Supporting Information

Walker et al. 10.1073/pnas.0913533107

SI Materials and Methods

High-Molecular-Weight Genomic DNA Preparation. Two cultures of *Nitrosopumilus maritimus* strain SCM1 were grown as previously described in 500 mL of media in 1-L flasks (1). Cells from both cultures were harvested in late-exponential phase using Sterivex filters for one culture and a 0.1-μm filter for the other. High-molecular-weight DNA was isolated as previously described using either agarose plugs (2) or a modified guanadinium thiocynate protocol (3). Cells from the Sterivex filter were resuspended in 1 mL of 2× STE buffer [1 M NaCl, 0.1 M EDTA (pH 8.0), 10 mM Tris (pH 8.0)], extracted from the filter, and mixed with 1 vol of 1% molten SeaPlaque LMP agarose (FMC). The mixture was cooled to 40 °C, immediately drawn into a 1-mL syringe, and placed on ice for 10 min. The agarose plug was mixed with 10 mL of lysis buffer, incubated at 37 °C for 1 h, and then transferred to 40 mL of ESP buffer (1% Sarkosyl–1 mg of Proteinase K per ml in 0.5 M EDTA). After incubation at 55 °C for 16 h, the solution was replaced with fresh ESP buffer and incubated at 55 °C for another hour. DNA was purified using phenol:chloroform:isoamyl alcohol (24:24:1) and recovered by precipitation with isopropanol.

Cells collected on the 0.1-μm filter were resuspended in 100 μL of Tris-EDTA (pH 8.0) and 100 mg/mL lysozyme before incubating at 37 °C for 30 min. Then, 3.0 mL of a solution containing 5 M guanidinium thiocyanate, 100 mM EDTA (pH 8.0), and 0.5% (vol/vol) sarkosyl was added. The solution was mixed gently for 15 min before being cooled on ice for 10 min. After cooling, an equal volume of cold 7.5 M ammonium acetate was added, and the solution was mixed gently and cooled on ice. Purification and precipitation of DNA were performed with chloroform:isoamyl alcohol (24:1) and isopropanol.

Genome Sequencing. A completely sequenced and closed genome of *N. maritimus* was obtained through collaboration with the Joint Genome Institute. Whole-genome shotgun sequencing of 3-, 8-, and 40-kb DNA libraries produced at least 8× coverage of the entire genome. Specifics of clone library generation, sequencing, and assembly strategies may be found at the DOE JGI website (www.jgi.doe.gov/sequencing/index.html).

Genome Sequence Analysis. Autoannotation of the closed genome sequence was performed by both the Computational Biology group at Oak Ridge National Laboratory (http://genome.ornl.gov/microbial/nmar/02jul07/) and the TIGR Autoannotation Service (now hosted by JCVI; details available from http://www.jcvi.org/cms/research/projects/annotation-service/overview/). The genome visualization software Manatee (release 2.4.1; latest version available from http://manatee.sourceforge.net/) was used for manual curation. Analysis of potential transporter genes was performed using the Transporter Automatic Annotation Pipeline (TransAAP) through the TransportDB genomic comparison tool (membranetransport.org).

Direct comparisons with the *C. symbiosum* genome and genome fragments were performed using the Artemis Comparison Tool (3) with a comparison library generated through WebACT (www.webact.org/WebACT/home). Orthologous genes shared between these two organisms were identified through reciprocal BLAST searches, with an expectation cutoff value of 10^{-4} and a minimum of 75% alignable *N. maritimus* sequence. Comparisons with the Global Ocean Sampling (GOS) and Sargasso Sea metagenomic datasets were performed using several single-copy universal archaeal genes to determine a count of ~15 archaeal genomes in the GOS dataset. An initial set of candidate *N. maritimus*-like proteins was found by BLASTP searches with each *N. maritimus* protein-coding gene, using a cutoff of 100 hits and a maximum expected value of $e = 10$. This set consisted of 125,326 peptides drawn from 107,223 scaffolds. Neighboring ORFs on any scaffold with two or more hits were added to make a total of 319,585 peptides, amounting to 5.2% of all ORFs in GOS. These sequences were filtered by BLAST alignment to four sequence datasets, containing 24 complete proteomes from diverse euryarchaeota, crenarchaeota, and bacteria. Sequences scoring more highly to *N. maritimus* and/or *C. symbiosum* proteins than any other entry in these datasets were retained, giving a filtered dataset of 21,278 ORFs. Coverage of the *N. maritimus* proteome was measured by bidirectional BLASTP of *N. maritimus* proteins against the filtered dataset to assign putative orthologs. The average coverage was ~11×, although some highly conserved genes had a much greater number of hits, probably due to recruitment of nonarchaeal homologs: 48 ORFs had >30 hits, of which most were highly conserved. Searches for CRISPR regions were performed using the Java-based CRISPR recognition tool (CRT) with least stringent settings (4).

