Supporting Information

Fan et al. 10.1073/pnas.1203287109

SI Results and Discussion

Marine Group I, Thaumarchaeota. Two groups of abundant Thaumarchaeota phylotypes belonging to the Marine Group I were present in *Stylissa* sp. 445, *Rhopaloeides odorabile*, and *Cymbastela concentrica*, including a dominant *Cenarchaeum*-like operational taxonomic unit (OTU) in *Stylissa* sp. 445 (Fig. 2). Marine Group I Thaumarchaeota [previously classified to phylum Crenarchaeota (1)] are often found in sponges (2) and can be subdivided into three clades, namely, Group C1a-α, Group C1a-Porifera A, and Group C1a-Porifera C (3). We constructed a phylogenetic tree for the four thaumarchaeal 16S rRNA gene sequences in these three sponges, including one from *Stylissa* sp. 445 (thaumarchaeal symbiont Subtype II) not in OTUs (Fig. S2*B*). The dominant thaumarchaeon Subtype I in *Stylissa* sp. 445 belonged to the sponge-specific Group C1a-Porifera C, which associates specifically with AXI2 sponges, including *Stylissa* sp. 445 (Fig. 1) (3). Association of a filamentous thaumarchaeon from this group within the collagen surrounding the siliceous spicules of three Mediterranean AXI2 sponges has been reported previously (4). The other three sequences all fell into Group C1a-α, which contains two sequenced taxa, *Nitrosopumilus maritimus* SCM1 (5) and Candidatus *Nitrosoarchaeum limnia* (6). This group contains clones from a diverse range of habitats, including hydrothermal vents, deep-sea sediments, sponges, and planktonic clones (3). No obvious host-clade specificity was found in this group (3). The thaumarchaeon in *R. odorabile* was found to be most abundant in the pinacoderm region (7) and can be transmitted vertically by sponge larvae (8). Nonetheless, the polyphyletic nature of Group C1a-α implies that at least some of its sponge-associated members may be facultative symbionts and can be free living or have a conditional association with sponge hosts. The conditional association of the thaumarchaea has been observed in *C. concentrica* (9, 10) and in sponges from Brazilian waters (11).

Community Profile Based on Single-Copy Genes. Because many bacterial and archaeal genomes contain more than one copy of the 16S rRNA gene, the real relative abundance of detected phylotypes is potentially biased (12, 13). To further quantify the community composition of our samples, the phylum-level profiles using both 16S rRNA gene sequences (assembled and unassembled; *SI Materials and Methods*) and single-copy gene (SCG)-based analyses were compared (Fig. S1 *A*–*C*). The analysis of the SCGs showed a consistent community composition between replicate samples; classification of unassembled 16S rRNA gene sequences showed greater variation, probably reflecting differences in copy number among species or strains. However, because of the limited reference database of SCGs, sequences belonging to the same ribotype can be assigned mistakenly to phylogenetically distant groups or even different phyla [e.g., the Thaumarchaeota population was assigned to both the Crenarchaeota/Thaumarchaeota and other archaeal phyla by MLTreeMap (Fig. S1 *A*–*C*)]. The assembly-based construction of 16S rRNA gene sequences gave more accurate classification for highly abundant taxa in the community, compared with direct classification of unassembled reads, which generally were too short for confident assignment. Nevertheless, all three methods confirmed that microbial populations were highly consistent within each sponge species and within seawater samples but were distinct between sample types.

Organisms Putatively Involved in Denitrification and Other Aspects of the Nitrogen Cycle in Sponge Symbionts. Several candidate organisms could perform the denitrification process in the sponge samples. *C. concentrica* contained a phylotype belonging to the family Phyllobacteriacea (Fig. 2). Members of the *Mesorhizobium* and *Nitratireductor* in this family are capable of fixing nitrogen and reducing nitrate to nitrite, respectively (14–16). Some of the NarG genes and an assembled NarGHIY gene cluster in this sponge could be assigned to the *Phyllobacteriacea* phylotype after genomic sequence binning (10). In *Scopalina* sp. and *Tedania anhelans*, two closely related, uncultured phylotypes of the family *Nitrosomonadaceae* (Betaproteobacteria) dominated the microbial communities (Fig. 2). These phylotypes also were related to *Nitrosomonas* spp. and *Nitrosospira* spp., both of which are ammonia-oxidizing bacteria (Fig. S2*A*). Species in these two genera also may contain NirK and cNorB, which are subjected to horizontal gene transfer (17, 18) and putatively are responsible for nitrous oxide production (19). Because ammonia monooxygenase was very rare in these two sponge metagenomes, but denitrification enzymes (i.e., NapA, NirK, and cNorB) were abundant (Fig. 5), these *Nitrosomonadaceae* phylotypes most likely are involved primarily in denitrification.

Oxidation of nitrite to nitrate might not be prevalent in the sponges investigated, because known nitrite-oxidizing bacteria, such as *Nitrospira*, were detected in only two *C. concentrica* samples and in very low abundance in *R. odorabile* (Fig. 2). However, the sequence of nitrite oxidoreductase subunit α is highly similar to NarG (20), so some of the sequences detected here still might be involved in the oxidation of nitrite. Anammox activity might be a rare feature in these six sponges, because homologs to the hydroxylamine-oxidizing enzyme (21) and sequences belonging to known anammox bacteria within the Planctomycete group were absent. Also, no gene for respiratory nitrite ammonification enzymes (e.g., NrfA, EC 1.7.2.2) was detected.

Glutamate dehydrogenase (GDH, PF05088) was overrepresented in *R. odorabile*, *Cymbastela coralliophila*, and *C. concentrica* (Fig. S3*C*). GDH has a potentially important role in nitrogen assimilation in pathogenic bacteria, such as *Mycobacterium smegmatis* (22). Ammonium assimilation through GDH requires much lower activation energy than the ubiquitous glutamine synthetase/glutamate synthase pathway and thus is used under conditions of nitrogen excess and energy preservation (22, 23). The distribution of GDH in sponge bacteria suggests ammonium excess is experienced by those host-associated taxa. However, GDH also can function in glutamate catabolism and therefore may act to release ammonia from natural glutamate sources, such as proteinaceous exudates from the host.

Photosynthesis and Photoprotection. Although sponges generally filter-feed to remove microbes or particulate organic matter from the surrounding seawater (24), phototrophy by microbial symbionts can make a substantial contribution to the host's growth, especially in low-nutrient and highly illuminated tropical waters (25, 26). To determine phototrophic populations in our sponge symbionts, the presence of ribulose-1,5-bisphosphate carboxylase oxygenase (RuBisCO) was investigated, and a phylogenetic analysis was performed using MLTreeMap (27). A high abundance of Form 1 and 4b RuBisCO was observed in the seawater samples (Fig. S9*A*), as is consistent with the potential for high rates of carbon fixation in marine surface waters (28). Among the six sponges, only the tropical species *C. coralliophila* and *Stylissa* sp. 445 possessed the highly abundant Form 1 RuBisCO,

mostly because of their cyanobacterial populations (Fig. 2 and Fig. S9*A*). However, the tropical sponge *R. odorabile* possessed only the RuBisCO-like proteins in the Form 4 clade, which catalyze the 2,3-diketo-5-methylthiopentyl-1-phosphate enolase reaction in the methionine-salvage pathway (29). The lack of phototrophy in *R. odorabile* has been reported previously; photorespirometry trials, photopigment analysis, and an absence of cyanobacteria in sponges from both inshore and offshore reefs clearly demonstrating that *R. odorabile* is not a photosynthetic species (30). These observations are consistent with the morphological properties of these three sponges. *C. coralliophila* and *Stylissa* sp. 445 are plate- and fan-shaped, respectively, and hence are structurally optimized to harvest light energy and couple it to carbon fixation. In contrast, *R. odorabile* is a massive 3D sponge with a dense canal system for filter-feeding (31). The bowl-shaped temperate sponge *C. concentrica* (32) may be morphologically optimized for phototropic growth but did not have a significant abundance of prokaryotic RuBisCO. However, it contains dense populations of symbiotic diatoms (33), which were mostly removed during prokaryotic cell enrichment in the present study (*SI Materials and Methods*). The data show that sponges with body shapes optimized for light harvesting conducted photosynthesis by phylogenetically diverse symbiotic populations (e.g., cyanobacterial populations vs. diatoms) that were distinct from the free-living populations (e.g., cyanobacteria and proteobacteria) (Fig. S9*A*).

High levels of illumination can result in photodamage, and, consistent with this condition, a high abundance of photolyases (PF03441, COG3046, COG0415, Subsystem: DNA Repair Bacterial Photolyase) and phytoene dehydrogenase, the key enzyme in carotenoid biosynthesis (COG1233, here comprising mostly CrtI-type phytoene dehydrogenase) were detected in planktonic samples (Fig. 4 and Fig. S3 *C* and *D*). Many phylogenetic divergent planktonic taxa had this photolyase protection mechanism (Fig. S9*B*). In contrast, photolyases were rare in the sponge communities, likely because of photoprotection provided by the sponge tissue and pigments (34). Nevertheless, some photostress still might occur, especially in the community of the tropical sponge *Stylissa* sp. 445, where a number of diverse photolyase sequences were found (Fig. S9*B*).

Regulation of Cellular Response. Functions in signal transduction and regulation are overrepresented in sponge samples (Fig. 4 and Fig. S3 *C* and *D*).

HAMP (for "present in histidine kinases, adenyl cyclases, methyl-accepting proteins and phosphatases")-containing proteins (PF00672) act as transmembrane modules of two-component signaling pathways for response to changing environmental conditions (35). The HAMP domains are chemoreceptors that couple the motions of transmembrane helices to the activity of a downstream cytoplasmic output domain (36). Their specific role in sponge symbionts is currently unclear.

Protein tyrosine kinase (PTK) (PF07714), which functions as an on/off switch in many cellular functions by modifying gene expression, was abundant in the sponge metagenomes (Fig. 6 *G* and *H*) (37). This canonical PTK family is found mostly in eukaryotes, whereas bacteria have developed several other types of enzymes that catalyze protein phosphorylation on tyrosine (38). Two eukaryotic-like PTKs have been found to function in signal transduction (39, 40). Why this eukaryotic-like PTK is present in sponge symbionts is not clear.

Protein homologs to the eukaryotic male sterility protein (PF07993) also were abundant in the sponge samples (41, 42). This protein is capable of lipid biosynthesis in prokaryotes (43). Synthetic lipids can be signals through which surface-associated microorganisms regulate motility for predatory feeding (43). Many sponge microorganisms live in the mesohyl between sponge cells (44), and thus surface motility might be important for their association with the sponge host.

Other regulatory functions might be provided by abundant ATPases (COG1373, COG4637, COG0464, and COG2865), including some specifically acting in gene regulation (COG0464 and COG2865) (Fig. 4).

Proteins belonging to the ribosome-binding GTPase superfamily (COG1217) were abundant in the seawater samples. This protein acts as a translational GTPase and as a global stress and virulence regulator. It has been found to be involved in diverse stress-resistant functions in different bacteria (45–50).

Enzymes Involved in DNA Recombination. The potential activities in genetic exchange and rearrangement in sponge microbial communities were further supported further by an overrepresentation of DNA recombination and repair enzymes, including RecD (COG0507), which is involved in dsDNA break repair, and DinG (COG1199), SSL2 (COG1061), and HepA (COG0553), which are crucial for DNA recombination repair and excision repair, and of protein families involved in general DNA-modifying activities, such as HNH endonuclease (PF01844) and the SNF2 family N-terminal domain (PF00176). These results are consistent with the genome of the sponge symbiont *Cenarchaeum symbiosum* (51), suggesting that these enzymes are essential for the stable insertion of mobile DNA into the chromosomes and repair of flanking regions in sponge symbionts.

