-peptidase homolog)
MG2_0311	2195	MER134379	4.70E-35	Bacteria	Anthranilate/para-aminobenzoate synthase component I (C26 family unassigned peptidase homolog)
MG2_0343	1259	MER037714	9.90E-12	Euryarchaeota	Cytosine deaminase or related metal-dependent hydrolase (M38 family non-peptidase homolog)
MG2_0387	6569	MER133814	6.00E-05	cellular organisms	regulator of chromosome condensation RCC1 with PRK-like repeat, fucose specific lectin, and pectin lyase domains (S8A subfamily non-peptidase homolog)
MG2_0514	1418	MER005767	6.40E-18	Bacteria	dihydroorotase (M38 family non-peptidase homolog)
MG2_0541	2114	MER077132	1.10E-16	Eukaryota	ABC-type phosphate/phosphonate transport system, periplasmic component (S60 family non-peptidase homolog)
MG2_0692	1538	MER003314	7.30E-32	Bacteria	amidophosphoribosyltransferase (C44 family)
MG2_0799	1334	MER060647	4.20E-61	cellular organisms	carbamoyl-phosphate synthase small subunit (C26 family non-peptidase homolog)
MG2_0857	1766	MER033254	2.20E-16	cellular organisms	asparagine synthase, glutamine-hydrolysing (C44 family non-peptidase homolog)
MG2_1094	3665	MER180647	1.80E-06	Bacteria	putative C-type lectin with I63 family peptidase inhibitor-like homology
MG2_1144	1358	MER033186	1.10E-48	cellular organisms	Imidazolonepropionase or related amidohydrolase (M38 family non-peptidase homolog)
MG2_1484	1820	MER003327	5.30E-55	Bacteria	glucosamine--fructose-6-phosphate aminotransferase, isomerizing (C44 family non-peptidase homolog)

Table S8

Putative lipid metabolism genes in the Marine Group-II Euryarchaeote genome. MEGAN classifications indicate the most specific taxonomic group with which the protein identified. Archaeal classifications are shaded for visibility