Maximum-Likelihood and Bayesian Trees of the Archaeal Domain. The trees are based on the concatenation of ribosomal proteins used by Brochier-Armanet et al. (5) but including sequences from *N. maritimus* and from "*Candidatus Korarchaeum cryptofilum*" OPF8. Sequences were aligned using MUSCLE (6). Resulting alignments were visually inspected and improved with the MUST software (7). Regions where homology between sites was doubtful were removed from further phylogenetic analyses. A total of 6,142 positions were kept for the phylogenetic analyses. The maximum-likelihood tree was computed with PHYML, using the WAG model corrected by a gamma law to take into account evolutionary rate among site variations (8). The parameter alpha of the gamma distribution as the proportion of invariable sites was estimated from the dataset. The robustness of each branch was estimated by the bootstrap procedure implemented in PHYML. A Bayesian tree analysis on a subset of 29 taxa was performed using MrBayes 3.2 (9) with a mixed model of amino acid substitution and a gamma distribution (eight discrete categories and an estimated proportion of invariant sites) to take into account among-site rate variation. MrBayes was run with four chains for 1 million generations and trees were sampled every 100 generations. To construct the consensus tree, the first 1,500 trees were discarded as "burn-in." The reduction of the taxonomic sampling was necessary to reduce the computation time.

1. Könneke M, et al. (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* 437:543–546.
2. Stein LY, et al. (2007) Whole-genome analysis of the ammonia-oxidizing bacterium, *Nitrosomonas eutropha* C91: Implications for niche adaptation. *Environ Microbiol* 9: 2993–3007.
3. Carver TJ, et al. (2005) ACT: The Artemis comparison tool. *Bioinformatics* 21:3422–3423.
4. Bland C, et al. (2007) CRISPR recognition tool (CRT): A tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209.
5. 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.
6. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
7. Philippe H (1993) MUST, a computer package of management utilities for sequences and trees. *Nucleic Acids Res* 21:5264–5272.
8. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
9. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.

