Anaerobic degradation of paraffins by thermophilic Actinobacteria under methanogenic conditions

Supporting Information

Yi-Fan Liu[1,4,5#], Jing Chen[1,4#], Zhong-Lin Liu[1,4], Li-Bin Shou[1,4], Dan-Dan Lin[1,4], Lei Zhou[1,4], Shi-Zhong Yang[1,4], Jin-Feng Liu[1,4], Wei Li[3,5], Ji-Dong Gu[2] and Bo-Zhong Mu[1,4*]

[#] These authors contributed equally to this work.

* Correspondence: Bo-Zhong Mu, State Key Laboratory of Bioreactor Engineering and School of of Chemistry and Molecular Engineering, East China University of Science and Technology, Shanghai 200237, 130 Meilong Road, Xuhui District, P. R. China.

E-mail: bzmu@ecust.edu.cn

[1] State Key Laboratory of Bioreactor Engineering and School of Chemistry and Molecular Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, P.R. China

[2] School of Biological Sciences, The University of Hong Kong, Pokfulam Road, Hong Kong, P. R. China.

[3] National Engineering Laboratory for Industrial Wastewater Treatment, School of Resources and Environmental Engineering, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, P. R. China

[4] Engineering Research Center of MEOR, East China University of Science and Technology, 130 Meilong Road, Shanghai 200237, P. R. China

[5] Shanghai Institute of Pollution Control and Ecological Security, Shanghai 200092, P.R. China

The file contains:

Pages: 14

Supplementary Methods

Supplementary Tables: 6

Supplementary Figures: 7

Supplementary Methods

Metagenomic analysis

For Binning using MaxBin2, both marker genes set of 107 and 40 for bacterial and archaeal lineages were used and kept for subsequently analysis. For MetaBAT2, each of the mapping files were summarized using jgi_summarize_bam_contig_depths and then metabat was run using the following settings: –minProb 75 –minContig 1500 –minContigByCorr 2000. For CONCOCT, scaffolds were cut into sequences of length 10000 bp using cut_up_fasta.py and sequencing coverage was produced using concoct_coverage_table.py. Original bins that generated from the three different tools were evaluated and either merged or split using DAS_Tool (v1.1.1) [1]. The resulted bins were refined by removing outlier scaffolds when setting percentile of divergent GC, sequencing coverage and tetranucleotide frequencies to 95 in RefineM (v0.0.23) [2]. Subsequent contamination removal was performed using anvio5 [3], briefly, bins were split as major clusters calculated based on splits coverage and GC content and then merged back if the separated clusters were placed parallel in the same position of a genome tree. And the statistics regarding the lineage, completeness, contamination and heterogeneity of refined bins were estimated by checkM lineage_wf (v1.0.16) [2]. Finally, bins that share ≥97% average nucleotide identity (ANI) and ≥90% overall alignment coverage, as calculated by FastANI [4], were dereplicated to generate non-redundant MAGs.

Phylogenetic analysis of MAGs

Firstly, the 16S rRNA genes in bins were identified in the MAGs using checkM (v1.0.16) [2] with 'ssu_finder' function and sequences longer than 400 bp were kept for further analysis. 16S rRNA gene sequences were searched against SILVA (SSU133) 16S rRNA database [5] using SINA function [6], and then classified with the least common ancestor (LCA) method based on the taxonomies hosted by SILVA. Additionally, these sequences were aligned using MAFFT [7] with iterative refinement methods 'G-INS-i' and then refined (retained columns with < 10% gaps) manually for phylogenomic tree building.

Secondly, in cases where MAGs lack 16S rRNA genes, GTDBtk prediction (https://github.com/Ecogenomics/GTDBTk) was used to predict their taxonomic affiliation by calculating a relative evolutionary distance metric with the genomes and MAGs deposited in its database. To construct the phylogenomic tree, all MAGs and reference genomes (listed in Table S5) were pooled into PhyloPhlAn (v0.99) [8], which extracts, aligns and concatenates 400 conserved protein

sequences from the genomes [8]. The concatenated alignments file was then trimmed for < 10% gaps. For both alignments of 16S rRNA gene sequences and concatenated proteins sequences, Maximum likelihood (ML) trees were reconstructed using IQ-tree (v1.6.7) under standard model selection with 1000 ultrafast bootstraps. The trees were further refined using Inkscape (v0.92.3) (inkscape.org) and visualized using iTOL [9].

Gene prediction and annotation

Genes within the MAGs were called using Prodigal v2.6 [10] in 'meta' mode. Then, amino acid files were submitted to GhostKOALA server [11] which utilizes more rapid GHOSTX algorithm in prokaryotic species database for assigning KO numbers. For MAGs described in this study, BlastKOALA server which utilizes more slow but more accurate BLAST search were used for assigning KO numbers [11]. Genes involved in anaerobic hydrocarbon degradation were identified using BLASTP against previously reported database [12] (coverage > 0.40; *E*-value < 1e-20; identity > 30%) and further confirmed with their phylogenies. Hydrogenases were firstly identified by HMMER (v3) search function [13] (*E*-value < 1e-20) using previous custom HMM models (Liu *et al.*, unpublished), and hits were further classified into different hydrogenase groups in HydDB web server [14].