Cyanophages in *Stylissa* sp. 445. Despite the general abundance of bacteriophages in the marine environment, no information is available on phage diversity in sponge systems. Here we noted 211 sequences encoding for a T4-like phage capsid-assembly protein G20: 107 from *Stylissa* sp. 445, 23 from *R. odorabile*, and 72 from and seawater samples. Phylogenetic analysis revealed that all 211 G20 sequences belonged to the cyanophage group (Fig. S7*C*). When normalized, the abundance of the G20 protein in *Stylissa* sp. 445 averaged seven copies per bacterial/archaeal genome (Fig. S7*B*). This abundance was correlated with the large population of cyanobacteria (mostly *Synechococcus*) in *Stylissa* sp. 445 (Fig. 2). T4 phages are capable of undergoing only a lytic and not the lysogenic lifecycle (52). Thus, it is predicted that the size of the cyanobacterial population in *Stylissa* sp. 445 is influenced strongly by lysis or that cyanobacterial metabolism is controlled by viral photosystem genes, as recently demonstrated in other systems (53).

The phylogenetic analysis also demonstrated that many G20 sequences formed distinct clusters and had no closely related homologs in the current National Center for Biotechnology Information (NCBI) nonredundant (NR) database (Fig. S7*C*). These cyanophage sequences showed no host specificity or apparent biogeography in the samples, as is consistent with the "everything is everywhere" notion of global phage distribution (54, 55).

Further Information on Selfish Genetic Elements. It is possible that mobile genetic elements (MGEs), such as plasmids or phages, acquire chromosomal restriction-modification (R-M) systems and hence become a stable part of the microbial cell. Such selfish and self-protecting features are recognized as an important mechanism for maintaining extrachromosomal elements (56, 57). Toxin–antitoxin (T–A) systems play a similar role in stabilizing selfish genetic elements. These systems generally are arranged with one toxin and one antidote and lead to postsegregational killing (58) or addiction (56) of the host cell. In all sponge metagenomes, we found a larger number of Type I (COG0286, COG0610, COG4096, PF02384, PF12161, and PF01420), Type II (COG0270, COG1743, COG0863, COG0338, COG4889, and PF00145), and Type III (COG2189 and PF04851) R-M systems and proteins from Doc/Phd family (COG3177 and PF02661),

VapI (HigA) of the HigAB system (COG3093), and other T–A systems [Subsystem: Toxin–antitoxin systems (other than RelBE and MazEF)], which would help to stabilize an array of MGEs. However, chromosomes also can unburden themselves from extrachromosomal "hitchhikers" by acquiring the same T–A or R-M systems as the MGEs. In turn, MGEs could acquire new T–A and R-M systems that would ensure their continuing propagation. This process would lead to an arms race between the chromosomes and the MGEs, and it has been hypothesized that it results in a higher number of T–A systems in bacterial species that have high rate of horizontal gene transfer (HGT) (59). The observation of abundant T–A and R-M systems therefore is consistent with the large number and diversity of MGEs observed as well as with the high frequency of HGT we postulate for sponge symbionts.

Potential Targets of Clustered, Regularly Interspaced, Short, Palindromic Repeats. To explore further the dynamics of the local phage populations, spacers were searched against the NCBI nucleotide (NT) and virus databases, but no hits were found; this result suggests host specificity in the local environment (60) and highlights the largely unknown viral diversity (61). Although the sample fractionation used in this study did not target viral particles specifically, it still was possible to identify 85 putative viral/phage sequences that matched 43 clustered, regularly interspaced, short, palindromic repeats (CRISPR) spacers from four sponge species, with a notable number of hits found in *C. concentrica* (*SI Materials and Methods* and Fig. S6*D*). All spacer–phage pairs were found exclusively within the same sponge species, again indicating a high degree of host specificity. Generally, spacers were in much lower abundance than their targets (Fig. S6*D*). Although the protocols in the present study enriched specifically for bacterial and archaeal cells, likely resulting in a considerable underestimation of viruses, the results showed that a large number of phage sequences were present in the sponge and were subject to potential defense by the CRISPR system.

SI Materials and Methods

Sample Collection. Sampling of *R. odorabile, Stylissa* sp. 445, and *C. coralliophila* occurred at Davies Reef on the Great Barrier Reef (18° 49′S, 147° 38′E), and sampling of *C. concentrica, T. anhelans*, and *Scopalina* sp. was done at Bare Island in Botany Bay, New South Wales (33° 59′S, 151° 14′E). All sponges were collected by SCUBA diving at depth of 7–10 m on the sampling days (Table 1) and were placed in ice-cold, filter-sterilized seawater. Further processing of the samples occurred in the laboratory within 15 min of collection.

Sponge Identification. Sponges were morphologically identified by Patricia Sutcliffe and Merrick Ekins at the Queensland Museum, Brisbane, Australia, and their phylogenetic relationships were investigated further using metagenomic-derived 18S rRNA gene sequences (see below).

Microbial Cell Enrichment, DNA Extraction, and Sequencing. Microorganisms were enriched from the sponges according to the methods described by Thomas et al. (9), except that the final filter cutoff used to remove eukaryotic cells was selected individually for each sponge species to meet the following criteria: (*i*) no preferential removal of microorganisms (checked by microscopy and denaturing gradient gel electrophoresis), and (*ii*) removal of as many eukaryotic cells and organelles as possible (checked by microscopy and 16S/18S rRNA gene comparative PCR). DNA extraction of the seawater samples from the 0.1-μm and 0.8-μm filters and from the cell pellets of sponge bacteria were performed as described by Thomas et al. (9). Shotgun libraries were constructed and sequenced on the Roche 454 Titanium platform at the J. Craig Venter Institute. The shotgun sequencing is available through the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis website (http://camera.calit2.net/) under project accession CAM_PROJ_BotanyBay. Dereplication of the raw reads was conducted by cd-hit-454 (62) with the similarity cutoff of 96% and the short-replicate read being covered by at least 95% of the longer replicate.

Small Subunit of rRNA Gene Reconstruction from Metagenomic Shotgun Data and Phylogenetic Analysis. Small subunit (SSU) (i.e., 16S and 18S) rRNA gene-containing reads were identified from the dereplicated sequence dataset using Metaxa (version 1.0.2) (63). Reads (>300 nt) from the three replicates of each sponge species or seawater samples were pooled and assembled separately with the GS De Novo Assembler 2.3 (454 Life Sciences) using the "cDNA" option. Parameters were set as 99% overlap identity, 43-nt overlap length, "Reads limited to one contig," and "Extending low depth overlaps." Contigs with coverage of more than 10 reads and length greater than 700 nt were selected and aligned to the SILVA SSURef 1.08 database (64) using the SINA aligner (v1.2.9) and were inserted into the SILVA SSURef tree using the specific pos_var_Archaea/Bacteria/Eukarya filters of the ARB package (65). The pair-wise phylogenetic distances of the sequences were calculated with an R script (66) and clustered by Mothur (average linkage and distance cutoff of 0.03). Representatives of these clusters then were defined as OTUs, and their abundance (number of reads contained) in each sample was calculated.

To establish the validity of this reconstruction procedure and to estimate the potential production of chimeric sequences, pyrosequencing reads (1,000,000 reads with read length of 350 nt) for artificial communities with low, medium, and high complexity (67) were simulated using GemSim (68). No chimeric OTUs for a phylogenetic distance cutoff of 0.03 were detected. Further details on this assembly-based SSU rRNA reconstruction and its evaluation will be presented elsewhere (69).

Community diversity of each sample was assessed by plotting richness curves based on OTU numbers and total phylogenetic distance, respectively, using QIIME (70). Taxonomic assignment of the OTUs was conducted manually based on their locations in the SILVA SSURef tree. A maximum-likelihood tree of the OTUs was constructed using RAxML (71) after alignment by SINA and removal of ambiguous positions by Gblocks (-t = d, -b4 = 5, -b5 = h) (72). The 16S rRNA gene profiles of the samples were clustered using the weighted Unifrac algorithm implemented in QIIME (73).

Assembly, Removal of Eukaryotic DNA, and Gene Prediction. Dereplicated reads of each sample were assembled separately using the GS De Novo Assembler "genomic" with the default settings. Contigs, singletons, and outliers were pooled, and sequences smaller than 100 nt were removed. During the microbial cell fractionation, eukaryotic cells, mitochondria, or plastids might not be removed sufficiently, and therefore the metagenomic data might be contaminated with eukaryotic sequences. To remove those contaminants, assembled sequences were searched against the NCBI NT database (accessed September 15, 2010), and the resulting files were parsed through the last common ancestor algorithm implemented in MEGAN (v3.9) (74). All sequences assigned to eukaryotic origin were removed according the procedure described by Thomas et al. (9). MLTreeMap was used to ensure few eukaryotic marker genes could be detected after the above procedures (version 2.05, "minimal sequence length after Gblocks" set to 35) (27). ORFs of coding genes were predicted from the filtered sequences with the MetaGeneAnnotator (75). The coverage of each gene was calculated from the average coverage of the contig to which the ORF belongs.

For the analysis of the genes encoding RuBisCO, photolyase, eukaryotic-like proteins (ELPs), and CRISPRs, an additional filtering process was added, involving taxonomic classification with PhymmBL V3.2 (76) using a custom-designed reference database. This database was based on the default PhymmBL reference dataset that includes all sequenced prokaryotic genomes in the NCBI RefSeq database (as of March 23, 2010) and was supplemented with the genomes of the sponge *Amphimedon queenslandica*, the round worms *Brugia malayi* and *Caenorhabditis briggsae*, the diatoms *Blastocystis hominis*, *Thalassiosira pseudonana*, and *Phaeodactylum tricornutum*, the hydrozoan *Hydra magnipapillata*, and all sequenced sponge mitochondria. Contigs and their corresponding protein sequences assigned as eukaryotic after PhymmBL analysis with default parameters were removed.

Functional Annotation. For functional annotation of the samples, predicted ORFs were translated to proteins and searched against the Clusters of Orthologous Group (COG) database (77) using rpsBlast, and against the Protein Family A (Pfam-A) database (v24.0) (78) using Hmmer 3 (79), both with an E-value cutoff of 10^{-10}. Proteins with multiple domains were counted separately; repeats of the same domain in a protein were counted once. Genes also were annotated to the SEED/Subsystems (80) using the online pipeline MG-RAST (v2) (81) with an E-value cutoff of 10^{-10}. Sample matrices for COG, Pfam, and Subsystem annotation were generated. The abundance of each function (e.g., a COG entry) in a sample was weighted by the coverage of the ORFs assigned to this function.

Normalization by SCGs. The average genome sizes potentially can be quite different for metagenomic samples and thus can bias the functional profile comparison (82). Several strategies have been proposed to predict average genome size (or genome copy) in metagenomic datasets (83–85). These approaches usually calculate the average coverage of conserved SCGs for normalization. A similar approach was used here by selecting 18 COGs (namely COG0048, COG0049, COG0087, COG0088, COG0091, COG0093, COG0094, COG0096, COG0097, COG0099, COG0100, COG0102, COG0184, COG0186, COG0256, and COG0522) from the 40 universal SCGs (86). These 18 COG entries were consistently abundant across all metagenomic samples, and thus functional matrices of COG, Pfam, and Subsystem annotation counts were normalized by the average abundance of the 18 COG entries in each sample.