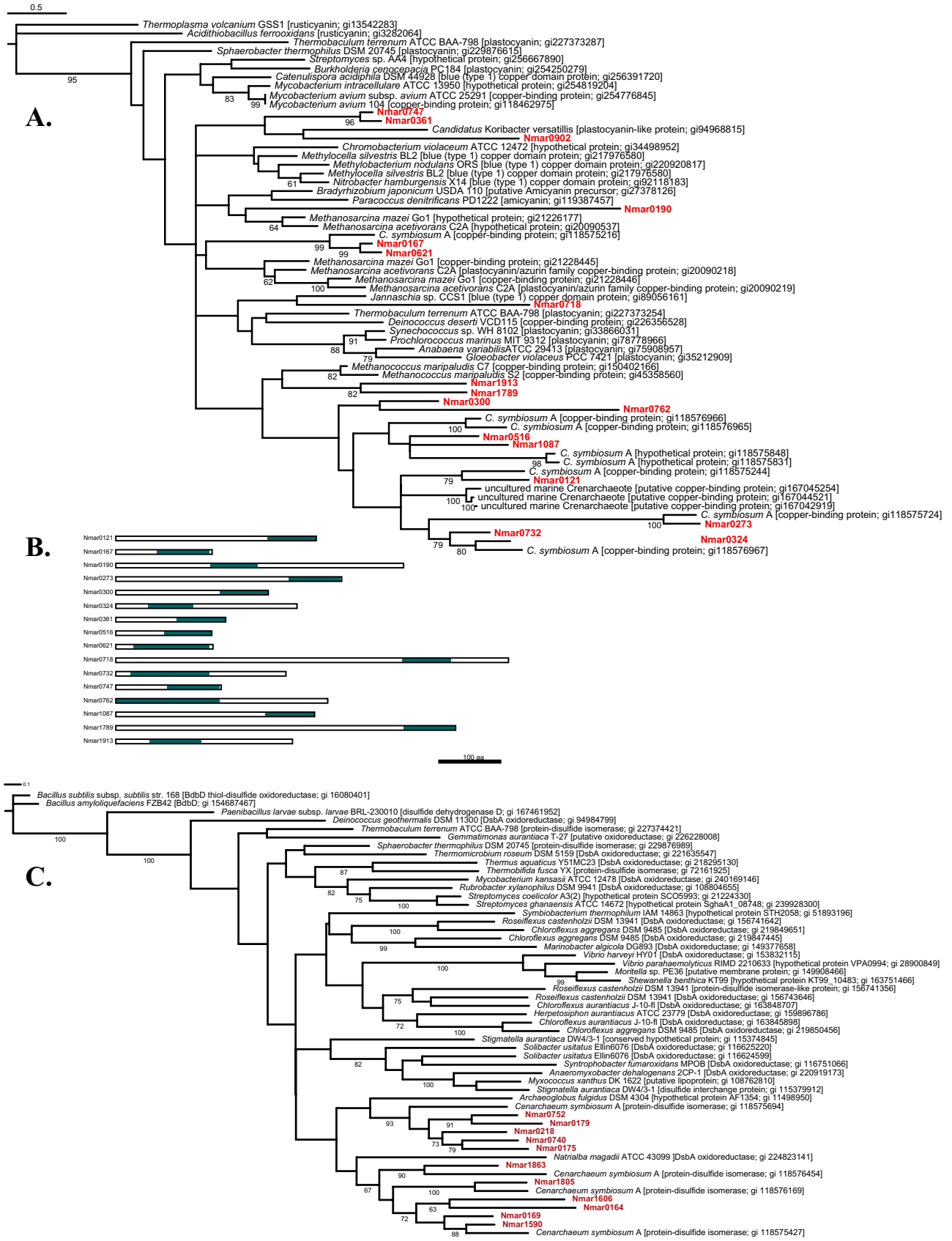

Fig. S1. Phylogeny of plastocyanin-like protein sequences. Sequences with significant matches to COG3794 (PetE: Plastocyanin [Energy production and conversion]) were used to query the non-redundant protein sequence database from NCBI. Sequences from the top 20–30 non-*N. maritimus* hits were retrieved and their match to the above conserved domain model verified. Sequences from experimentally characterized proteins were obtained from the available literature and included, aligned with ClustalW and then curated manually. Distance-based phylogenies were inferred in Phylip using the Neighbor-Joining algorithm and 100 bootstrap replicates. Bootstrap support values >60% are displayed. Nodes with <50% bootstrap support were collapsed.

