LBNL-62928
Preprint

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

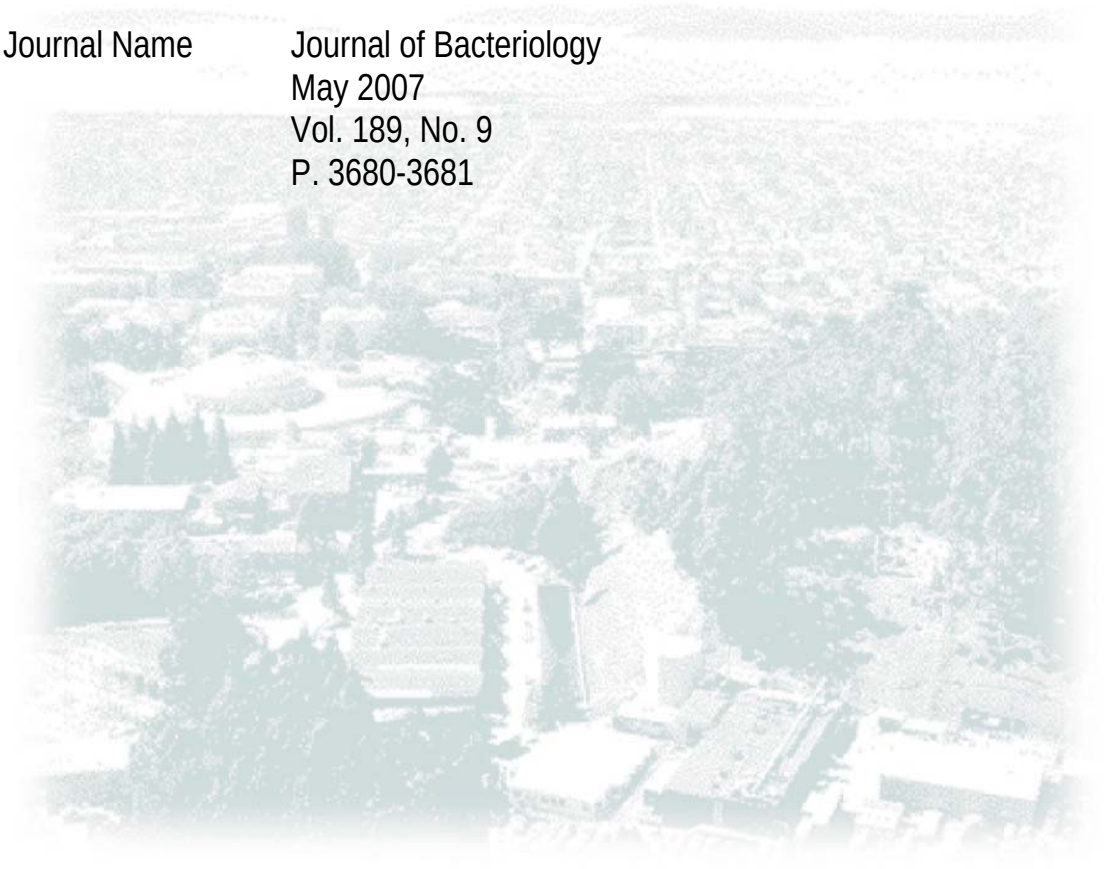
Title

The Complete Genome Sequence of *Bacillus thuringiensis* Al Hakam

Author(s), Jean F. Challacombe, Michael R. Altherr, et al

Division Genomics

Journal Name Journal of Bacteriology
May 2007
Vol. 189, No. 9
P. 3680-3681



Genomic analysis of the symbiotic marine crenarchaeon, *Cenarchaeum symbiosum*

Steven J. Hallam[1, 2], Konstantinos T. Konstantinidis[1], Celine Brochier[3], Nik Putnam[4], Christa Schleper[5], Yoh-ichi Watanabe[6], Junichi Sugahara[7], Christina Preston[8], José de la Torre[9], Paul M. Richardson[4], and Edward F. DeLong [1*]

[1]*Massachusetts Institute of Technology, Cambridge MA 02139, USA*

[2] *Current Address: University of British Columbia, Vancouver BC V6T 1Z3, Canada*

[3] *Evolution Génome Environnement, Université de Provence Aix-Marseille I, Cedex 3, France*

[4] *Joint Genome Institute, Walnut Creek CA 94598, USA*

[5] *Department of Biology, University of Bergen, Jahnebakken 5, N-5020 Bergen, Norway*

[6] *Department of Biomedical Chemistry, University of Tokyo, Tokyo, 113-0033, Japan*

[7] *Institute for Advanced Biosciences, Keio University, Tsuruoka, 997-0017, Japan*

[6] *Monterey Bay Aquarium Research Institute, Moss Landing CA 95069, USA*

[7] *University of Washington, Seattle WA 98195 USA*

Pages: Figures: 4, 5 supplementary
Tables: 1, 8 supplementary

Abstract: 246 words, 1611 characters
Text: 9,140 words, 52,515 characters

Running Title:

Cenarchaeum symbiosum Genome Sequence

* To whom correspondence should be addressed. E-mail: delong@mit.edu

Summary

Crenarchaea **are ubiquitous and abundant microbial constituents of soils, sediments, lakes and ocean waters, yet relatively little is known about their fundamental evolutionary, ecological, and physiological properties. To better describe the ubiquitous non-thermophilic** ***Crenarchaea,*** **we analyzed the genome sequence of one representative, the uncultivated sponge symbiont,** ***Cenarchaeum symbiosum.*** ***C. symbiosum*** **genotypes co-inhabiting the same host partitioned into two dominant populations, corresponding to previously described a- and b-type ribosomal RNA variants. Although syntenic, overlapping a- and b-type ribotypes harbored significant genetic variability. A single tiling path comprising the dominant a-type genotype was assembled, and used to explore the biological properties of C.** ***symbiosum*** **and its planktonic relatives. Out of a total of 2,066 predicted open reading frames, 36% were more highly conserved with other** ***Archaea*****. The remainder partitioned between bacteria (18%), eukaryotes (1.5%) and viruses (0.1%). A total of 525 open reading frames were more highly conserved with sequences derived from marine environmental genomic surveys, most probably representing orthologous genes found in free-living planktonic** ***Crenarchaea.*** **The remaining genes partitioned between functional RNAs (2.4%), and hypotheticals (42%) with limited homology to known functional genes. The latter category likely contains genes specifically involved in mediated archaeal-sponge symbiosis. Phylogenetic analyses placed** ***C. symbiosum*** **as a basal crenarchaeon, sharing specific genomic features in common with either** ***Crenarchaea*****,** ***Euryarchaea*****, or both. The genome sequence of** ***C. symbiosum*** **reflect a unique and unusual evolutionary, physiological, and ecological history, one remarkably distinct from that of any other previously known microbial lineage.**

Introduction

The *Archaea,* one of the three major domains of extant life*,* have traditionally been subdivided into two distinct kingdoms. The *Euryarchaea,* include cultured representatives of methanogens, extreme halophiles, sulfate-reducers, and thermoacidophiles, while the cultivated *Crenarchaea* are comprised mostly of thermophiles and hyperthermophiles (1, 2). Cultivation-independent surveys led to the discovery of nonthermophilic *Crenarchaea* in aquatic environments (3, 4), sediments (5-7) and soils (8, 9) around the globe, and have expanded our collective view of archaeal distribution, physiology, evolution and ecology. While the properties of cultivated *Archaea* project a picture of microbes that are restricted to very specialized, often extreme habitats, molecular phylogenetic surveys suggest that Life's third domain is much more cosmopolitan and widely distributed.

Planktonic *Crenarchaea* represent a significant component of marine microbial biomass, approximately 10^{28} cells total in the world's oceans (10). Although marine *Crenarchaea* span the depth continuum (11), they predominate in waters just below the photic zone (10, 12, 13). Isotopic analyses have indicated that marine *Crenarchaea* have the capacity for autotrophic carbon assimilation (14-17), and the recent isolation of *Nitrosopumilus maritimus,* a nonthermophilic marine crenarchaeote demonstrates conclusively that bicarbonate and ammonia can serve as sole carbon and energy sources respectively, for at least some members of this lineage (18). Consistent with these observations, comparative environmental genomic studies have confirmed the presence of a conserved crenarchaeal chemolithoautotrophic ammonia-oxidizing metabolism distributed across multiple oceanic provinces (19). The identification of crenarchaeal ammonia oxidation genes in soils also suggests the general importance of these

organisms in terrestrial systems (20-22). Given the ubiquity and abundance of nonthermophilic *Crenarchaea*, there is considerable interest in expanding our knowledge of their evolution, ecology and metabolism.

Cenarchaeum symbiosum is the sole archaeal symbiont of the marine sponge *Axinella mexicana* (23). Previous analysis of *C. symbiosum* small subunit ribosomal RNA sequences (SSU rRNA) place this archeaon within the same lineage that contains the ubiquitous and abundant planktonic marine *Crenarchaea* (22-24). Fluorescent *in situ* hybridization (FISH) of *Archaea* within tissues of the sponge *A. mexicana* indicate that *C. symbiosum* is an extracellular symbiont (24). Although yet-uncultivated, *C. symbiosum* can be harvested in significant quantities from host tissues, where it comprises up to 65% of the total prokaryotic biomass (23, 24). Fosmid libraries enriched for *C. symbiosum* genomic DNA were previously constructed and screened for phylogenetic and functionally informative genes (19, 25, 26). Considering the extensive conservation of predicted ORFs between *C. symbiosum* and free-living marine *Crenarchaea* (19), we undertook the assembly of the complete genome of *C. symbiosum,* to provide a useful reference for comparative genomic analyses within this archaeal lineage.

Results and dicussion

Genome assembly and population structure

The *C. symbiosum* genome was iteratively assembled from a set of 155 completed fosmid sequences selected from an environmental library enriched for *C. symbiosum* genomic DNA (see methods, Table S1, S2 and Figure S1) (19, 26). The fosmids AF083071 and AF083072 corresponding to previously described a- and b-type ribosomal variants (26) served as nucleation points for the separation of sequence variants into discrete genomic bins (Table S1, S2). Because

of the relatively small number of clones comprising this library (see methods), each fosmid is expected to derive from an independent donor genome. Given this expectation, the library contains a cross section of the naturally occurring genomic variation found within the population at the time of sampling. Therefore, any assembly derived from this sample must represent a composite of related, but non-identical genotypes, for the purposes of this study a population genome equivalent. Remarkably, a single tiling path containing the complete genomic complement of *C. symbiosum* could be assembled from this complex dataset, that corresponded to the a-type population of sequence variants (Table S3).

