Supplemental Material

Genome analysis of *Endomicrobium proavitum* reveals loss and gain of relevant functions during the evolution of intracellular symbionts

Hao Zheng, Carsten Dietrich, Andreas Brune

Description

SI_Figures_Zheng_Endomicrobia.pdf

Fig. S1 Phylogenetic analysis of the glucose 6-phosphate transporter UhpC.

Fig. S2 Phylogenetic analysis of the catalytic subunit of the ferredoxin/flavodoxin dependent 2-oxoacid oxidoreductases.

Fig. S3 Phylogenetic analysis of the catalytic subunit (EchE) of [NiFe]-hydrogenases.

Fig. S4 Phylogenetic analysis of isopropylmalate/citramalate/homocitrate synthases.

Fig. S5 Phylogenetic analysis of transporter proteins for aromatic (A) amino acids, (B) proline, and (C) serine.

Fig. S6 Phylogenetic analysis of amino acid sequences of ThiE (A) and ThiH (B).

SI_Dataset_S1-2_Zheng_Endomicrobia.xlsx

Dataset S1 Complete list of genes in the genome of *Endomicrobium proavitum*. Genes with orthologs (amino acid sequence similarity > 50%) in the genomes of "*Ca.* Endomicrobium trichonymphae" strain Rs-D17, *Elusimicrobium minutum*, or both genomes are marked.

Dataset S2 Genes of "*Ca.* Endomicrobium trichonymphae" strain Rs-D17 that have no homologs in *Endomicrobium proavitum* (amino acid sequence similarity < 50%). CDS with assigned functions are highlighted, and the amino acid sequence similarities to the top BLAST hits in sequenced genomes are given.

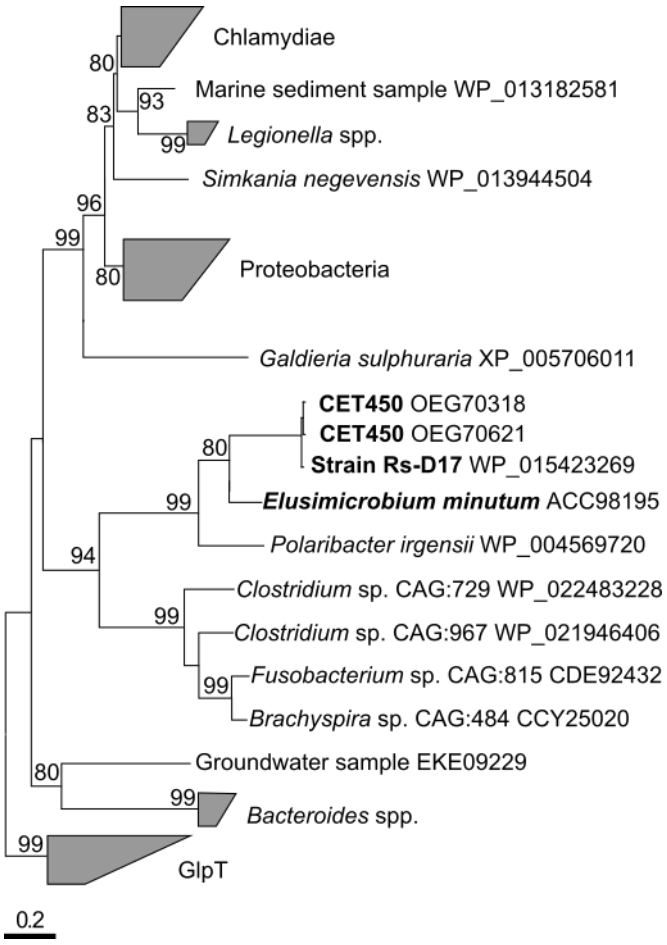

Fig. S1 Phylogenetic analysis of the glucose 6-phosphate transporter UhpC. The maximum-likelihood tree is based on an alignment of 525 amino acid positions of the top 150 BLAST hits of the ortholog in strain Rs-D17 against the UniRef90 database (1). The tree was inferred under the WAG+G+F model and rooted with the bacterial glycerol 3-phosphate transporter GlpT, a paralog of UhpC. Orthologs from the *Elusimicrobia* phylum are shown in bold.