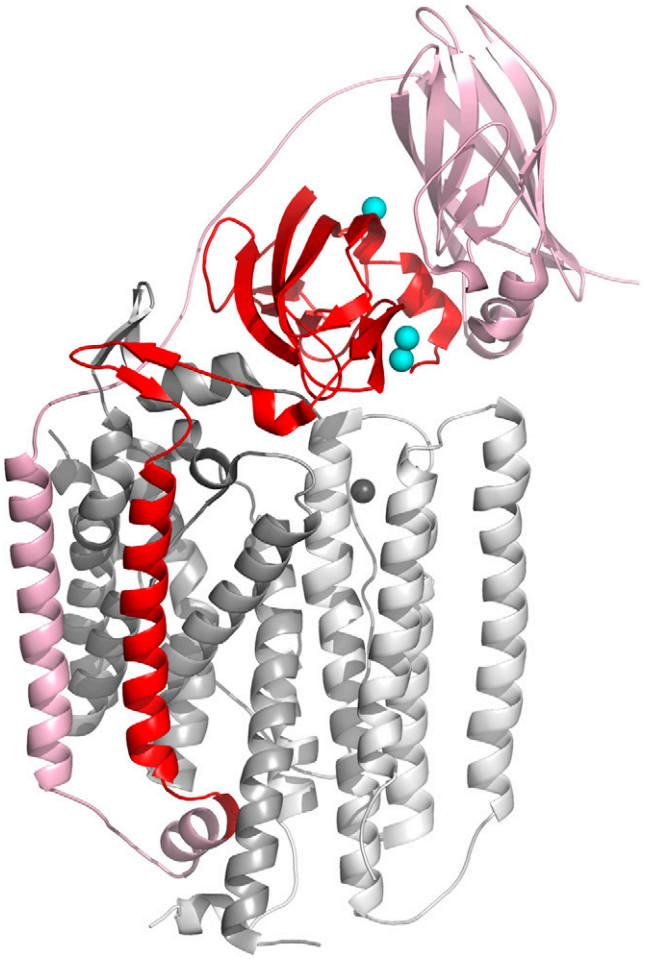

Fig. S2. Archaeal ammonia monooxygenase AmoB sequence mapped onto the crystal structure of the particulate methane monooxygenase (PDB accession code 1YEW). The pmoA subunit is shown in dark gray, the pmoC subunit in light gray, and the pmoB subunit in red and pink. The red part represents the region of pmoB conserved in the predicted *N. maritimus* AmoB. The transmembrane helix and C-terminal cupredoxin domain shown in pink are missing in the predicted *N. maritimus* AmoB. Cyan spheres represent copper ions. The grey sphere represents a zinc ion.

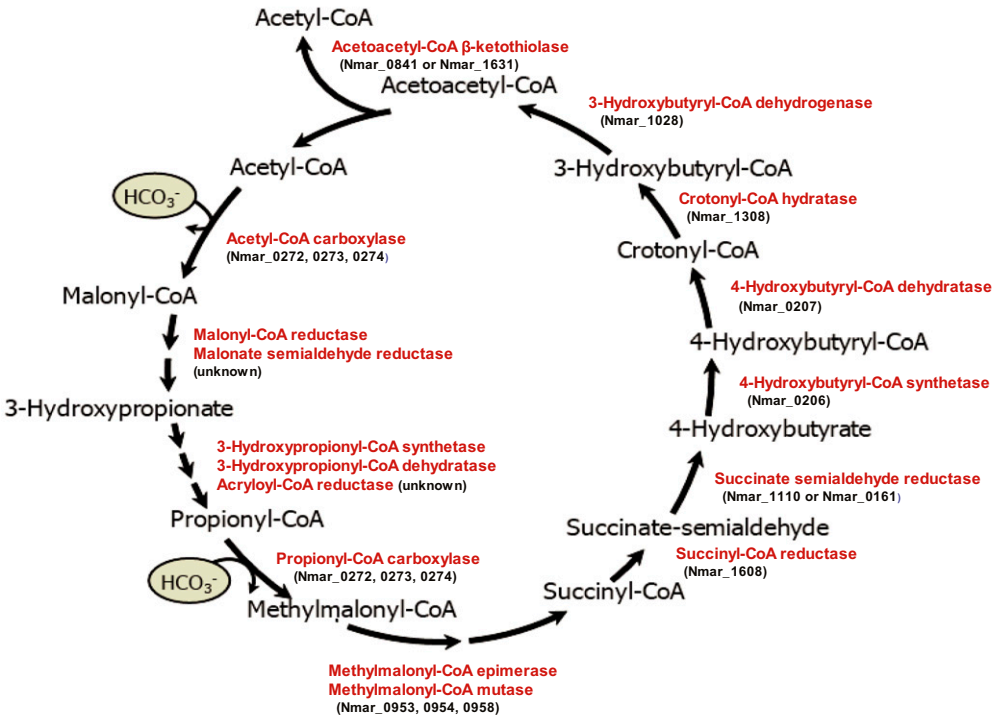

Fig. S3. Proposed 3-hydroxypropionate/4-hydroxybutyrate cycle for autotrophic carbon fixation by *N. maritimus*.