The carbohydrate-active genes were identified in the dbCAN web server [15] with cutoffs used in a previous study [12] (coverage > 0.40; *E*-value < 1e-18; identity > 30%). To characterize the mobile elements, the MAGs were searched for integrons, transposons and tRNAs. A local database of integrons was created from the nucleotide sequences for all integrases available in the database INTEGRALL v1.2.8414 (10533 records in total) [16]. A gene was recognized as an integron or insertion if the BLAST hit (blastn) had a minimum of 30% identity over 75% of the gene length, according to the previously published threshold [17]. Amino acid files of population genomes were submitted to ISfinder online server (updated on March, 4, 2019) [18] for searching for transposons using BLATP tool (identity > 30%, coverage > 75%, E-value < 1×10^{-5}). Nucleotide sequences of MAGs described in this study were uploaded to RAST for tRNA gene annotation using subsystem technology [19].

Determination of methanogenic metabolic type

We used two complementary methods to distinguish methanogens with different methanogenic metabolic type. Firstly, these methanogen-like MAGs were assigned with taxonomic information. Based on this, we could infer their methanogenic metabolic type.

Secondly, we directly look into these MAG contents and search for genes associated with hydrogenotrophic, acetoclastic or methylotrophic methanogenic pathway. Then, these information was combined to infer their methanogenic metabolic types.

Supplementary tables

Table S1 Detection of fatty acids using GC-MS (μM) in the *n*-alkanes amended enrichment cultures and sterile control samples. Long-chain fatty acids have been quantified using calibration curves of authentic standards wherein both the standards and metabolites were identified as their esterified derivatives. Volatile fatty acids (C_1-C_4) measured using ion chromatography (mM) were appended with respective names. Fatty acids that have not been detected in all cultures were not shown.

Fatty acids	Group A	Group E	Group O	Control	PW
C18	1.46 ± 1.95	0.06 ± 0.08	-	-	-
C16	4.32 ± 3.91	0.56 ± 0.77	-	-	-
C14	33.88 ± 23.87	0.54 ± 0.41	-	-	-
α,ω-C12	8.61 ± 10.39	0.33 ± 0.28	-	-	-
α,ω-C11	11.75 ± 6.88	-	-	-	-
α,ω-C10	67.03 ± 33.83	1.78 ± 1.67	-	-	-
α,ω-C9	47.23 ± 15.73	11.67 ± 9.16	-	-	-
α,ω-C8	12.25 ± 5.28	21.94 ± 17.68	-	-	-
α,ω-C7	1.24 ± 1.76	5.37 ± 4.66	-	-	-
α,ω-C6	13.36 ± 8.67	15.63 ± 13.07	-	-	-
α,ω-C5	10.42 ± 5.17	11.54 ± 8.78	-	-	-
C4/Butyrate	15.22 ± 7.32	2.29 ± 0.82	-	-	-
C3/Propionate	7.40 ± 1.69	2.03 ± 1.07	0.54 ± 0.11	0.07 ± 0.02	-
C2/Acetate	26.13 ± 6.71	8.01 ± 6.51	0.63 ± 0.07	0.02 ± 0.01	0.54
C1/Format	14.07 ± 5.42	3.19 ± 0.29	0.06 ± 0.02	-	0.17

Table S2: Summary statistics of the MAGs. Taxonomic classification of MAGs were mainly based on GTDBtk prediction (https://github.com/Ecogenomics/GTDBTk) and corrected/confirmed by polygenetic analysis of conserved protein sequences. In cases where 16S rRNA gene sequence (>400bp) have been found in the MAG, the taxonomy or the MAG was also corrected/confirmed by search 16S rRNA gene in SILVA server using multiple databases.

Table S3: Occurrence of metabolic markers in all lineages of Actinobacteria phylum. The numbers in the cells colored in red/blue indicate the proportion of genomes of one lineage that have a given marker (all = 1, none = 0). It is possible that some of the markers are present in the missing regions of the incomplete MAGs/genomes. For each MAG in *Ca.* Syntraliphaticia, genome completeness was listed instead of the number of genomes/MAGs of a specific order-level lineage.

Table S4: Presence and tFPKM values of genes involved in hydrocarbon degradation, fatty acid oxidation, glycolysis, Wood-Ljungdahl pathway and energy transfer. The threshold tFPKM values for top 20% are 1.39, 1.28, 0.25 and 0.80 in sample A, E, O and PW, respectively. Genes found in each MAG mentioned in this study were listed in separate spreadsheet.

Table S5: Genomes used to build the phylogenomic tree. The first spreadsheet shows the reference genomes/MAGs used to build the complete MAG tree; the second shows those for building Actinobacteria tree; the third s those for building Verstraetearchaeota tree.

Table S6: Metagenomic abundance, transcriptomic activity and relative activity of functional genes. Relative transcriptional activity (RTA) was calculated by dividing tFPKM values of metatranscriptome datasets by FPKM values of metagenome datasets. To fully explore the metabolic potential of the metagenomes, we analyzed both genes in the MAGs as well as those from the unbinned assembly fragments.