C. symbiosum population structure was evaluated by analyzing fosmid sequence variation over the length of the assembled tiling path (Figure 1). Overlapping fosmid sequences ranged between ~80-100% nucleotide identity, with the a- and b- type variants dominating at the extremes. Overlapping a- and b-type fosmids, while virtually indistinguishable at the level of gene content and organization, differed in average nucleotide identity by ~15% (Figure 1 and below). Average nucleotide identity within each set of overlapping a- or b-type fosmids was ~98% although the range of variation within the b-type population was considerably higher (Figure 1). To facilitate analyses, fosmid sequences were partitioned using a 93% identity cut-off, roughly corresponding to a standard demarcation of bacterial species based on whole genome analysis (27, 28). A small number of fosmid sequences fell on the margins of a- or b-type distributions, suggesting the presence of less abundant intermediate genotypes (Figure 1). In order to estimate the representation of a- and b-type donors in the fosmid library, a- and b-type sequences were queried against the set of end sequences ≥200 base pairs in length (see methods). This approach identified on average 11 ends per a-type fosmid, and 8 ends per b-type fosmid,

consistent with 60% representation of a-type donor genomes in the isolated genomic DNA (see supplementary online materials for further information).

In order to explore the coherence and diversity of donor genotypes within the a-type population, fosmid sequences exhibiting >95% nucleotide identity, were evaluated for nucleotide polymorphisms (see methods). On average, these overlapping sequences exhibited 25-30 nucleotide polymorphisms per 1 kbp. The majority of these cases involved variation within intergenic regions or synonymous changes within ORFs (77% synonymous compared to 23% non-synonymous changes). However, "hot-spots" of nucleotide variation were detected in some orthologous genes (>50-80 polymorphisms/Kb). These changes were often associated with the presence of a variant allele within one or more of the expanded gene families (see below), which typically originated from a donor genotype not covered by the sequenced fosmids. The frequency of highly variable alleles (indicated by large peaks in Figure 1), suggests that selective pressures act with variable intensity on different regions of the *C. symbiosum* genome.

Genome Features

The *C. symbiosum* genome sequence contains 2, 045, 089 base pairs with a 57.74% average G+C content (Table 1). Based upon the identification of gap spanning paired end sequences and subsequent fosmid walking excursions, it appears to exist as a singular circular chromosome (Figure 1, 3). No clear nucleotide transition pattern correlating with an origin of replication could be identified based upon the methods of cumulative G+C skew (29) or co-localization with genes predicted to encode components of the replication initiation complex (30) (data not shown). A total of 2,017 protein encoding genes were predicted in the genome sequence, as well as a single copy of a linked SSU-LSU ribosomal RNA (rRNA) operon, 1 copy

of a 5S rRNA, 45 predicted transfer RNAs (tRNA) (Table S4), and 1 copy of a 7S signal recognition particle RNA. Approximately 57% of all predicted protein encoding genes could be assigned to functional or conserved roles based upon homology searches (see methods). A more complete breakdown of genome features including ORF content and taxonomic distribution is provided in Table 1. The distribution of tRNAs was uneven, with the clear majority mapping to two distinct regions of the genome (Figure 2).

Expanded Gene Families

The *C. symbiosum* genome contains an estimated 79 expanded gene families accounting for over 25% of its coding potential (see methods, Table S5). The majority of families were predicted to encode hypothetical proteins with no more than 3 representatives. However, 15 families contained at least 4 representatives (Table S5). Many families, including the two largest (containing 34 and 15 members respectively), were predicted to encode hypothetical proteins with limited homology to surface-layer or extracellular matrix proteins. This suggests that *C. symbiosum* encodes a significant number of surface features with the potential to mediate cell envelope formation or contact with host tissue. Moreover, representatives of these families often contained high levels of nucleotide polymorphism, corresponding to "hot-spots" of allelic diversity (Figure S1, S3). The genomic distribution of the two largest expanded gene families was inversely related to the distribution of tRNA genes. As visualized in Figure 3, the most abundant of these families, whose proteins range in size between 2-35 Kbp, is predicted to encode numerous **b**ig **a**rchaeal **p**roteins (*bap*) of unknown function. As a percentage of coding potential, *bap* genes represent ~15% of the entire genome and 56% of the sequence covered by all expanded gene families.

Gene Content

Because *C. symbiosum* has resisted cultivation, there are currently few specific details known about its physiology, metabolism, nutritional versatility, or growth requirements. The genome sequence now provides the basis for comparative analyses, and for generating metabolic models and hypotheses about specific genes, pathways and enzymatic activities. The metabolic inventory of *C. symbiousm* genes, most of which are shared with planktonic *Crenarchaea*, provides a starting point for a better understanding of the ecology, evolution and physiology of this ubiquitous crenarchaeal lineage.

Central Metabolism

In previous analysis of individual fosmid sequences we identified multiple components of an autotrophic carbon assimilation pathway based on a modified 3-hydroxypropionate cycle, and a nearly intact oxidative tricarboxylic acid cycle (TCA), suggesting that *C. symbiosum* is likely a facultative autotroph, incorporating both carbon dioxide and organic carbon for cell growth (19). Accordingly, these metabolic subsystems are present in the assembled *C. symbiosum* genome sequence. The presence of a single operon encoding oxoacid:ferredoxin oxidoreductase subunits suggests that *C. symbiosum* utilizes a horseshoe version of the oxidative TCA cycle for the production of intermediates in cofactor and amino acid biosynthesis. However, an intact oxidative TCA cycle cannot be excluded in the absence of functional studies to determine the specificity of this enzyme complex towards pyruvate, oxoglutarate or ketoglutarate. With the exception of glucokinase (EC 2.7.1.2) and pyruvate kinase (2.7.1.40), *C. symbiosum* contains an intact form of the Embden-Meyerhof-Parnas (EMP) pathway for the metabolism of hexose

sugars. Several alternatives to glucokinase including 2 genes predicted to encode carbohydrate kinases of unknown specificity and 1 gene predicted to encode a ROK-family ribokinase were identified with the potential to mediate the first step in sugar activation. Similarly, a gene predicted to encode phosphoenolpyruvate synthase, an alternative to pyruvate kinase, mediating the interconversion of pyruvate and phosphoenolpyruvate was also identified. The absence of glucose 1-dehydrogenase (EC 1.1.1.47), gluconolactonase (EC 3.1.1.17), and 2-keto-3-deoxy gluconate aldolase (EC 4.1.2.-) homologues suggests that *C. symbiosum* does not utilize the Entner-Duodoroff (ED) pathway, an alternative to the EMP pathway, in the catabolism of hexose sugars. In addition to the EMP pathway, an intact non-oxidative pentose phosphate pathway was identified, providing a mechanism for production of NADPH and ribose sugars for nucleotide biosynthesis.

Energy Metabolism

Key genes associated with chemolithotrophic ammonia oxidation, including ammonia monooxygenase, ammonia permease, urease, a urea transport system, nitrite reductase and nitric acid reductase have been previously reported in *C. symbiosum* (19). Consistent with these analyses, the complete pathway for urea and ammonia transport and conversion, ammonia oxidation, and the metabolism of nitrogen oxides is present in the assembled genome sequence. Several loci, including the ammonia permease and ammonia monooxygenase subunit C were identified as members of expanded gene families (Table S5). However, homologues for hydroxylamine oxidoreductase (EC 1.7.3.4) and cytochromes c_{554} and c_{552} were not identified, suggesting that *C. symbiosum* employs alternative mechanisms for transferring electrons between hydroxylamine and the electron transport chain. Consistent with this hypothesis, 14 genes

predicted to encode domains related of the plastocyanin/azurin family of blue type (I) copper proteins were identified with the potential to substitue for cytochromes as mobile electron carriers (31, 32).

Three electron transport complexes, including a complete respiratory NADH dehydrogenase (complex I), succinate dehydrogenase (complex II), and cytochrome C oxidase (complex IV) were unambiguously identified. In addition, an operon predicted to encode a Rieske iron sulfur cluster protein, cytochrome b, and type (I) copper protein with the potential to function as a cytochrome c reductase (complex III) was identified. Finally, a gene cluster predicted to encode a complete archaeal ATP synthase (complex V) was identified, completing the aerobic respiratory circuit required for the indirect coupling of electron transfer to ATP synthesis. In the case of the NADH dehydrogenase (*nuo*), a gene cluster encoding the core respiratory complex *nuoABCDHIJKLMN* was identified. However, genes encoding the electron input module *nuoEFG*, involved in NADH binding and oxidation were not identified, reinforcing the hypothesis that *C. symbiosum* uses alternative electron carriers, potentially including type (I) copper proteins or ferredoxins. In addition to the proposed pathway of ammonia oxidation described above, 7 genes predicted to encode Fe-S cluster oxidoreductases and 5 genes predicted to encode ferredoxins were identified with the potential to contribute electrons into the respiratory chain, or act as low-potential electron donors for various enzymatic reactions.

Amino Acid and Cofactor Biosynthesis

C. symbiosum encodes the genes required to synthesize all 20 amino acids, with one exception. Genes predicted to encode pyrroline-5-carboxylate reductase (EC 1.5.1.2) and ornithine cyclodeaminase (EC 4.3.1.12) involved in the conversion of 1-pyrroline-5-carboxylate

and ornithine respectively, to L-proline were not identified. However, several genes predicted to encode aminopeptidases were identified, suggesting that *C. symbiosum* has the capacity to derive proline from dietary sources. Consistent with this hypothesis, an oligo-transport system containing linked permease and unlinked ATPase components with the potential to mediate uptake of free peptides was identified. A solute binding protein with the potential to interact with this transport complex was identified in close proximity to the predicted permease components. Previous studies of microbial host interactions have identified mechanisms of host-derived nutritional support, including amino acid provisioning (33) and energy transfer (34). The existence of a proline auxotrophy in *C. symbiosum* could help explain its commensal relationship with *A. mexicana*.

