

FIG S7. The evolutionary trajectory of six *Rickettsia* biosynthetic pathways that contain holes, or “missing enzymes”.

Our *Rickettsia* metabolic reconstruction identified six holes occurring in pathways for the synthesis of CMP-3-deoxy- β -D-manno-octulosonate (CMP-Kdo), Diaminopimelate (DAP), CDP-diacylglycerol (CDP-DG), terpenoid backbones, ubiquinone (CoQ₈), and queuosine (see manuscript for further details). These pathways were analyzed in the genomes of Rickettsiales and Holosporales to determine the presence/absence of the missing enzymes. If present, phylogenies were estimated to determine if the genes encoding these enzymes were present in the Rickettsiales/Holosporales common ancestor.

(A) *E. coli* enzymes corresponding to the holes in the *Rickettsia* CMP-Kdo (KdsC), DAP (DapC), CDP-DG (PlsX, PlsY or PlsB), terpenoid backbones (IspA), CoQ₈ (UbiC, Ubil), and queuosine (QueG) pathways were used as queries in blastp searches against the Rickettsiales and Holosporales databases at NCBI. For significant matches, only top hits with coverage to the query greater than 85% were considered. The four obtained rickettsial homologs were then used as queries in blastp searches against the *Alphaproteobacteria* (excluding Rickettsiales and Holosporales) and the NR (excluding *Alphaproteobacteria*) databases at NCBI. Again, for significant matches, only top hits with coverage to the query greater than 85% were considered.

(B) Phylogenomics analysis of CMP-Kdo, DAP, CDP-DG, terpenoid backbones, CoQ₈, and queuosine pathways across selected Holosporales and Rickettsiales genomes. The presence/absence of enzymes corresponding to the holes in the *Rickettsia* pathways are shown at right. Genome-based phylogeny was estimated on 105 orthologous groups (OGs) of proteins (single copy CDS in over 95% of genomes). We sampled 53 Rickettsiales/Holosporales and three outgroup *Alphaproteobacteria* genomes: IDs for sequences retrieved from PATRIC (P) and NCBI (N) are shown for each taxon. Several genomes were specifically annotated for this project: *Caedibacter varicaedens* (NCBI, NZ_BBVC00000000.1); “*Candidatus* Arcanobacter lacustris” str. SCGC (NCBI, JYHA00000000.1); “*Candidatus* Paracaedibacter symbiosus” (NCBI,

NZ_JQAK000000000.1); “*Candidatus Xenolissoclinum pacificiensis*” str. L6 (NCBI, AXCJ000000000.1); *Rickettsia tamurae* str. AT-1 (NCBI, CCMG01000000.1); “*Candidatus Hepatobacter penaei*” str. NHPB (NCBI, JQAJ000000000.1); *Rickettsia* endosymbiont of *Adalia bipunctata* (unpublished); *Rickettsia* endosymbiont of *Ichthyophthirius multifiliis* (unpublished); *Rickettsia* sp. MEAM1 str. *Bemisia tabaci* (unpublished). Data for Rickettsiales endosymbiont of *Trichoplax adhaerens* is described elsewhere (1). OGs were generated using FastOrtho, a modified version of OrthoMCL (2). Multiple sequence alignment of each OG was performed using MUSCLE (default parameters) (3), with regions of poor alignment (length heterogeneous regions) masked using Gblocks (4). All modified alignments were concatenated into a single dataset for phylogeny estimation. Using PhyloBayes MPI (5), we analyzed the dataset with the CAT model of substitution, a nonparametric method for modeling site-specific features of sequence evolution (6, 7). Given the strong base compositional bias of rickettsial genomes (~30 %GC), the ability of the CAT model to accommodate saturation due to convergences and reversions (8) is of substantial importance for estimating rickettsial phylogeny, as demonstrated by us and others (1, 9–11). Two independent Markov chains were run in parallel using PhyloBayes MPI v.1.2e under the CAT-GTR model, with the bipartition frequencies analyzed at various time points using the bpcomp program. For tree-building, appropriate burn-in values were determined by plotting the log likelihoods for each chain over sampled generations (time). Analyses were considered complete when the maximum difference in bipartition frequencies between the two chains was less than 0.1. Ultimately, a burn-in value of 1000, with sampling every 2 trees, was used to build a consensus tree.