Supplementary Figures

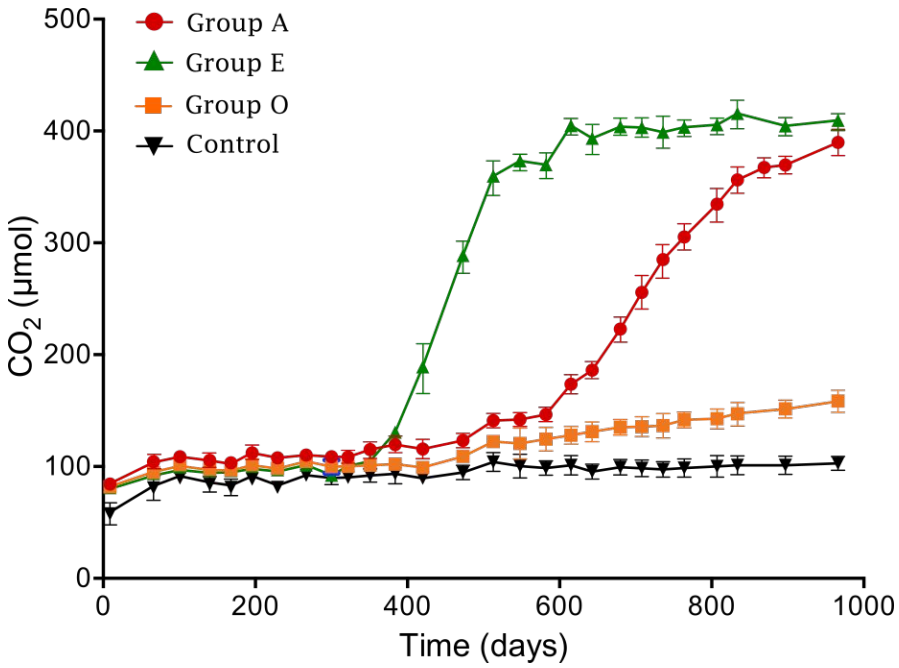

Figure S1. **Accumulative formation of CO_2** in serum bottles inoculated with production water amended with long-chain *n*-alkanes and substrate-free controls under methanogenic conditions during incubation.

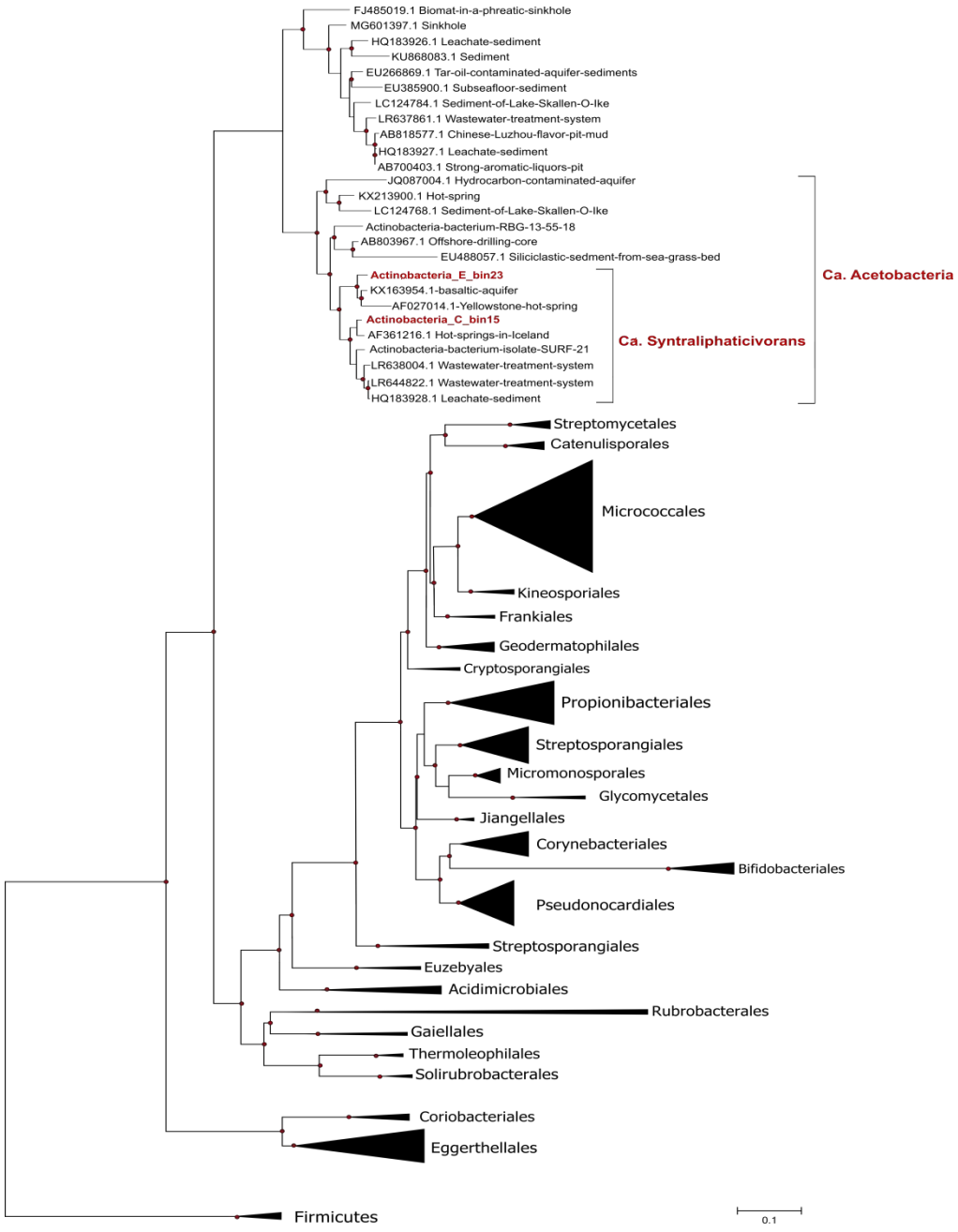

Figure S2. **Phylogenetic diversity and environmental distribution of *Ca*. Syntraliphaticium**. *Ca.* Syntraliphaticium MAGs were in red. The Maximum likelihood 16S rRNA tree was built using IQ-Tree with model GTR + F + I + G4 on 321 sequences (9 sequences of '*Ca*. Syntraliphaticivorans' and 6 sequences from other '*Ca*. Acetobacteria'). Node supports values were generated with 1,000 ultrafast bootstrap replicates, and only values above 0.80 for main branches are shown as red circles. The scale bar represents the average number of substitutions per site.