Pathways for cofactor biosynthesis are well represented in the *C. symbiosum* genome. Nearly complete sets of genes required for the *de novo* synthesis of biotin, vitamin B12, riboflavin, thiamine and pyridoxine were all identified. In the case of folic acid biosynthesis, genes encoding all steps for the conversion of the C1 carrier tetrahydrofolate (THF) to methyl-THF were identified. However, the absence of genes encoding dihydrofolate reductase (EC 1.5.1.3) or dihydrofolate synthetase (6.3.2.12) suggests that *C. symbiosum* is incapable of *de novo* folate biosynthesis, deriving 7,8-dihydrofolate (DHF) from dietary sources or regenerating THF during the conversion of homocysteine to methionine (35, 36).

Transporter Systems

The *C. symbiosum* genome encodes 47 genes comprising 18 different transport systems. The largest group consisted of the multisubunit ABC family and included transporters for Mn^{2+}/Zn^{2+}, Ni^{2+}, Fe^{3+} hydroxamate, phosphate/phosphonate, nitrate/sulfonate/bicarbonate,

branched chain amino acids, dipeptides, and multidrug resistance. A variety of genes encoding proteins involved in the transport or exchange of ammonia, urea, Na^{+}/H^{+}, K^{+}/H^{+}, Mg^{2+}/Co^{2+}, Mn^{2+}/Fe^{2+}, Ca^{+}/Na^{+} and unspecified divalent and heavy metal cations, were also identified.

Protein Translocation and Secretion

The *C. symbiosum* genome encodes a complete set of archaeal signal recognition particle (SRP) components for targeting secretary and membrane proteins, including *srp19*, *srp54*, a 7S RNA gene as well as an SRP receptor homologue (*ftsY*). The *ftsY* gene was found as part of a larger gene cluster containing additional secretory pathway components including a *secY* translocase homologue, and a second gene predicted to encode an integral membrane protein with similarity to the Sec accessory subunit YidC. Definitive homologues for *secE* and *secG,* two additional subunits of the archaeal Sec translocation pore complex, were not identified. Three copies of a conserved hypothetical gene with limited homology to eukaryotic vacuolar sorting factors and several components of the Sec-independent twin argine translocation (tat) system, including *tatA* and *tatC* were also identified in the genome sequence.

Signaling, Motility and Cell Surface features

C. symbiosum contains numerous genes predicted to encode cell surface features or associated membrane proteins mediating signaling, cell envelope formation, phospholipid binding and modification, glycosylation, and serine or metal-dependent protein degradation. The absence of genes encoding classical two-component sensory and motility systems suggests that *C. symbiosum* is a sedentary organism. A total of 6 genes predicted to encode signaling kinases, including 4 serine/threonine kinases, 1 signal transduction histidine kinase and 1 unusual protein

kinase of unknown function, were identified with the potential to transduce cell surface events or regulate cellular function. Consistent with the operation of these gene products in the development of reversible regulatory networks and signal integration, 2 genes predicted to encode protein phosphatases specific to serine/threonine and tyrosine respectively, were identified.

Information Processing and Chromatin dynamics

The *C. symbiosum* genome contains the full repertoire of genes necessary for chromosomal replication fork assembly and function, including components of the origin recognition complex (*cdc6*), two topoisomerases, single and double strand helicases, 3 copies of a predicted bacterial/archaeal-type DNA primase, a two-subunit eukaryal/archaeal DNA primase system, RNase H, sliding clamp, and DNA ligase (*cdc9*). In addition, genes encoding two distinct DNA polymerases, a single B family DNA polymerase I elongation subunit related to sequences derived from thermophilic *Crenarchaea* (37), and a second euryarchaeal-type polymerase II consisting of large and small subunits were both identified. Numerous genes involved in DNA repair were also present, including *recA* recombinase, nucleotidyltransferase, O-methyltransferase, and a complete *uvr* nucleotide excision repair system. In addition to DNA replication and repair systems, 3 genes predicted to encode ATPases typically associated with chromosome partitioning and maintenance, including a homologue of structural maintenance of chromosomes (*smc*), a membrane-associated ATPase (*minD*) and a gene predicted to encode the cell division protein *ftsZ*, were identified.

In previous analysis based on individual fosmid sequences, a eukaryotic-like histone homologue was identified in *C. symbiosum* (38). Accordingly, a single copy of a gene predicted

to encode a histone H3-H4 was identified in the assembled *C. symbiosum* genome sequence. Genes predicted to encode a histone H1 DNA binding protein involved in nucleosome packaging and 3 genes predicted to encode histone acetyltransferases typically associated with opening the nucleosome core to promote transcription, were also present. In addition, 27 helicase genes were identified with varying roles in DNA replication, transcription and repair, including 9 superfamily II members implicated in ATP-dependent chromatin remodeling. Moreover, 25 genes predicted to encode restriction modification methytransferases were identified with the potential to protect *C. symbiosum* from exogenous DNA or to modulate transcriptional activity by tightening the nucleosome core. Only 2 of these genes had close archaeal homologues. The remaining methyltransferase genes were affiliated with the domain *Bacteria*, including *Cyanobacteriales*, *Actinomycetes* and *Bacteroidetes*, and were possibly laterally acquired from members of the surrounding sponge microbiota. Five of the methyltransferases could be unambiguously linked to genes predicted to encode restriction endonucleases of untested specificity.

C. symbiosum contains a complete set of genes necessary for transcription initiation including preinitiation complex formation and RNA polymerase assembly. The presence of 3 divergent copies of the TATA binding protein (*tbp*) and 5 copies of the transcription initiation factor TFIIB (*tfb*) indicates that *C. symbiosum* has the potential to generate alternative preinitiation complexes. Over 40 genes predicted to encode transcriptional regulators with the capacity to modulate preinitiation complex formation were identified. The majority of these genes are predicted to encode members of the Lrp/AsnC family of transcriptional regulators. In two instances related groups of transcriptional regulators formed expanded gene families composed of 5 and 4 members, respectively (Table S5).

A total of 10 translation initiation and 4 elongation factors were identified in the *C. symbiosum* genome including 2 copies of the translation initiation factor 2B (eIF2B). In addition, a single gene predicted to encode a small bacterial-type cold shock protein (*cspB*) potentially involved in RNA binding and translational control was also identified. A highly conserved *cspB* gene was previously identified on a genomic DNA fragment derived from a planktonic marine crenarchaeote, 4B7 (39). So far, crenarchaeal *cspB* homologues have only been found in mesophiles or psychrophiles, a feature distinguishing *C. symbiosum* and cold-living relatives from other thermophilic lineages.

As mentioned earlier, the *C. symbiousm* genome contains 45 predicted tRNA genes. Ten of these predicted tRNAs contain putative introns (Table S4). Most of the exon-intron boundaries form the conserved bulge-helix-bulge motif (BHB), although several appear to adopt structurally divergent forms (40). Such divergent features have been previously correlated with the presence of two distinct copies of the splicing endonuclease (*endA*) (41-44). Consistent with these observations, the *C. symbiousm* genome encodes two copies of *endA,* corresponding to divergent crenarchaeal and euryarchaeal homologues respectively.

Aminoacyl-tRNA synthetases are encoded for every amino acid with the exception of glutamine. However, 4 subunits of a glutamyl-tRNA amidotranseferase (*gat*) encoded by separate *gatED* and *gatBA* operons likely mediated Glutamyl-tRNA activation. The genomic distribution of aminoacyl-tRNA synthetases was uneven, coinciding with the pattern observed for individual tRNAs (see previous and data not shown). Genes predicted to encode selenophosphate synthase or selenocysteine synthase both required for activation of selenocysteinyl-tRNA were not identified. However, the identification of a selenocysteine

specific elongation factor (*selB*) suggests that *C. symbiosum* retains the potential for co-translational insertion of selenocysteine residues into nascent peptide chains (45-47).

C. symbiosum harbors a number of genes predicted to encode chaperonins involved in cellular stress responses and protein folding and refolding processes. An operon predicted to encode a heat-shock or stress response complex composed of the genes *grpE,* hsp70 (*dnaK)*, and hsp40 (*dnaJ)* was identified. Three additional copies of genes containing *dnaK*-like domains, and 2 copies each of the small heat shock protein hsp20, the alpha and beta (*gimC*) subunits of prefoldin, and an hsp60 related thermosome alpha and beta subunits, were also identified. In addition to the classical chaperonin systems, 4 genes predicted to encode peptidyl-prolyl cis-trans isomerase, and at least 10 genes predicted to encode protein disulfide isomerase or thioredoxin, were identified with the potential to assist in *de novo* protein folding and oxidative stress responses.

Evolutionary Affinities of a Deeply Branching Archaeal Lineage

To better define the evolutionary relationships between *C. symbiosum* and other archaeal groups, a total of 57 genes encoding unique ribosomal proteins (r-proteins) identified in the assembled *C. symbiosum* sequence were analyzed. Several genes including rpl14e, rpl34e common to cultured *Crenarchaea* and basal *Euryarchaeota*, rpl13e, rps24e and rps25e common to cultured *Crenarchaea*, rpl20a common to most *Archaea* and rpl35ae common to *Thermococcales* and *Nanoarchaea*, were not identified in the assembled sequence, or the complete set of unassembled a-type and b-type fosmids (Table S6). The set of ribosomal proteins was supplemented with 17 additional taxonomically informative genes involved in translocation, information processing and DNA recombination and repair (Table S6). Given the small size of

most r-proteins, these analyses of individual genes have limited resolving power, reflected in weak bootstrap support (Table S6). To improve the phylogenetic signal individual r-proteins were combined into one concatenated alignment set, excluding r-proteins for which horizontal gene transfer is suspected (see methods) (48). The resulting phylogenetic trees place *C. symbiosum* in a deeply branching archaeal lineage (Figure 3, S3), similar in topology to previously reported relationships of fast evolving and/or deeply branching archaeal genomes (49, 50). Similar tree topologies were obtained in analyses of the SecY translocase (Figure S4), nucleotide repair and recombination protein RadA (Table S6**)** (51) the beta subunit of the DNA unwinding protein TOPO-VII (Table S6), and the elongation factor EF-1a (Table S6). In aggregate, the results support a basal position for *C. symbiosum* with respect to *Crenarchaea*. However, it remains unclear whether *C. symbiosum* truly represents a distinct phylum emerging near the root of the archaeal tree, or simply a deeply branching crenarchaeal lineage. Future analyses incorporating more diverse and basal archaeal representatives, including the Korarchaeota (52), may help resolve this uncertainty.