(C-F) Phylogeny estimations of DapC (C), PlsX (D), PlsY (E), and IspA (F) proteins. Datasets for each protein were constructed as follows: the rickettsial protein (panel A) was used in blastp queries against several taxon-specific databases: 1) “Rickettsiales”, 2) “Holosporales”, 3) “*Alphaproteobacteria* (minus Rickettsiales and Holosporales)”, 4) “*Proteobacteria* (minus *Alphaproteobacteria*)”, 5) “*Bacteria* (minus *Proteobacteria*)”, and 6) “minus *Bacteria*”. The top 5-

10 (query-dependent) subjects from each search resulting in significant (> 40 bits) alignments were all compiled and aligned using MUSCLE v3.8.31 (default parameters). Protein phylogenies were estimated under maximum likelihood with RAxML v8.2.4 (12), using a gamma model of rate heterogeneity and estimation of the proportion of invariant sites. Both the Lee and Gascuel (LG) and Blocks of Amino Acid Substitution Matrix (BLOSUM62) models of amino acid substitution were used, and branch support was assessed with 1,000 pseudo-replications.

REFERENCES

1. **Driscoll T, Gillespie JJ, Nordberg EK, Azad AF, Sobral BW.** 2013. Bacterial DNA sifted from the *Trichoplax adhaerens* (Animalia: Placozoa) genome project reveals a putative rickettsial endosymbiont. *Genome Biol Evol* **5**:621–45.
2. **Li L, Stoeckert CJ, Roos DS.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**:2178–89.
3. **Edgar RC.** 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**:1792–1797.
4. **Talavera G, Castresana J.** 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**:564–77.
5. **Lartillot N, Rodrigue N, Stubbs D, Richer J.** 2013. PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst Biol* **62**:611–5.
6. **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.
7. **Lartillot N, Philippe H.** 2006. Computing Bayes factors using thermodynamic integration. *Syst Biol* **55**:195–207.
8. **Lartillot N, Brinkmann H, Philippe H.** 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evol Biol* **7**

Suppl 1:S4.

9. **Rodríguez-Ezpeleta N, Embley TM.** 2012. The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. *PLoS One* **7**:e30520.
10. **Viklund J, Ettema TJG, Andersson SGE.** 2012. Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. *Mol Biol Evol* **29**:599–615.
11. **Gillespie JJ, Driscoll TP, Verhoeve VI, Utsuki T, Husseneder C, Chouljenko VN, Azad AF, Macaluso KR.** 2014. Genomic diversification in strains of *Rickettsia felis* isolated from different arthropods. *Genome Biol Evol.*
12. **Stamatakis A.** 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313.

A	Hole	Reference protein	Top Rick (% ID)	Top Alpha E value; % ID	Top non-Alpha E value; % ID
	KdsC	P0ABZ4: <i>E. coli</i> str. K12	---	---	---
	DapC	P18335: <i>E. coli</i> str. K12	WP_070064655: <i>Wolbachia pipientis</i> (42%)	4e-130; 50%	1e-122; 48%
	PlsX	P27247: <i>E. coli</i> str. K12	WP_011452016: <i>Neorickettsia sennetsu</i> (45%)	9e-94; 47%	4e-77; 45%
	PlsY	P60782: <i>E. coli</i> str. K12	KKB96285: " <i>Cand. Arcanobacter lacustris</i> " (40%)	5e-50; 49%	2e-36; 49%
	PlsB	P0A7A7: <i>E. coli</i> str. K12	---	---	---
	IspA	P22939: <i>E. coli</i> str. K12	WP_039459403: " <i>Cand. Jidaibacter acanthamoeba</i> " (44%)	7e-109; 59%	2e-72; 55%
	UbiC	P26602: <i>E. coli</i> str. K12	---	---	---
	Ubil	P25535: <i>E. coli</i> str. K12	---	---	---
	QueG	P39288: <i>E. coli</i> str. K12	---	---	---

