PNAS
www.pnas.org

Supporting Information (SI) Appendix

Form III RubisCO-mediated transaldolase variant of the Calvin cycle in a chemolithoautotrophic bacterium

Evgenii N. Frolov, Ilya V. Kublanov, Stepan V. Toshchakov, Evgenii A. Lunev, Nikolay V. Pimenov, Elizaveta A. Bonch-Osmolovskaya, Alexander V. Lebedinsky, Nikolay A. Chernyh

Corresponding author: Evgenii N. Frolov

E-mail: evgenii_frolov_89@mail.ru

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1904225116

SI Text

General genomic properties of *T. acidiphilum*. The genome of *T.acidiphilum* consists of one circular chromosome of 1,774,794 bp. The genome contains 1757 protein-coding genes, two identical rRNA operons and 46 tRNA genes. While the 16S rRNA genes of *T. acidiphilum* and those of *T. narugense* share 99.5% similarity, whole genome-genome ANI and AAI values were 86% and 90%, respectively, which is far below the species border (1). Analysis of the core *in silico* proteome and species-specific proteins of *T. acidiphilum* and *T. narugense* by OthoVenn server (2) showed that the proteins formed 1665 clusters of homologous genes, while 79 and 127 single copy protein-coding genes as well as 2 and 5 clusters of paralogous genes, were specific for *T. acidiphilum* and *T. narugense*, respectively. The increased number of species-specific proteins in *T.narugense* reflects higher genome mobility. Indeed, the *T. narugense* genome is 124 kb larger than that of *T.acidiphilum* and possesses greater number of mobile genetic elements (9 complete and 4 partial transposase genes in *T. narugense* vs. only 4 partial transposase genes in *T. acidiphilum* (Fig. S9)). Either higher level of genome mobility or viral load of *T. narugense*'s environment or both reasons resulted in obviously beneficial evolutionary acquisition of at least three CRISPR-Cas protein operons of types III-A, III-D and I-B, as well as corresponding CRISPR repeat clusters. At the same time, neither CRIPSR arrays nor functional CRISPR-Cas gene clusters were discovered in *T. acidiphilum* genome. In turn, the absence of genomic defense system in *T. acidiphilum* resulted in acquisition of several phage-related genes (Fig. S9). Search for genomic islands with three different tools (see Supplemental Materials and Methods), as well as analysis of tetranucleotide frequency biases, produced somewhat diverging results: none of the predicted HGT regions was univocally supported by all four methods (Figs. S9-S10). Despite their ambiguity, the HGT signatures observed in the region of the *cbb1* gene cluster and in 10-20 kb vicinity of the *cbb2* and *cbb3* gene clusters provide an opportunity to speculate that the *cbb* gene clusters, *cbb1* in particular, may have been acquired by distant lateral

transfer but the elapsed evolutionary time was sufficient to ameliorate the gene sequences, at least to a considerable extent.

