List of genes studied

Red: genes discarded because of insufficient taxonomic sample
Orange: tRNA synthetases and genes discarded because of too many paralogs, retained for fusion P4
Green: genes retained for fusion P1

Genes	number of species	number of positions	
efg	43	631	not retained in fusion P2
efp	44	177	too many paralogs
efts	45	234	too many paralogs
eftu	43	371	
fmt	45	213	not retained in fusion P2
glna	37	327	insufficient taxonomic sample
glnb	37	403	insufficient taxonomic sample
glnc	33	67	insufficient taxonomic sample
hpt	45	137	not retained in fusion P2
if-1	44	68	
if-2	45	488	not retained in fusion P2
if-3	45	140	
ipp	40	180	insufficient taxonomic sample
ksga	45	151	not retained in fusion P2
npt	42	534	
pdf	45	119	too many paralogs
rba	54	81	
rf1	42	243	
rf2	42	298	
rfr	43	164	
rpl1	44	219	
rpl2	44	264	
rpl3	43	181	
rpl4	44	136	
rpl5	45	176	not retained in fusion P2
rpl6	45	139	
rpl7	45	107	not retained in fusion P2
rpl9	45	76	not retained in fusion P2
rpl10	43	102	
rpl11	43	137	
rpl13	42	136	
rpl14	43	121	
rpl15	45	59	not retained in fusion P2
rpl16	45	133	
rpl17	45	96	
rpl18	45	98	not retained in fusion P2
rpl19	44	109	
rpl20	43	111	
rpl21	43	79	
rpl22	44	93	
rpl23	43	72	
rpl24	45	79	not retained in fusion P2
rpl27	44	77	
rpl28	44	70	too many paralogs
rpl29	45	57	not retained in fusion P2
rpl30	36	55	insufficient taxonomic sample
rpl31	45	63	too many paralogs
rpl32	40	42	insufficient taxonomic sample

Genes	number of species	number of positions	
rpl33	43	47	too many paralogs
rpl34	42	43	
rpl35	39	56	insufficient taxonomic sample
rpl36	43	37	too many paralogs
rps2	44	210	
rps3	45	201	
rps4	44	165	
rps5	45	150	not retained in fusion P2
rps6	45	71	
rps7	44	147	
rps8	45	112	
rps9	44	119	
rps10	38	100	insufficient taxonomic sample
rps11	45	117	
rps12	43	135	
rps13	45	115	
rps14	45	106	too many paralogs
rps15	44	69	
rps16	41	63	insufficient taxonomic sample
rps17	43	70	
rps18	43	62	
rps19	42	82	
rps20	42	71	
rps21	32	80	insufficient taxonomic sample
sp2	44	199	
trmd	44	180	
trua	44	164	
trub	40	164	insufficient taxonomic sample
tRNA synthetase			
ala	44	548	
arg	44	367	
asn	-	-	insufficient taxonomic sample
asp	44	485	
cys	44	316	
gln	-	-	insufficient taxonomic sample
glu	44	308	
gly	-	-	2 classes of sequences
his	43	278	
iso	45	668	
leu	44	667	
lys	-	-	2 classes of sequences
met	44	406	
phe	45	-	forgotten by mistake for the fusion P4
pro	45	344	
ser	44	401	
thr	44	530	
trp	43	307	
tyr	42	357	
val	44	675	

Fig. S1: Phylogeny based on the fusion P1 (8857 positions). This is the same tree as Fig. 1a, except that all the species names are given.