	Locus	Gene / EC number	KO/ProtClust	MEGAN classification	Annotation
Fatty acid synth					
	MG2_1652	6.3.4.14 / 6.4.1.2	K01961	cellular organisms	acetyl-CoA carboxylase, biotin carboxylase subunit
	MG2_0944/0945	FabH / 2.3.1.180	K00648	Actinomycetales	3-oxoacyl-[acyl-carrier-protein] synthase III
	MG2_0114	FabG / 1.1.1.100	K00059	cellular organisms	3-ketoacyl-(acyl-carrier-protein) reductase
	MG2_1373	FabG / 1.1.1.100	K00059	Bacteria	3-ketoacyl-(acyl-carrier-protein) reductase
	MG2_1640	FabI / 1.3.1.9	K00208	Bacteria	enoyl-[acyl-carrier protein] reductase I
	MG2_0458	4.2.1.17	K00074	Firmicutes	bifunctional 3-hydroxyacyl-CoA dehydrogenase/enoyl-CoA hydratase/isomerase family protein
	MG2_1647	3.1.2.20	K01076	cellular organisms	putative acyl-CoA hydrolase
PUFA Synth					
	MG2_1639	2.3.1.94	K10817	Chloroflexi	Polyketide synthase, beta-ketoacyl synthase domain
	MG2_1647	3.1.2.20	K01076	cellular organisms	putative acyl-CoA hydrolase
	MG2_1638	2.7.8.-	K06133	Alphaproteobacteria	putative 4'-phosphopantetheinyl transferase
	MG2_0653		CLSK946835	Myxococcales	thioesterase superfamily protein
	MG2_0738	1.14.19.1	K00507	Proteobacteria	steroyl-CoA desaturase (delta-9 fatty acid desaturase)
	MG2_1035	1.14.21.6	K00227	Bacteria	putative lathosterol oxidase, fatty acid hydroxylase superfamily
Ether-linked lipids					
	MG2_0160		CLSK2468350	Bacteroidetes	putative geranylgeranyl reductase
	MG2_1195		CLSK227774	Euryarchaeota	putative geranylgeranyl reductase
	MG2_1005	2.5.1.-	PRK09573	Euryarchaeota	4-hydroxybenzoate octaprenyltransferase
	MG2_0180		PRK04169	Euryarchaeota	putative geranylgeranylglyceryl phosphate synthase
	MG2_0580	1.4.1.-	K11646	cellular organisms	3-dehydroquinate synthase II
	MG2_1065	2.5.1.1 / 2.5.1.10	K13787	cellular organisms	geranylgeranyl diphosphate synthase
	MG2_0670	2.5.1.-	K02523	cellular organisms	octaprenyl diphosphate synthase / Geranylgeranylpyrophosphate synthase
Glycerophospholipids					
	MG2_1620	2.7.1.30???	CLSK865122	cellular organisms	Predicted Inorganic polyphosphate/ATP-NAD kinase
	MG2_1285	2.7.7.41	CLSK2789688	Euryarchaeota	CDP-diglyceride synthetase
	MG2_0037	2.7.1.107	K04718	Chlorophyta	sphingosine kinase
	MG2_1540	3.1.3.4	CLSK2770155	cellular organisms	putative membrane-associated phospholipid phosphatase
	MG2_1125	3.1.1.23	CLSK951691	no hit > 35 bits	alpha/beta hydrolase fold protein
	MG2_0104	2.3.1.15 / 2.3.1.42	K13507	cellular organisms	putative Lysophospholipid acyltransferase (LPLAT) involved in glycerophospholipid biosynthesis
	MG2_0568	2.3.1.15 / 2.3.1.42	K00655	cellular organisms	putative Phospholipid/glycerol acyltransferase
	MG2_0929	2.3.1.51	K00655	Bacteria	1-acyl-sn-glycerol-3-phosphate acyltransferase or related phospholipid/glycerol acyltransferase
	MG2_0546	1.1.1.94	K00057	Bacteria	NAD(P)H-dependent glycerol-3-phosphate dehydrogenase
	MG2_1675		CLSK903188	Plesiocystis pacifica SIR-1	putative integral membrane protein with PLC-likephosphodiesterase, TIM beta/alpha-barrel domain
	MG2_1349		PF03009 (Pfam)	Clostridiales	putative Glycerophosphoryl diester phosphodiesterase
	MG2_1350		CLSK752696	cellular organisms	putative Membrane-associated phospholipid phosphatase
	MG2_1352		CLSK2835513	Bacteria	Metal-dependent phosphoesterase (PHP family)
Fatty acid metabolism					
	MG2_1503	6.2.1.3	K01897	Myxococcales	long-chain acyl-CoA synthetase
	MG2_0560	1.3.99.3	K00249	cellular organisms	Acyl-CoA dehydrogenase
	MG2_0630	1.3.99.3	K00249	Bacteria	Acyl-CoA dehydrogenase
	MG2_1576	4.2.1.17	K01692	Bacteria	Enoyl-CoA hydratase / isomerase
	MG2_0458	4.2.1.17 / 1.1.1.211	K00074	Firmicutes	bifunctional 3-hydroxyacyl-CoA dehydrogenase/enoyl-CoA hydratase/isomerase family protein
	MG2_1061	2.3.1.16	K00632	cellular organisms	acetyl-CoA acetyltransferase Protein
	MG2_0518	2.3.1.9	K00626	Bacteria	acetyl-CoA C-acetyltransferase Protein
	MG2_0786	2.3.1.9	K00626	cellular organisms	acetyl-CoA C-acetyltransferase Protein
	MG2_0702	1.3.99.7	K00252	cellular organisms	Glutaryl-CoA dehydrogenase
Alkane degradation					
	MG2_0476	1.14.15.3	K00496	Bacteria	alkane 1-monooxygenase
	MG2_0934	1.1.1.1	K00001	cellular organisms	alcohol dehydrogenase-like protein
	MG2_1561	1.2.1.3	K00128	Halobacteriaceae	aldehyde dehydrogenase
	MG2_1600	1.2.1.3	K00128	Bacteria	aldehyde dehydrogenase

Table S9

Results of pairwise reciprocal blast analyses between proteins coded in seven archaeal genomes. Counts of homologous proteins shared between each pair of genomes, as determined by reciprocal best hits (RBHs), are shown above the diagonal. Below the diagonal is the fraction of proteins shared between each pair of genomes, calculated as percentage of shared homologs for the genome with the smallest number of total proteins. Percentages are shaded by decile for visibility.