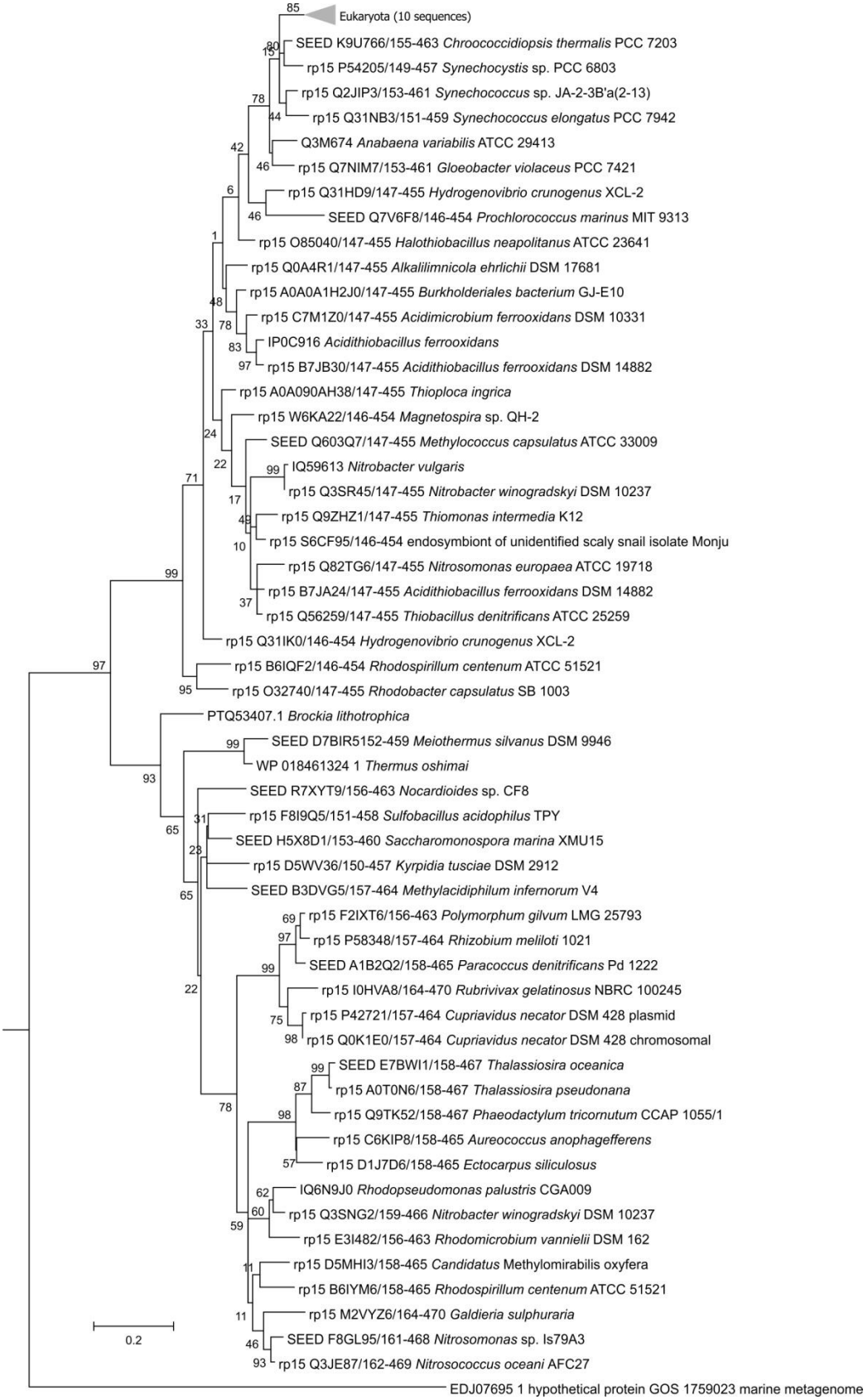


Fig. S1. RubisCO Form I phylogenetic subtree. For a collapsed view of the complete tree, see Fig. 1.

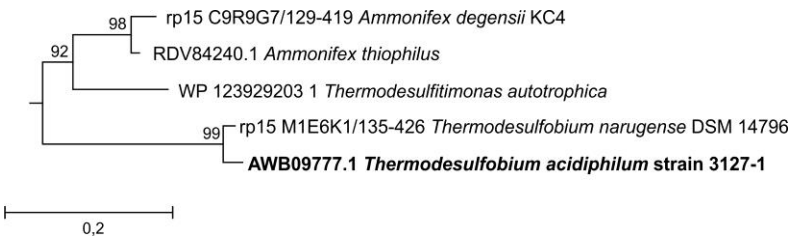


Fig. S2. RubisCO Form III deep branch comprised by *Thermodesulfobium* and *Ammonifex* sequences. For a collapsed view of the complete tree, see Fig. 1.

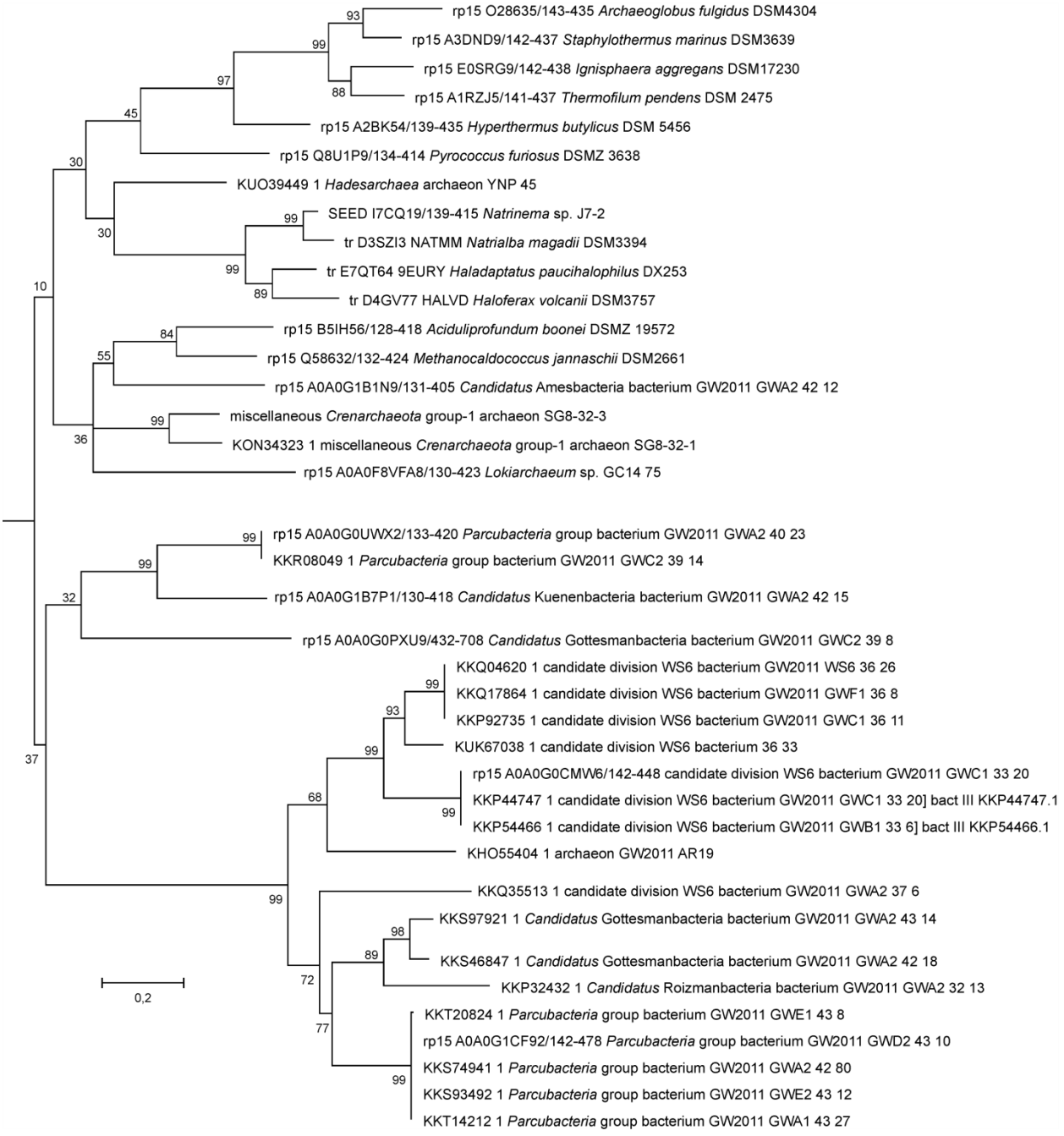


Fig. S3. Branch of RubisCO Form III. For a collapsed view of the complete tree, see Fig. 1.