Identification of Differential Abundance. Statistical pairwise comparisons of functional gene profiles of the sponge group and the seawater group were conducted using an R script modified from MetaStats (87). The MetaStats script handles two matrices, the original input counts table (Ctab) and the generated percentage table (Ptab). MetaStats uses the Ptab to run a *t* test and uses the Ctab to handle spare counts. In the modified script, the sample matrix without normalization was used as the Ctab, and the normalized matrix (see above) as the Ptab. Functional gene differences were defined if all of the following criteria were met: (*i*) the *P* value was less than 0.05; (*ii*) function counts were more than three times higher in one group than in the other group; (*iii*) for the group with higher abundance, the normalized count of the specific function was greater than one copy per genome. Differential gene functions annotated by COG, Pfam, and Subsystem, respectively, were used for sample clustering by PRIMER 6 (PRIMER-E Ltd). Heatmaps were generated using Cluster 3.0 (88) and Java TreeView (89).

ELP Analysis. Proteins were searched against the Pfam full profiles of the seven candidate ELPs (Fig. 6) using Hmmer 3 with a bit score cutoff of 25. Proteins from contigs of potential eukaryotic origin were removed as predicted by PhymmBL (see above). The abundance of a specific ELP in each sample was weighted by ORF coverage and normalized by genome copy (see above). The number of repeating ELP motifs in a protein was calculated from Hmmer search results. Secretion signals were predicted with the EFFECTIVE T3 software (90) for T3 and Sec secretory pathways. Only proteins with a complete ORF or an intact N terminus were included. If a given sequence was predicted to have both T3 and Sec signals, the prediction with the higher score was counted. To compare the diversity of the ELPs among samples in a sequence-similarity context, ELPs were clustered with a cutoff of 75% identity using BlastClust (91). Representative sequences were picked for each cluster, and a pairwise alignment was generated using ClustalW 2.0 (92). Proteins with homologous domains to ELPs were retrieved for amoebae [*Entamoeba histolytica* for ankyrin repeats (ANK), leucine-rich repeats (LRR), NHL, and PTK; *Entamoeba dispar* SAW760 for LRR and NHL; *Hartmannella vermiformis* for NHL; and *Polysphondylium pallidum* for fibronectin domain III (Fn3)], sponges [*A. queenslandica* for ANK and tetratricopeptide repeats (TPR); *Suberites domuncula* for ANK and PTK; *Geodia cydonium* and *Ephydatia fluviatilis* for Fn3], a tunicate (*Oikopleura dioica* for all seven protein classes), nematodes (*Caenorhabditis elegans* for LRR, TPR, and NHL and *B. malayi* for cadherin), a fruit fly (*Drosophila melanogaster* for all seven protein classes), and human (*Homo sapiens* for all seven proteins). Unweighted Unifrac clusters of samples were generated with QIIME.

Cyanophage Population Analysis Based on G20 Proteins. Representative sequences of the cyanophage capsid assembly protein G20 were obtained from the NCBI database along with some from noncyanophage T4-like phages as outgroup sequences (93). Redundancy of these reference proteins was removed by CD-Hit with a 99% identity cutoff (94). Nonredundant sequences with a length greater than 140 amino acids were searched against all predicted proteins in the 21 samples using PSI-Blast (blastpgp -b 0 -j 3 -h 0.002 -e 0.0001). Hits were searched against the NCBI NR database (as of September 15, 2010) using BlastP (E-value cutoff of 0.0001). The five best hits for each protein were obtained, and redundancy was removed by CD-Hit (identity cutoff 99%). All the sequences obtained from the above two Blast searches along with the cyanophage G20 reference sequences were clustered by Clans (95) with a *P* value cutoff of 10^{-30} (Fig. S7*A*). Three groups were formed, and portal vertex proteins (green and blue dots in Fig. S7*A*) were removed as false positives. Proteins indicated by red and black dots in Fig. S7*A* and outgroup proteins were aligned by Muscle (96), and ambiguous positions were removed using Gblocks (t = p, -b4 = 5, -b5 = h). A phylogenetic approximate maximum-likelihood tree was constructed using FastTree 2.1 (97).

CRISPR Analysis. CRISPR arrays were predicted from the contigs and singletons after assembly by the online prediction tool CRISPRFinder (98) followed by a series of quality-filtering steps. Specifically, candidate CRISPR arrays were predicted from the contigs and singletons by CRISPRFinder (98) using the default setting, except that "Allowed mismatch between DRs" was set to 5% and "Allowed mismatch for the degenerated DR" to 20%. Because of the complexity of the samples (short reads and DNA potentially originating from bacteria/archaea, phages, and eukaryotic sequence contaminants), the simple rule of "short interspacing repeat" to identify CRISPRs can generate many false positives, including hits to microsatellites and repeat proteins. To exclude those non-CRISPR repeat sequences, stringent filtering criteria were used during the CRISPR prediction. False positives generated from microsatellite regions were removed using the tandem repeat predictor Phobos (http://www.rub.de/spezzoo/cm/cm_phobos.htm) followed by a manual check. For CRISPRs containing more than two spacers, those whose longest and

shortest spacers had a difference in length of three nt or more were removed. CRISPRs containing two spacers with a difference in length exceeding one nucleotide were also removed. The remaining CRISPRs with more than one spacer were considered as positive multispacer CRISPRs. Because of the fragmented nature of the metagenomic sequences, many candidate CRISPRs predicted by CRISPRFinder contain only one spacer (monospacer CRISPRs). Only monospacer CRISPRs containing exactly the same repeat sequences found in positive, multispacer CRISPRs were accepted. This stringent filtering yielded 203 CRISPRs.

Repeats and spacers were extracted from the CRISPRs and clustered based on pairwise identity by BlastClust with an identity cutoff of 50% and the alignment region covering 80% of the shorter sequence. Samples were clustered according to the presence or absence of repeat/spacer clusters by Bray–Curtis similarity and group-average linkage implemented in PRIMER-6.

The NCBI NT database (as of September 15, 2010) and comprehensive viral databases downloaded from CAMERA (CAM_BroadPhage, BroadPhageGenomes, CBVIRIO, HFVirus, LakeLimnopolarVirome, MarineVirome, SalternMetagenome, TampaBayPhage, ViralSpring, and ViralStromatolite) (99) were used to examine the potential targets of the spacers in public databases by BlastN (-e 0.1, -W 7, -q 3, -r 1, -G 5, -E 2, -F F) with criteria allowing one gap and one mismatch in the query sequence. To identify the potential targets of these CRISPRs in the present metagenomic samples, all spacers and repeats were searched using BlastN (-e 0.1, -W 7, -q 3, -r 1, -G 5, -E 2, -F F) against the contigs/singletons containing no CRISPR loci. Only contigs/singletons with matched spacers but not repeats were taken as potential targets of the CRISPRs. Plotting of spacers and their potential targets was conducted in Cytoscape 2.8.1 (100).

CRISPR-Associated Protein Analysis. All ORFs were searched against the Tigrfam HMM profiles for CRISPR-associated (CAS) proteins (101) using Hmmer 3 (79) with a cutoff score of 25. Raw counts were normalized by genome copy (see above) for each sample.

Because the Csn1 profile (TIGR01865) may pick up non-Csn1 proteins containing HNH domains, protein clustering based on pairwise identity was used to remove false positives (Fig. S8*A*). Specifically, the Tigrfam model (TIGR01865) for the multidomain protein Csn1 from the Nmeni subtype potentially could pick up other proteins with the HNH domains (e.g., proteins belonging to the R-M system). Therefore, TIGR01865 profile hits with canonical Csn1 proteins were clustered further and visualized based on pairwise sequence identity using Clans (95) with the *P* value cutoff of 10^{-20} (-blastpath 'blastall -p blastn -W 7, -q 3, -r 1, -G 5, -E 2, -F F') (Fig. S8*A*). Most proteins from TIGR01865 (black dots in Fig. 7*A*) formed a single group. Some of the proteins from samples in the present study formed two adjacent groups (blue and green dots in Fig. 8*A*), whereas the other proteins (red dots in Fig. S8*A*) generally showed high sequence variance from each other. None of the sequences in green or blue had close homology to Csn1 in the NCBI database (all hits belonged to other HNH endonuclease domains and proteins from R-M systems) and therefore were removed. Many of the sequences represented as red dots in Fig. S8*A* have Csn1 as their best homology (by BlastP search against the NCBI NR database) and are considered candidate Csn1 proteins. Their positions in Fig. S8*A* indicate their variation from each other and from the canonical Csn1 proteins (black dots).

1. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic Crenarchaeota: Proposal for a third archaeal phylum, the Thaumarchaeota. *Nat Rev Microbiol* 6:245–252.
2. Taylor MW, Radax R, Steger D, Wagner M (2007) Sponge-associated microorganisms: Evolution, ecology, and biotechnological potential. *Microbiol Mol Biol Rev* 71:295–347.
3. Holmes B, Blanch H (2007) Genus-specific associations of marine sponges with group I crenarchaeotes. *Mar Biol* 150:759–772.
4. Margot A, Toril A, Puentes F (2002) Consistent association of crenarchaeal Archaea with sponges of the genus Axinella. *Mar Biol* 140:739–745.
5. Walker CB, et al. (2010) Nitrosopumilus maritimus genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea. *Proc Natl Acad Sci USA* 107:8818–8823.
6. Blainey PC, Mosier AC, Potanina A, Francis CA, Quake SR (2011) Genome of a low-salinity ammonia-oxidizing archaeon determined by single-cell and metagenomic analysis. *PLoS ONE* 6:e16626.
7. Webster NS, Watts JE, Hill RT (2001) Detection and phylogenetic analysis of novel crenarchaeote and euryarchaeote 16S ribosomal RNA gene sequences from a Great Barrier Reef sponge. *Mar Biotechnol (NY)* 3:600–608.
8. Steger D, et al. (2008) Diversity and mode of transmission of ammonia-oxidizing archaea in marine sponges. *Environ Microbiol* 10:1087–1094.
9. Thomas T, et al. (2010) Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J* 4:1557–1567.
10. Liu M, Fan L, Zhong L, Kjelleberg S, Thomas T (2012) Metaproteogenomic analysis of a community of sponge symbionts. *ISME J* 2, 10.1038/ismej.2012.1.
11. Turque AS, et al. (2010) Environmental shaping of sponge associated archaeal communities. *PLoS ONE* 5:e15774.
12. Pei AY, et al. (2010) Diversity of 16S rRNA genes within individual prokaryotic genomes. *Appl Environ Microbiol* 76:3886–3897.
13. Maslunka C, Carr E, Gürtler V, Kämpfer P, Seviour R (2006) Estimation of ribosomal RNA operon (rrn) copy number in Acinetobacter isolates and potential of patterns of rrn operon-containing fragments for typing strains of members of this genus. *Syst Appl Microbiol* 29:216–228.
14. Kaneko T, et al. (2000) Complete genome structure of the nitrogen-fixing symbiotic bacterium Mesorhizobium loti. *DNA Res* 7:331–338.
15. Labbé N, Parent S, Villemur R (2004) Nitratireductor aquibiodomus gen. nov., sp. nov., a novel alpha-proteobacterium from the marine denitrification system of the Montreal Biodome (Canada). *Int J Syst Evol Microbiol* 54:269–273.
16. Kim KH, et al. (2009) Nitratireductor basaltis sp. nov., isolated from black beach sand. *Int J Syst Evol Microbiol* 59:135–138.
17. Casciotti KL, Ward BB (2005) Phylogenetic analysis of nitric oxide reductase gene homologues from aerobic ammonia-oxidizing bacteria. *FEMS Microbiol Ecol* 52:197–205.
18. Garbeva P, Baggs EM, Prosser JI (2007) Phylogeny of nitrite reductase (nirK) and nitric oxide reductase (norB) genes from Nitrosospira species isolated from soil. *FEMS Microbiol Lett* 266:83–89.
19. Wrage N, Velthof GL, Beusichem MLV, Oenema O (2001) Role of nitrifier denitrification in the production of nitrous oxide. *Soil Biol Biochem* 33(12–13): 1723–1732.
20. Kirstein K, Bock E (1993) Close genetic relationship between Nitrobacter hamburgensis nitrite oxidoreductase and Escherichia coli nitrate reductases. *Arch Microbiol* 160:447–453.
21. Schmid MC, et al. (2008) Environmental detection of octahaem cytochrome c hydroxylamine/hydrazine oxidoreductase genes of aerobic and anaerobic ammonium-oxidizing bacteria. *Environ Microbiol* 10:3140–3149.
22. Belanger AE, Hatfull GF (1999) Exponential-phase glycogen recycling is essential for growth of Mycobacterium smegmatis. *J Bacteriol* 181:6670–6678.
23. Harper C, Hayward D, Wiid I, van Helden P (2008) Regulation of nitrogen metabolism in Mycobacterium tuberculosis: A comparison with mechanisms in Corynebacterium glutamicum and Streptomyces coelicolor. *IUBMB Life* 60:643–650.
24. Yahel G, Sharp JH, Marie D, Häse C, Genin A (2003) In situ feeding and element removal in the symbiont-bearing sponge Theonella swinhoei: Bulk DOC is the major source for carbon. *Limnol Oceanogr* 48(1):141–149.
25. Steindler L, Beer S, Ilan M (2002) Photosymbiosis in intertidal and subtidal tropical sponges. *Symbiosis* 33(3):263–273.
26. Wilkinson CR (1983) Net primary productivity in coral reef sponges. *Science* 219: 410–412.
27. Stark M, Berger SA, Stamatakis A, von Mering C (2010) MLTreeMap—accurate Maximum Likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* 11:461.
28. Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* 281:200–207.
29. Ashida H, et al. (2003) A functional link between RuBisCO-like protein of Bacillus and photosynthetic RuBisCO. *Science* 302:286–290.
30. Bannister RJ, et al. (2011) Incongruence between the distribution of a common coral reef sponge and photosynthesis. *Mar Ecol Prog Ser* 423:95–100.
31. Thompson JE, Murphy PT, Berquist PR, Evans EA (1987) Environmentally induced variation in diterpene composition of the marine sponge Rhopaloeides odorabile. *Biochem Syst Ecol* 15(5):595–606.
32. Roberts DE, Cummins SP, Davis AR, Pangway C (1999) Evidence for symbiotic algae in sponges from temperate coastal reefs in New South Wales, Australia. *Mem Queensl Mus* 44:493–498.
33. Taylor MW, Schupp PJ, Dahllöf I, Kjelleberg S, Steinberg PD (2004) Host specificity in marine sponge-associated bacteria, and potential implications for marine microbial diversity. *Environ Microbiol* 6:121–130.
34. Sara M (1971) Ultrastructural aspects of the symbiosis between two species of the genus Aphanocapsa (Cyanophyceae) and Ircinia variabilis (Demospongiae). *Mar Biol* 11(3):214–221.
35. Szurmant H, White RA, Hoch JA (2007) Sensor complexes regulating two-component signal transduction. *Curr Opin Struct Biol* 17:706–715.

36. Hazelbauer GL, Falke JJ, Parkinson JS (2008) Bacterial chemoreceptors: High-performance signaling in networked arrays. *Trends Biochem Sci* 33:9–19.
37. Mauro LJ, Dixon JE (1994) 'Zip codes' direct intracellular protein tyrosine phosphatases to the correct cellular 'address'. *Trends Biochem Sci* 19:151–155.
38. Grangeasse C, Cozzone AJ, Deutscher J, Mijakovic I (2007) Tyrosine phosphorylation: An emerging regulatory device of bacterial physiology. *Trends Biochem Sci* 32:86–94.
39. Thomasson B, et al. (2002) MglA, a small GTPase, interacts with a tyrosine kinase to control type IV pili-mediated motility and development of Myxococcus xanthus. *Mol Microbiol* 46:1399–1413.
40. Zhao X, Lam JS (2002) WaaP of Pseudomonas aeruginosa is a novel eukaryotic type protein-tyrosine kinase as well as a sugar kinase essential for the biosynthesis of core lipopolysaccharide. *J Biol Chem* 277:4722–4730.
41. Aarts MG, et al. (1997) The Arabidopsis MALE STERILITY 2 protein shares similarity with reductases in elongation/condensation complexes. *Plant J* 12:615–623.
42. Arbeitman MN, Fleming AA, Siegal ML, Null BH, Baker BS (2004) A genomic analysis of Drosophila somatic sexual differentiation and its regulation. *Development* 131: 2007–2021.
43. Curtis PD, Geyer R, White DC, Shimkets LJ (2006) Novel lipids in Myxococcus xanthus and their role in chemotaxis. *Environ Microbiol* 8:1935–1949.
44. Vacelet J, Donadey C (1977) Electron microscope study of the association between some sponges and bacteria. *J Exp Mar Biol Ecol* 30(3):301–314.
45. Owens RM, et al. (2004) A dedicated translation factor controls the synthesis of the global regulator Fis. *EMBO J* 23:3375–3385.
46. Grant AJ, et al. (2003) Co-ordination of pathogenicity island expression by the BipA GTPase in enteropathogenic Escherichia coli (EPEC). *Mol Microbiol* 48:507–521.
47. Farris M, Grant A, Richardson TB, O'Connor CD (1998) BipA: A tyrosine-phosphorylated GTPase that mediates interactions between enteropathogenic Escherichia coli (EPEC) and epithelial cells. *Mol Microbiol* 28:265–279.
48. Barker HC, Kinsella N, Jaspe A, Friedrich T, O'Connor CD (2000) Formate protects stationary-phase Escherichia coli and Salmonella cells from killing by a cationic antimicrobial peptide. *Mol Microbiol* 35:1518–1529.
49. Reva ON, et al. (2006) Functional genomics of stress response in Pseudomonas putida KT2440. *J Bacteriol* 188:4079–4092.
50. Kiss E, Huguet T, Poinsot V, Batut J (2004) The typA gene is required for stress adaptation as well as for symbiosis of Sinorhizobium meliloti 1021 with certain Medicago truncatula lines. *Mol Plant Microbe Interact* 17:235–244.
51. Hallam SJ, et al. (2006) Genomic analysis of the uncultivated marine crenarchaeote Cenarchaeum symbiosum. *Proc Natl Acad Sci USA* 103:18296–18301.
52. Tarahovsky YS, Ivanitsky GR, Khusainov AA (1994) Lysis of Escherichia coli cells induced by bacteriophage T4. *FEMS Microbiol Lett* 122:195–199.
53. Sharon I, et al. (2009) Photosystem I gene cassettes are present in marine virus genomes. *Nature* 461:258–262.
54. Short CM, Suttle CA (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Appl Environ Microbiol* 71:480–486.
55. Breitbart M, Rohwer F (2005) Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* 13:278–284.
56. Yarmolinsky MB (1995) Programmed cell death in bacterial populations. *Science* 267: 836–837.
57. Garcia-Pino A, et al. (2008) Doc of prophage P1 is inhibited by its antitoxin partner Phd through fold complementation. *J Biol Chem* 283:30821–30827.
58. Gerdes K, Rasmussen PB, Molin S (1986) Unique type of plasmid maintenance function: Postsegregational killing of plasmid-free cells. *Proc Natl Acad Sci USA* 83:3116–3120.
59. Cooper TF, Paixão T, Heinemann JA (2010) Within-host competition selects for plasmid-encoded toxin-antitoxin systems. *Proc Biol Sci* 277:3149–3155.
60. Tyson GW, Banfield JF (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* 10:200–207.
61. Thurber RV (2009) Current insights into phage biodiversity and biogeography. *Curr Opin Microbiol* 12:582–587.
62. Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. *BMC Bioinformatics* 11:187.
63. Bengtsson J, et al. (2011) Metaxa: A software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek* 100:471–475.
64. Pruesse E, et al. (2007) SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 35:7188–7196.
65. Ludwig W, et al. (2004) ARB: A software environment for sequence data. *Nucleic Acids Res* 32:1363–1371.
66. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20:289–290.
67. Mavromatis K, et al. (2007) Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* 4:495–500.
68. McElroy K, Luciani F, Hui J, Rice S, Thomas T (2011) Bacteriophage evolution drives Pseudomonas aeruginosa PAO1 biofilm diversification. *BMC Bioinformatics* 12(Suppl 12):A2.
69. Fan C, McElroy K, Thomas T Reconstruction of ribosomal RNA genes from metagenomic data. *PLoS One*, in press.
70. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7:335–336.
71. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
72. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552.
73. Lozupone C, Lladser ME, Knights D, Stombaugh J, Knight R (2011) UniFrac: An effective distance metric for microbial community comparison. *ISME J* 5:169–172.
74. Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. *Genome Res* 17:377–386.
75. Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: Detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* 15:387–396.
76. Brady A, Salzberg SL (2009) Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 6:673–676.
77. Tatusov RL, et al. (2003) The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* 4:41.
78. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38 (Database issue):D211–D222.
79. Finn RD, Clements J, Eddy SR (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res* 39(Web Server issue):W29-37.
80. Overbeek R, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702.
81. Meyer F, et al. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 9:386.
82. Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ (2010) Average genome size: A potential source of bias in comparative metagenomics. *ISME J* 4:1075–1077.
83. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P (2007) Prediction of effective genome size in metagenomic samples. *Genome Biol* 8:R10.
84. Angly FE, et al. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLOS Comput Biol* 5:e1000593.
85. Frank JA, Sørensen SJ (2011) Quantitative metagenomic analyses based on average genome size normalization. *Appl Environ Microbiol* 77:2513–2521.
86. Ciccarelli FD, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.
87. White JR, Nagarajan N, Pop M (2009) Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLOS Comput Biol* 5:e1000352.
88. de Hoon MJ, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics* 20:1453–1454.
89. Saldanha AJ (2004) Java Treeview—extensible visualization of microarray data. *Bioinformatics* 20:3246–3248.
90. Jehl MA, Arnold R, Rattei T (2011) Effective—a database of predicted secreted bacterial proteins. *Nucleic Acids Res* 39(Database issue):D591–D595.
91. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
92. Larkin MA, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.
93. Wang G, Asakawa S, Kimura M (2011) Spatial and temporal changes of cyanophage communities in paddy field soils as revealed by the capsid assembly protein gene g20. *FEMS Microbiol Ecol* 76:352–359.
94. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659.
95. Frickey T, Lupas A (2004) CLANS: A Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 20:3702–3704.
96. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
97. Price MN, Dehal PS, Arkin AP (2009) FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26:1641–1650.
98. Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35(Web Server issue):W52-7.
99. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A community resource for metagenomics. *PLoS Biol* 5:e75.
100. Shannon P, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504.
101. Haft DH, Selengut J, Mongodin EF, Nelson KE (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLOS Comput Biol* 1:e60.