	MG-II	*A. boonei*	*P. torridus*	*T. volcanium*	*A. fulgidus*	*P. abyssi*	*N. maritimus*
MG-II		506	472	476	505	468	440
A. boonei	33%		655	668	702	770	487
P. torridus	30%	42%		1047	620	578	535
T. volcanium	30%	43%	68%		636	599	512
A. fulgidus	28%	45%	40%	41%		819	583
P. abyssi	26%	50%	37%	38%	43%		538
N. maritimus	25%	31%	34%	33%	32%	30%	

Table S10

Assignment of non-homologous proteins among three comparison genomes to NCBI taxonomic groups. The table background is shaded by domain level classification (corresponding with the color scheme of Figure S20.) Top sub-domain groups are shown underlined and taxonomic groups within them that received classified proteins are shown below, within the same box. Counts represent the number of proteins classified by MEGAN at that taxonomic level. Sums are calculated for each domain and sub-domain. Totaled sums in bold type correspond with the counts shown on Figure S19C. The genomes used are: Marine Group-II *Euryarchaeote*, *Aciduliprofundum boonei* (NC_013926) and *Nitrosopumilus maritimus* (NC_010085).

	Mar. Group-II		*A. boonei*		*N. maritimus*		**Total**	
	Count	Sum	Count	Sum	Count	Sum	**Count**	**Sum**
root	4	1126	1	848	2	1161	7	**3135**
cellular organisms	128	544	140	625	186	647	454	**1816**
Bacteria	180	332	77	119	177	262	434	**713**
Actinobacteria	0	14	0	0	0	6	0	**20**
Actinobacteria (class)	0		0		6		6	
Actinobacteridae (subclass)	1		0		0		1	
Actinomycetales (order)	13		0		0		13	
Bacteroidetes/Chlorobi group	0	17	5	5	0	0	5	**22**
Bacteroidetes (phylum)	11		0		0		11	
Flavobacteria (class)	6		0		0		6	
Chloroflexi	6	11	0	0	14	14	20	**25**
Chloroflexi (class)	5		0		0		5	
Cyanobacteria	4	9	0	0	7	7	11	**16**
Chroococcales (order)	5		0		0		5	
Firmicutes	9	15	2	17	6	15	17	**47**
Bacillales (order)	0		0		3		3	
Bacillus (genus)	0		0		6		6	
Clostridia (class)	6		7		0		13	
Clostridiales (order)	0		8		0		8	
Planctomycetes	0	9	0	0	0	0	0	**9**
Planctomycetaceae (family)	4		0		0		4	
Planctomyces (genus)	5		0		0		5	
Proteobacteria	18	77	3	14	17	43	38	**134**
Alphaproteobacteria (class)	11		0		7		18	
Betaproteobacteria (class)	6		0		8		14	
delta/epsilon subdiv. (subphylum)	0		2		0		2	
Deltaproteobacteria (class)	11		9		0		20	
Myxococcales (order)	6		0		0		6	
Gammaproteobacteria (class)	25		0		11		36	
Thermotogae	0	0	0	6	0	0	0	**6**
Thermotogales (order)	0		1		0		1	
Thermotogaceae (family)	0		5		0		5	
Archaea	6	64	51	366	40	187	97	**617**
Crenarchaeota	0	5	0	22	0	44	0	**71**
Thermoprotei (class)	5		2		17		24	
Desulfurococcales (order)	0		3		0		3	
Desulfurococcaceae (family)	0		5		8		13	
Sulfolobaceae (family)	0		5		4		9	
Sulfolobus (genus)	0		0		6		6	
Thermoproteales (order)	0		7		3		10	
Thermoproteaceae (family)	0		0		6		6	
Euryarchaeota	32	53	83	288	47	94	162	**435**
Archaeoglobaceae (family)	0		8		6		14	
Archaeoglobus (genus)	0		8		0		8	
Halobacteriaceae (family)	15		5		10		30	
Methanobacteriales (order)	0		1		0		1	
Methanobacteriaceae (family)	0		7		0		7	
Methanococcales (order)	0		7		0		7	
Methanomicrobia (class)	6		12		11		29	
Methanomicrobiales (order)	0		7		0		7	
Methanosarcinales (order)	0		0		2		2	
Methanosaeta thermophila	0		6		0		6	
Methanosarcinaceae (family)	0		11		2		13	
Methanosarcina (genus)	0		0		5		5	
Thermococcaceae (family)	0		79		6		85	
Thermococcus (genus)	0		54		0		54	
Thermoplasmatales (order)	0		0		5		5	
Korarchaeota	0	0	0	5	0	9	0	**14**
Candidatus Korarchaeum cryptofilum	0		5		9		14	
Eukaryota	12	20	0	0	6	12	18	**32**
Fungi/Metazoa group	0	0	0	0	1	6	1	**6**
Metazoa (kingdom)	0		0		5		5	
Viridiplantae	2	8	0	0	0	0	2	**8**
Chlorophyta (phylum)	6		0		0		6	
no hits	578	578	222	222	512	512	1312	**1312**

