Supporting Information

Zhaxybayeva et al. 10.1073/pnas.0901260106

SI Methods

Genome Sequences. Lyophilized cells of *Thermotoga petrophila* RKU-1, *Tt. lettingae* TM, *Thermosipho melanesiensis* BI429, and *Fervidobacterium nodosum* Rt17-B1 were obtained from the Deutsche Sammlung von Mikroorganismen und Zellkulturen. Liquid cultures of each were prepared and an aliquot was removed from each culture and streaked on solid medium to obtain single colonies. A single colony of each organism was used to grow cells for DNA purification. Chromosomal DNA was isolated from cells using a standard cetyl trimethylammonium bromide-based protocol.

DNA sequencing was performed at the U.S. Department of Energy's Joint Genome Institute Production Genomics Facility in Walnut Creek, CA. Sequencing was finished at the Joint Genome Institute (JGI) Bioscience Division, Los Alamos National Laboratory. Automated annotation was performed at Oak Ridge National Laboratory. Genome sequences were deposited in GenBank under the following accession numbers: CP000702 (*Tt. petrophila* RKU-1), CP000812 (*Tt. lettingae* TMO), CP000716 (*Ts. melanesiensis* BI429), and CP000771 (*F. nodosum* Rt17-B1). The *Tt. maritima* strain MSB8 genome sequence was obtained from GenBank [accession no. AE000512.1 (ref. 1)].

Reassignment of the *Tt. maritima* Strain MSB8 Nucleotide Positions and Assessment of Synteny. The *Tt. maritima* strain MSB8 genome sequence available in GenBank does not begin at the site of the *dnaA* gene as do the newly available Thermotogales genome sequences. To facilitate comparison of the *Tt. maritima* genome sequence with the other sequences the order of its ORFs was rearranged such that its *dnaA* gene was placed as its first ORF. The new first nucleotide of the genome sequence was placed 114 bp upstream of *dnaA* (position 945,687 in the AE000512.1 record). This repositioning of the start necessitated exchanging the designations of the forward and reverse strands relative to the ones in the GenBank record. To translate from the GenBank genome positions to those used in this report, the following transformations were used: $y = 945{,}687 - x$, if $x < 945{,}687$, and $y = 2{,}806{,}412 - x$, if $x \geq 945{,}687$, where x is the position of a nucleotide in AE000512.1 and y is the position of the same nucleotide in the *Tt. maritima* genome sequence considered here. Synteny among genomes was determined using MUMMER 3.20 (2) with the *mum*, *b*, and *c* flags. The rearranged nucleotide sequence file for *Tt. maritima* and the fna files of the other genomes retrieved from NCBI were used as input sequences for the alignments.

Identification of IS Elements and CRISPR Elements. Fragmented IS elements were identified using a PSI-TBLASTN approach (3, 4). Checkpoint files were created using the amino acid sequences of transposases annotated in the 5 Thermotogales genomes as queries in PSI-BLAST searches in 2 iterations against the GenBank *nr* database, using the BLASTPGP program. Separate checkpoint files were made for the IS605, IS110, IS6, and IS3 families of insertion sequences. These checkpoint files were used in TBLASTN searches of each individual genome, using the original amino acid sequences of the annotated transposases. Top-scoring hits with an *E*-value $<10^{-5}$ were then used in a BLASTN search against their original genomes to determine if they had previously been annotated. Sequences were designated as putative or fragmentary IS elements if they were not annotated as a transposase or were not assigned a putative product. In each Thermotogales genome the CRISPR elements were predicted using the CRISPR Recognition Tool version 1.1 (5).

Relationships Between RuBisCO-like protein sequences. Sequences of selected RuBisCO-like proteins belonging to group IV-Deep Ykr (6) were aligned using the default parameters of ClustalX 2.0 and the maximum-likelihood tree was reconstructed in the PhyML program (under the JTT+PINVAR substitution model). Bootstrap values from 1,000 replicates are shown. Sequences (GI numbers) used for this analysis were *Tt. lettingae* (Tlet_1684), *Beggiatoa* sp. PS (153872960), *Ochrobactrum anthropi* (153012125), *Oceanicola granulosus* (89067677), *Synechococcus* sp. CC9605 (78196756), *Rhodobacter capsulatus* (1710033), *Alkalilimnicola ehrlichei* (114320324), *Archaeoglobus fulgidus* (11499182), *Halorhodospira halophila* (121997270), *Heliobacillus mobilis* (111075030), *Rhodopseudomonas palustris* strains (115525677, 90423687, 91976710, 39649090, 86750338), *Rhodospirillum rubrum* (83576247), marine metagenome 1 (143308646, 143312839, 142156458), marine metagenome 2 (143826546), *Geobacillus kaustophilus* (56379330), *Bacillus* species (42783150, 29897680, 51974924, 56909783, 52003120, 52347782, 49333072), and *Exiguobacterium sibiricum* (68055023).

Gross Comparisons of Gene Composition. Partitioning of the Thermotogales' pan-genome was determined by clustering all of the ORFs encoded in the 5 genomes into orthologous gene families, using all-against-all BLASTP searches with an *E*-value cutoff of 10^{-20} followed by the BRANCHCLUST algorithm (7), which sorts out paralogs using phylogenetic information. We identified families of homologous genes present in the same genome but absent in other genomes (genome-specific expansions) using BRANCHCLUST.

Quartet Decomposition of 5 Thermotogales Genomes. The analyses were performed according to the quartet decomposition method described in (8). In brief, 1,115 sets of orthologous genes (gene families) in the 5 Thermotogales genomes were detected through all-against-all BLASTP searches with an *E*-value cutoff of 10^{-20} followed by BRANCHCLUST (7). A total of 944 of 1,115 families were present in all 5 genomes (core gene families), while the remaining 171 were present in any 4 of 5 genomes. The amino acid sequences of members of each gene family were aligned in ClustalW v. 1.83 (9) with default parameters and cleaned with GBLOCKS v. 0.91b (10) [with parameters $-b1$ = (50% of the number of sequences + 1), $-b2$ = (50% of the number of sequences + 1), $-b3 = 20$, $-b4 = 2$, $-b5 = h$]. Gene families were tested for compositional heterogeneity, using a χ^2 test implemented in TREE-PUZZLE v. 5.2 (11). A phylogenetic tree for each gene family was calculated, using RAxML v. 7.0.0 (12) under the JTT+Γ model with 100 bootstrap replicates. Embedded quartets (8) in each gene tree were evaluated. Plurality topology was calculated from quartets supported by plurality of gene families, using the MRP matrix obtained with CLANN v. 2.0.2 (13). Gene families were identified as conflicting with plurality if they contained at least 1 embedded quartet contradicting plurality with at least 80% bootstrap support.

Phylogenetic Tree Based on Number of Genomic Rearrangements. Five Thermotogales genomes were aligned pairwise using the MAUVE program (match seed weight = 9, minimum island size = 10, backbone size = 200, other settings default). The information on gene order from syntenic blocks in MAUVE alignments was submitted to the GRIMM server (14), calculating the number of rearrangements required to convert gene order in one genome to the order in the other one. The number of rearrangements was normalized to the number of aligned nucleotides and used as a distance measure. The phylogenetic tree was reconstructed from the distance matrix using the FITCH program of the PHYLIP package (15).

Phylogenetic Tree Based on Gene Presence/Absence. The gene content tree was built by randomly selecting 21,902 proteins from the *nr* database and searching 5 Thermotogales, 30 bacterial, and 8 archaeal genomes for the presence of homologous matches, using BLASTP searches with a bit score cutoff of 50. A matrix with presence of matches coded as "A" and absence of matches coded as "C" was constructed. The matrix contained 7,435 variable and 14,467 constant sites. A maximum-likelihood tree was reconstructed from the data in the matrix using the PhyML program version 2.4.4 (16) under the F81+Γ model (with 4 rate categories) (17, 18) and with 100 bootstrap replicates.

Taxonomic Affiliations of Top-Scoring BLAST Hits of Thermotogales ORFs. Top-scoring BLAST hits were determined by performing a BLASTP search of an ORF against the *nr* database (downloaded on August 29, 2008). The highest-ranking match that is not a member of the Thermotogales and has an *E*-value $<10^{-4}$ was retrieved. The taxonomic assignments were done using the NCBI Taxonomy database. To test if the low number of top-scoring BLAST hits between the Thermotogales and the Aquificales is an artifact of underrepresentation of Aquificales genes in GenBank, we created a reduced *nr* database by removal of all Clostridiales protein sequences and addition of 2 randomly chosen Clostridiales genomes [10 replicates were generated, see Table S3]. The assignment of taxonomic affiliations of top-scoring hits to ORFs in the *Tt. maritima* genome was repeated for each database replicate as described above.