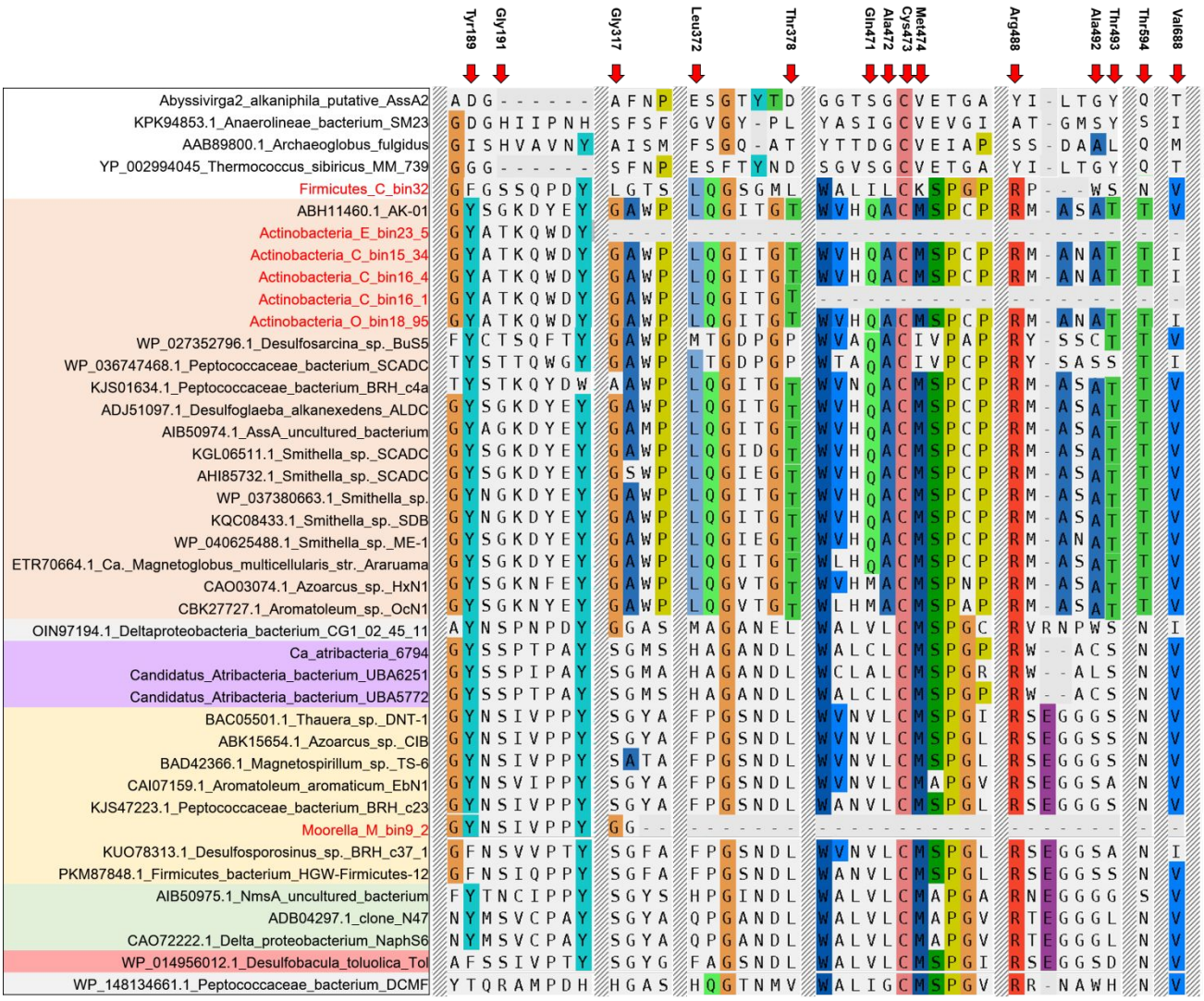

Figure S3. **Conserved amino acids in the active sites among FaeA paralogs.** The alignment has been trimmed for better view. The residue numbering system follows Heider *et al*. [20]. PflD proteins were shaded in white background; AssA proteins were shaded in pink; Atribacterial FaeA proteins were shaded in purple; BssA proteins were shaded in yellow; NmsA proteins proteins were shaded in green; HbsA proteins were shaded in red and others were in grey. MAGs that were constructed in this study were marked in red.