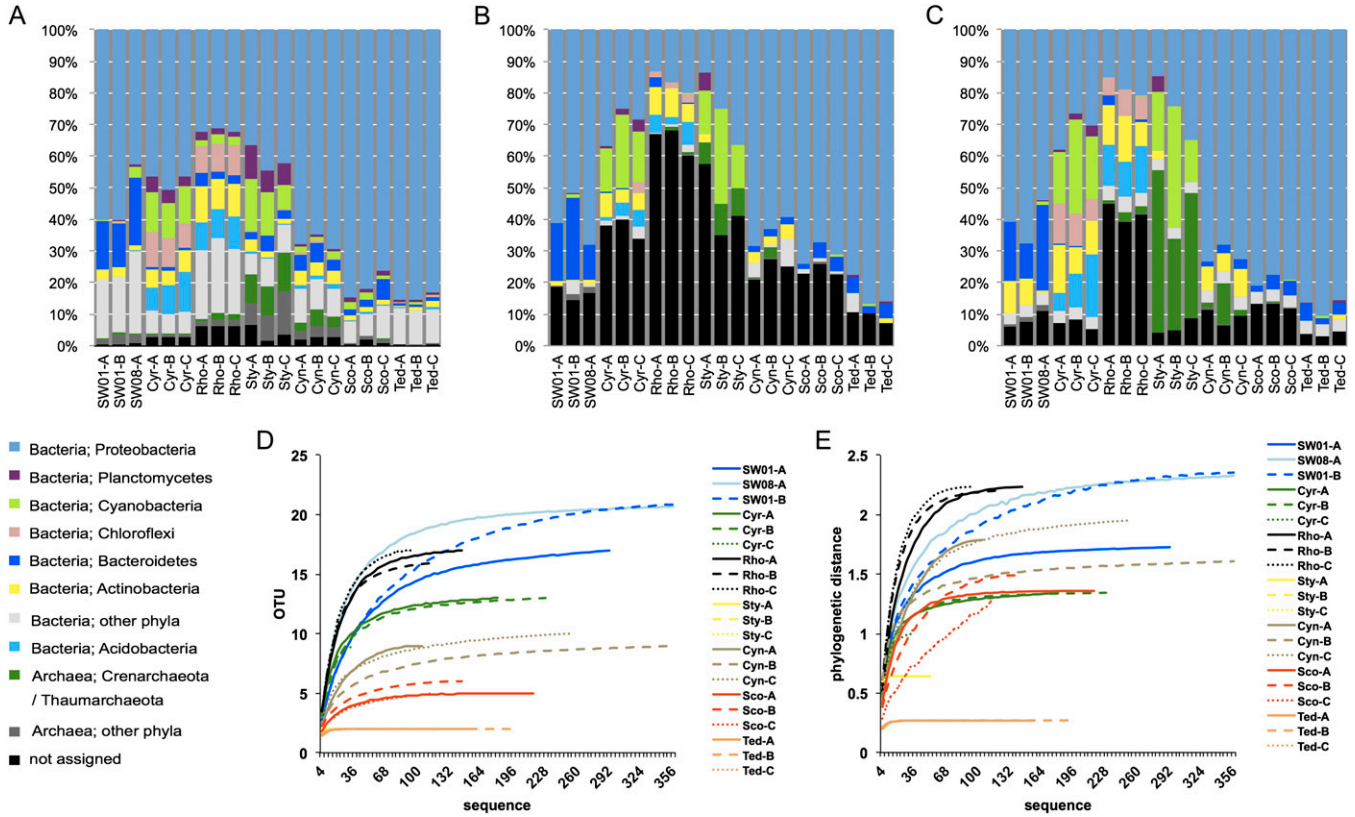

Fig. S1. Community composition at phylum level and community diversity. Species richness is based on 16S rRNA gene OTUs reconstructed from metagenomic data. Rarefaction curves are generated with means of 1,000 rounds of Jackknife subsampling. (*A*) Classification based on SCGs. (*B*) Classification by 16S rRNA gene sequences without assembly (80% confidence). (*C*) Classification by 16S rRNA gene OTUs constructed in the present study together with unassembled 16S rRNA gene sequences (80% confidence). (*D*) Species richness calculation based on observed OTU number. (*E*) Species richness calculation based on phylogenetic distance.

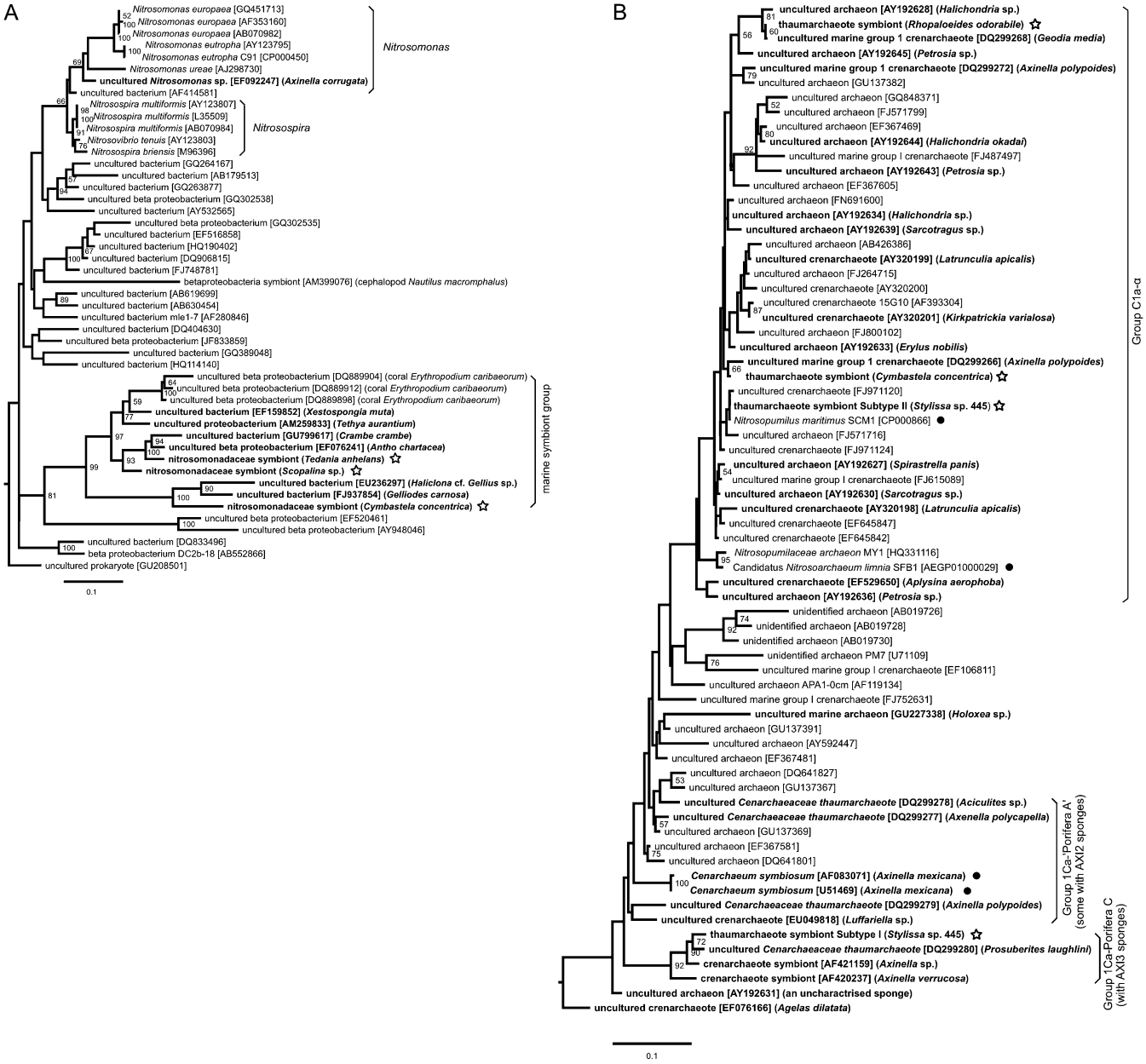

Fig. S2. The 16S rRNA gene maximum-likelihood tree of Nitrosomonadaceae and Marine Group 1 Thaumarchaeota. Percentage bootstrapping values (1,000 replications) greater than 50% are shown. Sponge-derived sequences are shown in bold. Pentagram-marked sequences are from the present study. (*A*) Tree of Nitrosomonadaceae. The tree is rooted to *Petrobacter succinatimandens* (AY219713). (*B*) Tree of Marine Group 1 Thaumarchaeota. The tree is rooted to *Thermofilum pendens* (X14835). Groups were named according to Holmes and Blanch (3). Species/strains marked by a solid circle have complete/draft genomes available.

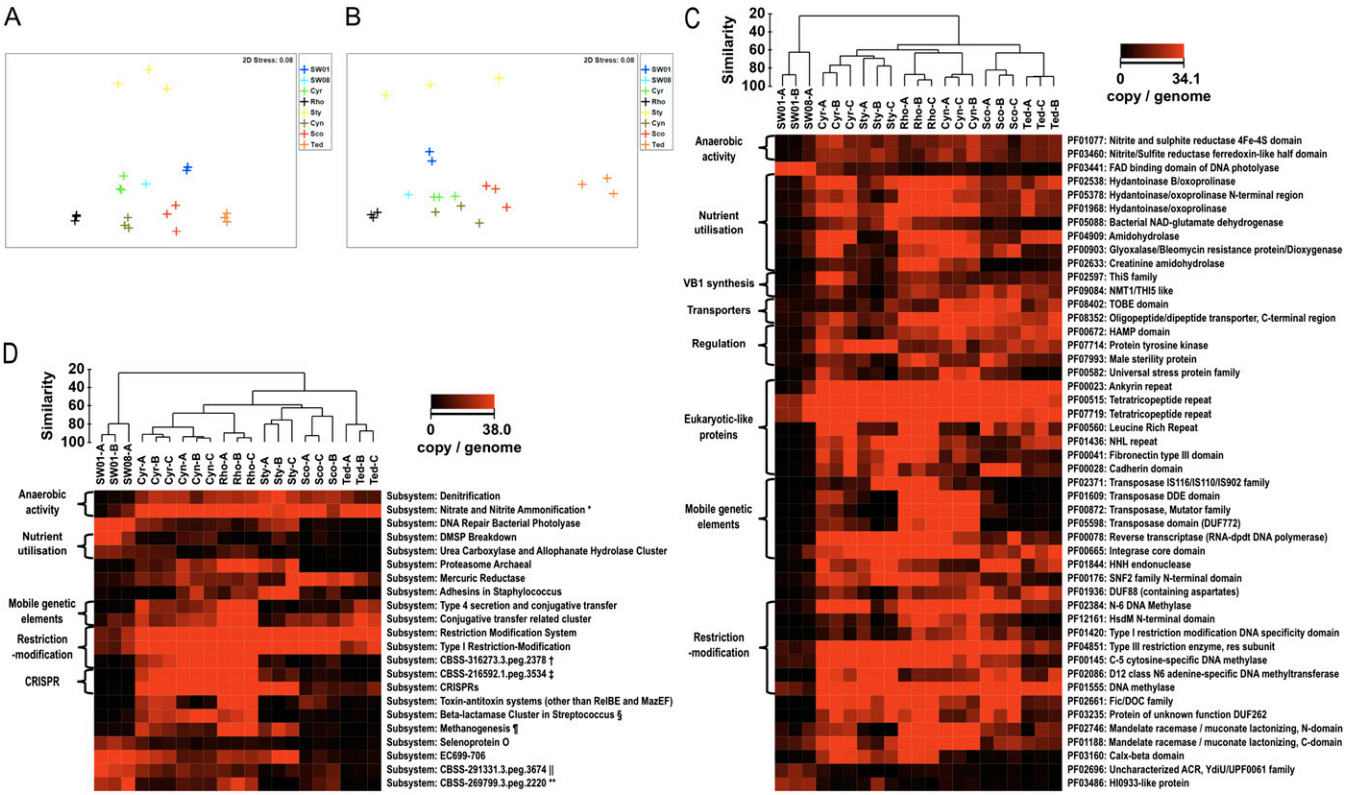

Fig. S3. Multidimensional scaling (MDS) plots of samples by Bray–Curtis similarity and sponge/seawater sample specific functions according to Pfam and Subsystem functional annotation. (*A*) Sample MDS plots by Pfam annotation. (*B*) Sample MDS plots by Subsystem annotation. (*C*) Abundance of sponge- or seawater-specific functions by Pfam annotation. (*D*) Abundance of sponge- or seawater-specific functions by Subsystem annotation. The brightness (red) in the heatmap reflects abundance (copies per genome) of a particular function in a sample. Samples are clustered by Bray–Curtis similarity and average general algorithm. *Mostly nitrate reductase for respiration. †R-M system component Yee. ‡CAS protein Cas1. §Mostly RecD-like DNA helicase YrrC. ¶Mostly F420-dependent n(5)n(10) methylenetetrahydromethanopterin reductase (EC 1.5.99.11). ||Unknown function. **Unknown function.