Phylogenetic Assessment of Relationships Between Genes in Thermotogales Genomes and Their Homologs in Archaea, Aquificales, and Clostridia. Six genomes representing Archaea (*Sulfolobus solfataricus* P2 and *Pyrococcus furiosus* DSM3638), Clostridia (*Clostridium thermocellum* ATCC27405 and *Thermoanaerobacter tengcongensis* MB4, both thermophilic), and Aquificales (*Aquifex aeolicus* VF5 and *Sulfurihydrogenibium* sp.YO3AOP1) were used as a database in BLASTP searches [with *E*-value $<10^{-4}$ and database size ($-z$ option) set to 20 million] of individual ORFs from each of 5 Thermotogales genomes. The top-scoring hit from each genome was retained. Only genes with homologs in all 3 groups (Archaea, Aquificales, and Clostridia) were used in further analyses. The data sets (varying in size from 4 to 7 homologous sequences) were aligned in ClustalW version 1.83 and phylogenetic trees were reconstructed in NEIGHBOR, using distances obtained with TREE-PUZZLE 5.2 (11) under a JTT+Γ model (100 bootstrap replicates). For each data set, embedded quartets reflecting the relationships between Thermotogales ORFs and their homologs in 3 taxonomic groups were evaluated. Each scenario received a score calculated as a ratio of the number of embedded quartets supporting the scenario with at least 80% bootstrap support to the total number of evaluated quartets (if a data set supported quartets for different scenarios, the score of each scenario would be less than one.

Thermoadaptation of Proteins. Absolute differences between charged (KRDE) and polar (NQST) amino acid residues (CvP bias) (19) and IVYWREL amino acid bias (20) of predicted proteins in each of 5 genomes were calculated using in-house Perl scripts. Only proteins with fewer than two predicted transmembrane helices were used, as determined using the TMAP program of the EMBOSS package (21).

Ancestral Protein Reconstruction. For each core gene family in Thermotogales, top-scoring hits in BLASTP searches with an *E*-value cutoff of 10^{-10} (4) were obtained from *A. fulgidus*, *S. solfataricus* P2, *A. aeolicus*, and *C. phytofermentans* genomes (available at NCBI's FTP site) and these homologs were added as an outgroup, forming extended gene families. These extended gene families were aligned in ClustalW v. 1.83 (9) with default parameters, and phylogenetic trees were constructed using TREE-PUZZLE v. 5.2 (11) (JTT+Γ model) and the NEIGHBOR program of the PHYLIP package (15). A total of 482 gene families that contained at least 2 outgroup sequences (with additional requirement that the outgroup formed a monophyletic group with at least 50% bootstrap support) were further used to determine ancestral protein sequence of the most recent common ancestor of the Thermotogales for each gene. Ancestral sequences were reconstructed using two methods: marginal reconstruction as implemented in PAML v. 4.0 (22) and ANCESCON reconstruction (with $-O$ option and no optimization of π vector) (23). A total of 423 of 482 gene families did not have family members with more than 1 predicted transmembrane helix and were used to calculate the CvP value (as described above) for ancestral proteins.

GC Content of Ancestral rRNA Gene Sequences. A 16S rRNA gene alignment, based on secondary structure containing several Thermotogales lineages and outgroup sequences, was obtained from the European ribosomal RNA database (24). Additional Thermotogales sequences were added to the alignment manually. 23S rRNA sequences were aligned using the MUSCLE program (25). Poorly aligned regions and gaps in both alignments were deleted. 16S and 23S rRNA alignments were concatenated. For taxa without 23S rRNA sequence data available, the 23S rRNA portion was annotated as missing data. A phylogenetic tree was calculated in PhyML v. 2.4.5 (16) under the HKY85+Γ+PINVAR model with 100 bootstrap samples and ancestral sequences at all nodes were determined using marginal reconstruction as implemented in PAML v. 4.0 (22). The GC content of ancestral sequences was calculated using an in-house Perl script. The tree branches were colored according to GC content, using FigTree v.1.1.2 (http://tree.bio.ed.ac.uk/software/figtree/).

Phylogenetic Analysis of Concatenated Ribosomal Proteins Data Set. Sequences for 29 universally conserved ribosomal proteins were collected from the GenBank (26) database for 41 completed bacterial genomes with a wide phylogenetic distribution. The MUSCLE program (25) was used to perform multiple sequence alignments for each ribosomal protein data set, which were then concatenated. A maximum-likelihood tree was reconstructed in the PhyML program (16) under the JTT+Γ substitution model, with 4 rate categories, estimated α, and 100 bootstrapped replicates. Fitch–Margoliash tree reconstruction was performed using the program FITCH from the PHYLIP package (one jumble, all other settings default) (15), using parameter values generated in TREE-PUZZLE (11) (α = 1.13). One hundred bootstrap replicates of the data set were produced using the SEQBOOT application of the PHYLIP program package (15). Distance matrices were then generated using TREE-PUZZLE. Trees were combined into a consensus tree with support values, using the CONSENSE application of the PHYLIP package. Bayesian analysis was performed using the PhyloBayes program version 2.3 under the CAT+Γ model (27), with posterior probabilities taken from the consensus trees of 2 convergent chains (1,000-tree burn-in; the remaining 3,291 trees from chain 1 and 3,255 trees from chain 2 reported maxdiff = 0.095). For all tree reconstructions, support values were mapped to the maximum-likelihood tree at congruent nodes.

Phylogenetic Analyses of Individual Ribosomal Proteins. Phylogenetic trees of individual ribosomal proteins were reconstructed under the JTT+Γ (4 rate categories, α estimated) + PINVAR model in the PhyML program (16). Bipartitions with at least 70% bootstrap support in each of 29 individual gene trees were compared to those from the concatenated tree (see above).

Slow–Fast Analyses of Concatenated Ribosomal Protein Alignment. The alignment was divided into 3 overlapping subsets, depending on the site conservation: sites that vary by at least 25% in amino acid composition (i.e., highly conserved sites excluded, 2,407 sites), at least 50% (429 sites), and at least 75% (i.e., only most variable sites included, 7 sites). The latter subset was not further analyzed due to the low number of sites. Phylogenetic trees were then reconstructed in PhyML under the WAG+Γ (4 rate categories, estimated α) + PINVAR model and in GARLI v. 0.96 (http://garli.nescent.org/) under the WAG+Γ+PINVAR model with 15,000 generations.

Slow–Fast Analyses of Concatenated Genes That Group Thermotogales Closer to Clostridia. A total of 138 gene families that supported the position of *F. nodosum* closer to Clostridia in at least one embedded quartet (Fig. 3) were concatenated. The alignment was divided into 2 nonoverlapping subsets, one containing sites with no more than 2 different amino acids per site (slow sites) and the remainder of the alignment (fast sites). Phylogenetic trees (100 bootstrap replicates) for 2 alignment subsets were reconstructed in the NEIGHBOR program from the PHYLIP package, using ML distances obtained from TREE-PUZZLE under the JTT+Γ model.

1. Nelson KE, et al. (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of Thermotoga maritima. *Nature* 399:323–329.
2. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
3. Natale DA, et al. (2000) Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). *Genome Biol.* 1:RESEARCH0009.
4. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
5. Bland C, et al. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209.
6. Tabita FR, et al. (2007) Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. *Microbiol Mol Biol Rev* 71:576–599.
7. Poptsova MS, Gogarten JP (2007) BranchClust: a phylogenetic algorithm for selecting gene families. *BMC Bioinformatics* 10:120.
8. Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT (2006) Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. *Genome Res* 16:1099–1108.
9. Thompson J, Higgins D, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22:4673–4680.
10. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552.
11. Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* 18:502–504.
12. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
13. Creevey CJ, McInerney JO (2005) Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* 21:390–392.
14. Tesler G (2002) GRIMM: genome rearrangements web server. *Bioinformatics* 18:492–493.
15. Felsenstein J (2005) *PHYLIP (Phylogeny Inference Package).* Distributed by the author (Department of Genome Sciences, University of Washington, Seattle).
16. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
17. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* 17:368–376.
18. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* 39:306–314.
19. Suhre K, Claverie JM (2003) Genomic correlates of hyperthermostability, an update. *J Biol Chem* 278:17198–17202.
20. Zeldovich KB, Berezovsky IN, Shakhnovich EI (2007) Protein and DNA sequence determinants of thermophilic adaptation. *PLoS Comput Biol* 3:e5.
21. Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet* 16:276–277.
22. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24:1586–1591.
23. Cai W, Pei J, Grishin NV (2004) Reconstruction of ancestral protein sequences and its applications. *BMC Evol Biol* 4:33.
24. Wuyts J, Perriere G, Van de Peer Y (2004) The European ribosomal RNA database. *Nucleic Acids Res* 32:D101–D103.
25. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
26. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank. *Nucleic Acids Res* 34:D16–D20.
27. Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* 21:1095–1109.