Fig. S2 Phylogenetic analysis of the catalytic subunit of the ferredoxin/flavodoxin dependent 2-oxoacid oxidoreductases encoded in the genomes of *E. proavitum* and *Ca.* E. trichonymphae strain Rs-D17. The unrooted maximum-likelihood tree is based on an alignment of amino acid sequences of representative orthologs in other bacteria, including those with the top 100 bitscores displayed using the *homology toolkit* of the IMG/ER platform (2). Orthologs from the *Elusimicrobia* phylum are shown in bold.

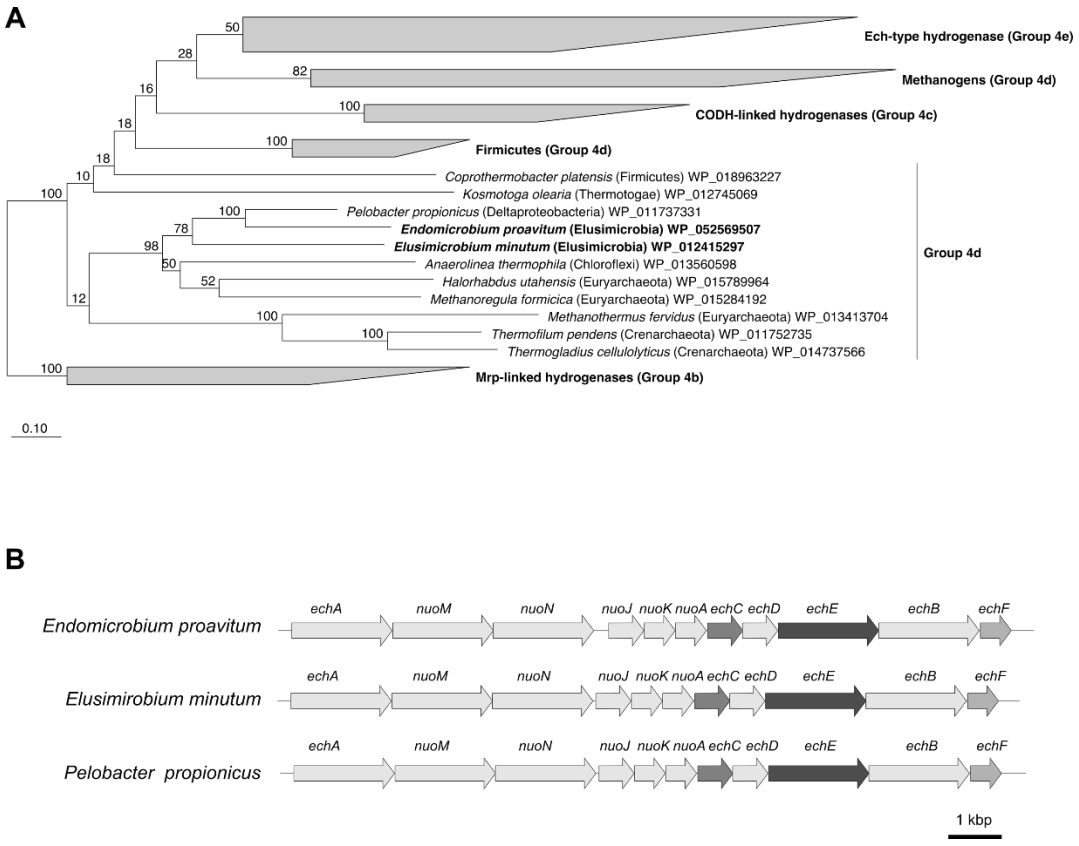

Fig. S3 (A) Phylogenetic analysis of the catalytic subunit (EchE) of [NiFe]-hydrogenases (Group 4), focusing on the relationship between orthologs from *E. proavitum* and *El. minutum* (marked in bold) and other bacteria and archaea. The maximum-likelihood tree is based on the comprehensive dataset reported by Greening *et al.* (3). (B) Organization of the gene sets encoding the [NiFe]-hydrogenases in *E. proavitum* and bacteria with the most closely related orthologs of the *echE* gene.