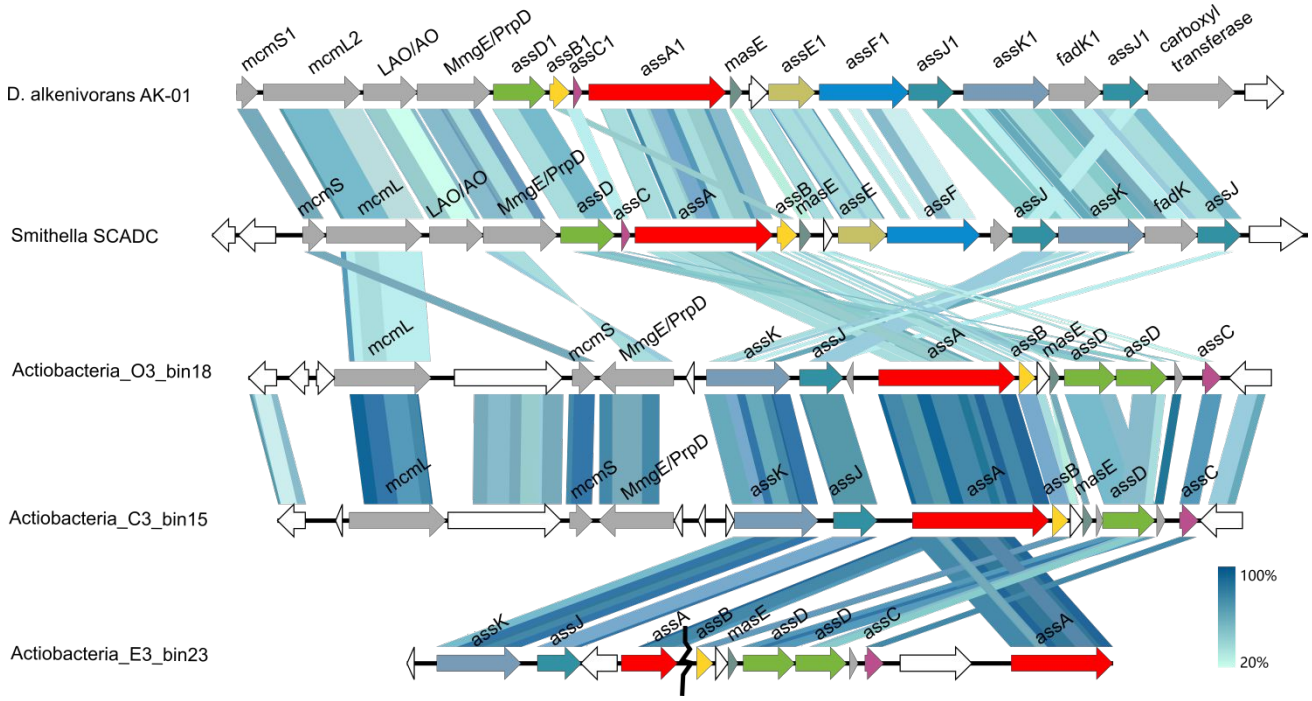

Figure S4. **Analysis of *ass* operons in actinobacterial MAGs.** TBLASTx comparison of the *ass* gene clusters of actinobacterial MAGs with *ass* operons from previously described *Smithella* SCADC and *Desulfatibacillum alkenivorans* AK-01. For the comparison, an cutoff of E-value < 1×10^{-10} was used, and visualization of the gene clusters was done with the program Easyfig [21]. In Actinobacteria_E_23, *ass* operon were broke into separate scaffolds due to fragmentary assembly, and in this case, borders between the scaffolds are marked with black lines.