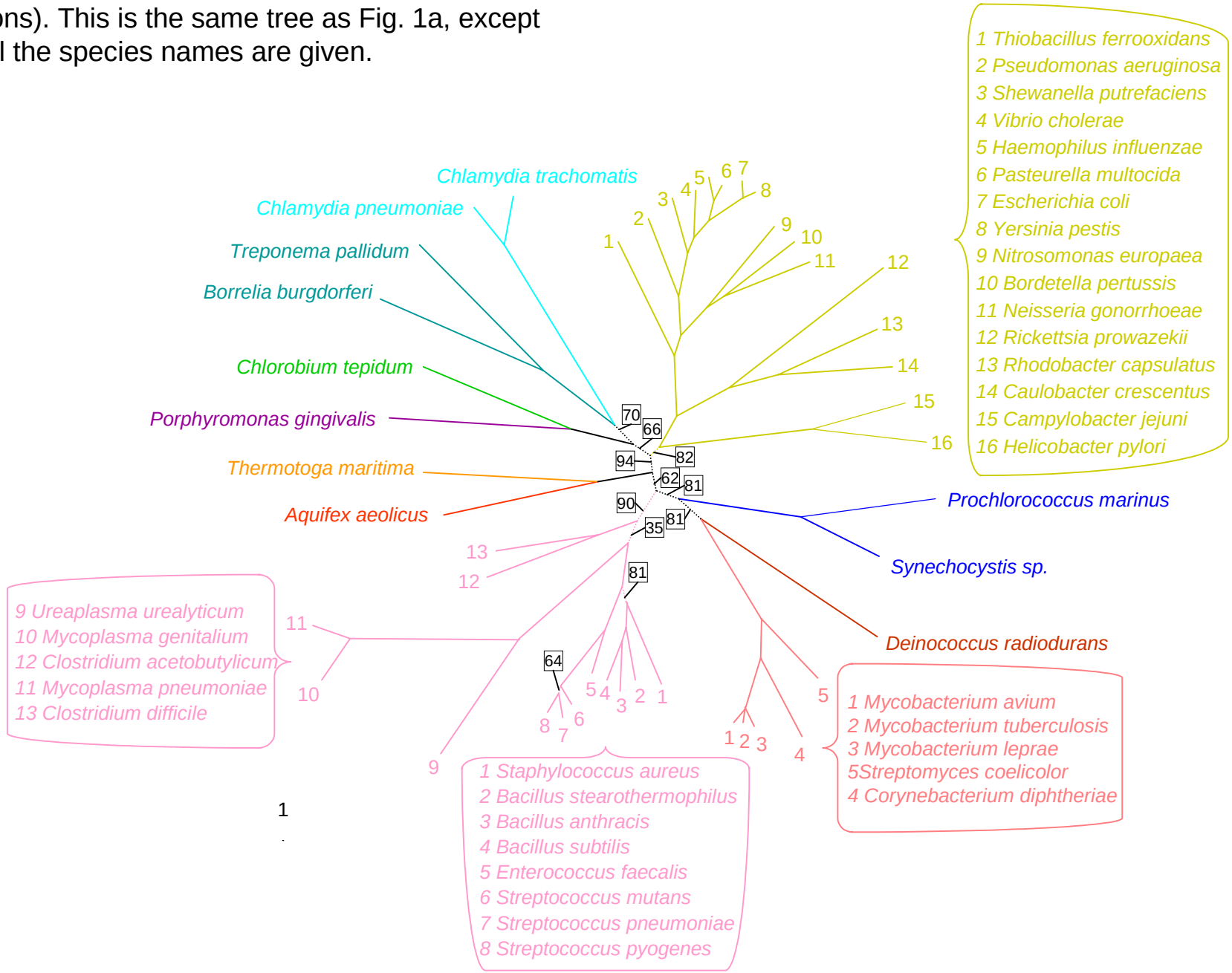


Fig. S2: Phylogeny based on the concatenated SSU and LSU ribosomal RNA (3704 positions). This is the same tree as Fig. 1b, except that all the species names are given.