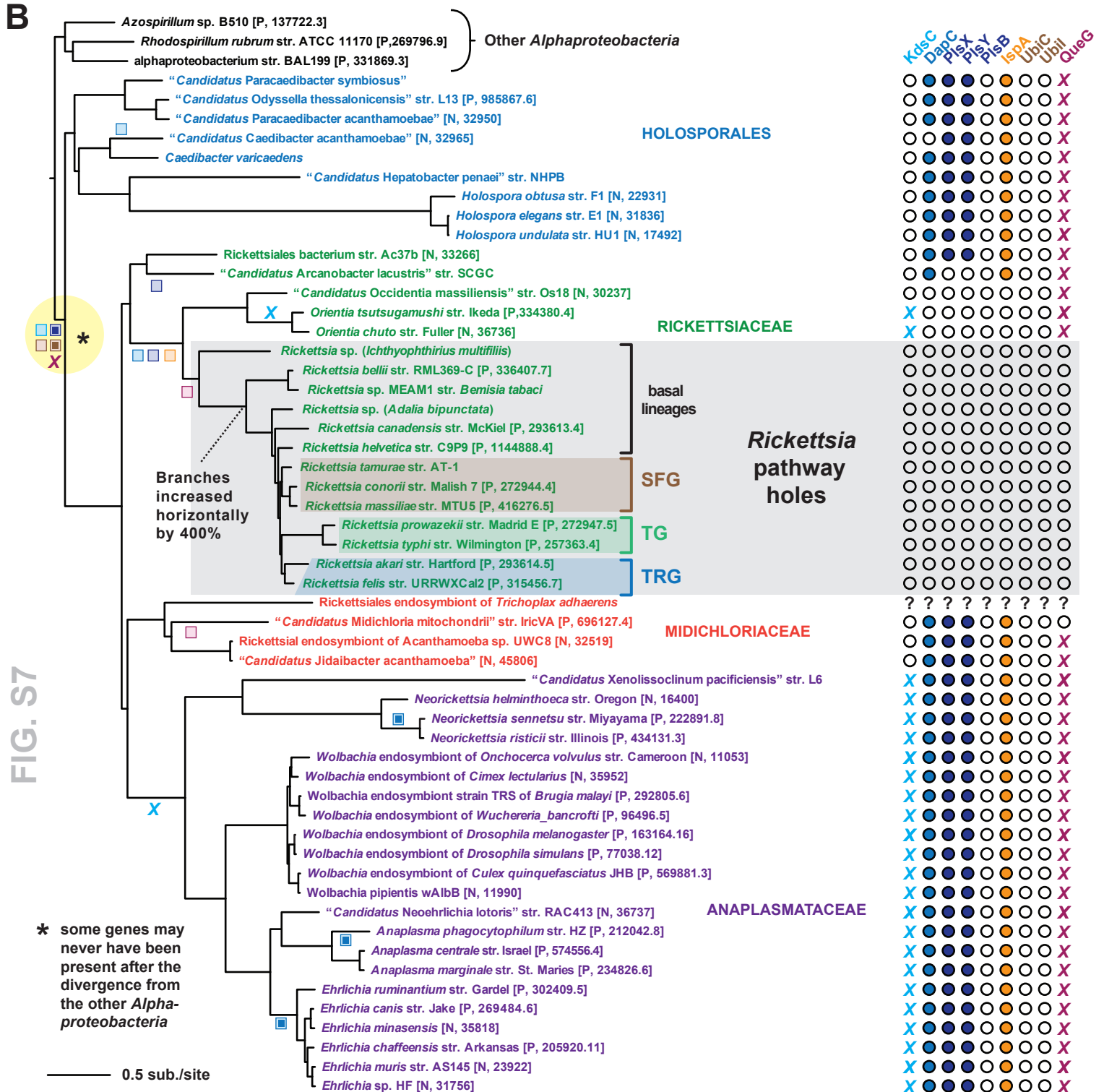


FIG. S7

C

■ Rickettsiales ■ other *Alphaproteobacteria* ■ other bacteria
■ Holosporales ■ other Proteobacteria ■ non-bacteria