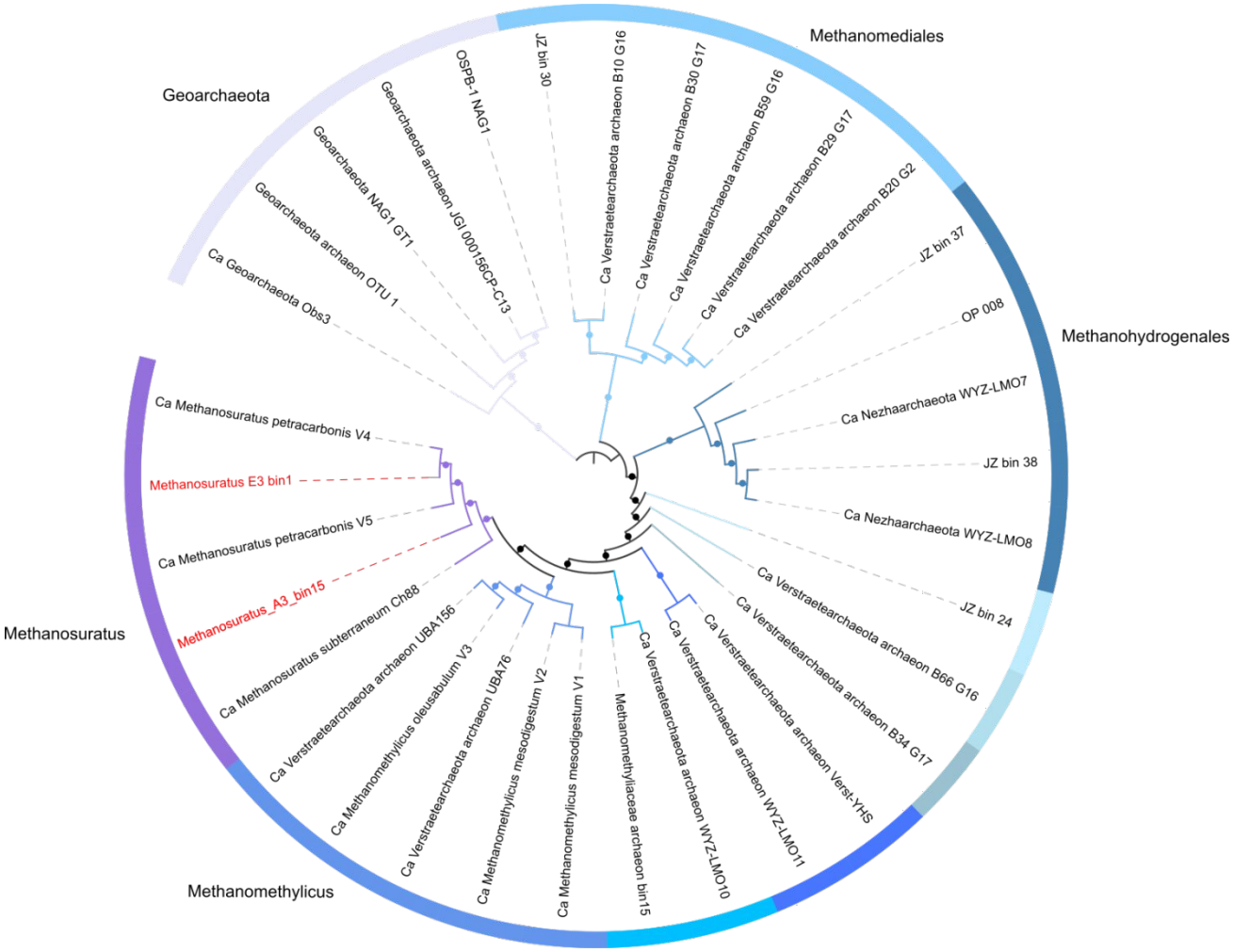

Figure S5. Maximum likelihood phylogenetic trees of the *Ca*. Methanosuratus MAGs, together with 31 MAGs across all lineages established so far within the phylum of *Ca*. Verstraetearchaeota. MAGs that contain divergent MCR were in red. The trees were built using IQ-Tree with model VT + F + G4. Node supports values were generated with 1,000 ultrafast bootstrap replicates, and values above 0.80 were shown as black dots.

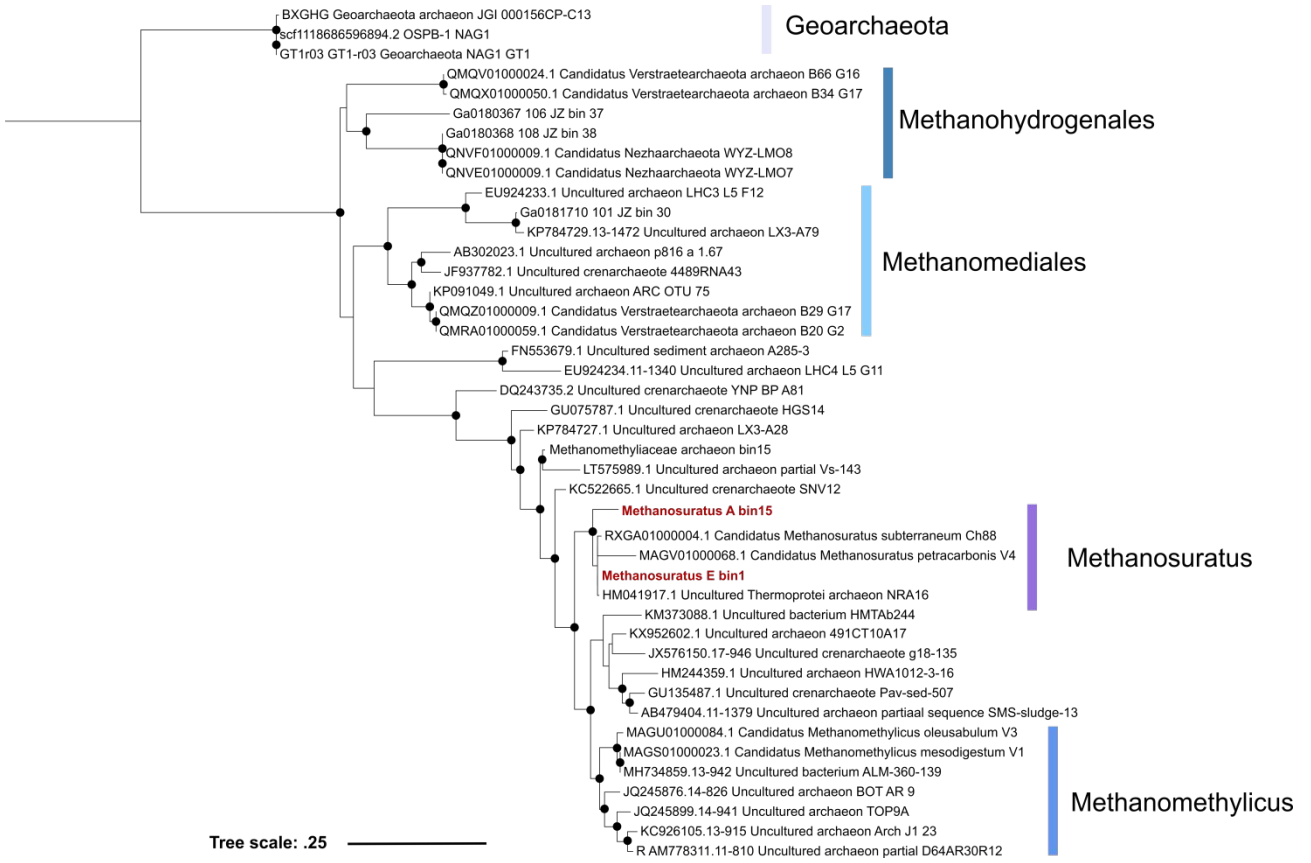

Figure S6. **Phylogenetic affiliation of the *Ca.* Methanosuratus MAGs and representative *Ca.* Verstraetearchaeota MAGs based on their 16S rRNA genes.** The tree were built by the IQ-Tree method with the model TIM3 + F + I + G4 with 1,000 bootstrap replicates on 39 sequences. The phylogenetic tree was rooted at the *Ca*. Geoarchaeota and all lineages were assigned with the same background colors to **Fig. 2b**. Divergent MCR-containing MAGs found in this study were in red. The black dots at each node corresponds to bootstrap values >0.80. The scale bar represents the average number of substitutions per site.