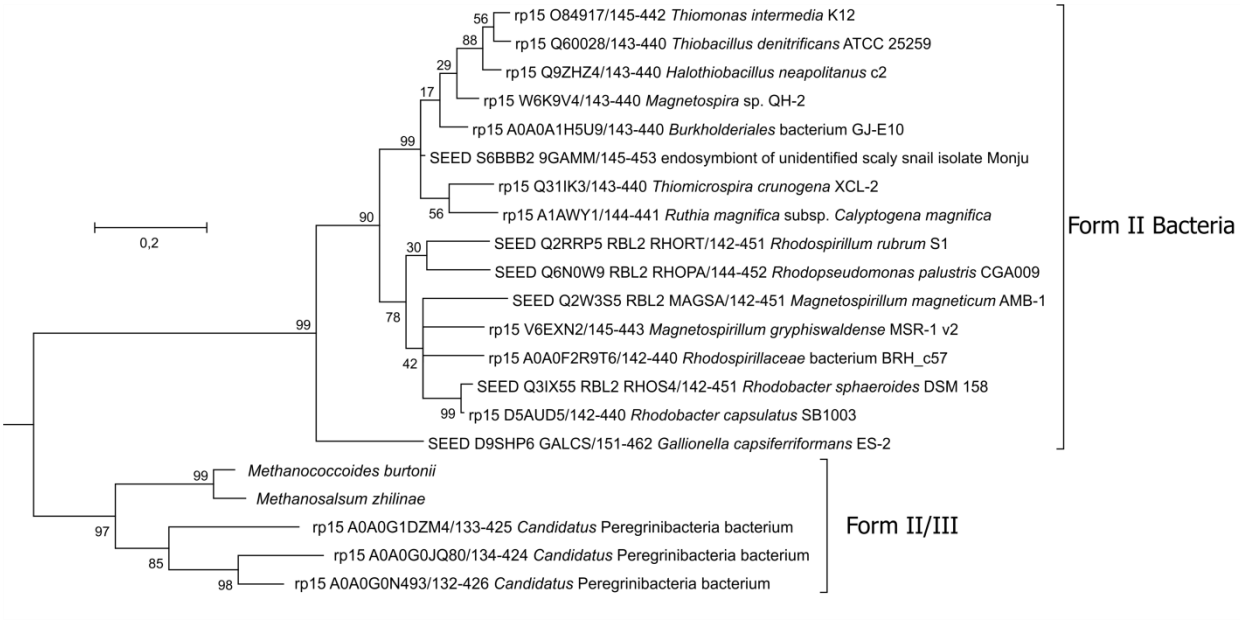


Fig. S4. RubisCO Form II. For a collapsed view of the complete tree, see Fig. 1.