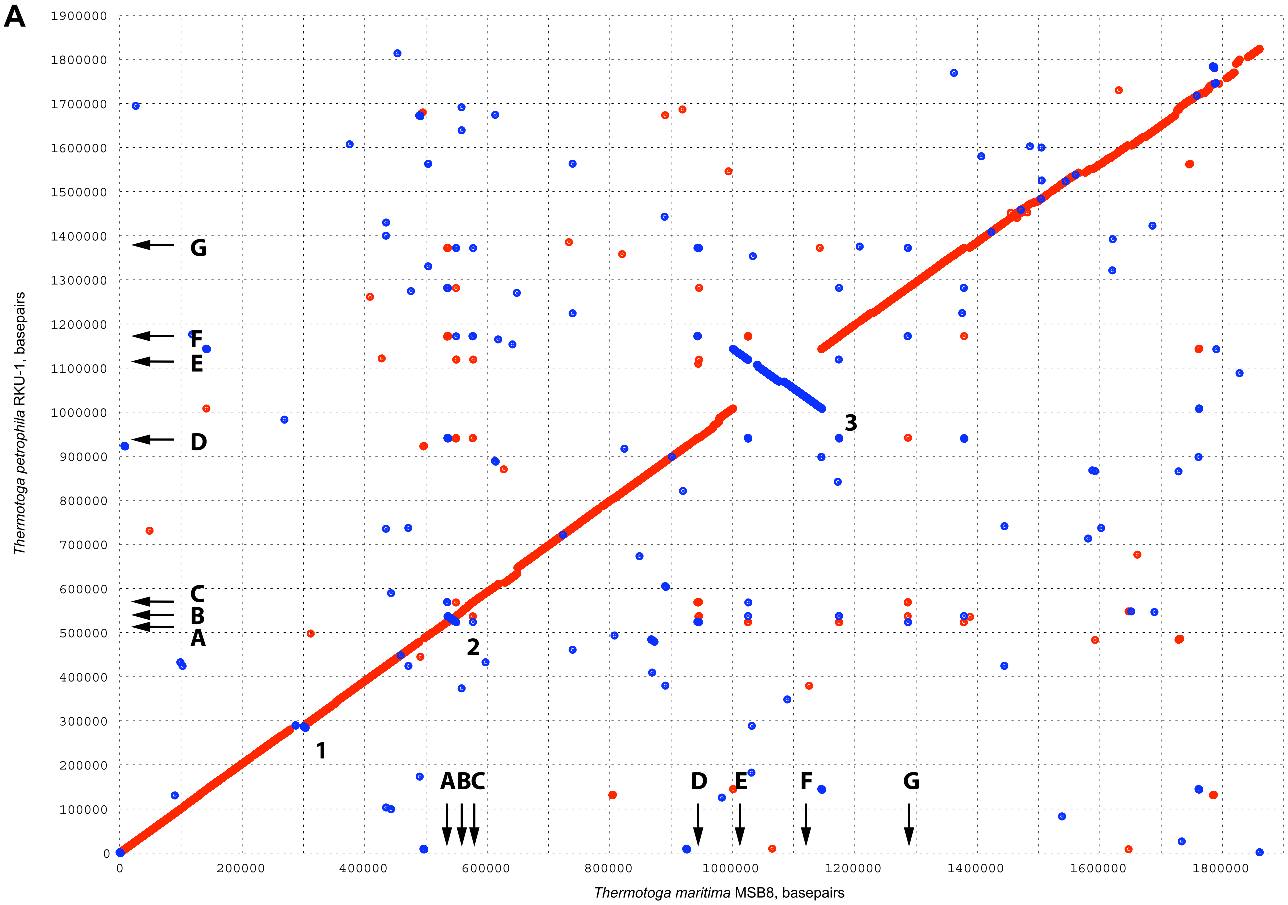

Fig. S1. Synteny among Thermotogales genomes. (*A*) Synteny between *Tt. maritima* and *Tt. petrophila* genomes. A dot plot was generated using the MUMmer program. The 3 largest inversions are marked by numbers and the locations of CRISPR elements in each genome are designated with letters (see Table S2). Inversion 1 contains 9 and 16 genes in *Tt. maritima* and *Tt. petrophila*, respectively. Inversion 2 contains 11 and 10 genes, respectively, and is flanked by CRISPR sequences in both genomes (indicated as A and B). Inversion 3 is the largest of the 3 inversions, 138 and 158 genes, respectively, and is flanked in both organisms by transposases from the IS605 family with amino acid sequences at least 98% identical. An IS200 family transposon is downstream of inversion 3 in the *Tt. petrophila* genome, while an IS200 transposon is upstream of inversion 3 in *Tt. maritima*. Red color indicates matches on the same strand in both genomes and blue those on different strands. (*B*) Gene dot plots showing remaining pairwise comparisons of Thermotogales genomes. ORFs were obtained from RefSeq records at NCBI's FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/). The axes are scaled with respect to nucleotide positions of the genomes, and each ORF is plotted at its central nucleotide position. Red +'s give matches with *E*-values $<10^{-21}$, and blue x's indicate top-scoring BLAST hits (*E*-value cutoff of 10^{-21}), if the ORF indicated on the *x*-axis was used to search the genome represented on the *y*-axis. BLAST searches were performed using blastall v. 2.2.17. Dot plots were generated with gnuplot 4.0 (http://www.gnuplot.info), using in-house Perl scripts.

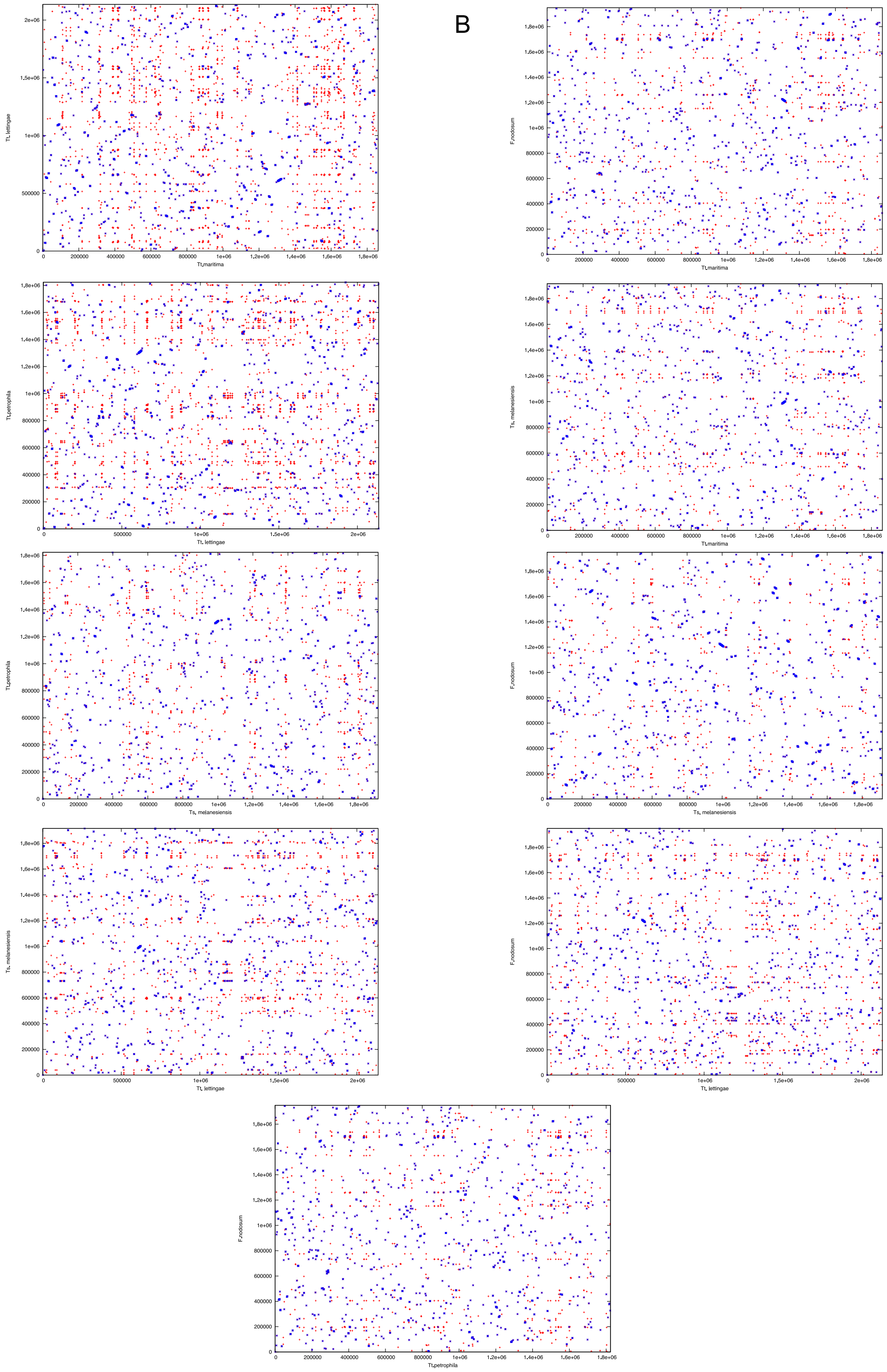

Fig. S1. (*Continued*).