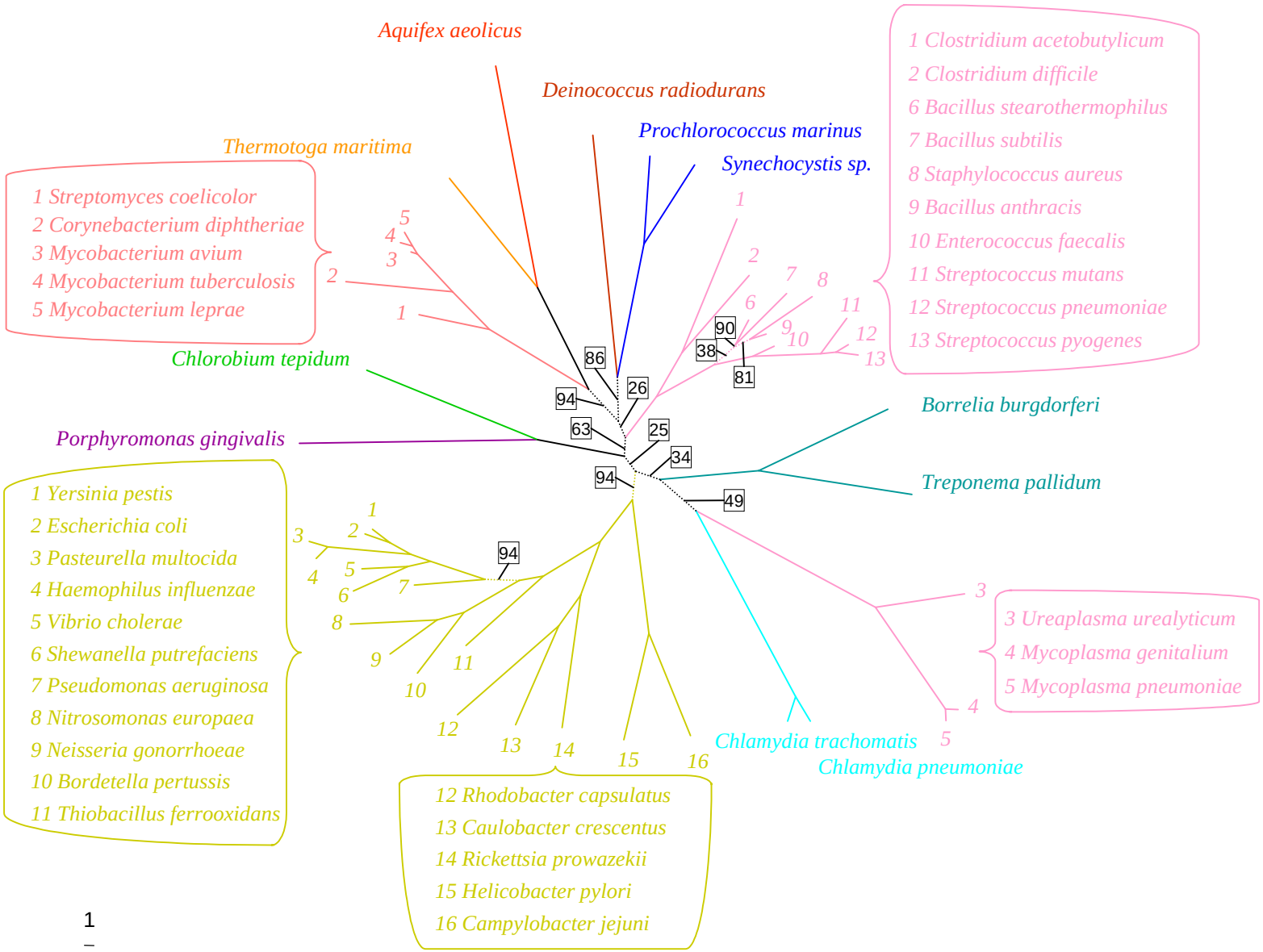
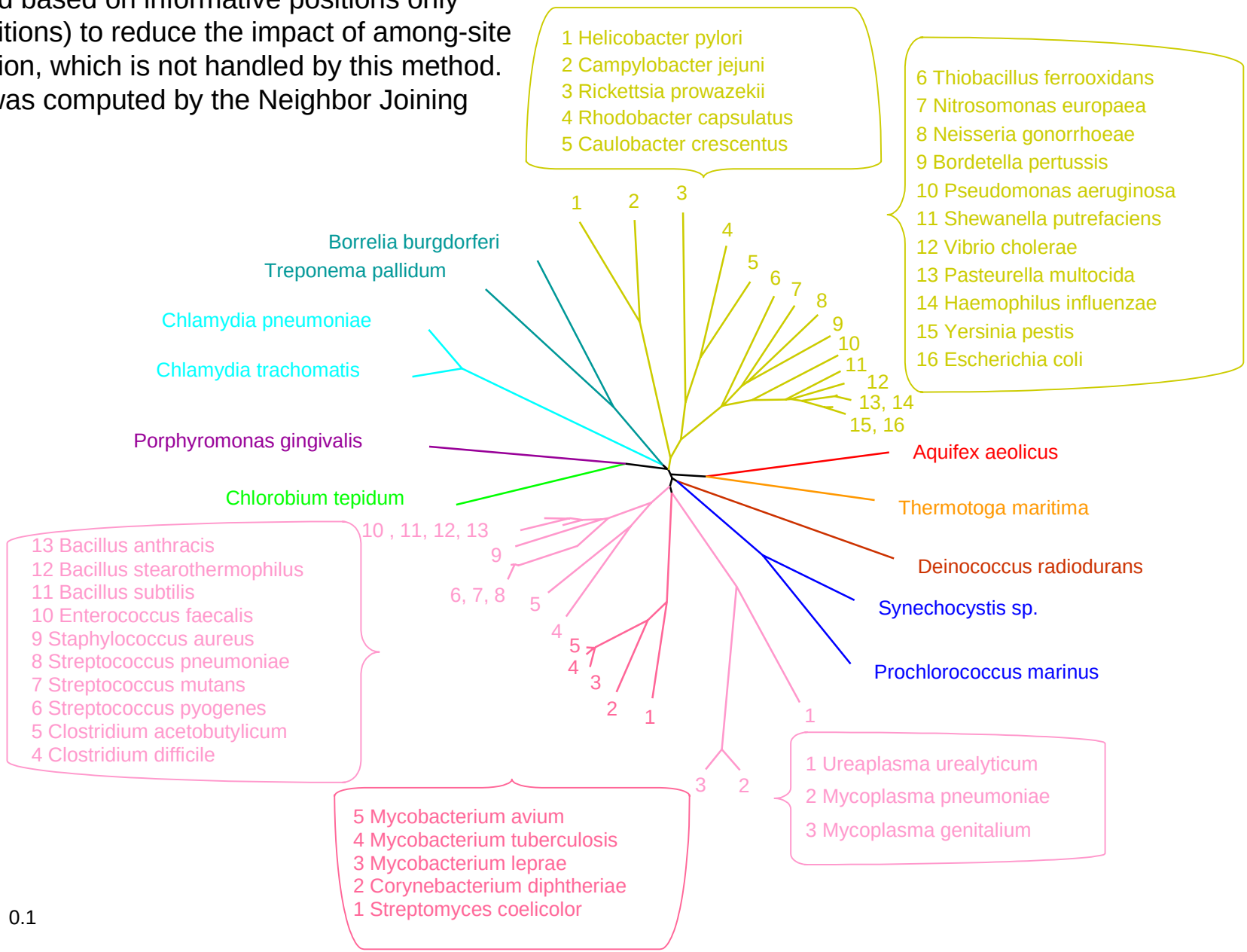