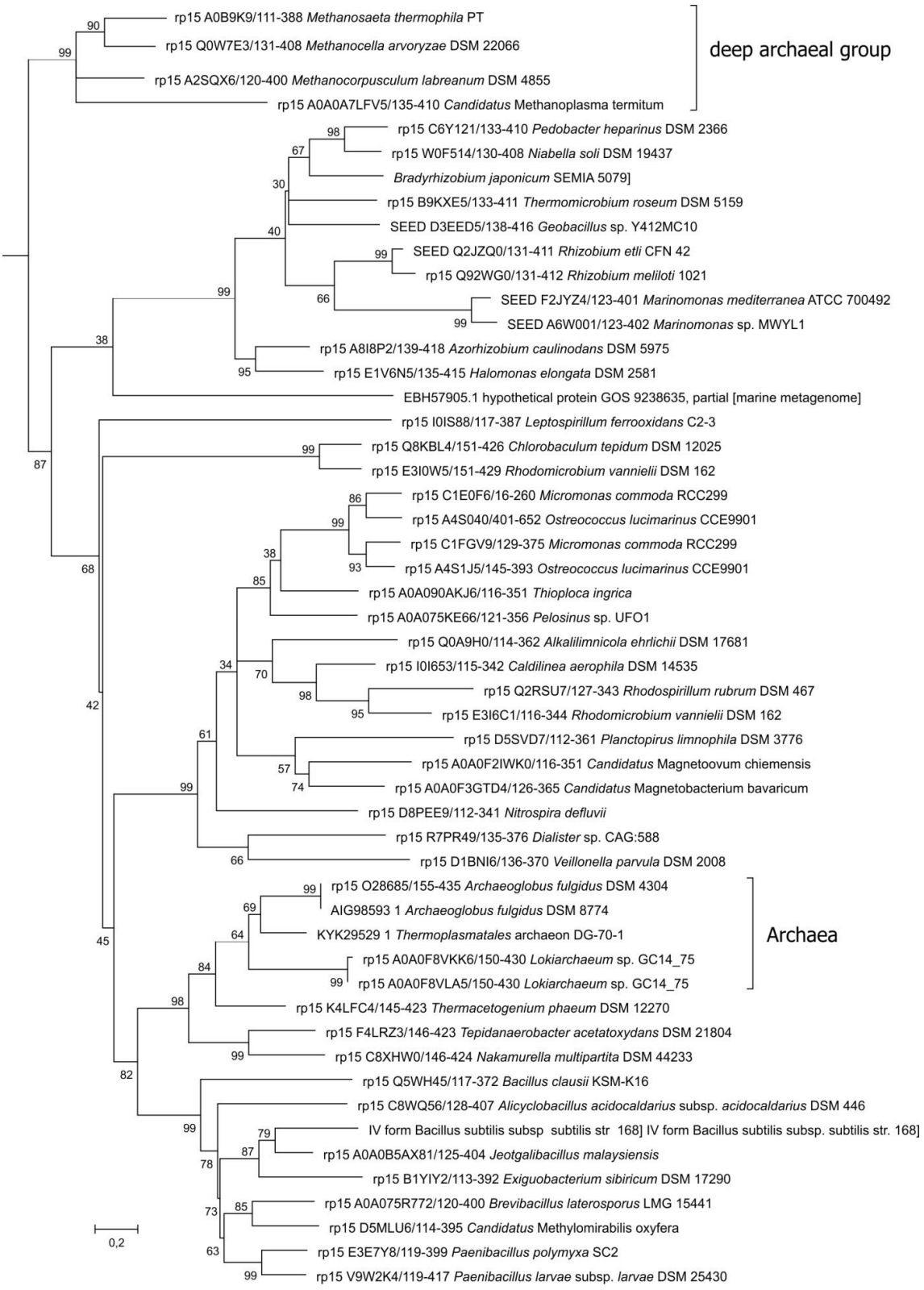


Fig. S5. RubisCO Form IV and deep archaeal branch. For a collapsed view of