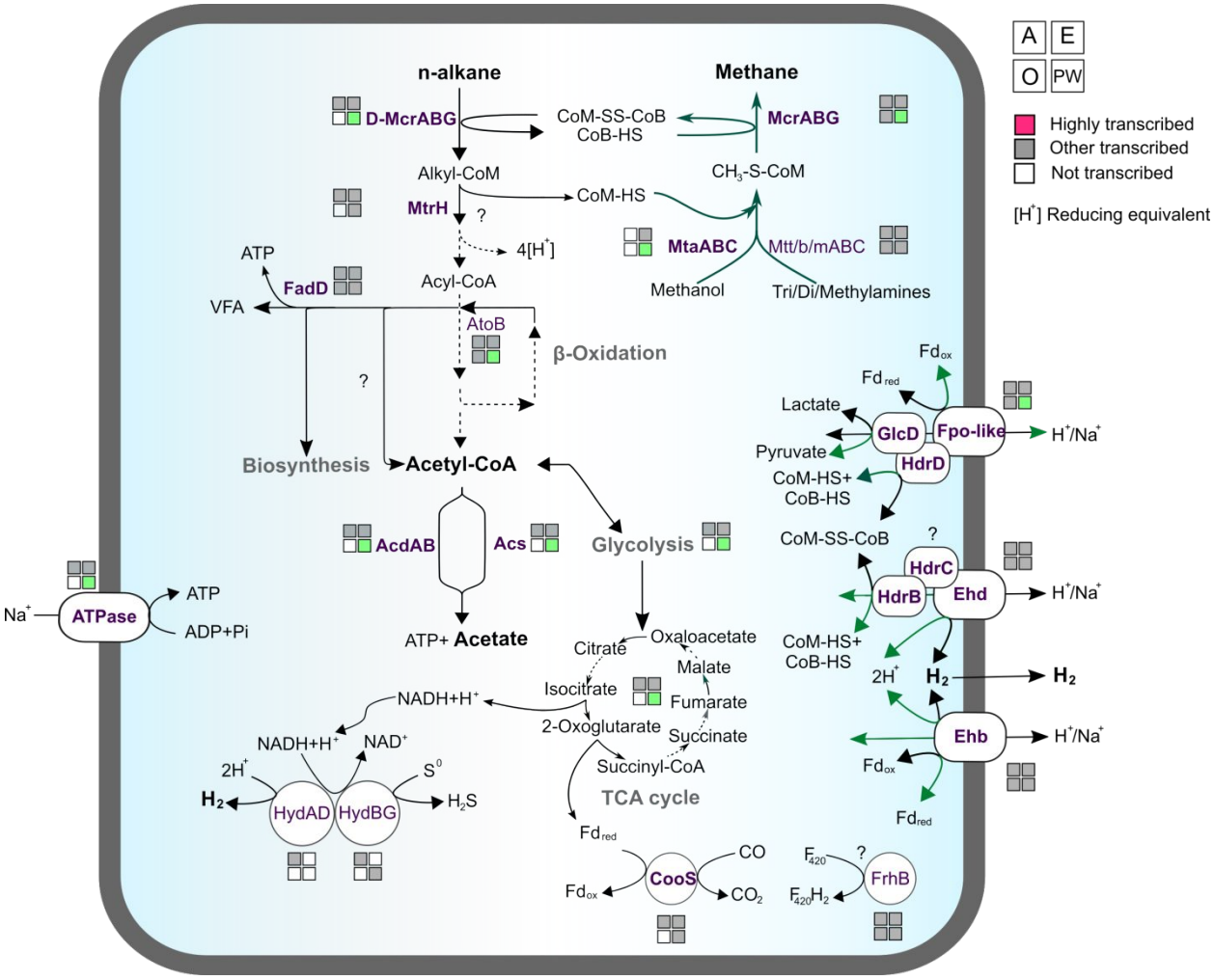

Figure S7 Predicted short-chain alkane metabolism in *Ca.* Methanosuratus. Enzymes shown in bold are present in all MAGs, those shown in light are present in a partial MAGs. 'D-Mcr' represent the divergent Mcr. The presence/absence of genes mentioned and their full names could be found in Table S4. The tFPKM values listed here represent the average values of genes found in described MAGs. Genes with tFPKM values ranked in upper 20th percentile were considered to be highly transcribed. For enzyme complexes and pathways of glycolysis and (r)TCA cycle, the average tFPKM values of genes encoding the subunits or every genes associated were shown.

References

(1) Sieber, C. M. K.; Probst, A. J.; Sharrar, A.; Thomas, B. C.; Hess, M.; Tringe, S. G.; Banfield, J. F. Recovery of Genomes from Metagenomes via a Dereplication, Aggregation and Scoring Strategy. *Nat. Microbiol.* **2018**, *3* (July), 1–8. https://doi.org/10.1038/s41564-018-0171-1.

(2) Parks, D. H.; Imelfort, M.; Skennerton, C. T.; Hugenholtz, P.; Tyson, G. W. CheckM : Assessing the Quality of Microbial Genomes Recovered from Isolates , Single Cells , and Metagenomes. *Genome Res.* **2015**, *25* (7), 1043–1055. https://doi.org/10.1101/gr.186072.114.Freely.