Comparative environmental genomics

To explore the shared coding potential between *C. symbiosum* and its planktonic relatives, sequences from individual Sargasso Sea (SAR) whole genome shotgun DNA libraries (53) were aligned to the assembled *C. symbiosum* genome (Figure 4, see methods). The distribution of sequences encoding marine crenarchaeal homologues over the entire length of the *C. symbiosum* genome varied considerably between samples. Whole genome shotgun coverage was highest in the SAR3 sample providing over 4,000 unique reads averaging 65% amino acid identity and 78% amino acid similarity over the length of the aligning read. This represents

~1.25% of the total sequence population within the SAR3 sample, encompassing over 4 Mb of WGS sequence (approximately 2 genome equivalents based upon an estimated 2 Mb genome size). The depth of sequence coverage over the *C. symbiosum* genome was uneven, varying between 1 and >20 fold between homologous intervals. More than 20% of the aligning sequences were derived from mate pairs mapping within the average range of insert sizes (3-6 Kb), suggesting that gene order is conserved between *C. symbiosum* and its planktonic relatives over short syntenic intervals. Numerous gaps in sequence coverage were also identified indicating that a significant proportion of *C. symbiosum* genes are absent or not well conserved within planktonic *Crenarchaea* (Figure 4).

To investigate functional implications associated with the observed variation in sequence coverage, all protein encoding sequences predicted in the *C. symbiosum* genome were queried against a local database containing the Sargasso whole genome shotgun data as well as the set of public genomes (see methods). The resulting alignments were compared using the BLAST score ratio (BSR) to identify highly conserved genes shared between *C. symbiosum*, SAR and public genomes (Figure 2). The bar heights in circles 3 and 4 of Figure 3 span a range of BSR values between 30-100, with 100 representing a perfect match and 30 representing the lower cut-off. The cut-off corresponds to approximately 30-40% amino acid identity. A total of 65 genes with a BSR ≥30 were more highly conserved between *C. symbiosum* and the public genomes (Table S7). Of these, 43 fell into defined COG categories, including 10 genes associated with DNA replication, recombination and repair (L), 9 genes associated with amino acid transport and metabolism (E), and 6 genes associated with posttranslational modification, protein turnover, and chaperones (O). The distribution of genes within these three categories was far from random. For instance, within the first category, 7 genes were most similar to bacterial associated DNA

modification methyltransferases, and within the third category, 5 genes were homologous to serine protease inhibitors (serpins). A total of 525 genes with a BSR ≥30 were more highly conserved between *C. symbiosum* and the Sargasso Sea, corresponding to ~26% of all predicted protein encoding genes in the *C. symbiosum* genome (Table S8). This set of shared genes spanned the complete spectrum of COG categories, with the highest representation in energy production and conversion (C), amino acid transport and metabolism, (E), translation, ribosomal structure and biogenesis (J), transcription (K), and DNA replication, recombination and repair (L). The remaining gene predictions were either shared equally between the SAR and public genomes or were not well conserved at all (Figure 2 and data not shown). The latter case, represented by gaps in both the circular genome map (Figure 2) and coverage plots (Figure 4), encompassed over 800 genes with a BSR <30, corresponding to ~39% of all predicted protein encoding genes in the *C. symbiosum* genome (data not shown). Many of these sequences represented hypothetical genes or belonged to expanded gene families that appear unique to the *C. symbiosum* genome, and are likely involved in sponge-associated processes.

Conclusion

Population Structure and Genomic Coherence

The *C. symbiosum* genome reported here, represents a composite sequence assembled from individual, highly related sympatric donor genotypes. As such, the genetic variability and population structure of *C. symbiosum* cells residing in an individual sponge host is partly reflected in the sequence. The partitioning of syntenic a- and b-type sequences during assembly suggests that genetic or physicochemical barriers exist within the host environment. The average nucleotide divergence between a- and b-type populations was comparable to the evolutionary

distance between *Escherichia coli* and *Salmonella* spp., estimated to have diverged approximately 100 million years ago (54). The majority of a-type and b-type donor genomes therefore, have likely been evolving separately for some time. Given that gene content, order and orientation in overlapping a- and b-type fosmids is the same, selective forces may be acting on individual genes, yielding specific physiological adaptations, perhaps allowing colonization of different niches within the host tissue. It is possible that periodic selection could separate tissue-specific ecotypes into distinct sequence clusters or monophyletic groups over time (55, 56). Fine scale determination of a- and b-type genotype spatial distributions, and deeper exploration of allelic variation in *C. symbiosum* populations is required to further test this hypothesis.

Functional and Metabolic Relationships

The results presented here in combination with previous studies (18, 19), further substantiate that *C. symbiousm* and its planktonic marine relatives are nitrifiers, deriving cellular energy from ammonia oxidation, and carbon from CO_2. This suggests a major role for *Crenarchea* in nitrogen cycling in the marine environment. Comparison with the Sargasso Sea whole genome shotgun dataset indicates that the majority of metabolic subsystems identified in *C. symbiosum* are conserved within planktonic *Crenarchaeota*. In addition to core information processing systems, the *C. symbiosum* genome encodes complete or nearly complete subsystems for a wide variety of biosynthetic and housekeeping functions including glycolysis, gluconeogenesis, pentose phosphate conversion, TCA cycle, cofactor and vitamin metabolism, amino acid biosynthesis, oxidative phosphorylation, ATP synthesis, and variant pathways for autotrophic carbon assimilation and chemolithotrophic ammonia oxidation. Closely related homologues within each of these subsystems are found in free-living planktonic marine

Crenarchaeota. The conservation of oxidative TCA, 3-hydroxypropionate cycle, and genes involved in ammonia oxidation, strongly support the hypothesis that *C. symbiosum* (and its planktonic relatives) are aerobic, facultative ammonia oxidizing chemolithoautotrophs (19). Given the biogeochemical significance of these metabolic pathways, it will be interesting to determine whether conserved or divergent forms exist in other *Crenarchaea* dwelling in diverse habitats such as marine sediments and the deep subsurface (57, 58).

Evolutionary Relationships

Over 60% of all conserved protein encoding genes predicted within the *C. symbiosum* genome partition within the archaeal domain. Although placed within the *Crenarchaea* based on rRNA sequence analyses (22, 23), the majority of these sequences exhibit a strong euryarchaeal signal. Curiously, over 30% of all conserved protein encoding genes predicted in the *C. symbiosum* genome are most closely related to bacterial counterparts identified in the public databases. Whether these observations reflect current biases in database representation, or bona fide lateral gene transfer between domains, remains to be determined. The answer may be particularly important with regard to the potential adaptive radiation of this lineage into "nonextreme" environments, and the acquisition of new physiological traits enabling niche expansion.

Comparison of 57 concatenated ribosomal protein sequences placed *C. symbiosum* deep within the archaeal lineage. If the hypothesis of deep emergence is true, interesting questions about the hyperthemophilic nature of the common archaeal ancestor arise. However, it seems clear that low and high temperature thermal adaptations have occurred multiple times within the *Crenarchaea* (52). Both low and high temperature adaptation may therefore represent a

homeoplastic trait in the crenarchaeal lineage. Strikingly, *C. symbiosum* and relatives share numerous characters (including nearest neighbor matching of *C. cymbiosum* ORFs, individual r-protein analyses, and the presence of a histone homologue (38), two DNA polymerase II subunits, and the cell division control genes *ftsZ*, *mind,* and the heat shock protein complex hsp70) with *Euryarchaea*, but not with other *Crenarchaea*. Taking these features into account, and consistent with rRNA analyses, *C. symbiosum* (and planktonic marine relatives) seem best viewed as a highly divergent sister taxon of cultivated hyperthermophilic *Crenarchaea*. While some of the phylogenetic signal observed may be due to long branch attraction, the shared gene content observed between *C. symbiosum* and members of the *Euryarchaeota* support a deep split of the *Crenarchaea* into two distinct lineages, one major lineage represented by *C. symbiosum* and free-living planktonic relatives. Whether these sister taxa share a common mesophilic or thermophilic ancestor remains unclear.

Symbiosis

Little is known about the specific nature of the symbiosis between *C. symbiosum* and the marine sponge *A. mexicana*. However, the *C. symbiosum* genome sequence does provide several mechanistic clues regarding potential trophic interactions and hints at regulatory mechanisms necessary for extracellular contact, communication and defense. In the case of trophism, the host may provide dietary sources of proline and folate in exchange for the removal of urea and ammonia waste products by the archaeal symbiont. The sponge cortex is populated by a variety of commensal and opportunistic microorganisms. A portion of this microbiota forms the basis of the sponge's own nutrition. As a nonmotile extracellular symbiont *C. symbiosum* has likely developed or acquired mechanisms to inhibit or evade host consumption and defend against

bacterial or viral predation. A significant number of predicted genes encode domains reminiscent of cell surface, regulatory or defense mechanisms, including numerous restriction modification systems to protect against foreign DNA, autotransporter adhesins potentially involved in mediating cell-cell contact, proteases with the potential to modify or degrade extracellular matrix proteins, glycosyltransferases involved in cell wall biogenesis, and secreted seine protease inhibitors with the potential to mediate evasion of innate host defense systems. Many of these features do not appear to be encoded in the genomes of *C. symbiosum's* planktonic marine relatives and likely constitute specific adaptive alleles mediating the symbiotic life style. Given the relative dearth of archaeal/metazoan symbioses, the *C. symbiosum* genome provides an unprecedented opportunity to explore the genetic features mediating host contact, communication and trophic exchange.

Coda

The *C. symbiosum* genome now provides the basis for a wide variety of functional and comparative studies relating to both free-living and symbiotic life styles. It also provides a resource for heterologous expression of archaeal proteins that work at moderate temperatures, facilitating archaeal/eukaryal *in vitro* functional testing of replication, transcription or translation properties. As well, the *C. symbiosum* genome provides a valuable reference for identifying key genes involved in carbon and nitrogen cycling within mesophilic *Crenarchaea*, and a useful phylogenetic reference point for inferring the evolutionary relationships among various members of the archaeal domain. Future comparisons of the *C. symbiosum* genome with the recently isolated ammonia-oxidizing crenarchaeon, *Nitrosopumilus marina* (18), should be particularly enlightening.