DapC: N-succinylidiaminopimelate aminotransferase

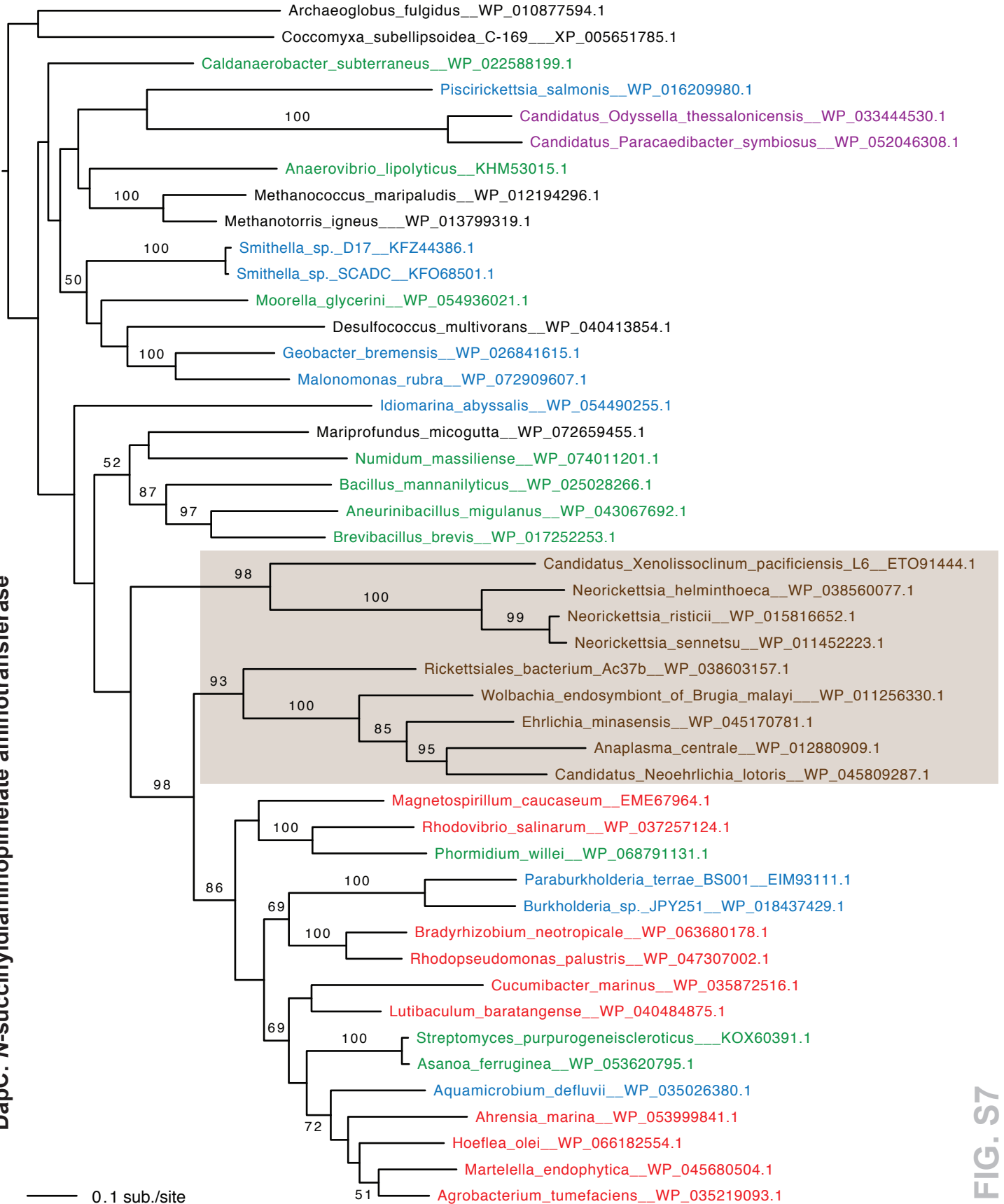
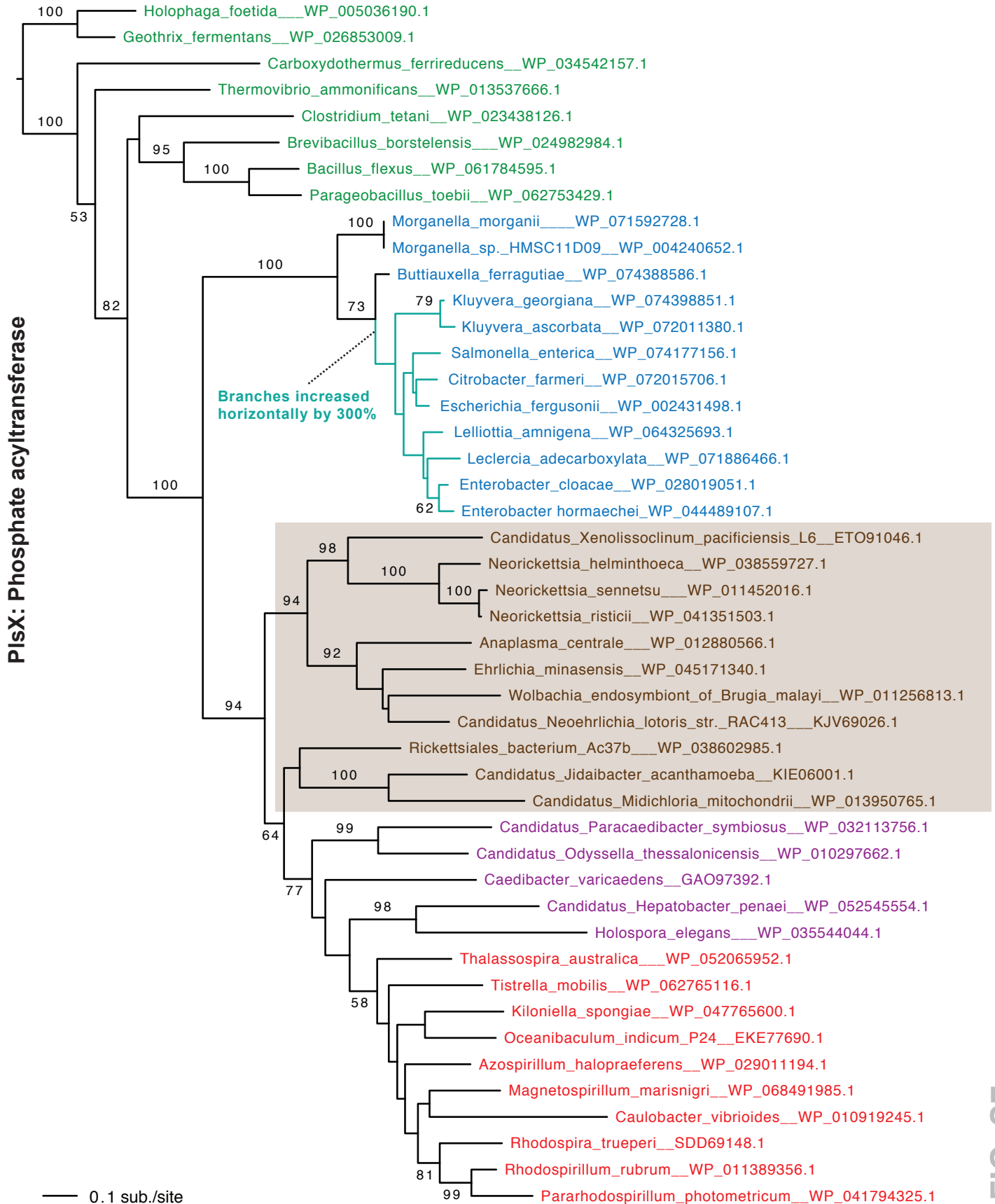