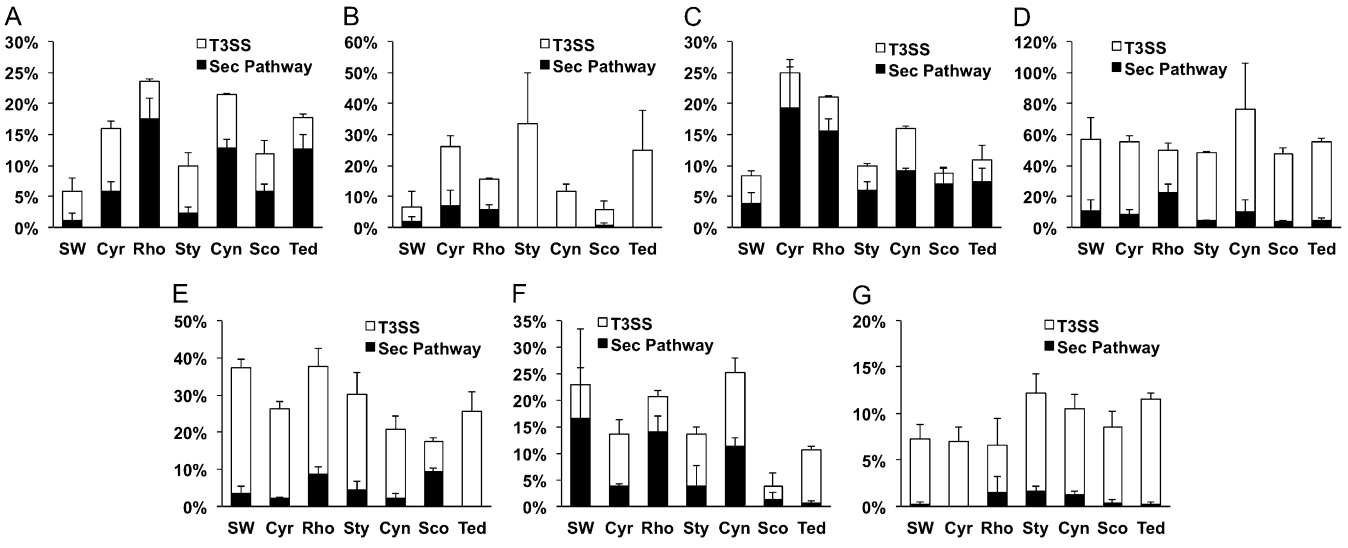

Fig. S4. T3SS and Sec pathway secretion prediction of ELPs. SDs are shown. (*A*) ANK. (*B*) LRR. (*C*) TPR. (*D*) Fn3. (*E*) Cadherin. (*F*) NHL. (*G*) PTK.

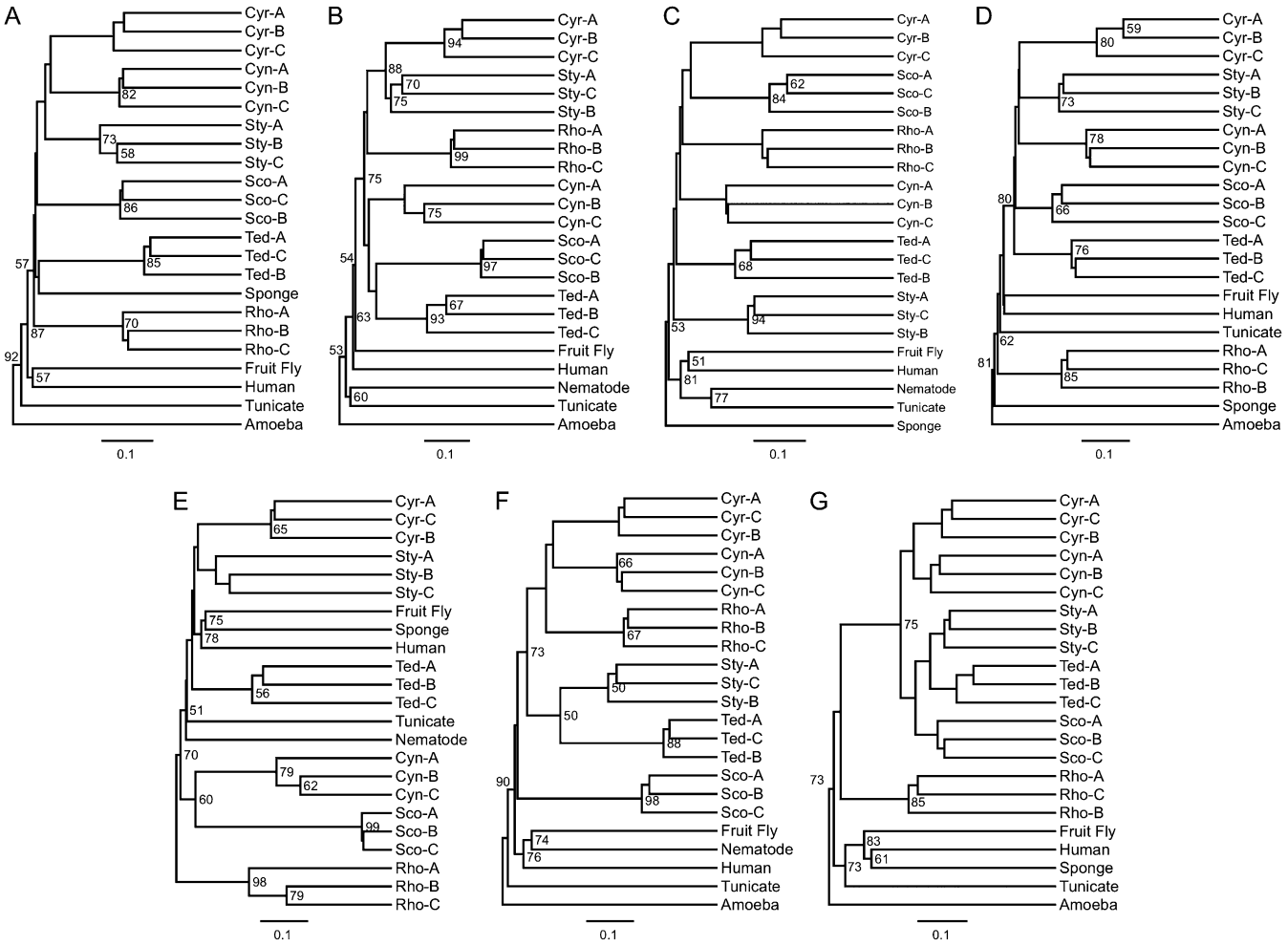

Fig. S5. Sample clustering based on ELP sequence similarity. Samples are clustered using a weighted Unifrac algorithm. Supporting values (in percentage) greater than 50% of 1,000 replications of Jackknife subsampling are marked. (*A*) ANK. (*B*) LRR. (*C*) TPR. (*D*) Fn3. (*E*) Cadherin. (*F*) NHL. (*G*) PTK.

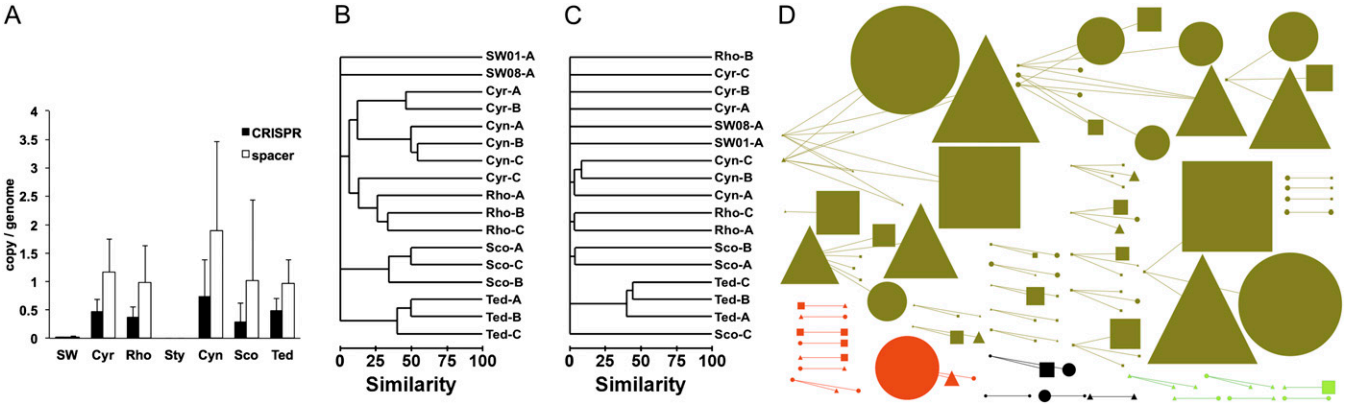

Fig. S6. CRISPR abundance and local specificity. (*A*) Abundance of CRISPR loci and spacers. SDs are shown. (*B*) Sample clustering by CRISPR repeats. Clustering is based on Bray–Curtis similarity (presence/absence) by average general algorithm. (*C*) Sample clustering by CRISPR spacers. (*D*) Local specificity and abundance of CRISPR and their potential targets. Edges indicate the connection between the spacers and their potential targets. The object on the left side of an edge represents the spacer and the one on the right side represents the target matched by this spacer. The size of objects indicates their abundance in samples (from 0.0087 to 0.443 copy per genome). Circles represent replicate A, triangles represent replicate B, and rectangles represent replicate C. *C. concentrica* samples are shown in olive, *Scopalina* sp. samples in red, *R. odorabile* samples in black, and *C. coralliophila* samples in lime.