the complete tree, see Fig. 1.

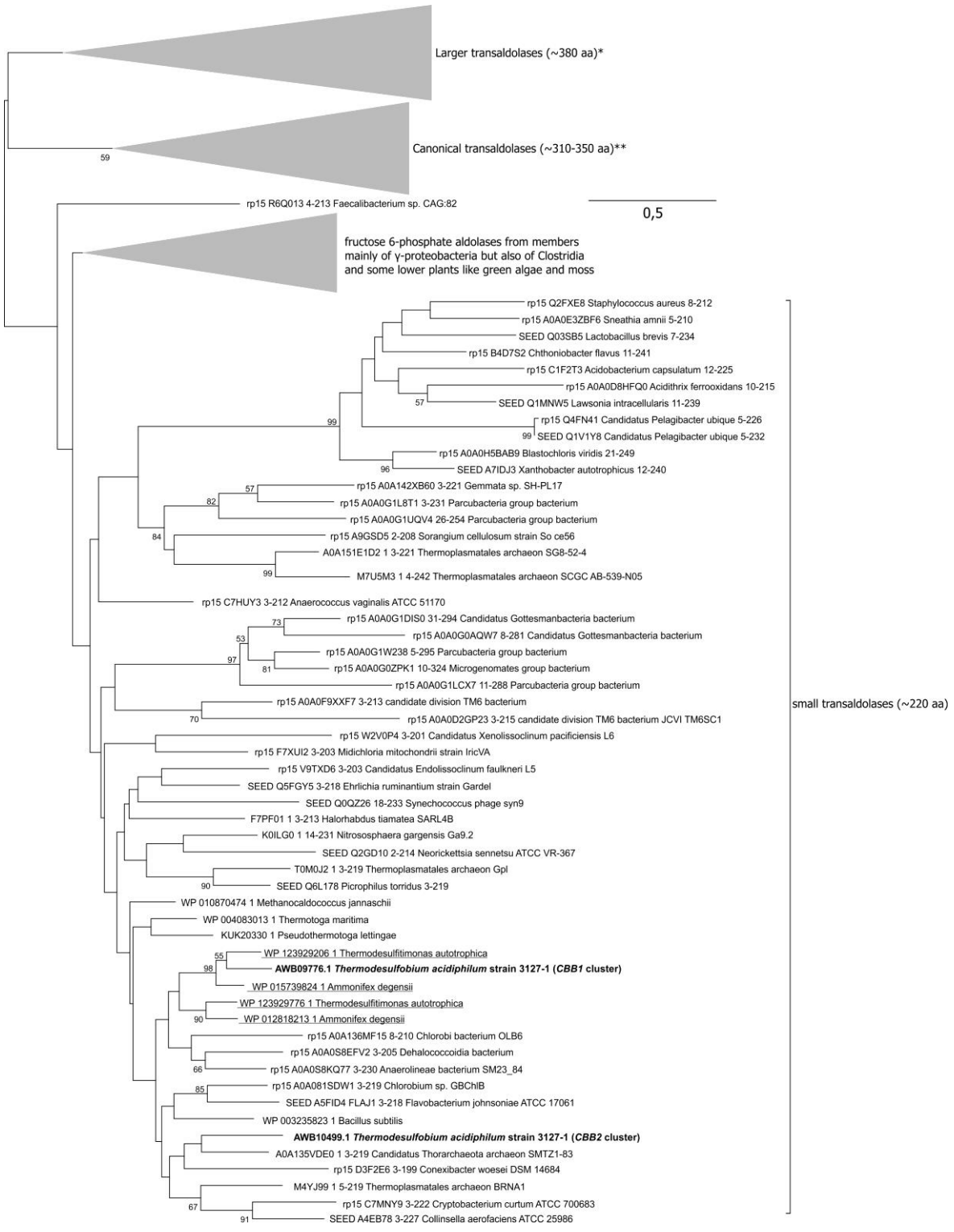
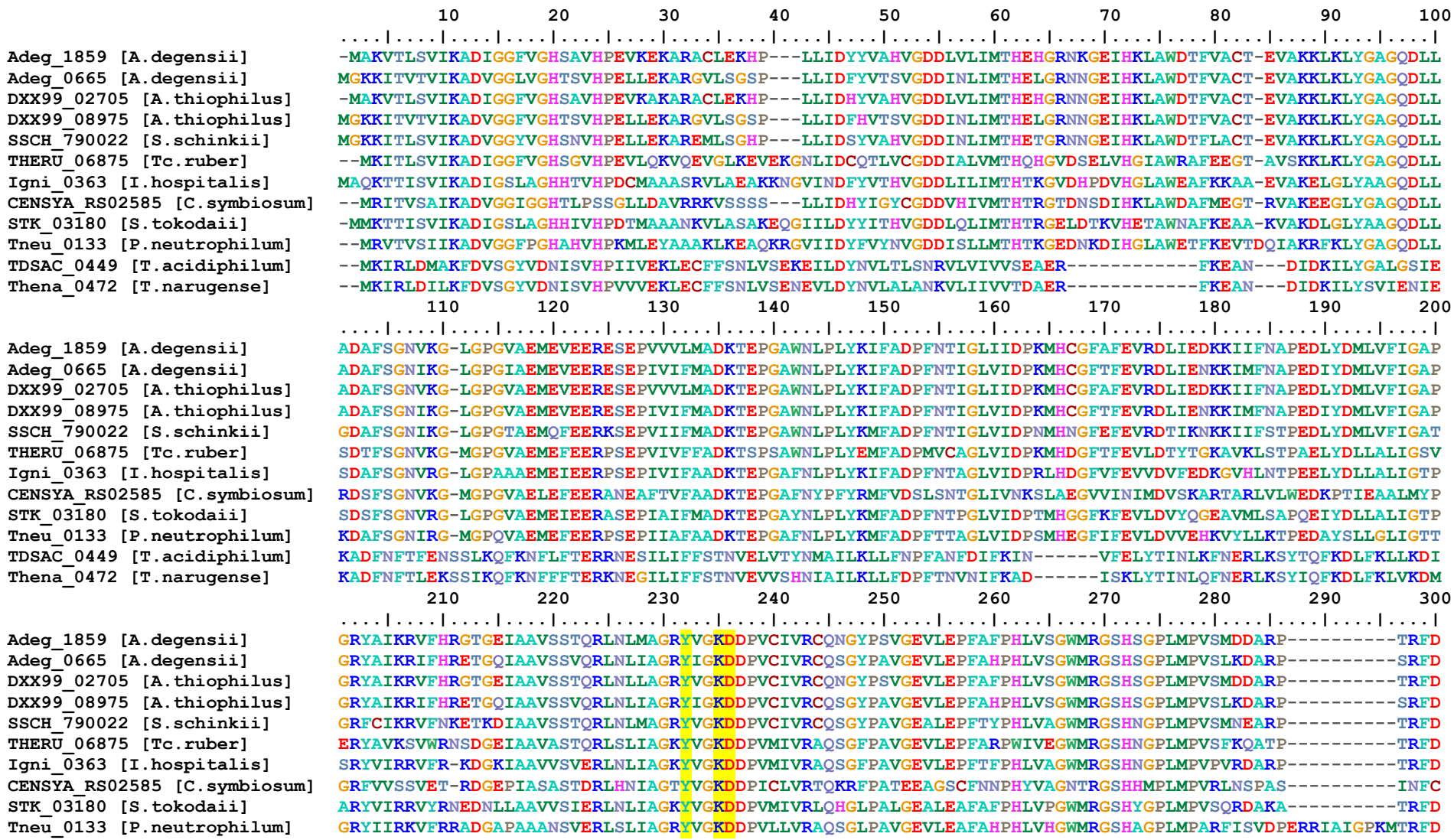