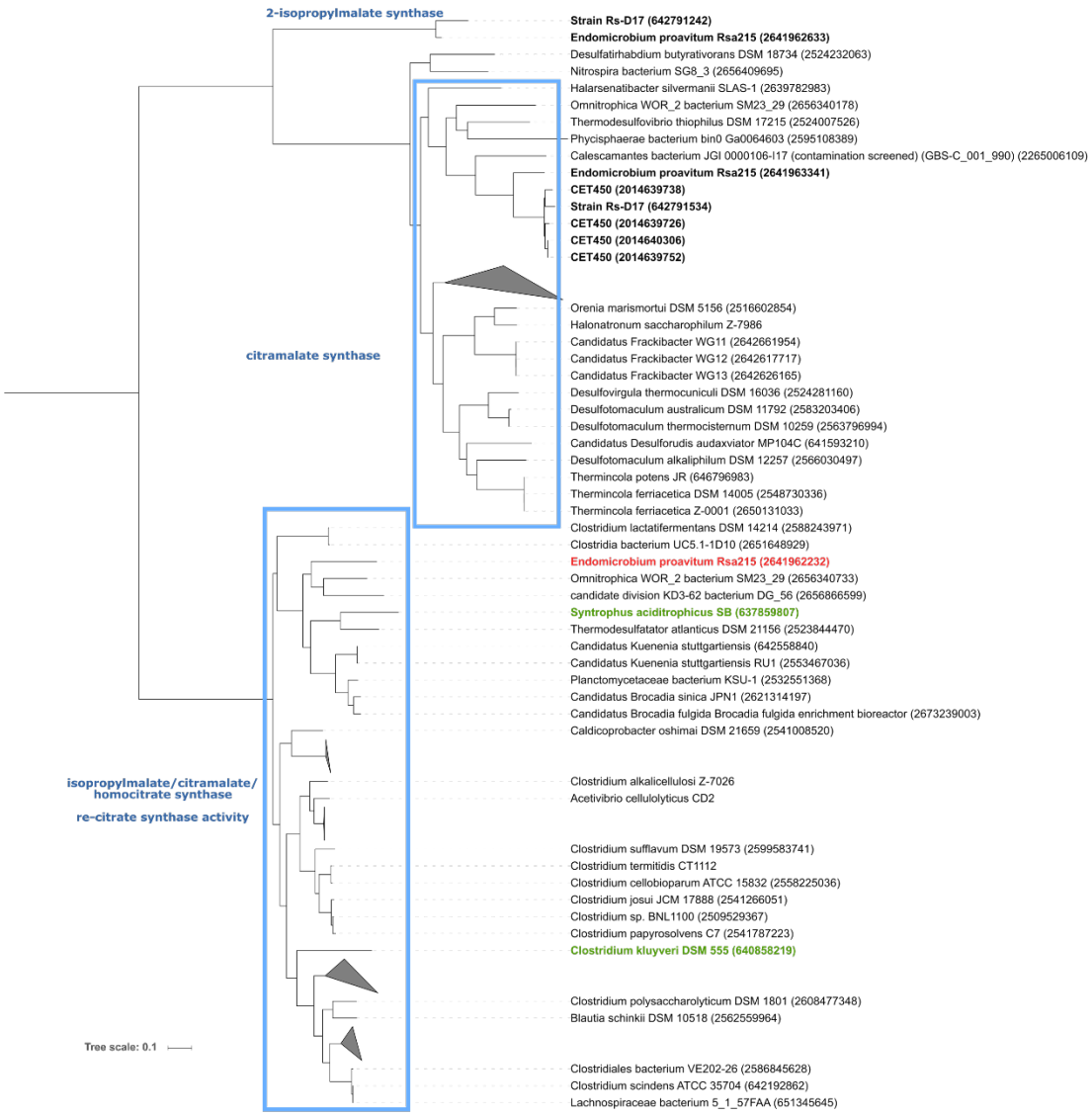

Fig. S4 Phylogenetic analysis of isopropylmalate/citramalate/homocitrate synthases encoded in the genomes of *E. proavitum* and strain Rs-D17. The unrooted maximum-likelihood tree is based on an alignment of amino acid sequences of representative orthologs in other bacteria, including those with the top 100 bitscores displayed using the *homology toolkit* of the IMG/ER platform (2). The orthologs with *re*-citrate synthase activity (green), the putative *re*-citrate synthase of *E. proavitum* (red), and orthologs annotated as *bona fide* isopropylmalate and citramalate synthases in *Endomicrobia* (black) are highlighted.