Materials and Methods

Library construction, specifications and sequencing

C. symbiosum cell enrichment, DNA extraction from sponge tissue, and fosmid library construction have been previously described (19, 59) The fosmid library used in the present study contains 2100 clones arrayed in twenty-two 96-well plates, with an estimated average insert size ~40 thousand base pairs (Kb) per clone. Prior to fosmid selection, end sequencing generated 2,779 non-redundant reads greater than 200 base pairs (bp) per read, averging 500 bp per read. Of the set of non-redundant reads, 1,041 clones were represented by paired-ends covering both the 5' and 3' ends. Overall, 66.2% of the library was represented by at least one end sequence, and 49.6% was represented by paired-end sequences.

Sequencing and assembly approach

Five successive phases of fosmid selection, sequencing and assembly, were conducted over a four-year period. Initial selection was based on the following three criteria: (1) paired end sequences predicted to contain ORFs most similar to archaeal genes, (2) linkage with previously reported fosmids harboring phylogenetic anchors including genes encoding the small subunit ribosomal RNA, *radA* recombinase (51) and DNA polymerase subunits (37)and sets of paired ends, assembled in opposing orientations and predicted to contain ORFs homologous to two or more archaeal genes. Two previously described fosmids, AF083071 and AF083072, harboring a-type and b-type SSU ribosomal RNA genes respectively were included as seeds for fosmid walking excursions (26) (Figure S1). Refer to supporting online material for information relating

to the JGI sequencing pipeline and assembly of raw trace files, and information related to tiling path construction can be found in the supplementary on-line material (SOM).

DNA Sequence Analysis

Contigs were annotated using the FGENESB pipeline for automatic annotation of bacterial genomes from Softberry (http://www.softberry.com/berry.phtml, Mount Kisco, NY) using the following parameters and cut-offs: open reading frame size = 100 aa, Expectation = 1e-10. Predicted ORFs were queried against the KEGG, COG and GenBank non-redundant (NR) databases. SSU and LSU rRNA genes were identified by blastn query against NR with expectation cut-offs of $1X10^{-8}$. Automated FGENESB annotation of fosmids was manually refined and corrected using the genome annotation and visualization tool Artemis (http://www.sanger.ac.uk/Software/Artemis) (60). Putative tRNA genes were identified using the program tRNAscan-SE 1.21 (http://www.genetics.wustl.edu/eddy/tRNAscan-SE) (61) set to the archaeal tRNA covariance model, and SPLITS 1.0 (http://splits.iab.keio.ac.jp) (J. Sugahara, N. Yachie, Y. Sekine, A. Soma, M. Matsui, M. Tomita, and A. Kanai, submitted) set to the following parameters; -d 2 -p 0.65 -F 5. The putative tRNA genes containing possible intron(s) detected with only tRNAscan-SE 1.21 were verified by the prediction of the bulge-helix-bulge motif for the hallmark of the exon-intron boundaries (Marck, C. and Grosjean, H. (2003), RNA, 9, 1516-1531) by using SPLITS 1.0. The 5S rRNA sequence was initially identified on the basis of blastn searches of the unassembled fosmid sequences using representative archaeal reference sequences. The 7S RNA sequence was initially identified on the basis of blastn searches using a 17-nt conserved region (AGG(C/T)CCGGAAGGGAGCA) deduced from archaeal 7S sequences. The tentative 5' and 3' ends of the gene were assigned from the conservation between a- and b-type genome sequences. The similarity of the predicted secondary structures to the consensus

folding of 7S RNA and the presence of conserved motifs (Zwieb C, van Nues RW, Rosenblad MA, Brown JD, Samuelsson T. (2005) RNA. 11, 7-13) were also considered. The complete tiling path was assembled in Artemis by sequential addition of ordered and oriented contigs and gap spanning fosmids with accompanying feature tables, and visualized with the program GenomeViz http://www.uniklinikum-giessen.de/genome/genomeviz/download.html (62).

Nucleotide polymorphism determination

To identify orthologous regions (defined here as the reciprocal best matches using the blastn algorithm (63) and a minimum cut-off of 50% identity over a minimal 700 base pair interval), the complete genomic scaffold was divided into 1,000 base pair (1 kbp) long, consecutive fragments, and searched against the set of fosmids. The same analysis was performed at the gene level as well, using the set of genes annotated on the genomic scaffold as reference sequences. When the total length of ortholgous sequences (gene-level comparisons) between a fosmid and the genomic scaffold was longer than 5 kbp, the fosmid was considered to derive from *C. symbiosum* and the average nucleotide identity between the fosmid and the genome was calculated directly from the resulting blastn output. Ortholgous regions between *C. symbiosum* fosmids (1-kbp window comparisons) were subsequently aligned using clustalw (64). The number of invariable and variable bases in the 1 kbp fragments, which is shown in Fig.1 was calculated directly from the clustalw alignments for the fosmids that showed >95% average nucleotide identity to the genomic scaffold.

Comparative analysis between C. symbiosum, public genomes and Sargasso Sea

The Sargasso Sea database contained the complete set of unassembled, vector-trimmed, whole genome shotgun sequences (53), while the public genomes database included all whole-genome sequences accessible through NCBI's ftp site as of December of 2006 (260 genomes in total). Since the Sargasso average read length is only ~818 bases, *C. symbiosum* proteins longer than 300 amino acids were split into 300 amino acid long consecutive fragments, which where then queried against the Sargasso database. Use of the whole-genome sequences (as opposed to annotated protein sequences for the public genomes) avoided inconsistencies between differently annotated genomes, thereby facilitating comparison to the unannotated Sargasso database. Evaluation of gene conservation was based upon analysis of BLAST score ratios (BSR) between *C. symbiosum*, Sargasso Sea and public genomes (65) using tblastn. The BSR represents the ratio of the bit score for the set of predicted *C. symbiosum* proteins queried against the Sargasso Sea or public genomes database divided by the bit score of *C. symbiosum* queried against itself (self-match). Application of the BSR reduces biases associated with database size and the length of matching segments by normalizing the bit scores derived from blast algorithms.

Coverage plots relating the set of whole genome shotgun (WGS) reads from individual Saragsso Sea (SAR) sample bins, SAR1-7 (http://www.venterinstitute.org/sargasso/) (53) to the *C. symbiosum* genomic scafffold were generated using the Promer program implemented in MUMmer3.18 (66). Promer generates amino acid alignments based on the translation of both query and subject sequences in all six ORFs. The following parameters and cut-offs were used: breaklength = 60, minimum cluster length = 20, and match length = 10. Test alignments intended to explore the specificity and depth of coverage of SAR alignments to the *C. symbiousm* genome were performed with the following archaeal reference genomes: *Archaeoglobus fulgidus* (NC_000917), *Methanothermobacter thermautotrophicus* (NC_000916), *Thermoplasma*

volcanium (NC_002689), *Pyrobaculum aerophilum* (AE009441) and *Sulfolobus solfataricus* (NC_002754). Resulting delta files were converted into coordinate files for plotting and sequence analysis using the show-coords program and visualized in graphical format (coverage plot) using the mummerplot program.

Phylogenetic Analysis

Phylogenetic analyses were performed using maximum-likelihood (ML) methods implemented in PHYML (http://atgc.lirmm.fr/phyml/) (67). For the set of ribosomal protein s (Rbp) individual analyses of each RBp was performed prior to the construction of a concatenated alignment to assess the relative phylogenetic signal of each protein. Resulting trees were left unrooted to maximize the number of useful positions in the alignments and to limit the risk of long-branch attraction (LBA). Several r-proteins, including Rpl2e, Y and Z were excluded from the concatenated dataset based on prior studies indicating lateral gene transfer between archaeal groups (48). Additional information related to the phylogenetic analysis can be found in the supplementary online material (SOM).

Supporting Online Material

All supporting information is available on the XXXX web site (http://www.XXXX.org). Complete genome annotation files are available through the Joint Genome Institute's **I**ntegrated **M**icrobial **G**enomes system (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi) and through the NCBI web portal (http://www.ncbi.nlm.nih.gov/). Individual fosmid sequences can be obtained from GenBank under the accession numbers DQ397540-DQ397640 and DQ397827-DQ397878

corresponding to a- and b-type population bins respectively. The complete a-type genome sequenced can be obtained from GenBank the accession number XXXX.

Acknowledgements

Special thanks to Asuncion Martinez, Tsultrim Palden, Tracy Mincer, Matthew Sullivan and Maureen Coleman at MIT, Jarod Chapman, Sam Pitluck, Chris Detter, Krishna Palaniappan and the entire JGI staff for computational and technical assistance, and J. S and Y. W. thanks to Nozomu Yachie, Masaru Tomita and Akio Kanai at Keio University for tRNA analysis using SPLITS. This study was supported by NSF# MCB-0509923, the Gordon and Betty Moore Foundation, and the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program and the University of California, Lawrence Livermore National Laboratory, under contract no W-7405-ENG-48, Lawrence Berkeley National Laboratory contract no. DE-AC03-765F00098, and Los Almos National Laboratory contract no. W-7405-ENG-36.

Figures

Tables

Table 1. *C. symbiosum* genome features

Figure Legends

Figure 1. *C. symbiosum* fosmid population structure and mosaic genome assembly

(Top) Fosmids partition into two distinct population bins corresponding to a-type and b-type ribosomal variants. Average nucleotide identity of each fully sequenced fosmid is plotted against the position of each fosmid in the assembled a-type scaffold. Blue lines represent the set of fosmids falling within the a-type population and red lines indicate the set of fosmids falling within the b-type population. The inset histograms represent the overall sequence divergence among and between overlapping fosmids using the in BLAST algorithm. The distribution of observed sequence similarity (percent identity) in high-scoring segment pairs (HSPs) for alignments (left) between fosmid clones assigned to population "a", (middle) between "a+b" populations, and (right) between fosmid clones assigned to population "b". (Bottom) Number of nucleotide polymorphisms per 1Kb of orthologous sequence shared between overlapping fosmids within the a-type population exhibiting >95% nucleotide identity to the genomic scaffold. Gaps in the distribution represent genomic intervals covered by a single fosmid clone (see methods).