Fig. S6. Phylogenetic tree of transaldolase family proteins. *T. acidiphilum* proteins are in bold.


10 20 30 40 50 60 70 80 90 100
Adeg_1859 [A.degensii]
Adeg_0665 [A.degensii]
DXX99_02705 [A.thiophilus]
DXX99_08975 [A.thiophilus]
SSCH_790022 [S.schinkii]
THERU_06875 [Tc.ruber]
Igni_0363 [I.hospitalis]
CENSYA_RS02585 [C.symbiosum]
STK_03180 [S.tokodaii]
Tneu_0133 [P.neutrophilum]
TDSAC_0449 [T.acidiphilum]
Thena_0472 [T.narugense]
110 120 130 140 150 160 170 180 190 200
210 220 230 240 250 260 270 280 290 300

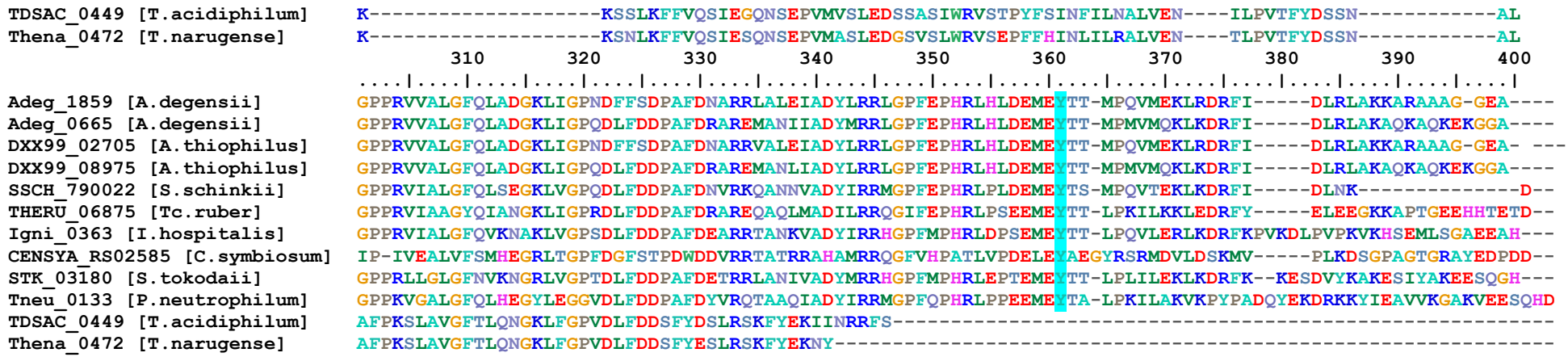


Fig. S7. Multiple sequence alignment of ten fructose-1,6-bisphosphate aldolase/phosphatases (FBPAPs) and the two FBPAP-like proteins of *Thermodesulfobium* spp. Yellow and blue highlighting respectively mark active site residues essential for the aldolase (Tyr-229, Lys-232, Asp-233, numbering by the *Sulfolobus tokadaii* enzyme) and phosphatase (Tyr-348) activities according to Say and Fuchs and Du et al. (3, 4). The alignment was made by using CLUSTAL W implemented within BioEdit software (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

Source organisms for the amino acid sequences are as follows:

TDSAC_0449 is from *T. acidiphilum*

Thena_0472 is from *T. narugens*e;

Adeg_1859 and Adeg_0665 are from *A. degensii*;

DXX99_02705 and DXX99_08975 are from *A. thiophilus* SR;

THERU_06875 is from *Thermocrinis ruber*; second best blastp hit of TDSAC_0449 (22% identities, 42% positives) in NCBI nr among cultivated microorganisms (second best hit after Thena_0472);

SSCH_790022 is from *Syntrophaceticus schinkii*; fourth best blastp hit of Adeg_1859 in NCBI nr among cultivated microorganisms (fourth best hit after Adeg_0665, DXX99_02705 and DXX99_08975).

The other sequences belong to fructose-1,6-bisphosphate aldolases/phosphatases studied and/or discussed in the works on the FBPAP active site (3, 4):

Igni_0363 is from *Ignicoccus hospitalis*;

CENSYA_RS02585 is from *Cenarchaeum symbiosum*;

STK_03180 is from *Sulfurisphaera tokodaii* (formerly, *Sulfolobus tokodaii*);

Tneu_0133 is from *Pyrobaculum neutrophilum* (formerly, *Thermoproteus neutrophilus*).

Fig. S8. Phylogenetic tree of class II aldolase proteins. The *T. acidiphilum* protein (TDSAC_1156) is in bold.

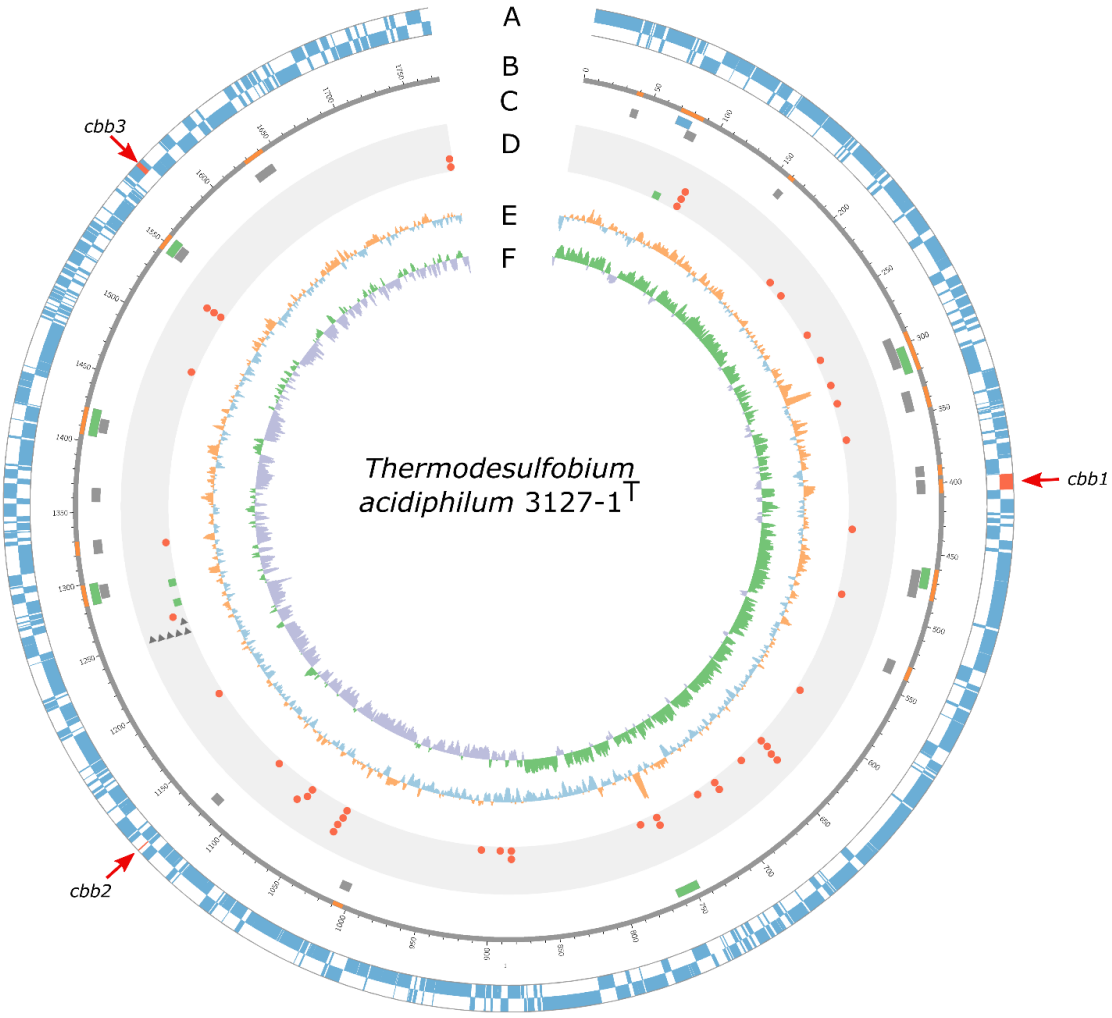


Fig. S9. Chromosome map of *T.acidiphilum*.