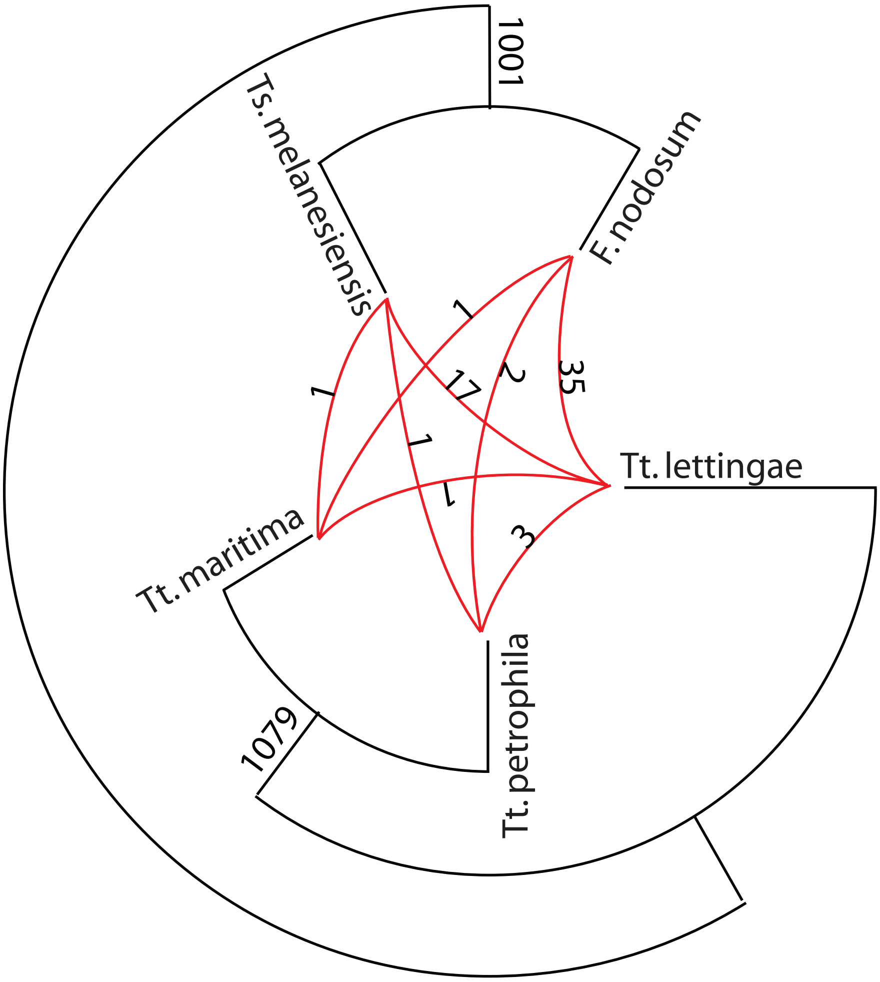

Fig. S2. Phylogenetic relationships among 5 Thermotogales genomes. The backbone topology is a tree reconstructed from embedded quartets supported by a plurality of gene families. The numbers on the node indicate how many gene families supported that branch at ≥80% bootstrap support. Lines connecting branches indicate how many gene families supported alternative branching (with ≥80% bootstrap support). The list of conflicting gene families is available upon request. The same topology was obtained from gene presence–absence and analyses based on the number of genomic rearrangements (not shown). The strong tree-like signal was not a surprising result, given that 2 of the 5 genomes are very closely related compared to the other 3.

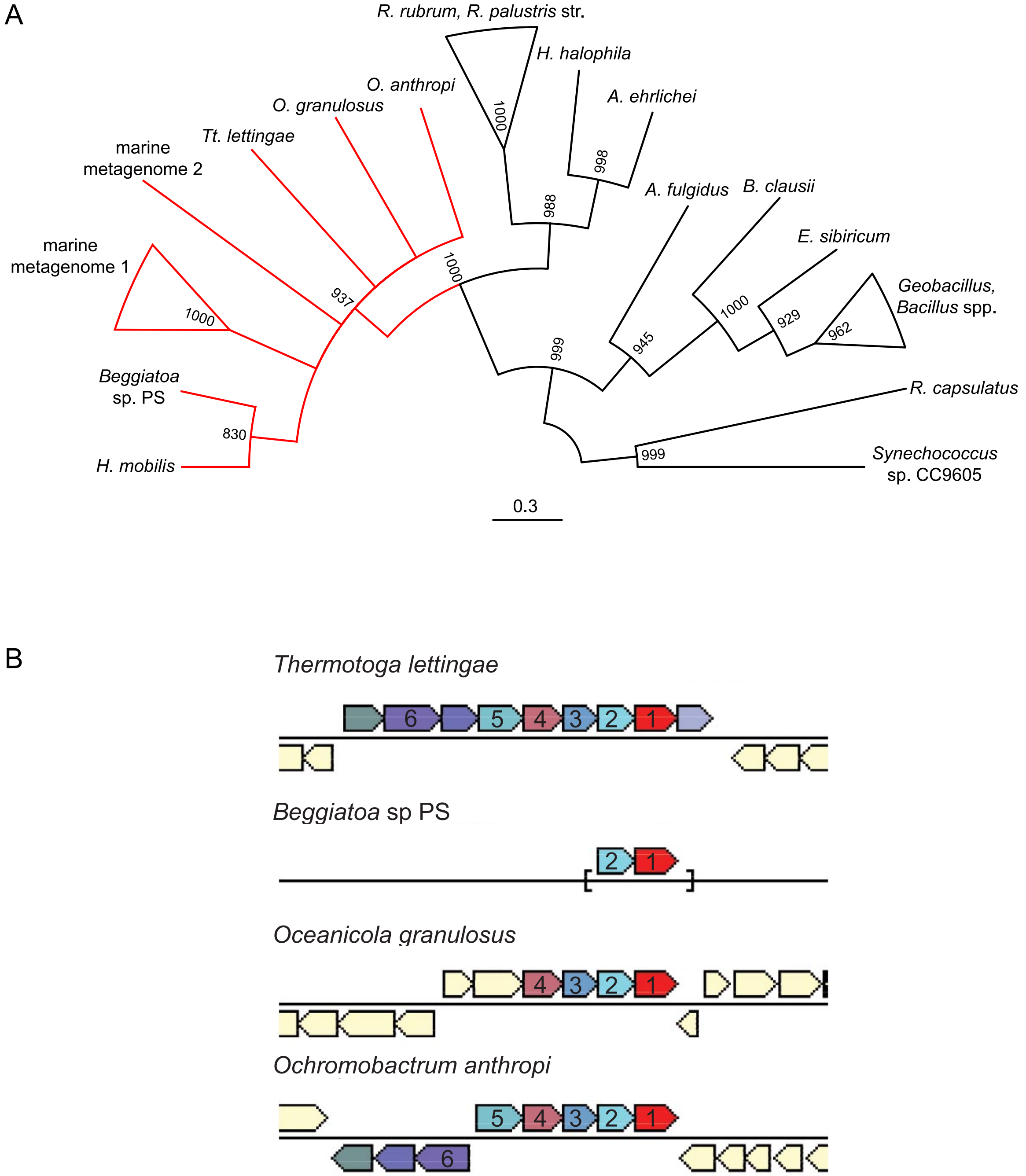

Fig. S3. Evolution of RuBisCO-like protein sequences. (*A*) Phylogenetic tree showing the relationships between group IV-Deep Ykr RuBisCO-like protein sequences. Sequences selected from the group IV-Deep Ykr cluster [Tabita FR, et al. (2007) Function, structure, and evolution of the RuBisCO-like proteins and their RuBisCO homologs. *Microbiol Mol Biol Rev* 71:576–599] were aligned and a maximum-likelihood tree was reconstructed as described in *SI Methods*. The IV-Deep Ykr clade is highlighted in red. (*B*) Genome neighbors of homologous RuBisCO-like protein genes. Gene neighborhoods were created at the JGI IMG system (http://img.jgi.doe.gov/). ORFs were annotated as encoding the following: 1, RuBisCO-like protein; 2, transketolase central region; 3, transketolase domain; 4, methylthioribose-1-phosphate isomerase; 5, 5-methylthioribose kinase; and 6, ABC transporter-related protein.