Figure 2. The *Cenarchaeum symbiosum* genome

Nested circles from outermost to innermost represent the following information: 1) Gene content predicted on the forward strand. 2) Gene content predicted on the reverse strand. The color of predicted open reading frames (ORFs) is based on COG functional categories (see key for color designations). ORFs were assigned to COG categories according to their top match against the COGs database, using the blastp algorithm set to a minimum cut-off of 30% amino acid identity

over at least 70% of the length of the gene. 3) Conservation of predicted *C. symbiosum* genes in the set of published and completed microbial genomes (see methods). 4) Conservation of predicted *C. symbiosum* genes in the unassembled set of whole genome shotgun data from the Sargasso Sea. The height of the bars in circles 3 and 4 indicates the BLAST score ratio (BSR) for the set of predicted *C. symbiosum* proteins queried against the public genomes and Sargasso Sea respectively using a baseline cut-off ≥30 (see methods). 5) The extent of polymorphisms within the type-a population shown in figure 1 mapped on the genome (see figure 1 for details). 6) Expanded gene families (discussed in the text, see key for color designations and Table S5 additional information) ; notice that high number of polymorphisms (circle 5) frequently coinside with the expanded protein families, i.e., these proteins families are frequently hot-spots of diversity within the type-a population. 7) tRNA and rRNA gene positions. 8) G+C content deviation from the mean (57.5%) in 1, 000 base pair windows.

Figure 3. Phylogenetic position of *C. symbiosum*

Bayesian phylogenetic tree constructed using MrBayes v3_0b4 (68) with a mixed model of amino acid substitution and a Γ-law (8 discrete categories plus a proportion of invariant sites) to take into account among site rate variation. Numbers in bold associated with each branch represent the posterior probabilities from the Bayesian analysis performed with Mr. Bayes whereas other numbers are the bootstrap values from the maximum likelihood analysis performed with PHYML (JTT model and Γ-law (8 discrete categories plus a proportion of invariant sites) (67). The scale bar represents the average number of substitutions per site.

Figure 4. Comparative analysis of *C. symbiosum* and Saragasso Sea sample bins

Coverage plots for individual SAR sample bins aligned to the *C. symbiosum* genomic scaffold (see methods). Each vertical bar represents an individual whole genome shotgun (WGS) read. The average percent amino acid identity (top) and similarity (bottom) for SAR WGS reads aligning to the *C. symbiosum* genome is shown to the far right of each coverage plot. For simplified visualization of gaps in the alignment, all matches are replotted near the base of the x-axis to form a normalized 1D plot spanning the reference sequence. To illustrate the gene content of gaps within the distribution of aligning WGS reads several regions corresponding to expanded gene families 1-3 (Table S3) unique to the *C. symbiosum* genome are highlighted. Sampling parameters, including the pore size of pre-filtration and collection filters used for each sample bin: (0.8-0.1 μm for SAR sample 1 and 7, 0.8-0.22 μm for SAR samples 2-4, 20-3.0 μm for SAR sample 5, and 3-0.8 μm for SAR sample 6) and collection dates (2/25/2003 for SAR samples 3-4, 2/26/2003 for SAR samples 1-2 and 5/15/2003 for SAR samples 5-7) (53). The late February collection dates for SAR samples 1-4 were at a time of deeper nutrient-rich water mixing with surface waters.

References

1. **Woese, C. R., Kandler, O. & Wheelis, M. L. (1990)** ***Proc Natl Acad Sci U S A*** **87, 4576-9.**
2. **Woese, C. R., Magrum, L. J. & Fox, G. E. (1978)** ***J Mol Evol*** **11, 245-51.**
3. **DeLong, E. F. (1992)** ***Proc Natl Acad Sci U S A*** **89, 5685-9.**
4. **Fuhrman, J. A., McCallum, K. & Davis, A. A. (1992)** ***Nature*** **356, 148-9.**
5. **MacGregor, B. J., Moser, D. P., Alm, E. W., Nealson, K. H. & Stahl, D. A. (1997)** ***Appl Environ Microbiol*** **63, 1178-81.**
6. **Munson, M. A., Nedwell, D. B. & Embley, T. M. (1997)** ***Appl Environ Microbiol*** **63, 4729-33.**
7. **Vetriani, C., Jannasch, H. W., MacGregor, B. J., Stahl, D. A. & Reysenbach, A. L. (1999)** ***Appl Environ Microbiol*** **65, 4375-84.**
8. **Bintrim, S. B., Donohue, T. J., Handelsman, J., Roberts, G. P. & Goodman, R. M. (1997)** ***Proc Natl Acad Sci U S A*** **94, 277-82.**
9. **Buckley, D. H., Graber, J. R. & Schmidt, T. M. (1998)** ***Appl Environ Microbiol*** **64, 4333-9.**
10. **Karner, M. B., DeLong, E. F. & Karl, D. M. (2001)** ***Nature*** **409, 507-10.**
11. **Massana, R., Murray, A. E., Preston, C. M. & DeLong, E. F. (1997)** ***Appl Environ Microbiol*** **63, 50-6.**
12. **Damste, J. S., Schouten, S., Hopmans, E. C., van Duin, A. C. & Geenevasen, J. A. (2002)** ***J Lipid Res*** **43, 1641-51.**
13. **DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N. U., Martinez, A., Sullivan, M. B., Edwards, R., Brito, B. R., Chisholm, S. W. & Karl, D. M. (2006)** ***Science*** **311, 496-503.**
14. **Pearson, A., McNichol, A. P., Benitez-Nelson, B. C., Hayes, J. M. & Eglinton, T. I. (2001)** ***Geochemica et Cosmochimica Acta*** **65, 3123-3137.**
15. **Wuchter, C., Schouten, S., Boschker, H. T. & Sinninghe Damste, J. S. (2003)** ***FEMS Microbiol Lett*** **219, 203-7.**
16. **Herndl, G. J., Reinthaler, T., Teira, E., van Aken, H., Veth, C., Pernthaler, A. & Pernthaler, J. (2005)** ***Appl Environ Microbiol*** **71, 2303-9.**
17. **Ingalls, A. E., Shah, S. R., Hansman, R. L., Aluwihare, L. I., Santos, G. M., Druffel, E. R. & Pearson, A. (2006)** ***Proc Natl Acad Sci U S A*** **103, 6442-7.**
18. **Konneke, M., Bernhard, A. E., de la Torre, J. R., Walker, C. B., Waterbury, J. B. & Stahl, D. A. (2005)** ***Nature*** **437, 543-546.**
19. **Hallam, S. J., Mincer, T. J., Schleper, C., Preston, C. M., Roberts, K., Richardson, P. M. & DeLong, E. F. (2006)** ***PLoS Biol*** **4, e95.**
20. **Treusch, A. H., Kletzin, A., Raddatz, G., Ochsenreiter, T., Quaiser, A., Meurer, G., Schuster, S. C. & Schleper, C. (2004)** ***Environ Microbiol*** **6, 970-80.**
21. **Treusch, A. H., Leininger, S., Kletzin, A., Schuster, S. C., Klenk, H. P. & Schleper, C. (2005)** ***Environ Microbiol*** **7, 1985-95.**
22. **Schleper, C., Jurgens, G. & Jonuscheit, M. (2005)** ***Nat Rev Microbiol*** **3, 479-88.**
23. **Preston, C. M., Wu, K. Y., Molinski, T. F. & DeLong, E. F. (1996)** ***Proc Natl Acad Sci U S A*** **93, 6241-6.**
24. **Preston, C. M. (1998) in** ***Ecology, Evolution, and Marine Biology*** **(University of California Santa Barbara, Santa Barbara), pp. 210.**

25. Schleper, C., Holben, W. & Klenk, H. P. (1997) *Appl Environ Microbiol* 63, 321-3.
26. Schleper, C., DeLong, E. F., Preston, C. M., Feldman, R. A., Wu, K. Y. & Swanson, R. V. (1998) *J Bacteriol* 180, 5003-9.
27. Konstantinidis, K. T. & Tiedje, J. M. (2005) *J Bacteriol* 187, 6258-64.
28. Konstantinidis, K. T. & Tiedje, J. M. (2005) *Proc Natl Acad Sci U S A* 102, 2567-72.
29. Zhang, R. & Zhang, C. T. (2005) *Archaea* 1, 335-46.
30. Kelman, Z. (2000) *Trends Biochem Sci* 25, 521-3.
31. Mattar, S., Scharf, B., Kent, S. B., Rodewald, K., Oesterhelt, D. & Engelhard, M. (1994) *J Biol Chem* 269, 14939-45.
32. Scharf, B. & Engelhard, M. (1993) *Biochemistry* 32, 12894-900.
33. Graf, J. & Ruby, E. G. (1998) *Proc Natl Acad Sci U S A* 95, 1818-22.
34. Kohl, D. H., Schubert, K. R., Carter, M. B., Hagedorn, C. H. & Shearer, G. (1988) *Proc Natl Acad Sci U S A* 85, 2036-40.
35. Maden, B. E. (2000) *Biochem J* 350 Pt 3, 609-29.
36. White, R. H. (1997) *J Bacteriol* 179, 3374-7.
37. Schleper, C., Swanson, R. V., Mathur, E. J. & DeLong, E. F. (1997) *J Bacteriol* 179, 7803-11.
38. Cubonova, L., Sandman, K., Hallam, S. J., Delong, E. F. & Reeve, J. N. (2005) *J Bacteriol* 187, 5482-5.
39. Beja, O., Koonin, E. V., Aravind, L., Taylor, L. T., Seitz, H., Stein, J. L., Bensen, D. C., Feldman, R. A., Swanson, R. V. & DeLong, E. F. (2002) *Appl Environ Microbiol* 68, 335-45.
40. Marck, C. & Grosjean, H. (2003) *Rna* 9, 1516-31.
41. Tocchini-Valentini, G. D., Fruscoloni, P. & Tocchini-Valentini, G. P. (2005) *Proc Natl Acad Sci U S A* 102, 8933-8.
42. Yoshinari, S., Fujita, S., Masui, R., Kuramitsu, S., Yokobori, S., Kita, K. & Watanabe, Y. (2005) *Biochem Biophys Res Commun* 334, 1254-9.
43. Calvin, K., Hall, M. D., Xu, F., Xue, S. & Li, H. (2005) *J Mol Biol* 353, 952-60.
44. Yoshinari, S., Itoh, T., Hallam, S. J., Delong, E. F., Yokobori, S. I., Yamagishi, A., Oshima, T., Kita, K. & Watanabe, Y. I. (2006) *Biochem Biophys Res Commun*.
45. Forchhammer, K., Rucknagel, K. P. & Bock, A. (1990) *J Biol Chem* 265, 9346-50.
46. Bock, A., Forchhammer, K., Heider, J., Leinfelder, W., Sawers, G., Veprek, B. & Zinoni, F. (1991) *Mol Microbiol* 5, 515-20.
47. Rother, M., Resch, A., Wilting, R. & Bock, A. (2001) *Biofactors* 14, 75-83.
48. Matte-Tailliez, O., Brochier, C., Forterre, P. & Philippe, H. (2002) *Mol Biol Evol* 19, 631-9.
49. Brochier, C., Gribaldo, S., Zivanovic, Y., Confalonieri, F. & Forterre, P. (2005) *Genome Biol* 6, R42.
50. Brochier, C., Forterre, P. & Gribaldo, S. (2005) *BMC Evol Biol* 5, 36.
51. Sandler, S. J., Hugenholtz, P., Schleper, C., DeLong, E. F., Pace, N. R. & Clark, A. J. (1999) *J Bacteriol* 181, 907-15.
52. Barns, S. M., Delwiche, C. F., Palmer, J. D. & Pace, N. R. (1996) *Proc Natl Acad Sci U S A* 93, 9188-93.
53. Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R.,

Baden-Tillson, H., Pfannkoch, C., Rogers, Y. H. & Smith, H. O. (2004) *Science* 304, 66-74.