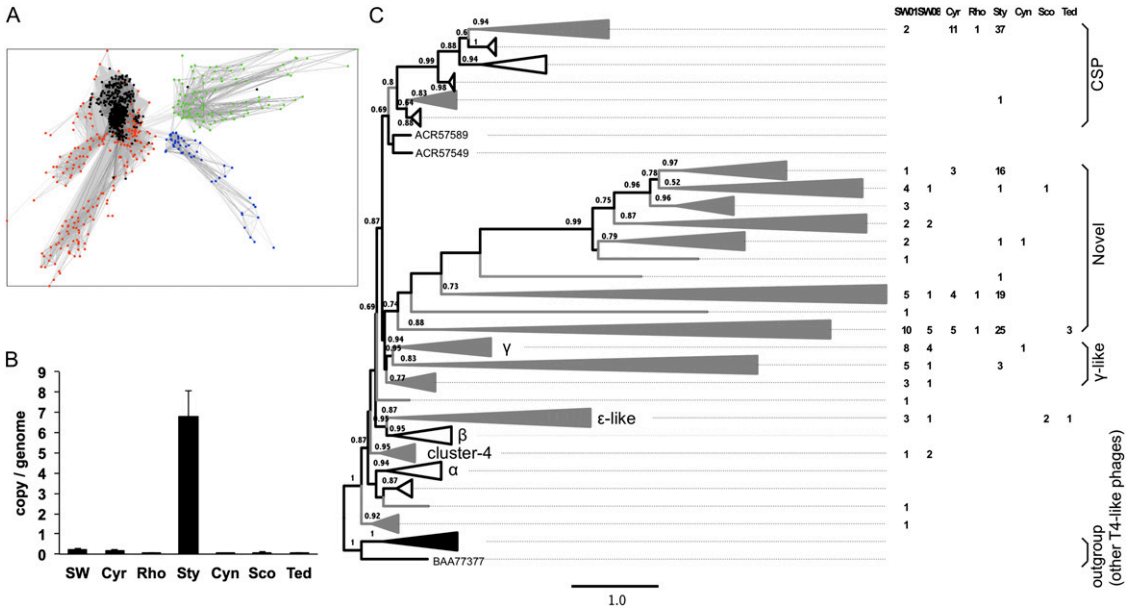

Fig. S7. Cyanophage abundance based on G20 protein analysis. (*A*) Clustering of proteins found by searching cyanophage G20 proteins. Black dots represent canonical cyanophage G20 reference sequences; red dots represent candidate cyanophage G20 proteins in the present study; green dots and blue dots represent false positives (portal vertex proteins). (*B*) Abundance of cyanophage based on G20 protein number. SDs are shown. (*C*) T4-like phage populations. The approximately maximum-likelihood tree of T4 phage G20 proteins was constructed and supported by 1,000 rounds of FastTree local support values (values >0.5 are marked). The tree is rooted to the outgroup consisting of noncyanophage T4-like phages. Each grouped clade in gray contains at least one protein detected across samples. The number of proteins detected in each sample type is shown. Clades are named according to Wang et al. (92).

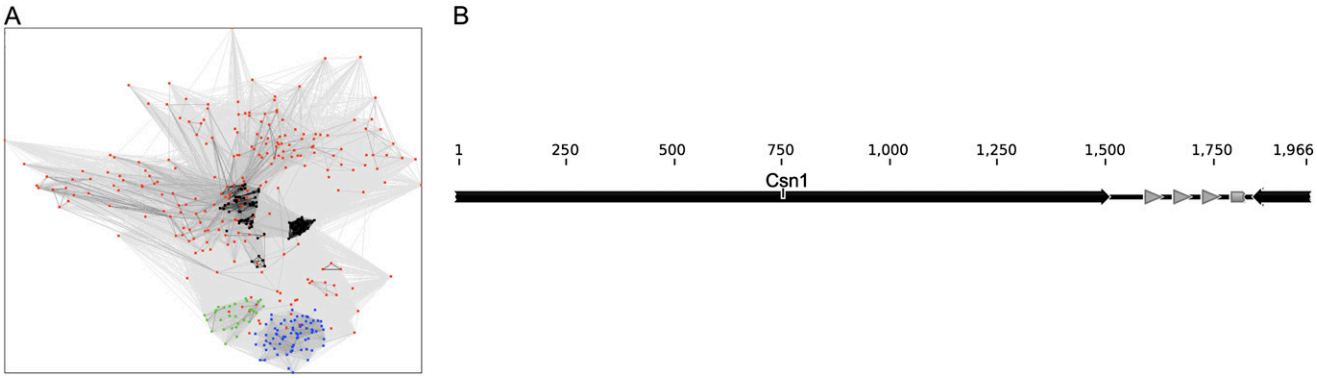

Fig. S8. Csn1 analysis. (*A*) Clustering of Csn1 and related proteins (TIGR01865). Black dots represent canonical Csn1 reference sequences; red dots represent candidate Csn1 proteins found in sponge samples; green dots and blue dots represent false positives (HNH endonuclease domain-containing proteins and proteins from R-M systems). (*B*) A Csn1 arrangement in a CRISPR cassette. Contig layout was generated using Geneious 4.86 (http://www.geneious.com). Gray arrows indicate the CRISPR array.

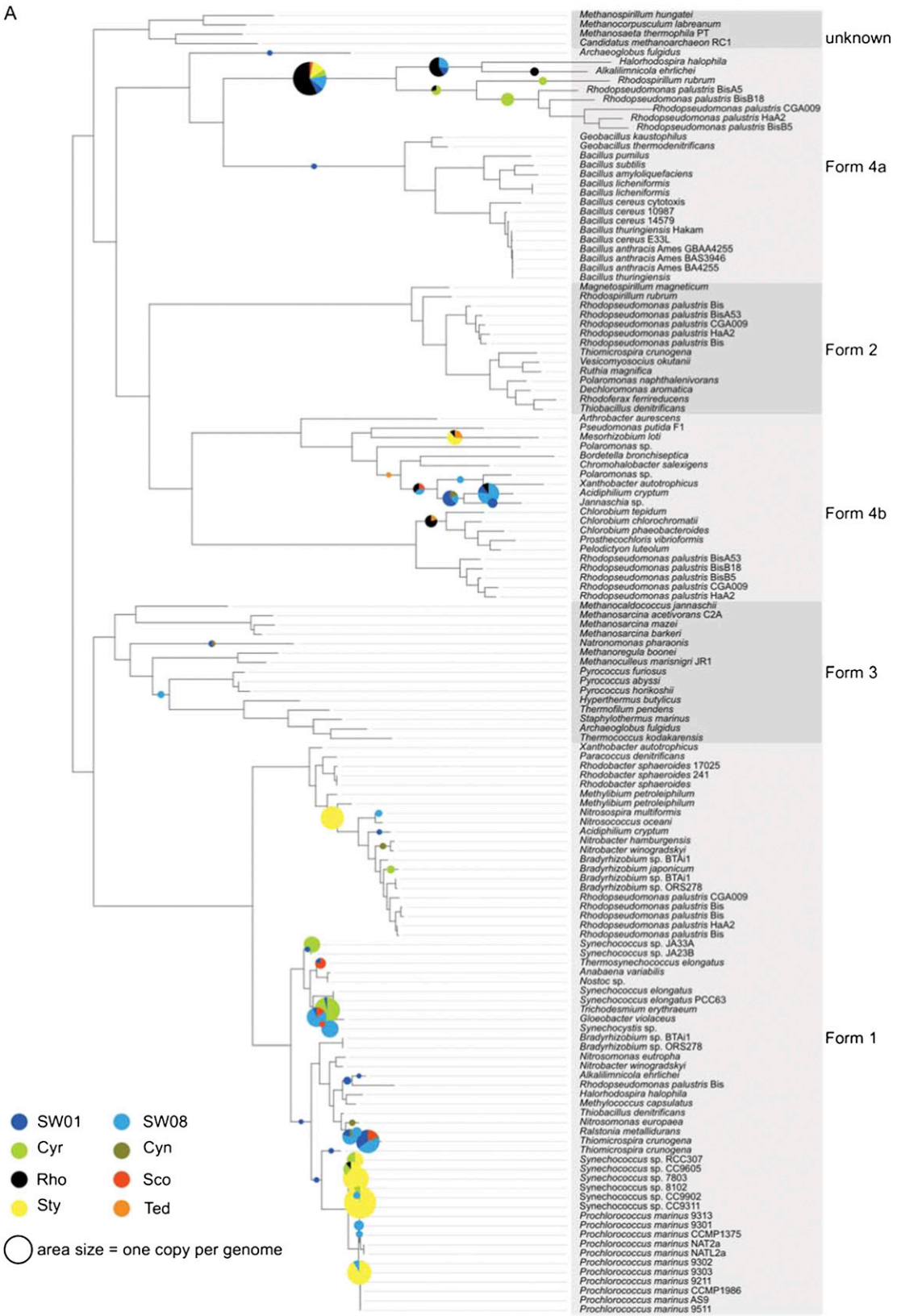

Fig. S9. (Continued)

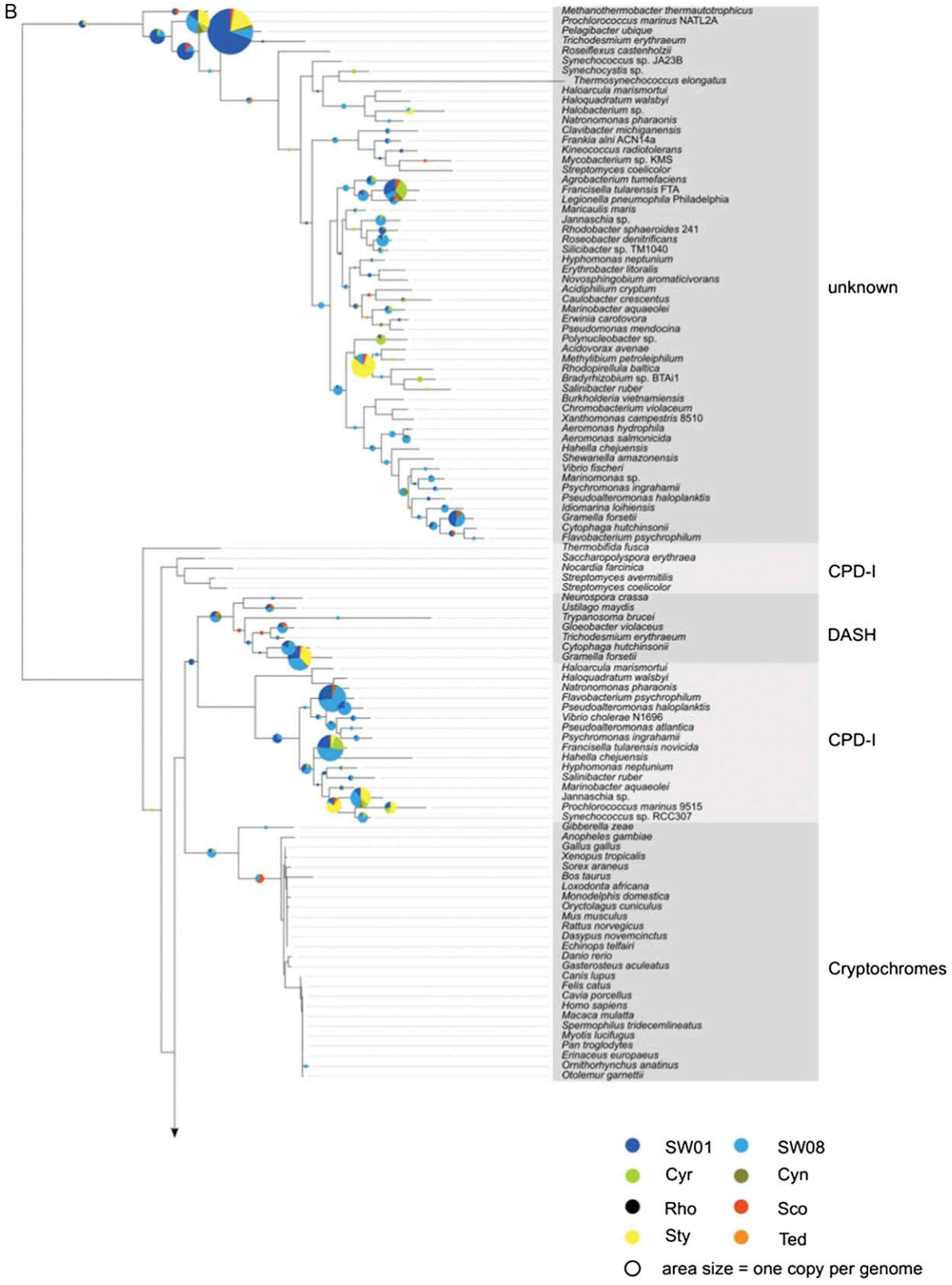

Fig. S9. (Continued)