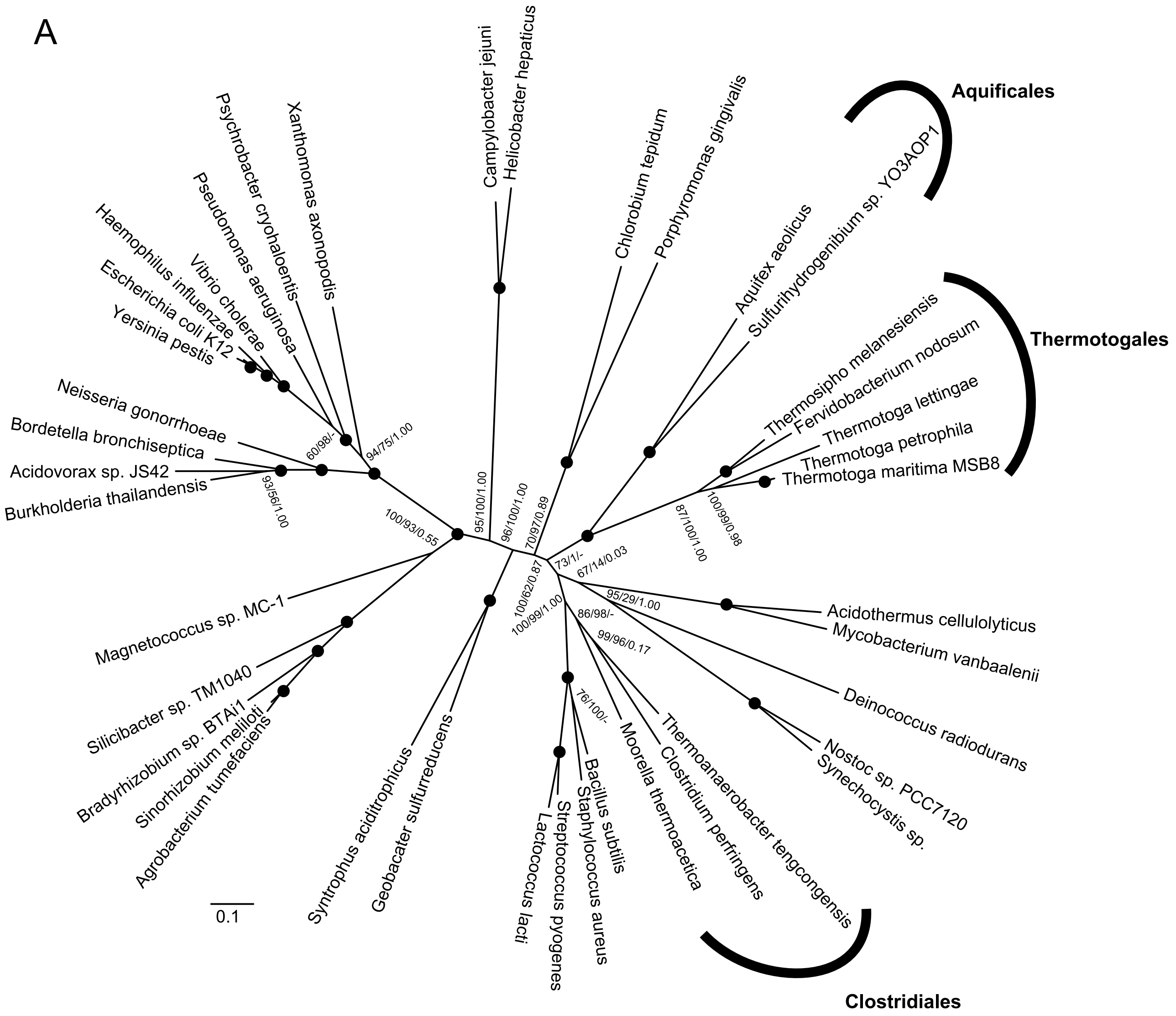

Fig. S4. Phylogenetic analyses of the ribosomal proteins data set. (*A*) Maximum-likelihood tree of concatenated bacterial ribosomal proteins. Values at each branch correspond to bootstrap support from PhyML, bootstrap support from FITCH, and posterior probabilities from PHYLOBAYES analyses, respectively. Black circles indicate branches with 100% (bootstrap), 100% (bootstrap), and 1.00 (posterior probability) support. Monophyly of Aquificales and Thermotogales is strongly supported. (*B*) Incongruence of individual ribosomal protein trees with the tree from the concatenated data set. Backbone topology is based on the concatenated data set with branches with <70% bootstrap support collapsed (branch lengths are not scaled with respect to substitutions). Red numbers at each branch show how many individual ribosomal proteins (of 29 analyzed) support the branch with at least 70% bootstrap support, while green numbers show the corresponding number of conflicts. The alternative groupings from individual ribosomal trees are shown as purple lines with names of ribosomal proteins listed. Ribosomal protein S14 supported one more bipartition that could not be easily depicted. (*C*) Phylogenetic tree reconstructed from a subset of concatenated ribosomal proteins alignment containing only sites varying by at least 50% in amino acid composition. See *SI Methods* for details.

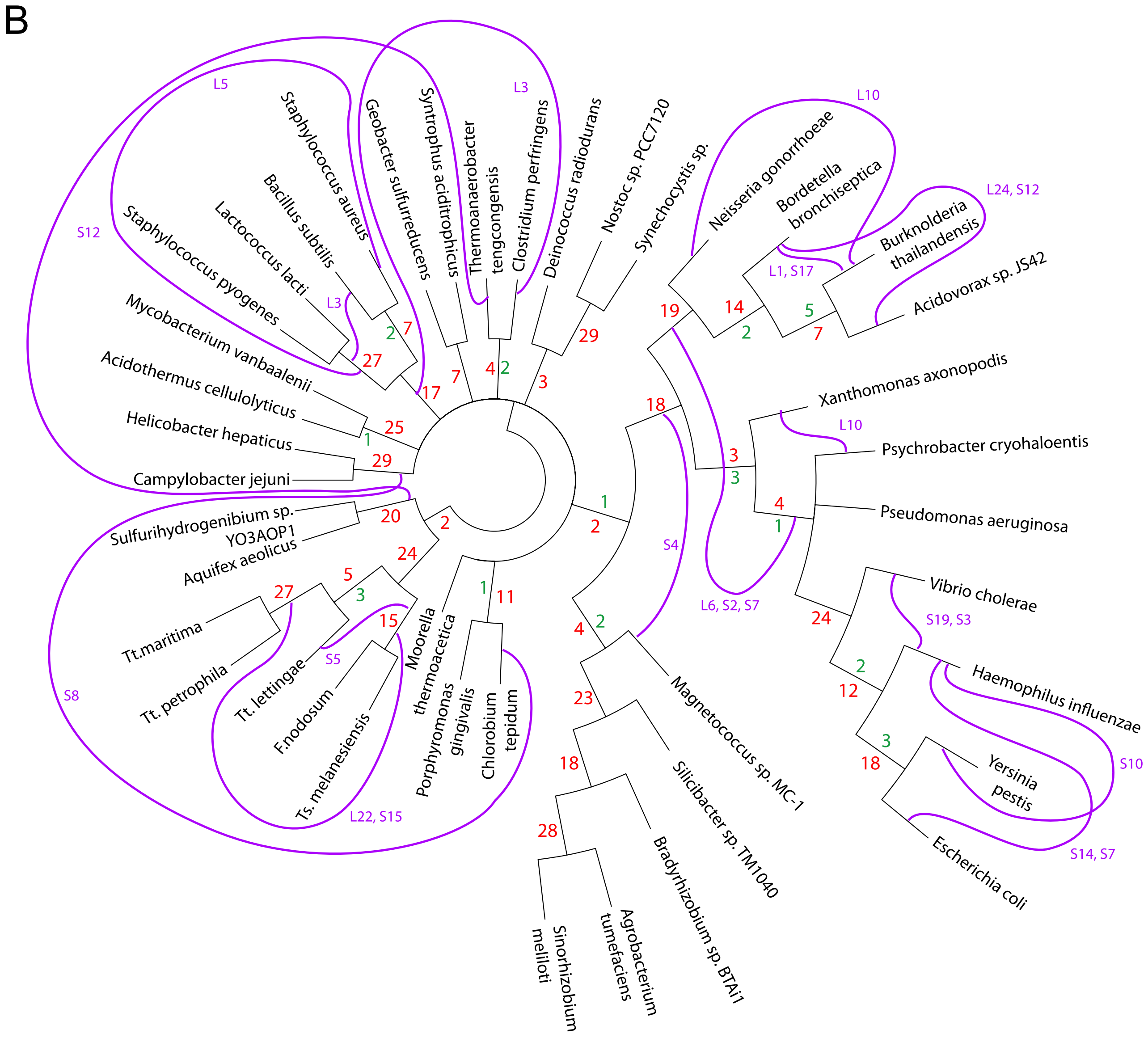

Fig. S4. (*Continued*).