54. Lawrence, J. G. & Ochman, H. (1998) *Proc Natl Acad Sci U S A* 95, 9413-7.

55. Cohan, F. M. (2001) *Syst Biol* 50, 513-24.

56. Cohan, F. M. (2002) *Annu Rev Microbiol* 56, 457-87.

57. Biddle, J. F., Lipp, J. S., Lever, M. A., Lloyd, K. G., Sorensen, K. B., Anderson, R., Fredricks, H. F., Elvert, M., Kelly, T. J., Schrag, D. P., Sogin, M. L., Brenchley, J. E., Teske, A., House, C. H. & Hinrichs, K. U. (2006) *Proc Natl Acad Sci U S A* 103, 3846-51.

58. Inagaki, F., Nunoura, T., Nakagawa, S., Teske, A., Lever, M., Lauer, A., Suzuki, M., Takai, K., Delwiche, M., Colwell, F. S., Nealson, K. H., Horikoshi, K., D'Hondt, S. & Jorgensen, B. B. (2006) *Proc Natl Acad Sci U S A* 103, 2815-20.

59. Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. (1996) *J Bacteriol* 178, 591-9.

60. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A. & Barrell, B. (2000) *Bioinformatics* 16, 944-945.

61. Lowe, T. & Eddy, S. (1997) *Nucleic Acids Research* 25, 955-964.

62. Ghai, R., Hain, T. & Chakraborty, T. (2004) *BMC Bioinformatics* 5, 198.

63. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J Mol Biol* 215, 403-10.

64. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res* 22, 4673-80.

65. Rasko, D. A., Myers, G. S. & Ravel, J. (2005) *BMC Bioinformatics* 6, 2.

66. Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C. & Salzberg, S. L. (2004) *Genome Biol* 5, R12.

67. Guindon, S. & Gascuel, O. (2003) *Syst Biol* 52, 696-704.

68. Ronquist, F. & Huelsenbeck, J. P. (2003) *Bioinformatics* 19, 1572-4.

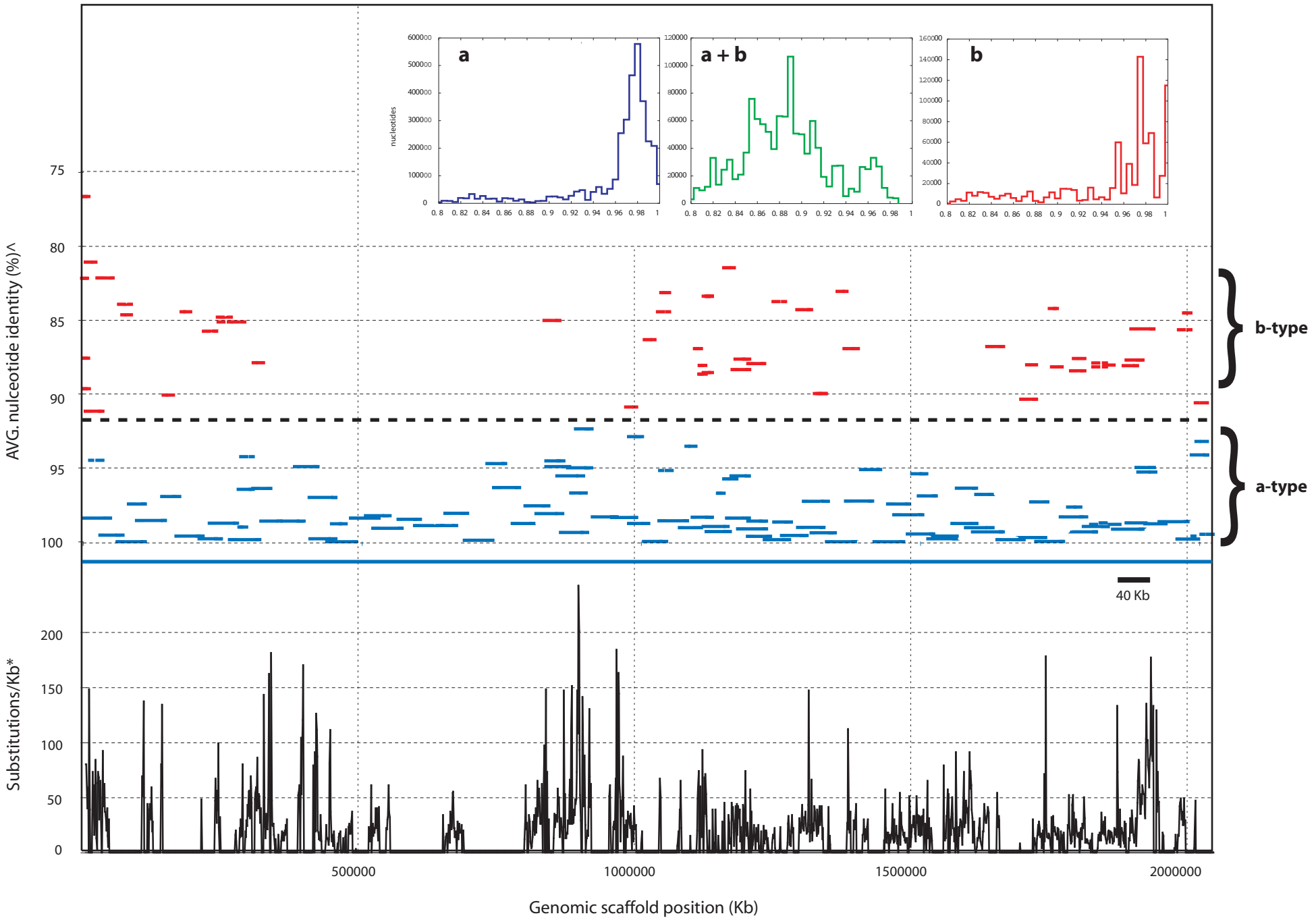
Table 1. *C. Symbiosum* genome features

Specifications*		
Size (bp)		2,045,089
Average G+C Content (%)		57.74
Predicted open reading frames (ORFs)		2,066
ORF Density (gene/Kb)		0.986
Average ORF length (bp)		924
Coding percentage (%)		91.2
ORF Content		
Predicted functional		1,070
Predicted functional in COGs		1,024
Conserved hypothetical		86
Hypothetical		861
RNA genes		
	16S-23S rRNA operon	1
	5S rRNA	1
	tRNAs	45
	7S RNA	1
Expanded gene families †		
Number of families		79
Number of genes in families		263
Coding percentage (%)		26.78
Taxonomic Distribution of functional and conserved ORFs^		
Archaea		64.41
Bacteria		32.68
Eukarya		2.87
Virus		0.04

* see methods for fosmid assembly parameters

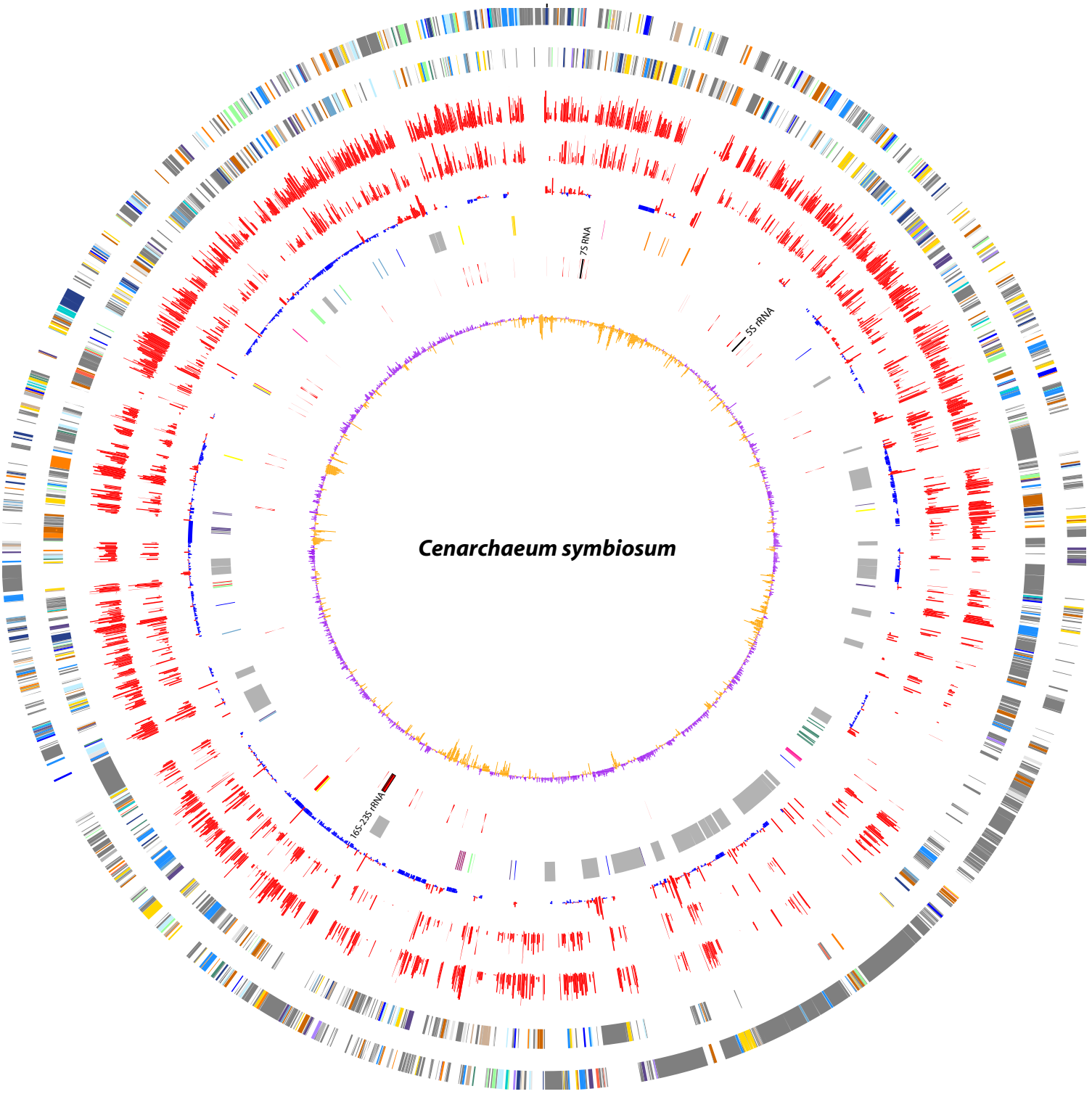
† based on following cutoffs: expection ≥ 1e-20, bitscore ≥ 100, identity ≥ 40% and overlap ≥ 100 aa