References and Notes

1. D. M. Karl, Microbial oceanography: Paradigms, processes and promise. *Nat. Rev. Microbiol.* **5**, 759 (2007). doi:10.1038/nrmicro1749 Medline

2. F. Azam, F. Malfatti, Microbial structuring of marine ecosystems. *Nat. Rev. Microbiol.* **5**, 782 (2007). doi:10.1038/nrmicro1747 Medline

3. J. T. Staley, A. Konopka, Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu. Rev. Microbiol.* **39**, 321 (1985). doi:10.1146/annurev.mi.39.100185.001541 Medline

4. M. S. Rappé, S. J. Giovannoni, The uncultured microbial majority. *Annu. Rev. Microbiol.* **57**, 369 (2003). Medline

5. P. G. Falkowski, T. Fenchel, E. F. Delong, The microbial engines that drive Earth's biogeochemical cycles. *Science* **320**, 1034 (2008). doi:10.1126/science.1153213 Medline

6. R. Stepanauskas, M. E. Sieracki, Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 9052 (2007). doi:10.1073/pnas.0700496104 Medline

7. O. Béjà, To BAC or not to BAC: Marine ecogenomics. *Curr. Opin. Biotechnol.* **15**, 187 (2004). doi:10.1016/j.copbio.2004.03.005 Medline

8. J. C. Venter *et al*., Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66 (2004); 10.1126/science.1093857. doi:10.1126/science.1093857 Medline

9. E. F. DeLong *et al*., Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496 (2006). doi:10.1126/science.1120250 Medline

10. D. B. Rusch *et al*., The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007). doi:10.1371/journal.pbio.0050077 Medline

11. S. Yooseph *et al*., The Sorcerer II Global Ocean Sampling expedition: Expanding the universe of protein families. *PLoS Biol.* **5**, e16 (2007). doi:10.1371/journal.pbio.0050016 Medline

12. D. A. Walsh *et al*., Metagenome of a versatile chemolithoautotroph from expanding oceanic dead zones. *Science* **326**, 578 (2009). doi:10.1126/science.1175309 Medline

13. D. B. Rusch, A. C. Martiny, C. L. Dupont, A. L. Halpern, J. C. Venter, Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16184 (2010). doi:10.1073/pnas.1009513107 Medline

14. E. R. Mardis, The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133 (2008). doi:10.1016/j.tig.2007.12.007 Medline

15. M. Hess *et al*., Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463 (2011). doi:10.1126/science.1200387 Medline

16. Materials and methods are available as supporting material on *Science* Online.

17. K. D. Pruitt, T. Tatusova, D. R. Maglott, NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61 (2007). doi:10.1093/nar/gkl842 Medline