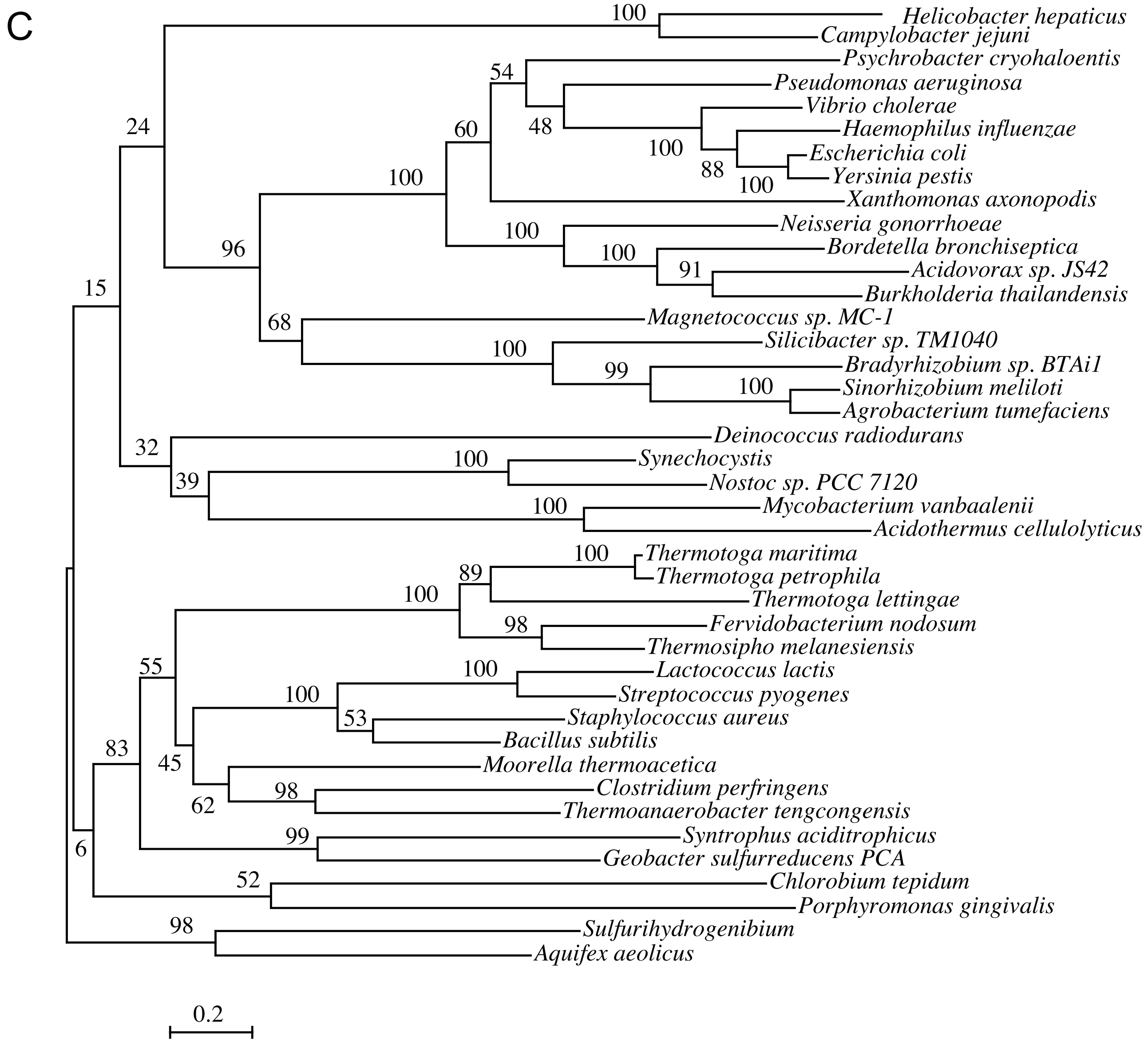

Fig. S4. (*Continued*).

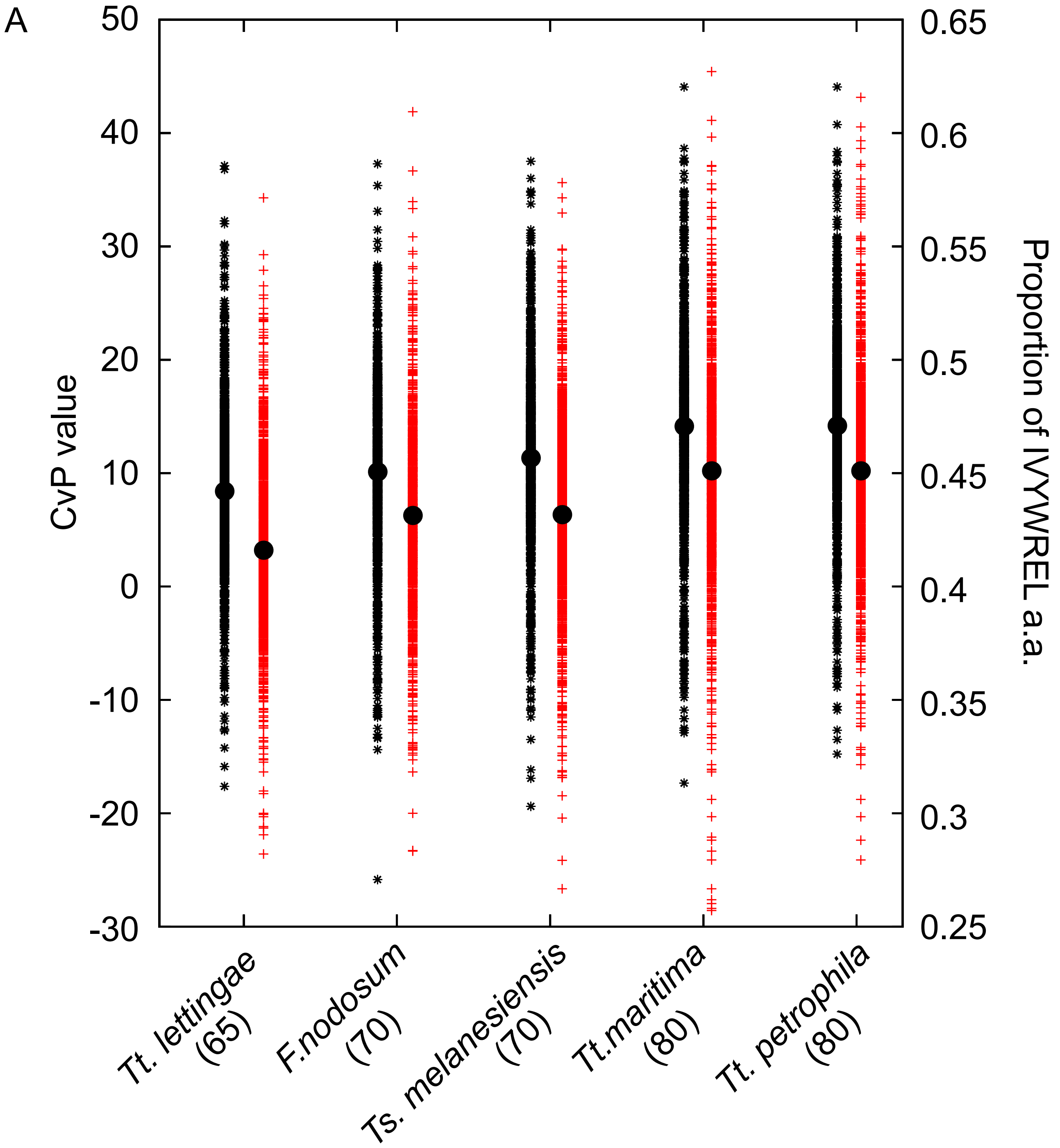

Fig. S5. Thermoadaptation of genes in Thermotogales. (*A*) Each column shows the distribution of CvP values (black) and IVYWREL values (red) among proteins predicted in a genome (see *SI Methods* on how proteins were selected for the analyses). Numbers in parentheses on the *x*-axis show optimal growth temperature, in degrees Celsius. Black circles represent median values for all used proteins per genome. CvP values 10.62 are suggested to indicate thermophily. (*B*) Distribution of CvP values of ORFs within each genome. The same data as in *A* are shown, but summarized differently. *B* shows that the distribution of the CvP values is unimodal, with the majority of genes having a CvP value close the mean CvP value. (*C*) GC ''thermometer'' of combined 16S and 23S rRNA gene sequences. The tree branches are color coded according to GC content in each branch's child node (color coding of GC content in percent is shown on the *Left*). The most recent ancestor of Thermotogales rRNA has predicted GC content of 60%, which suggests its host to be thermophilic. The tree should be considered unrooted but shown as rooted using Clostridia sequences as an outgroup. For details on tree and ancestral sequences reconstruction see *SI Methods*. Five Thermotogales genomes analyzed in this paper are marked with asterisks. Bootstrap values <80% are not shown.