^ GenBank queries based on blastp searches constrained to expectation cut-off ≥ 1e-10



^ Dashed line represents identity cut-off (~92% average nucleotide identity) for type-a fosmids used in tiling path construction

* Sequence divergence within type-a fosmids based on comparison of fosmids sharing >95% average nucleotide identity



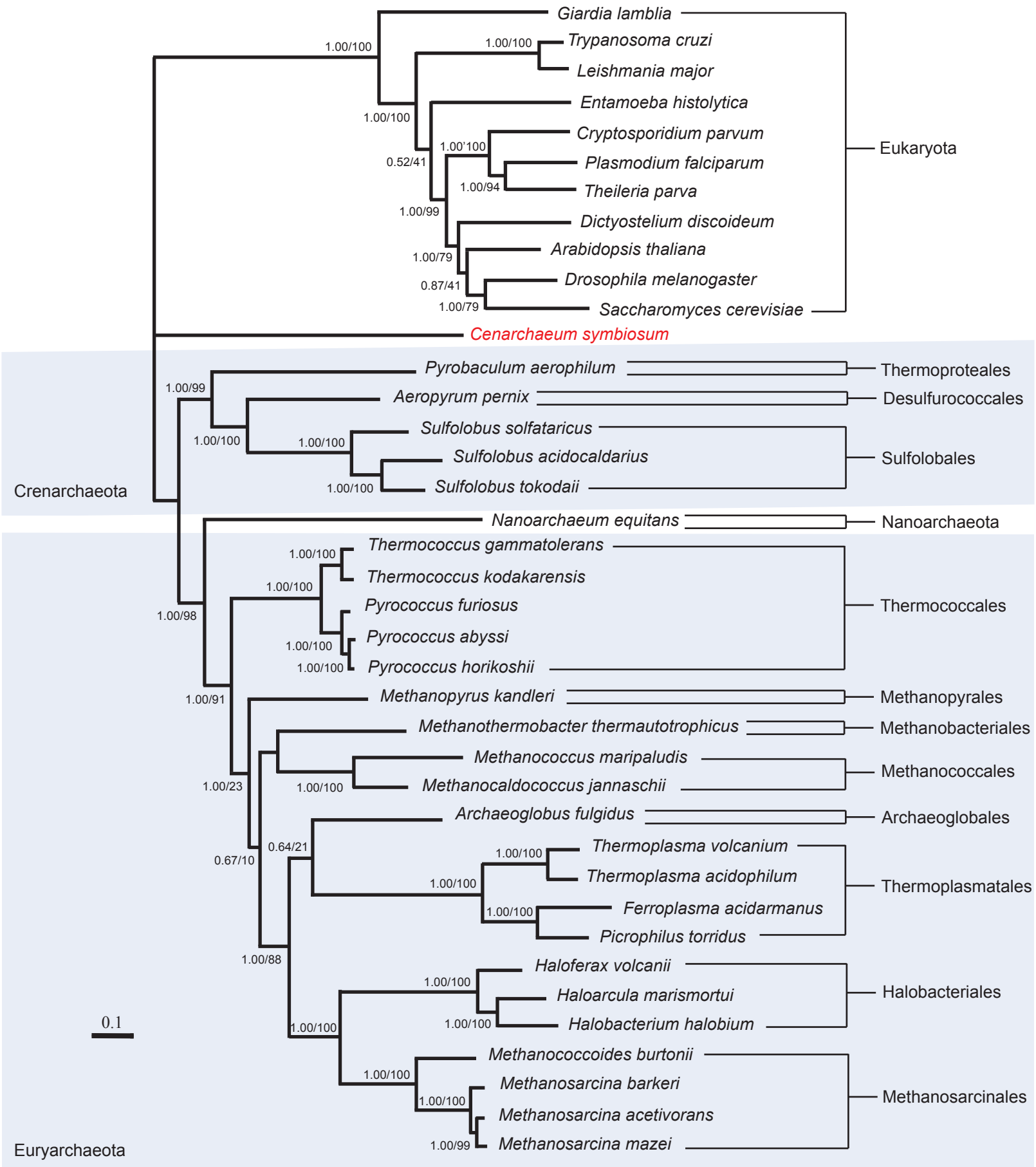
COG Category

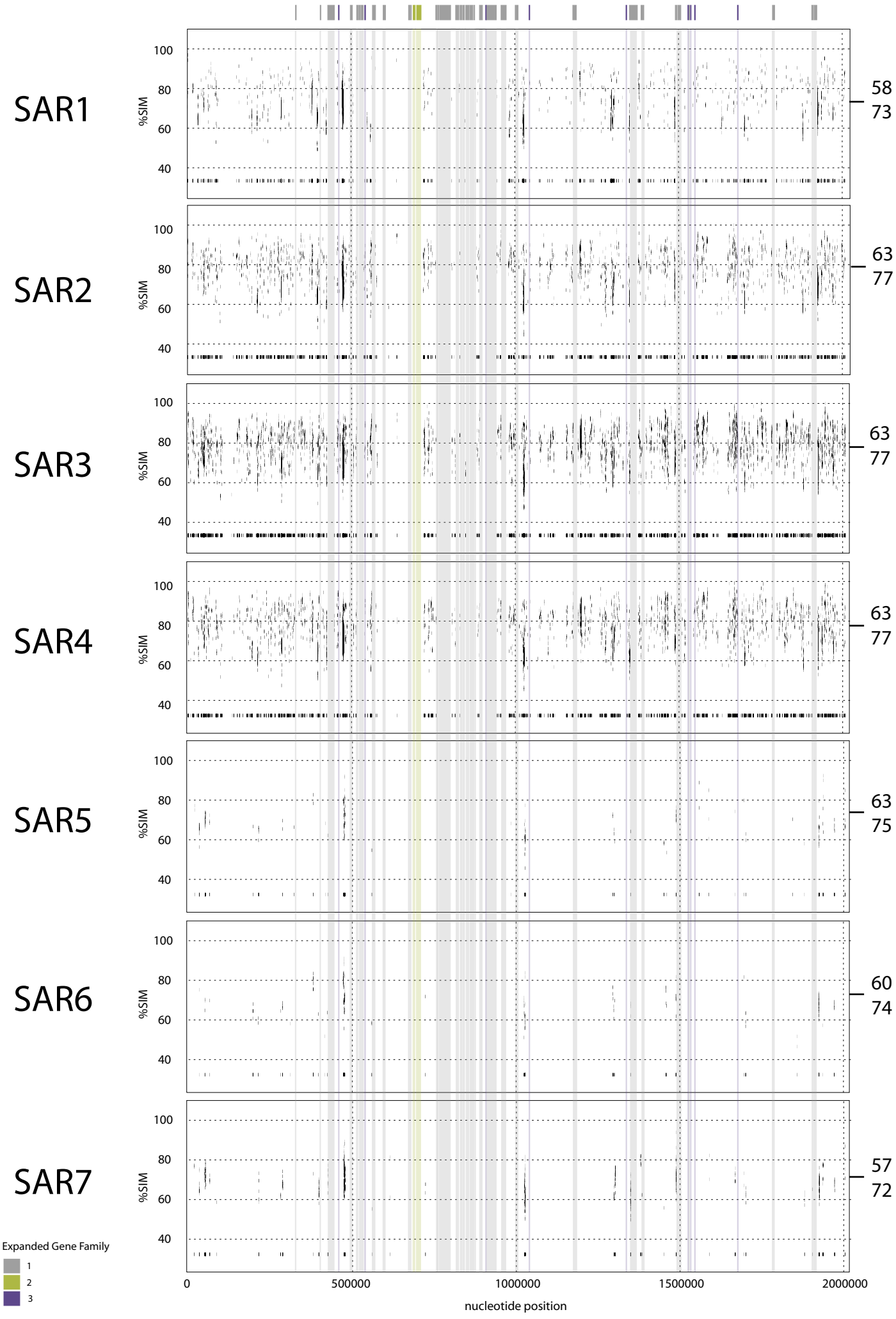
J	Translation,ribosomal structure and biogenesis		G	Carbohydrate transport and metabolism
K	Transcription		E	Amino acid transport and metabolism
L	DNA replication, recombination and repair		F	Nucleotide transport and metabolism
D	Cell division and chromosome partitioning		H	Coenzyme metabolism
T	Signal transduction mechanisms		I	Lipid metabolism
M	Cell envelope biogenesis, outer membrane		P	Inorganic ion transport and metabolism
N	Cell motility and secretion		Q	Secondary metabolites biosynthesis, transport and catabolism
U	Intracellular trafficking and secretion		R	General function prediction only
O	Posttranslational modification, protein turnover, chaperones		S	Function unknown
C	Energy production and conversion		-	Not in COGs

Expanded Gene Family

1	3	5	7	9 and RNAs	11	13
2	4	6	8	10	12	

Ribosomal Protein Concatenation (4,569 positions)




SAR1
SAR2
SAR3
SAR4
SAR5
SAR6
SAR7
%SIM
100
80
60
40
58
73
63
77
63
77
63
77
63
75
60
74
57
72
0
500000
1000000
1500000
2000000
nucleotide position
Expanded Gene Family
1
2
3