(3) Eren, A. M.; Esen, Ö. C.; Quince, C.; Vineis, J. H.; Morrison, H. G.; Sogin, M. L.; Delmont, T. O. Anvi'o: An Advanced Analysis and Visualization Platform for 'omics Data. *PeerJ* **2015**, *3*, e1319. https://doi.org/10.7717/peerj.1319.

(4) Jain, C.; Rodriguez-R, L. M.; Phillippy, A. M.; Konstantinidis, K. T.; Aluru, S. High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries. *Nat. Commun.* **2018**, *9* (1), 1–8. https://doi.org/10.1038/s41467-018-07641-9.

(5) Quast, C.; Pruesse, E.; Yilmaz, P.; Gerken, J.; Schweer, T.; Yarza, P.; Peplies, J.; Glöckner, F. O. The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools. *Nucleic Acids Res.* **2013**, *41* (D1), 590–596. https://doi.org/10.1093/nar/gks1219.

(6) Pruesse, E.; Peplies, J.; Glöckner, F. O. SINA: Accurate High-Throughput Multiple Sequence Alignment of Ribosomal RNA Genes. *Bioinformatics* **2012**, *28* (14), 1823–1829. https://doi.org/10.1093/bioinformatics/bts252.

(7) Yamada, K. D.; Tomii, K.; Katoh, K. Application of the MAFFT Sequence Alignment Program to Large Data - Reexamination of the Usefulness of Chained Guide Trees. *Bioinformatics* **2016**, *32* (21), 3246–3251. https://doi.org/10.1093/bioinformatics/btw412.

(8) Segata, N.; Börnigen, D.; Morgan, X. C.; Huttenhower, C. PhyloPhlAn Is a New Method for Improved Phylogenetic and Taxonomic Placement of Microbes. *Nat. Commun.* **2013**, *4*, 2304. https://doi.org/10.1038/ncomms3304.

(9) Letunic, I.; Bork, P. Interactive Tree Of Life (ITOL): An Online Tool for Phylogenetic Tree Display and Annotation. *Bioinformatics* **2007**, *23* (1), 127–128. https://doi.org/10.1093/bioinformatics/btl529.

(10) Hyatt, D.; Chen, G.-L.; Locascio, P. F.; Land, M. L.; Larimer, F. W.; Hauser, L. J. Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification. *BMC Bioinformatics* **2010**, *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

(11) Kanehisa, M.; Sato, Y.; Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **2016**, *428* (4), 726–731. https://doi.org/10.1016/j.jmb.2015.11.006.

(12) Dong, X.; Greening, C.; Rattray, J. E.; Chakraborty, A.; Chuvochina, M.; Mayumi, D.; Dolfing, J.; Li, C.; Brooks, J. M.; Bernard, B. B.; et al. Metabolic Potential of Uncultured Bacteria and Archaea Associated with Petroleum Seepage in Deep-Sea Sediments. *Nat. Commun.* **2019**, *10* (1), 1816. https://doi.org/10.1038/s41467-019-09747-0.

(13) Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7* (10). https://doi.org/10.1371/journal.pcbi.1002195.

(14) Sondergaard, D.; Pedersen, C. N. S.; Greening, C. HydDB: A Web Tool for Hydrogenase Classification and Analysis. *Sci. Rep.* **2016**, 061994. https://doi.org/10.1101/061994.

(15) Yin, Y.; Mao, X.; Yang, J.; Chen, X.; Mao, F.; Xu, Y. DbCAN: A Web Resource for Automated Carbohydrate-Active Enzyme Annotation. *Nucleic Acids Res.* **2012**, *40* (W1), 445–451. https://doi.org/10.1093/nar/gks479.

(16) Moura, A.; Soares, M.; Pereira, C.; Leitão, N.; Henriques, I.; Correia, A. INTEGRALL: A Database and Search Engine for Integrons, Integrases and Gene Cassettes. *Bioinformatics* **2009**, *25* (8), 1096–1098. https://doi.org/10.1093/bioinformatics/btp105.

(17) Wang, Y.; Wegener, G.; Hou, J.; Wang, F.; Xiao, X. Expanding Anaerobic Alkane Metabolism in the Domain of Archaea. *Nat. Microbiol.* **2019**, *4* (April), 595–602. https://doi.org/10.1038/s41564-019-0364-2.

(18) Siguier, P. ISfinder: The Reference Centre for Bacterial Insertion Sequences. *Nucleic Acids Res.* **2006**, *34* (90001), D32–D36. https://doi.org/10.1093/nar/gkj014.

(19) Aziz, R. K.; Bartels, D.; Best, A. A.; DeJongh, M.; Disz, T.; Edwards, R. A.; Formsma, K.; Gerdes, S.; Glass, E. M.; Kubal, M.; et al. The RAST Server: Rapid Annotations Using Subsystems Technology. *BMC Genomics* **2008**, *9* (1), 75. https://doi.org/10.1186/1471-2164-9-75.

(20) Heider, J.; Szaleniec, M.; Martins, B. M.; Seyhan, D.; Buckel, W.; Golding, B. T. Structure and Function of Benzylsuccinate Synthase and Related Fumarate-Adding Glycyl Radical Enzymes. *J. Mol. Microbiol. Biotechnol.* **2016**, *26* (1–3), 29–44. https://doi.org/10.1159/000441656.

(21) Sullivan, M. J.; Petty, N. K.; Beatson, S. A. Easyfig: A Genome Comparison Visualizer. *Bioinformatics* **2011**, *27* (7), 1009–1010. https://doi.org/10.1093/bioinformatics/btr039.