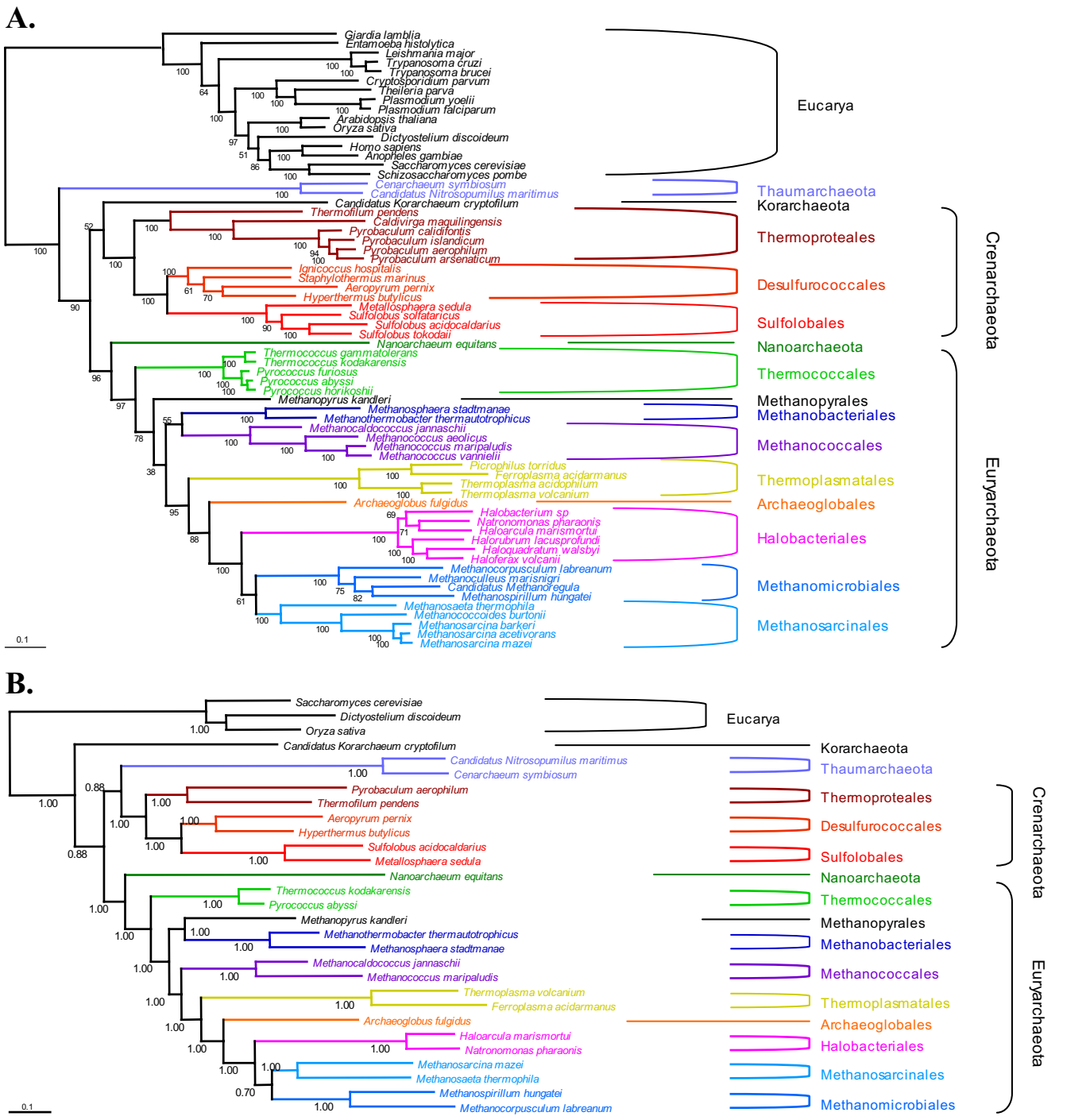

Fig. S4. (*A*) Maximum-likelihood phylogeny of Group 1 *Archaea*. The phylogeny was inferred using an alignment of concatenated R-proteins (66 taxa, 6,142 positions). WAG+Inv+Gamma (4 classes); 100 replicates. (*B*) Bayesian tree of mesophilic Group 1 *Archaea* inferred using an alignment of concatenated R-proteins (29 taxa, 6,142 positions). Mixed model + Gamma (4 classes); 100 replicates.

Other Supporting Information Files

Table S1 (DOC)
Table S2 (DOC)
Table S3 (DOC)
Table S4 (DOC)
Table S5 (DOC)
Table S6 (DOC)