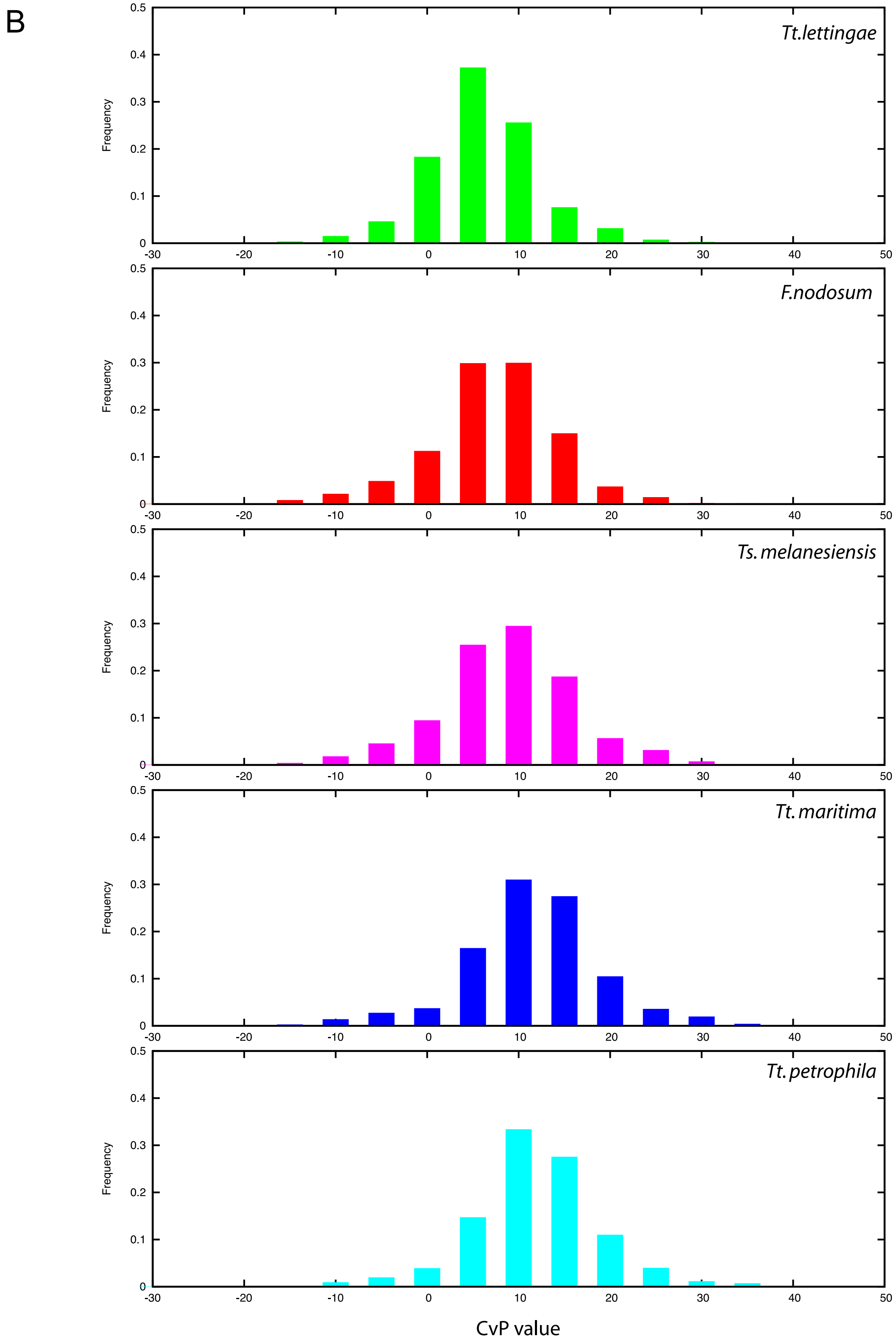

Fig. S5. (*Continued*).

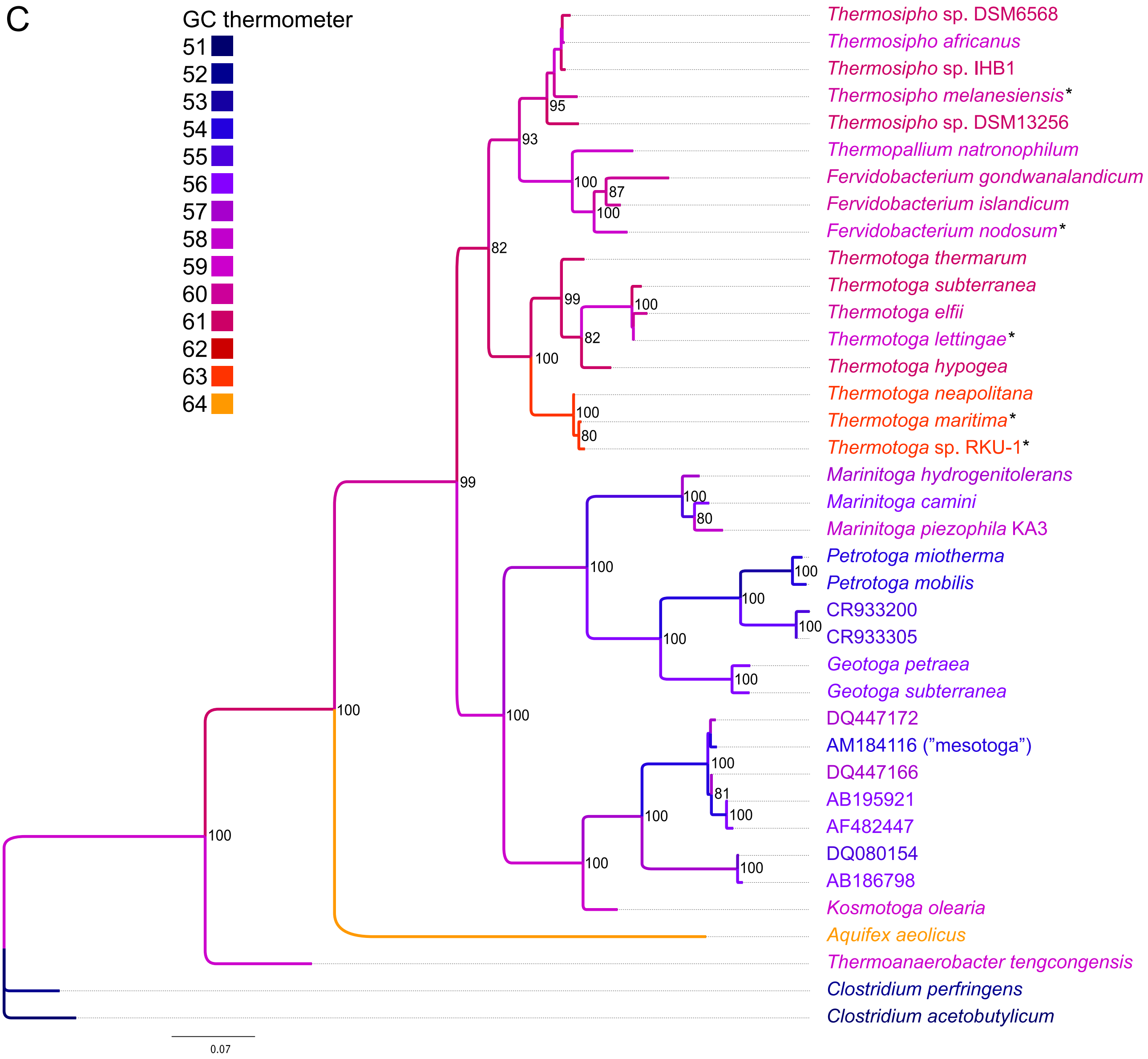

Fig. S5. (*Continued*).