18. A. Pernthaler, C. M. Preston, J. Pernthaler, E. F. DeLong, R. Amann, Comparison of fluorescently labeled oligonucleotide and polynucleotide probes for the detection of pelagic marine bacteria and archaea. *Appl. Environ. Microbiol.* **68**, 661 (2002). doi:10.1128/AEM.68.2.661-667.2002 Medline

19. L. Herfort *et al*., Variations in spatial and temporal distribution of Archaea in the North Sea in relation to environmental variables. *FEMS Microbiol. Ecol.* **62**, 242 (2007). doi:10.1111/j.1574-6941.2007.00397.x Medline

20. O. Béjà *et al*., Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2**, 516 (2000). doi:10.1046/j.1462-2920.2000.00133.x Medline

21. D. Moreira, F. Rodríguez-Valera, P. López-García, Analysis of a genome fragment of a deep-sea uncultivated Group II euryarchaeote containing 16S rDNA, a spectinomycin-like operon and several energy metabolism genes. *Environ. Microbiol.* **6**, 959 (2004). doi:10.1111/j.1462-2920.2004.00644.x Medline

22. N.-U. Frigaard, A. Martinez, T. J. Mincer, E. F. DeLong, Proteorhodopsin lateral gene transfer between marine planktonic Bacteria and Archaea. *Nature* **439**, 847 (2006). doi:10.1038/nature04435 Medline

23. A.-B. Martin-Cuadrado *et al*., Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J.* **2**, 865 (2008). doi:10.1038/ismej.2008.40 Medline

24. L. E. Petrovskaya *et al*., Predicted bacteriorhodopsin from *Exiguobacterium sibiricum* is a functional proton pump. *FEBS Lett.* **584**, 4193 (2010). doi:10.1016/j.febslet.2010.09.005 Medline

25. R. Ghai *et al*., Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing. *ISME J.* **4**, 1154 (2010). doi:10.1038/ismej.2010.44 Medline

26. P. E. Galand, C. Gutiérrez-Provecho, R. Massana, J. M. Gasol, E. O. Casamayor, Inter-annual recurrence of archaeal assemblages in the coastal NW Mediterranean Sea (Blanes Bay Microbial Observatory). *Limnol. Oceanogr.* **55**, 2117 (2010). doi:10.4319/lo.2010.55.5.2117

27. A.-L. Reysenbach, G. E. Flores, Electron microscopy encounters with unusual thermophiles helps direct genomic analysis of *Aciduliprofundum boonei*. *Geobiology* **6**, 331 (2008). doi:10.1111/j.1472-4669.2008.00152.x Medline

28. A. Ruepp *et al*., The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* **407**, 508 (2000). doi:10.1038/35035069 Medline

29. Y. Koga, H. Morii, Biosynthesis of ether-type polar lipids in archaea and evolutionary considerations. *Microbiol. Mol. Biol. Rev.* **71**, 97 (2007). doi:10.1128/MMBR.00033-06 Medline

30. J. M. Walter, D. Greenfield, C. Bustamante, J. Liphardt, Light-powering *Escherichia coli* with proteorhodopsin. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2408 (2007). doi:10.1073/pnas.0611035104 Medline

31. C. E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379 (1948).

32. J. R. Cole *et al*., The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141 (2009). doi:10.1093/nar/gkn879 Medline

33. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**, 5261 (2007). doi:10.1128/AEM.00062-07 Medline

34. C. B. Walker *et al*., *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 8818 (2010). doi:10.1073/pnas.0913533107 Medline

35. R. M. Morris *et al*., SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**, 806 (2002). doi:10.1038/nature01240 Medline

36. R. M. Morris *et al*., Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J.* **4**, 673 (2010). doi:10.1038/ismej.2010.4 Medline

37. W. Ludwig *et al*., ARB: A software environment for sequence data. *Nucleic Acids Res.* **32**, 1363 (2004). doi:10.1093/nar/gkh293 Medline

38. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589 (2010). doi:10.1093/bioinformatics/btp698 Medline

39. R. Seshadri, S. A. Kravitz, L. Smarr, P. Gilna, M. Frazier, CAMERA: A community resource for metagenomics. *PLoS Biol.* **5**, e75 (2007). doi:10.1371/journal.pbio.0050075 Medline