(A) Predicted CDS of *T. acidiphilum* genome. Positive strand CDS are shown on the outer circle, negative strand CDS are on the inner circle. *cbb* gene clusters are highlighted with red;

(B) Chromosome coordinates; regions corresponding to predicted genomic islands are highlighted with orange;

(C) Regions corresponding to genomic islands predicted by IslandViewer (blue), SeqWordSniffer (green) and AlienHunter (gray);

(D) Genomic island-associated genomic features: tRNA (orange circles), transposase genes and pseudogenes (green rectangles), phage-related proteins (gray triangles);

(E) GC-bias. Estimated in a window of 2000 nt with the step size of 200 nt. Values are shown relative to average GC-content of *T. acidiphilum* genome. Values inferior to the average GC are highlighted by light blue, values superior to average GC are in light orange.

(F) GC-skew. Estimated in a window of 2000 nt with step size of 200 nt. Negative values are highlighted by purple, positive values are highlighted by green.

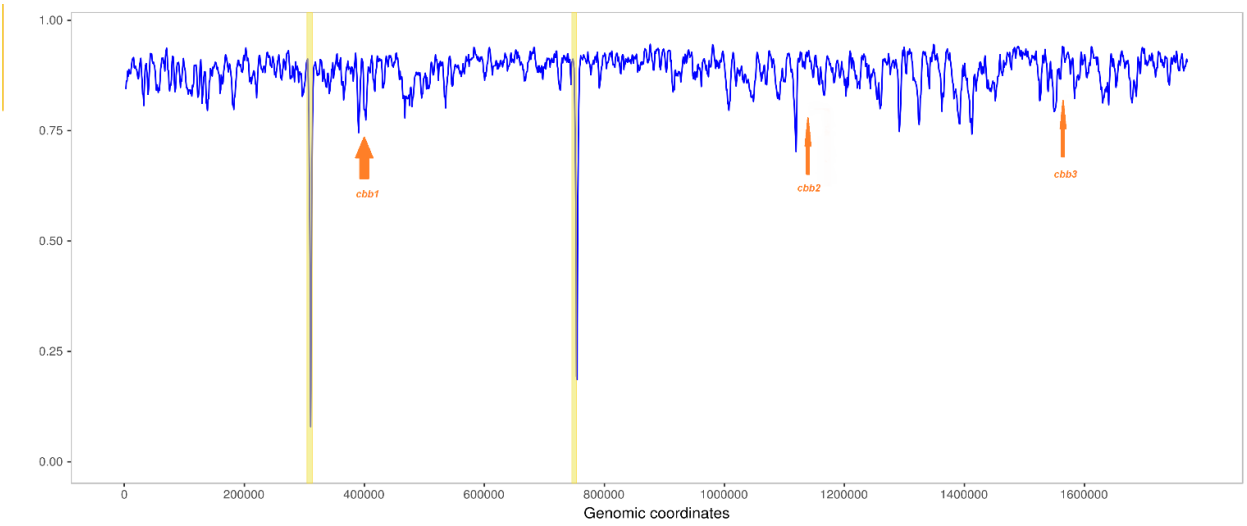


Fig. S10. Tetranucleotide frequency bias of *T.acidiphilum* genome. Correlation of tetranucleotide frequencies against genome-wide tetranucleotide signature. The positions of key CBB cycle gene clusters are indicated by arrows. Regions with significant deviation of tetranucleotide patterns are highlighted by yellow rectangles.

177 **Table S1.** Predicted genes of *cbb* gene clusters in *T. acidiphilum* genome