FIG. S7

D

■ Rickettsiales ■ other *Alphaproteobacteria* ■ other bacteria
■ Holosporales ■ other Proteobacteria

**FIG. S7**

■

■ Rickettsiales ■ other *Alphaproteobacteria* ■ other bacteria
■ Holosporales ■ other Proteobacteria ■ non-bacteria

PlsY: glycerol-3-phosphate acyltransferase

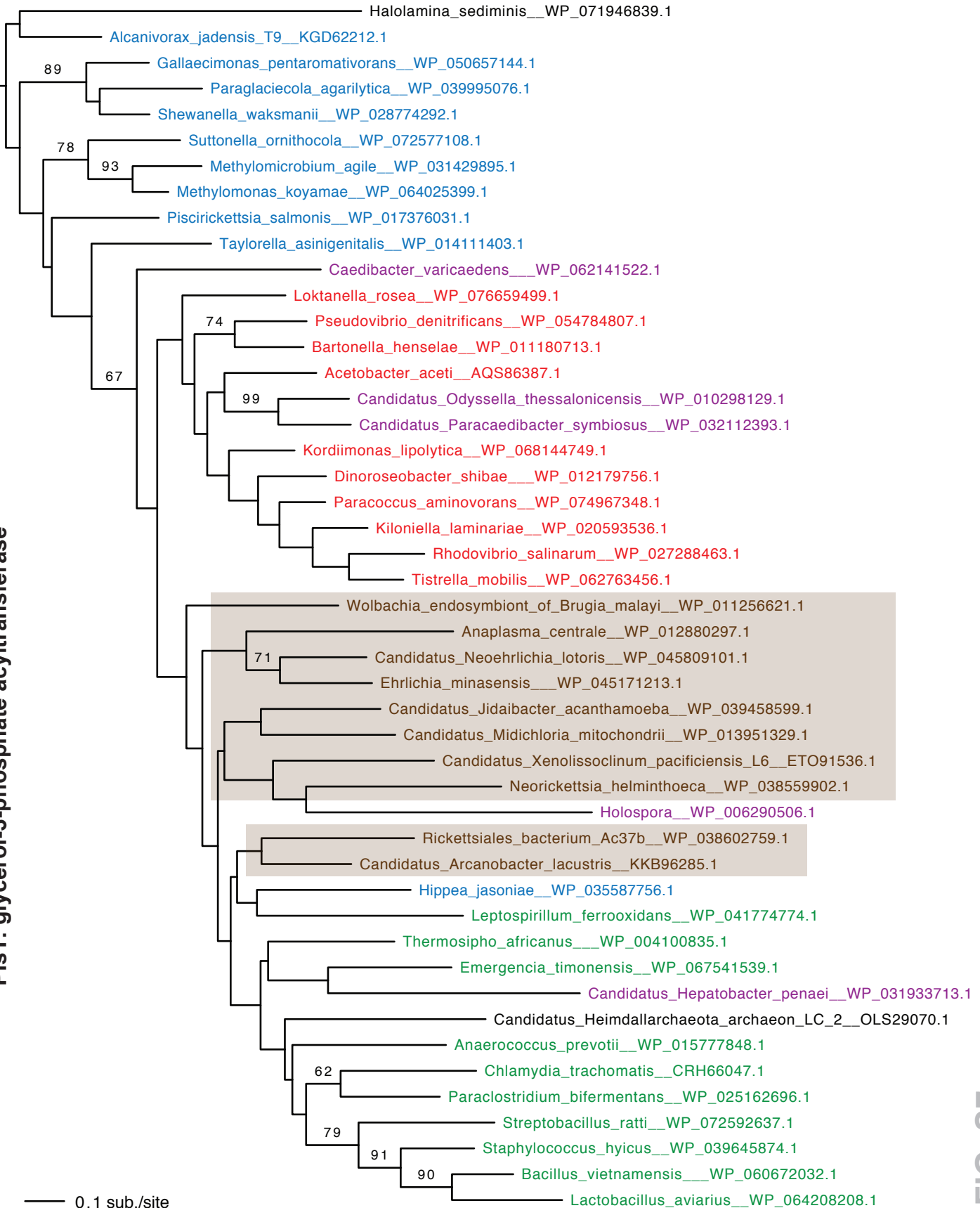


FIG. S7

F

■ Rickettsiales ■ other *Alphaproteobacteria* ■ other bacteria
■ Holosporales ■ other Proteobacteria ■ non-bacteria