40. H. Li *et al*., The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078 (2009). doi:10.1093/bioinformatics/btp352 Medline

41. J. R. Cole *et al*., The ribosomal database project (RDP-II): Introducing myRDP space and quality controlled public data. *Nucleic Acids Res.* **35**, D169 (2007). doi:10.1093/nar/gkl889 Medline

42. A. J. Drummond *et al*., www.geneious.com.

43. H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851 (2008). doi:10.1101/gr.078212.108 Medline

44. D. R. Zerbino, E. Birney, Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821 (2008). doi:10.1101/gr.074492.107 Medline

45. E. Gansner, S. North, An open graph visualization system and its applications to software engineering. *Softw. Pract. Exper.* **30**, 1203 (2000). doi:10.1002/1097-024X(200009)30:11<1203::AID-SPE338>3.0.CO;2-N

46. H. Teeling, A. Meyerdierks, M. Bauer, R. Amann, F. O. Glöckner, Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ. Microbiol.* **6**, 938 (2004). doi:10.1111/j.1462-2920.2004.00624.x Medline

47. S. Kurtz *et al*., Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004). doi:10.1186/gb-2004-5-2-r12 Medline

48. R. K. Aziz *et al*., The RAST Server: Rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008). doi:10.1186/1471-2164-9-75 Medline

49. K. Lagesen *et al*., RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100 (2007). doi:10.1093/nar/gkm160 Medline

50. J. Besemer, M. Borodovsky, GeneMark: Web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* **33**, W451 (2005). doi:10.1093/nar/gki487 Medline

51. A. L. Delcher, K. A. Bratke, E. C. Powers, S. L. Salzberg, Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23**, 673 (2007). doi:10.1093/bioinformatics/btm009 Medline

52. E. Quevillon *et al*., InterProScan: Protein domains identifier. *Nucleic Acids Res.* **33**, W116 (2005). doi:10.1093/nar/gki442 Medline

53. A. Marchler-Bauer *et al*., CDD: A Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* **39**, D225 (2011). doi:10.1093/nar/gkq1189 Medline

54. S. F. Altschul *et al*., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389 (1997). doi:10.1093/nar/25.17.3389 Medline

55. R. L. Tatusov, M. Y. Galperin, D. A. Natale, E. V. Koonin, The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33 (2000). doi:10.1093/nar/28.1.33 Medline

56. K. S. Makarova, A. V. Sorokin, P. S. Novichkov, Y. I. Wolf, E. V. Koonin, Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol. Direct* **2**, 33 (2007). doi:10.1186/1745-6150-2-33 Medline

57. M. Kanehisa *et al*., From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **34**, D354 (2006). doi:10.1093/nar/gkj102 Medline

58. W. Klimke *et al*., The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.* **37**, D216 (2009). doi:10.1093/nar/gkn734 Medline

59. N. D. Rawlings, A. J. Barrett, A. Bateman, MEROPS: The peptidase database. *Nucleic Acids Res.* **38**, D227 (2010). doi:10.1093/nar/gkp971 Medline

60. S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696 (2003). doi:10.1080/10635150390235520 Medline

61. B. Gao, R. S. Gupta, Phylogenomic analysis of proteins that are distinctive of Archaea and its main subgroups and the origin of methanogenesis. *BMC Genomics* **8**, 86 (2007). doi:10.1186/1471-2164-8-86 Medline

62. L. Guy, J. R. Kultima, S. G. E. Andersson, genoPlotR: comparative gene and genome visualization in R. *Bioinformatics* **26**, 2334 (2010). doi:10.1093/bioinformatics/btq413 Medline

63. F. Ronquist, J. P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572 (2003). doi:10.1093/bioinformatics/btg180 Medline

64. M. C. Rivera, R. Jain, J. E. Moore, J. A. Lake, Genomic evidence for two functionally distinct gene classes. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6239 (1998). doi:10.1073/pnas.95.11.6239 Medline

65. D. H. Huson, A. F. Auch, J. Qi, S. C. Schuster, MEGAN analysis of metagenomic data. *Genome Res.* **17**, 377 (2007). doi:10.1101/gr.5969107 Medline