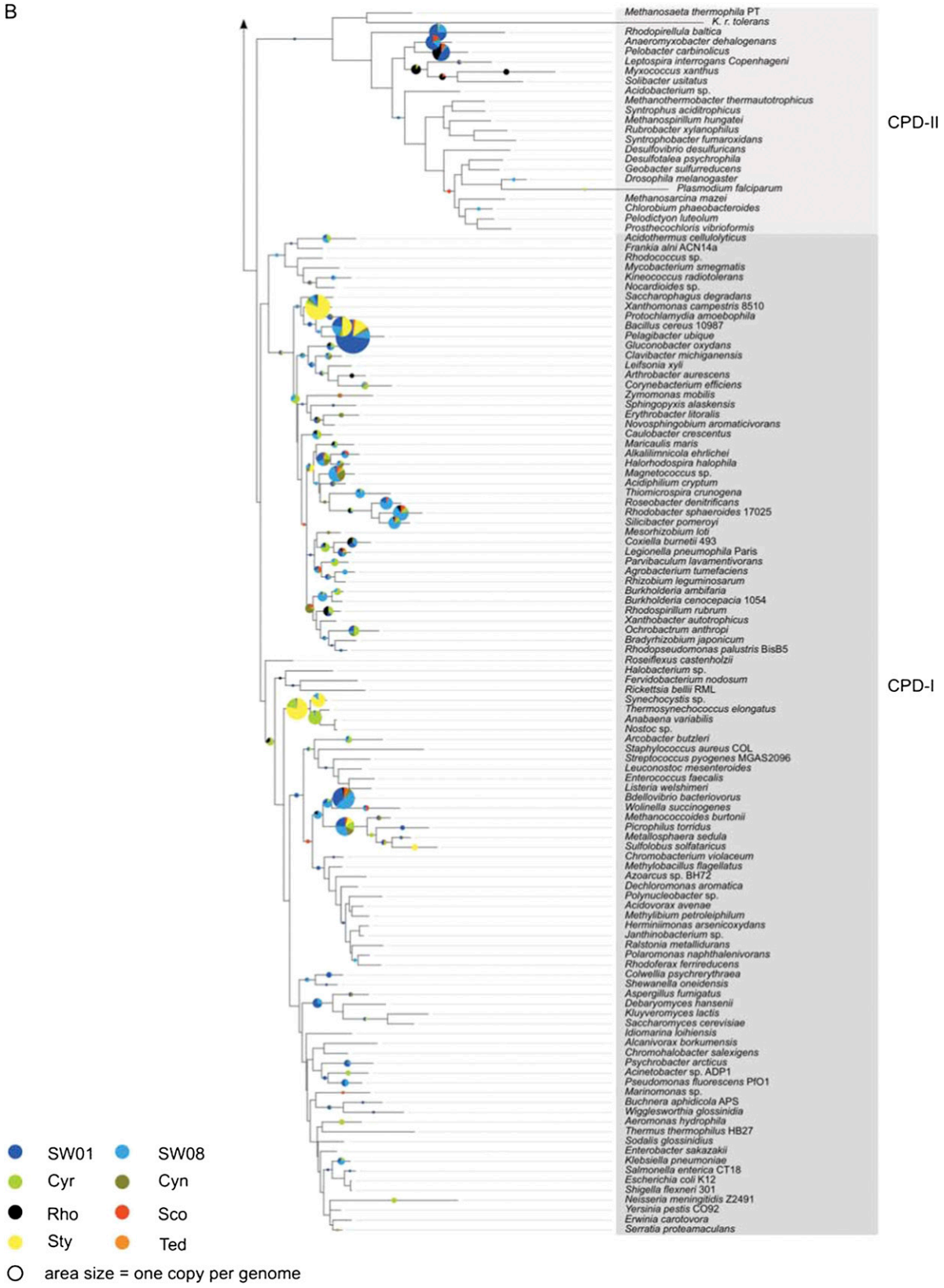

Fig. S9. Abundance and diversity of Rubisco/Rubisco-like proteins (*A*) and proteins in the photolyase/cryptochrome family (*B*) with taxonomic annotation. Size of the circles reflects abundance in gene copy per genome.

Table S1. Sample information in read processing, 16S rRNA gene-containing sequences, assembly, decontamination, and functional annotation

Sample	SW01-A	SW01-B	SW08-A	Cyr-A	Cyr-B	Cyr-C	Rho-A	Rho-B	Rho-C	Sty-A	Sty-B	Sty-C	Cyn-A	Cyn-B	Cyn-C	Sco-A	Sco-B	Sco-C	Ted-A	Ted-B	Ted-C
Dataset ID	BBAY01	BBAY02	BBAY01SM	BBAY31	BBAY32	BBAY33	BBAY34	BBAY35	BBAY36	BBAY37	BBAY38	BBAY39	BBAY40	BBAY41	BBAY42	BBAY43	BBAY44	BBAY45	BBAY62	BBAY63	BBAY64
Raw reads	580,055	746,735	1,073,717	897,408	971,976	888,127	949,133	583,576	519,285	626,753	464,320	514,969	678,263	1,169,872	1,323.699	647,813	1,015,654	609,105	59,168	611,468	363,686
Avg read size (nt)	354.9	338.1	372.5	387.6	353.2	276.8	402.0	377.3	401.6	319.5	275.0	285.5	358.0	408.1	392.8	349.0	264.3	350.7	356.9	310.1	302.7
Unique reads	572,731	737,462	1,028,840	859,525	898,161	788,662	904,146	505,357	506,761	590,822	436,017	457,425	660,869	1,004,075	1,111,093	629,438	908,792	499,316	472,310	575,079	331,761
16S rRNA gene -containing reads	450	651	855	244	306	56	298	198	201	73	20	22	188	467	350	277	208	184	211	232	128
Constructed 16S seqs	22	28	27	13	13	9	19	18	19	3	3	3	11	11	12	5	6	5	4	4	4
Reads in 16S seqs	294	440	508	180	228	37	147	120	98	52	14	12	107	368	252	218	144	118	157	196	103
Min length of 16S rRNA gene seqs (nt)	435	435	435	1,218	1,218	1,218	781	781	781	620	620	620	493	493	493	666	666	861	822	822	822
Max length of 16S rRNA gene seqs (nt)	1,541	1,541	1,541	1,535	1,535	1,535	1,542	1,542	1,542	1,459	1,459	1,459	1,516	1,517	1,517	1,525	1,539	1,539	1,518	1,518	1,518
Avg length of 16S rRNA gene seqs (nt)	1,285	1,251	1,269	1,418	1,418	1,416	1,308	1,307	1,308	1,042	1,042	1,042	1,227	1,264	1,251	1,207	1,263	1,382	1,194	1,194	11,94
Aligned reads	183,126	229,604	336,186	434,407	471,683	259,531	460,513	198,840	180,893	212,513	142,256	169,357	347,438	704,702	679,422	394,585	507,543	239,533	316,595	369,205	182,101
%	32.0	31.1	32.7	50.5	52.5	32.9	50.9	39.4	35.7	36.7	33.2	37.5	53.8	72.0	63.3	62.7	55.9	48.0	67.0	64.2	54.9
Contigs > 500 nt	8,117	11,058	23,064	23,889	18,203	14,309	44,408	17,286	19,306	7,117	2,124	3,023	11,417	11,398	26,612	6,922	8,902	8,636	4,356	3,622	2,404
Average size of contigs >500 nt (nt)	918	890	1,038	1,022	1,285	710	902	900	836	941	711	742	1,185	1,459	1,119	1,160	1,215	1,030	1,005	1,095	1,265
N50 size of contigs >500 nt (nt)	859	843	1,050	993	1,614	682	909	909	838	894	674	713	1,345	2,262	1,192	1,279	1,461	1,046	1,030	1,200	1,598
Maximum size of contigs (nt)	59,504	56,677	33,749	101,234	107,941	9,624	52,481	16,414	16,419	15,909	7,128	7,136	28,780	323,086	39,109	21,389	28,627	17,816	15,827	14,797	17,891
Contigs >100 nt	16,740	22,129	40,022	54,483	39,079	48,688	71,479	29,504	32,819	23,857	14,243	16,470	22,865	22,168	56,410	19,077	22,936	22,414	11,737	11,167	8,027
Singletons >100 nt	362,835	466,957	656,384	393,881	372,735	405,414	416,488	284,341	296,611	329,277	248,978	251,623	275,175	251,668	360,031	203,477	254,563	233,729	131,480	148,917	115,509
Prokaryotic-originned contigs and singletons	312,189	401,518	541,809	306,688	296,277	277,660	458,148	290,530	303,412	211,574	147,447	149,459	212,117	200,120	275,351	148,431	174,793	166,156	92,981	107,452	80,526
%	82.2	82.1	77.8	68.4	71.9	61.1	93.9	92.6	92.1	59.9	56.0	55.7	71.2	73.1	66.1	66.7	63.0	64.9	64.9	67.1	65.2
Unique proteins	397,993	498,626	627,609	320,969	312,616	232,380	566,378	345,825	369,637	184,763	121,107	115,832	215,220	229,113	281,686	152,849	176,185	175,117	92,709	109,457	78,864
Total proteins	503,724	625,286	778,971	488,668	475,089	282,226	787,435	427,733	449,663	235,211	140,259	140,770	342,235	550,559	555,480	343,903	323,773	262,430	26,009	255,111	144,937
Proteins annotated by COG (1e-10)	147,095	175,547	225,095	110,188	120,192	34,505	224,936	116,837	119,903	20,576	5,998	6,397	103,220	223,797	180,529	99,959	96,012	56,501	93,834	85,932	41,350
%	29.2	28.1	28.9	22.5	25.3	12.2	28.6	27.3	26.7	8.7	4.3	4.5	30.2	40.6	32.5	29.1	29.7	21.5	36.1	33.7	28.5
Proteins annotated by Pfam (1e-10)	135,897	160,948	220,514	123,056	138,496	35,542	229,853	119,095	121,332	24,034	6,584	6,595	113,803	260,250	206,000	112,534	114,479	61,817	105,494	100,697	48,512
%	27.0	25.7	28.3	25.2	29.2	12.6	29.2	27.8	27.0	10.2	4.7	4.7	33.3	47.3	37.1	32.7	35.4	23.6	40.6	39.5	33.5
Proteins annotated by SEED (1e-10)	313,252	369,728	415,554	171,709	183,220	56,199	312,420	162,140	168,854	36,410	11,874	11,798	156,775	319,211	262,973	154,869	150,679	93,418	142,113	127,306	61,999
%	62.2	59.1	53.3	35.1	38.6	19.9	39.7	37.9	37.6	15.5	8.5	8.4	45.8	58.0	47.3	45.0	46.5	35.6	54.7	49.9	42.8