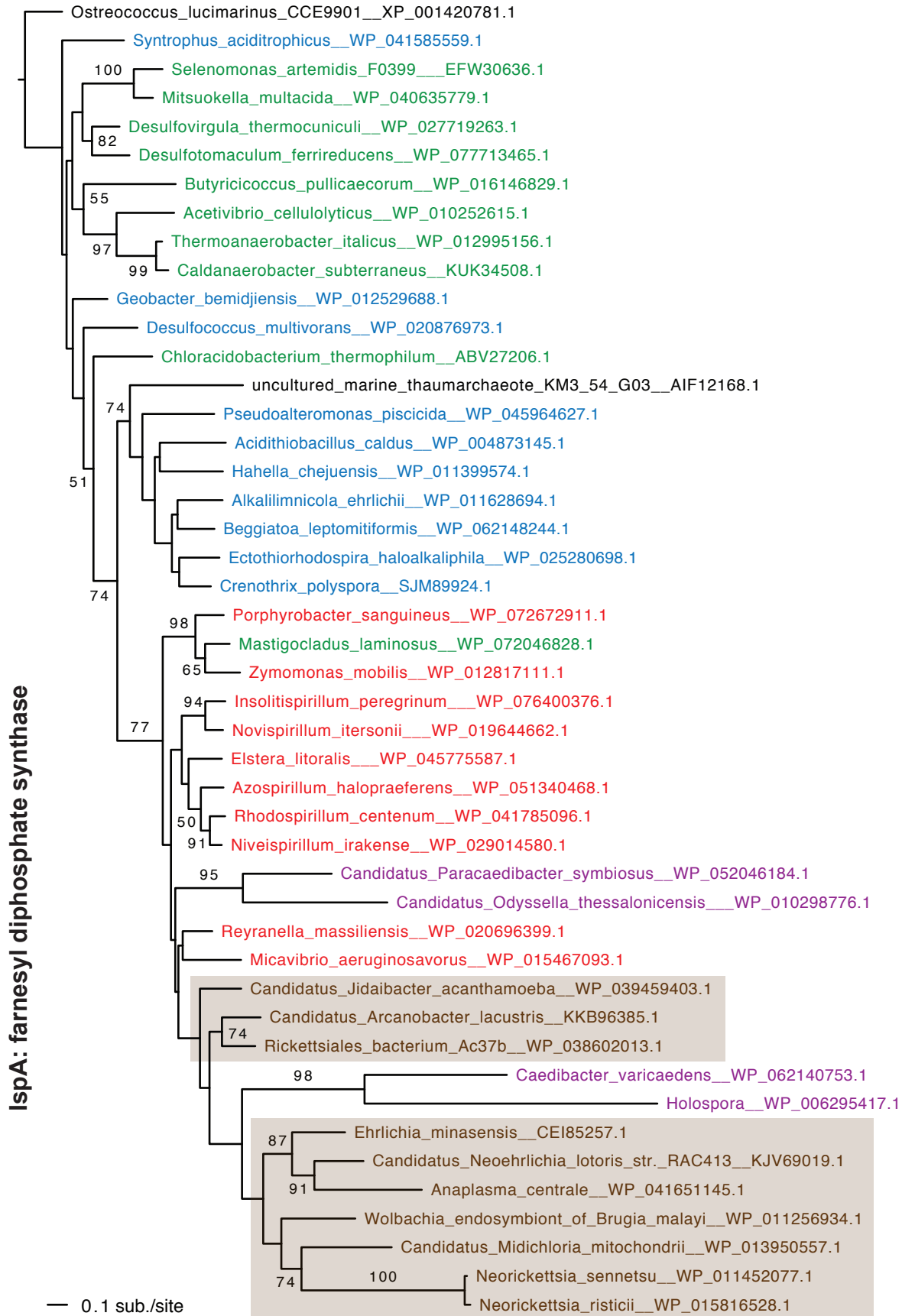


FIG. S7