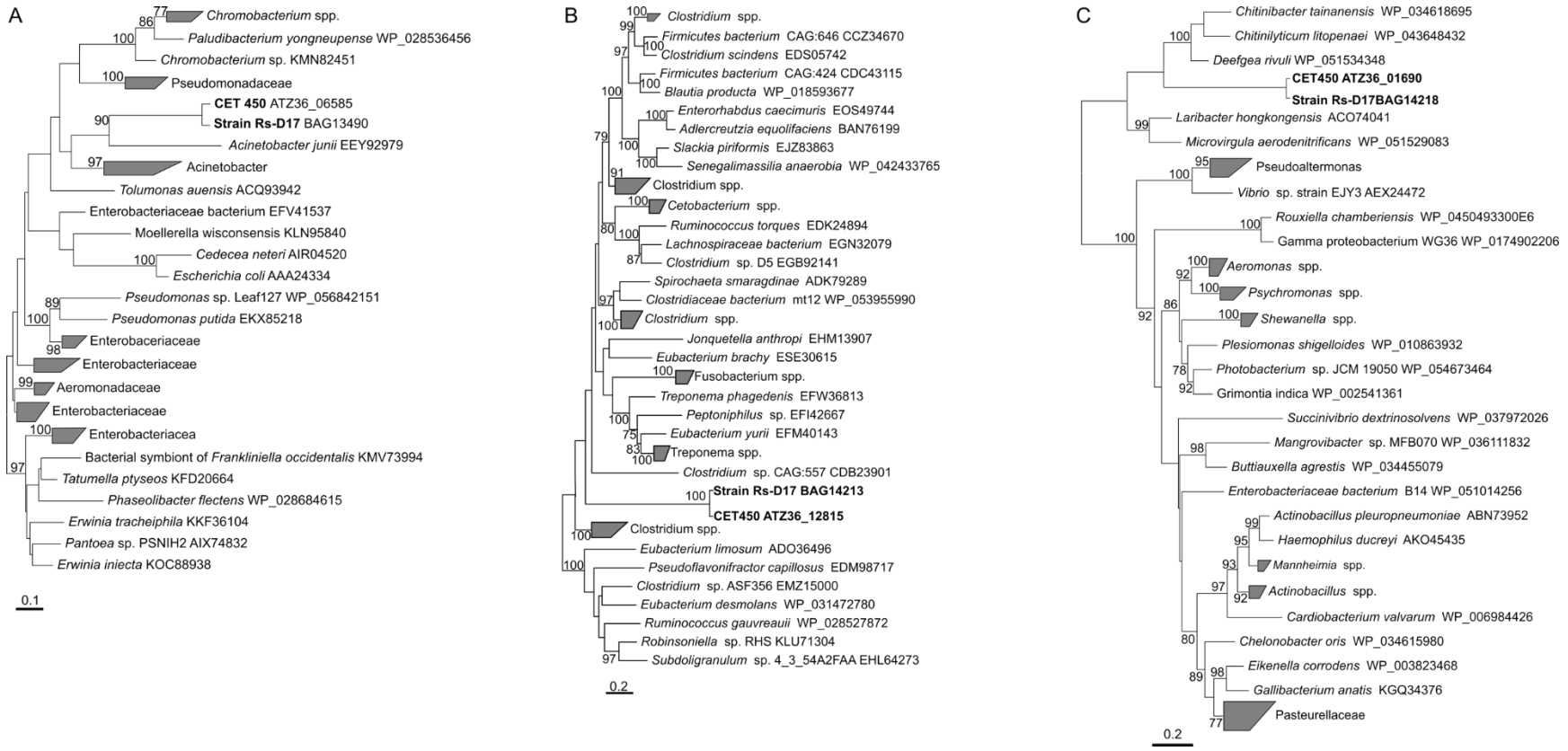

Fig. S5 Phylogenetic analysis of transporter proteins for (A) aromatic amino acid (AroP), (B) proline (ProT), and (C) serine (SdaC). The unrooted maximum-likelihood trees are based on alignments of amino acid sequences of the top 50 BLAST hits of the orthologs in strain Rs-D17 against the UniRef90 database (1). The trees were inferred under the WAG+G+I+F model. Orthologs in *Endomicrobia* endosymbionts are shown in bold.

Fig. S6 Phylogenetic analysis of amino acid sequences of ThiE (A) and ThiH (B). The unrooted maximum-likelihood trees are based on alignments of amino acid sequences of the top BLAST hits of the orthologs in strain Rs-D17 against the UniRef90 database (1). The trees were inferred under the WAG+G+I+F model. Orthologs from *Endomicrobia* are shown in bold

References

1. **Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH.** 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23:**1282-1288.

2. **Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC.** 2009. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics **25:**2271–2278.

3. **Greening C, Biswas A, Carere CR, Jackson CJ, Taylor MC, Stott MB, Cook GM, Morales SE.** 2016. Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. ISME J **10:**761-777.