Locus tags	Gene name	Predicted function	Abbreviations for enzymes	Best BlastP hit*	% identity	Score	E-value	2nd BlastP hit*	% identity	Score	E-value	COGs and Pfam domains
***cbb1* gene cluster**												
TDSAC_0400	*tal1*	Transaldolase [EC 2.2.1.2]	TAL1	*T. narugense* (Thena_0422)	97	432	2e-153	*Ammonifex degensii* (Adeg_1864)	70	321	6e-109	COG0176, pfam00923
TDSAC_0401	*cbbL-III*	Form III ribulose bisphosphate carboxylase [EC 4.1.1.39]	RubisCO	*T. narugense* (Thena_0423)	95	845	0.0	*A. degensii* (Adeg_1863)	68	578	0.0	COG1850, pfam00016, pfam02788
TDSAC_0402	*cbbP*	Phosphoribulokinase [EC 2.7.1.19]	PRK	*T. narugense* (Thena_0424)	97	601	0.0	*Desulfotomaculum putei* (BUB67_RS15390)	50	295	2e-96	COG0572, pfam00485
TDSAC_0403	*cbbT-n*	Transketolase, N-terminal section [EC 2.2.1.1]	TK-N	*T. narugense* (Thena_0425)	97	588	0.0	*A. degensii* (Adeg_1861)	72	425	3e-148	COG3959, pfam00456
TDSAC_0404	*cbbT-c*	Transketolase, C-terminal section [EC 2.2.1.1]	TK-C	*T. narugense* (Thena_0426)	94	676	0.0	*A. degensii* (Adeg_1860)	62	418	1e-143	COG3958, pfam02779, pfam02780
TDSAC_0405	*cbbG1*	NAD-dependent glyceraldehyde-3-phosphate dehydrogenase [EC 1.2.1.12]	GAPDH1	*T. narugense* (Thena_0427)	91	629	0.0	*T. narugense* (Thena_1627)	62	420	2e-144	COG0057, pfam02800 pfam00044
TDSAC_0406	*cbbK1*	Phosphoglycerate kinase [EC 2.7.2.3]	PGK1	*T. narugense* (Thena_0428)	95	784	0.0	*T. narugense* (Thena_1628)	65	558	0.0	COG0126, pfam00162
TDSAC_0407	*cbbF1*	Fructose-1,6-bisphosphatase, type I [EC 3.1.3.11]	FBPase1	*T. narugense* (Thena_0429)	93	621	0.0	*Thermodesulfobac-terium hydrogeniphilum* (CC87_RS03320)	68	474	3e-166	COG0158, pfam00316
TDSAC_0408	*cbbI*	Ribose 5-phosphate isomerase B [EC 5.3.1.6]	RPI	*T. narugense* (Thena_0430)	97	291	1e-99	*Clostridiales bacterium* (AYC61_RS17895)	62	176	5e-54	COG0698, pfam02502
TDSAC_0409	*cbbE*	Ribulose-phosphate 3-epimerase [EC 5.1.3.1]	RuPE	*T. narugense* (Thena_0431)	97	413	6e-146	*Campylobacter peloridis* (CPEL_1132)	44	167	1e-48	COG0036, pfam00834
***cbb2* gene cluster**												
TDSAC_1154	*tal2*	Transaldolase [EC 2.2.1.2]	TAL2	*T. narugense* (Thena_1180)	91	400	2e-140	*Pseudothermotoga lettingae* (Tlet_1124)	55	251	2e-81	COG0176, pfam00923
TDSAC_1155	*cbbF2*	Fructose-1,6-bisphosphatase, type I [EC 3.1.3.11]	FBPase2	*T. narugense* (Thena_1181)	97	648	0.0	bacterium BMS3Bbin07 (BMS3Bbin07_00863)	69	474	3e-166	COG0158, pfam00316
TDSAC_1156	*cbbA*	Aldolase II class [EC 4.1.2.13]	FBPA	*T. narugense* (Thena_1182)	94	923	0.0	*Sulfurihydrogenibium* sp. YO3AOP1 (SYO3AOP1_0025)	69	650	0.0	COG0191, pfam01116, pfam00596

***cbb3* gene cluster**												
TDSAC_1598	*cbbG3*	NAD-dependent glyceraldehyde-3-phosphate dehydrogenase [EC 1.2.1.12]	GAPDH3	*T. narugense* (Thena_1627)	97	669	0.0	*T. narugense* (Thena_0427)	63	427	3e-147	COG0057, pfam02800 pfam00044
TDSAC_1599	*cbbK3*	Phosphoglycerate kinase [EC 2.7.2.3]	PGK3	*T. narugense* (Thena_1628)	99	796	0.0	*T. narugense* (Thena_0428)	98	559	0.0	COG0126, pfam00162
TDSAC_1600	*tpi*	Triosephosphate isomerase [EC 5.3.1.1]	TPI	*T. narugense* (Thena_1629)	93	461	1e-155	*Clostridium sp.* (HMPREF1092_01812)	48	189	2e-56	COG0149, pfam00121

*When analyzing results of NCBI blasp searches, hits to sequences from the metagenomic assemblies PNIV01000000 (*Fervidicoccus fontis* ARK-12) and PNIY01000000 (*Thermodesulfobium narugense* ARK-09) were neglected, because these are evidently misassemblies. The former one contains 962 (66.5%) genes 99-100% identical to genes of *Fervidicoccus fontis* Kam940^T and 175 (12% genes) 99-100% identical to genes of *T. acidiphilum* 3127-1^T. In the cases we checked, the *Fervidicoccus fontis* ARK-12 genes related to *Thermodesulfobium* genes were in short contigs of the missassembly, and these contigs lacked any genes related to *Fervidicoccus fontis* Kam940^T genes. On the other hand, the latter above-mentioned assembly PNIY01000000 (*T. narugense* ARK-09), originating from the same metagenome, missed some of its native genes, attributed to *F. fontis* ARK-12.

References

1. E. N. Frolov *et al., Thermodesulfobium acidiphilum* sp. nov., a new thermoacidophilic sulfate-reducing chemoautotrophic bacterium from a Kamchatkan thermal site. *Int J Syst Evol Microbiol* **67,** 1482-1485 (2017).
2. Y. Wang, D. Coleman-Derr, G. Chen, Y. Q. Gu, OrthoVenn: a web server for genome wide comparison and annotation of orthologous clusters across multiple species. *Nucleic Acids Res* **43(W1)**, W78-W84 (2015).
3. Say, R. F. & Fuchs, G. Fructose 1,6-bisphosphate aldolase/phosphatase may be an ancestral gluconeogenic enzyme. *Nature* **464**, 1077–1081 (2010).
4. J. Du, F. R. Say, W. Lü, G. Fuchs, O. Einsle, O. Active-site remodelling in the bifunctional fructose-1,6-bisphosphate aldolase/phosphatase. *Nature* **478**, 534-537 (2011)