Table S1. Features of 5 Thermotogales genomes

	Tt. maritima	*Tt. petrophila*	*Tt. lettingae*	*F. nodosum*	*Ts. melanesiensis*
			General features		
Chromosome size (Mbp)	1.86	1.82	2.14	1.95	1.92
% GC	46	46	38	34	31
No. of protein-coding genes	1,858	1,785	2,040	1,750	1,879
No. of tRNAs	49	52	52	59	65
Optimal growth temp. (°C)	80	80	65	70	70
Source	Vulcano, Italy	Japanese oil reservoir	Methanol-fed hot bioreactor	New Zealand hot spring	Deep sea southwest Pacific vent
			Repeated elements*		
IS110/IS204	0	0	0	19	2
IS6	0	0	0	26	0
IS200/IS605	12	10	3	1	3
IS3/911	0	0	0	3	1
Total	12	10	3	49	6

*Values are the sum of IS elements identified by annotation of putative transposases (as in GenBank) and those identified here using PSI-BLAST searches (see *SI Methods*).

Table S2. CRISPR elements in Thermotogales genomes

	CRISPR	*N*	Sequence of repeat	Left end	Right end
Tt. maritima	A	25	GTTTCCATACCTCTAAGGAATTATTGAAAC	534,764	536,392
	B	9	GTTTCAATAATTCCTTAGAGGTATGGAAAC	548,554	549,119
	C	9	GTTTCAATAATTCCTTAGAGGTATGGAAAC	575,956	576,517
	D	41	GTTTCAATACTTCCTTAGAGGTATGGAAAC	942,582	945,275
	E	13	GTTTCCATACCTCTAAGGAAGTATTGAAAC	1,025,711	1,026,537
	F	4	GTTTCCATACCTCTAAGGAAGTATTGAAAC	1,174,396	1,174,622
	G	9	GTTTCAATACTTCCTTAGAGGTATGGAAAC	1,285,938	1,286,500
Tt. petrophila	A	5	GTTTCAATAGTTCCTTAGAGGTATGGAAAC	523,936	524,234
	B	13	GTTTCCATACCTCTAAGGAACTATTGAAAC	536,387	537,215
	C	8	GTTTCCATACCTCTAAGGAACTATTGAAAC	568,240	568,733
	D	19	GTTTCCATACCTCTAAGGAATTATTGAAAC	940,034	941,251
	E	41	GTTCATATTCCTCTTAGGAAGATAAAAAC	1,107,494	1,110,144
	F	13	GTTTCAATAATTCCTTAGAGGTATGGAAAC	1,171,790	1,172,614
	G	7	GTTTCAATAATTCCTTAGAGGTATGGAAAC	1,372,310	1,372,743
Tt. lettingae		44	GTTTCCATCCCTCTAAGGTTCGATTGAAAC	205,528	208,415
		54	GTTTCAATCGAACCTTAGAGGGATGGAAAC	1,131,044	1,134,585
Ts. melanesiensis		11	GTTTCTACCTTACCTTGGAGGAATTGAAAC	137,986	138,684
		8	ATTTCAATTCCTCCAAGGTAAGGTAAAAAC	360,035	360,533
		52	ATTTCTATTCCTCATAGGTAGATTCTAAAC	754,563	758,169
		16	GTTTAGAATCTACCTATGAGGAATGGAAAC	1,638,704	1,639,789
		12	GTTTCCATTCCTCATAGGTAGATTCTAAAC	1,651,052	1,651,851
F. nodosum		179	GCTTTTAGCATACCTATTAGGGATTGAAAC	1,019,473	1,031,294
		15	GTTTTAGAAGTGACTATGAGGGATGGAAAC	1,506,355	1,507,353

CRISPR notations for *Tt. maritima* and *Tt. petrophila* (letters) correspond to those shown in *SI* Fig. S1*A*. *N* is the number of repeats of the sequence between the indicated sequence positions of the left and right ends.

Table S3. Taxonomic affiliations of top-scoring BLAST hits of *Tt. maritima* genes against complete and modified *nr* databases

	nr	1*	2	3	4	5	6	7	8	9	10	11	12
Bacteria	**1,391**	**1,294**	**1,300**	**1,297**	**1,298**	**1,301**	**1,317**	**1,327**	**1,300**	**1,298**	**1,304**	**1,319**	**1,313**
Firmicutes	825	365	397	412	410	412	486	515	406	410	444	484	483
Clostridia	328	0	63	78	76	75	24	53	65	76	132	20	25
Thermoanaerobacterales	273	0	0	0	0	0	161	184	0	0	0	160	166
Bacilli (Bacillales)	118	305	316	278	279	281	249	232	284	279	262	252	241
Proteobacteria	215	360	349	343	343	348	322	312	345	343	332	321	324
Aquificae	46	62	62	62	62	62	60	55	62	62	62	60	60
Chloroflexi	62	88	85	74	86	85	82	83	86	86	80	83	81
Deinococcus-Thermus	38	55	54	55	55	53	47	50	55	55	51	47	46
Bacteroidetes	42	67	86	65	64	62	58	72	65	64	60	58	75
Cyanobacteria	43	64	62	61	60	59	56	57	60	60	62	57	57
Actinobacteria	26	48	46	45	45	45	44	43	45	45	47	45	45
Planctomycetes	22	36	36	34	35	35	33	33	35	35	33	33	33
Acidobacteria	10	16	16	16	15	16	16	15	16	15	15	16	15
Spirochaetes	10	20	18	18	17	17	15	18	17	17	16	16	15
Archaea	**204**	**285**	**279**	**283**	**282**	**279**	**266**	**256**	**280**	**282**	**276**	**264**	**268**
Euryarchaeota	171	232	227	230	229	227	220	205	227	229	226	217	218
Thermococcales	95	117	115	117	116	116	114	108	115	116	115	112	114
Archaeoglobales	18	22	22	22	22	22	22	19	22	22	21	22	22
Methanococcales	18	24	22	24	24	24	22	22	23	24	22	21	22
Methanosarcinales	20	39	39	37	37	35	34	29	37	37	38	34	35
Crenarchaeota	27	43	42	43	43	42	37	42	43	43	40	38	40
Thermoproteales	12	21	20	21	21	21	16	20	21	21	20	16	19
Desulfurococcales	8	10	10	10	10	9	9	10	10	10	8	10	9
Sulfolobales	5	10	10	10	10	10	10	10	10	10	10	10	10
Unclassified Archaea	6	10	10	9	10	10	9	9	10	10	10	9	10
Eukaryotes	**17**	**20**	**20**	**20**	**20**	**20**	**20**	**19**	**20**	**20**	**19**	**20**	**20**
Viruses	**1**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**	**2**
Others	**6**	**8**	**8**	**7**	**7**	**7**	**8**	**8**	**7**	**7**	**8**	**8**	**8**
Thermotogales specific	**239**	**249**	**249**	**249**	**249**	**249**	**245**	**246**	**239**	**249**	**249**	**245**	**247**

*Replicate 1: *nr* database with all Clostridiales sequences excluded.
Replicate 2: Replicate 1 database plus *C. perfringens* str. 13 and *C. acetobutylicum* ATCC824 genomes.
Replicate 3: Replicate 1 database plus *C. beijerinckii* NCIMB8052 and *C. botulinum* A ATCC19397 genomes.
Replicate 4: Replicate 1 database plus *C. botulinum* A ATCC3502 and *C. kluyveri* DSM555 genomes.
Replicate 5: Replicate 1 database plus *C. botulinum* F str. Langeland and *C. tetani* E88 genomes.
Replicate 6: Replicate 1 database plus *C. tetani* E88 and *Thermoanaerobacter* sp. X514 genomes.
Replicate 7: Replicate 1 database plus *Thermoanaerobacter tengcongensis* MB4 and *Alkaliphilus metalliredigens* QYMF genomes.
Replicate 8: Replicate 1 database plus *C. botulinum* A3 str. Loch Maree and *C. perfringens* SM101 genomes.
Replicate 9: Replicate 1 database plus *C. botulinum* A ATCC19397 and *C. kluyveri* DSM555 genomes.
Replicate 10: Replicate 1 database plus *C. thermocellum* ATCC27405 and *C. beijerinckii* NCIMB8052 genomes.
Replicate 11: Replicate 1 database plus *C. perfringens* str. 13 and *Thermoanaerobacter* sp. X514 genomes.
Replicate 12: Replicate 1 database plus *Thermoanaerobacter pseudethanolicus* ATCC33223 and *C. novyi* NT genomes.