Fig. S3: Phylogeny based on the fusion P1. The evolutionary distances are computed using the log-det method based on informative positions only (7234 positions) to reduce the impact of among-site rate variation, which is not handled by this method. The tree was computed by the Neighbor Joining method.



Selection of topologies for Principal Component Analysis

For 45 species, there are 10^{64} possible topologies. It is therefore impossible to compute the likelihood of all the trees. We tried three different approaches to have a sampling of topologies that correctly represent the tree space (i.e. the 10^{64} topologies) in order to discriminate the genes. First, we used the 2000 best topologies obtained through the quick search option of the protml program with the sequences of the fusion P1. However, the genes were not well separated in the PCA because the 2000 topologies appeared to be very similar (i.e. they corresponded to a single island of the tree space). Second, we used 100 random topologies (generated with the software MacClade). Similarly, the discriminatory power was weak. This is due to the fact that random trees are systematically so far from the best tree of each gene that they are always strongly rejected by the data (and similarly by all the genes). Third, we used the best topologies obtained for each gene and each fusion with protml software (for computing time reason, we used the star decomposition algorithm followed by local rearrangement heuristic to compute the ML tree). However, since only 32 genes have 45 species, only 35 topologies can be obtained (32 + P1 + P4 + rRNA). To increase the number of topologies, we also used topologies inferred from random sub-samples of fusions P1, P4 and R.

In summary, 375 tree topologies for the 45-species sample were chosen to represent the tree space. They included the most likely ones obtained from the ML analyses of (i) the rRNA fusion (all positions or with invariant positions excluded), (ii) the protein fusions P1 and P4, and (iii) all the individual proteins for which the sequences for the 45 bacterial species were available (namely, 22 proteins without *a priori* LGT and 10 genes with *a priori* LGT or duplications). In addition, the topologies derived from (i) 285 random sub-samples of fusion P1, (ii) 37 random sub-samples of fusion P4, and (iii) 22 random sub-samples of the rRNA fusion with identical sizes to those of the 22 proteins without *a priori* LGT, were also employed.

To verify that our sampling of tree topologies, we performed PCA only on the first fifty topologies. This further allowed us to include the fusion P1 with likelihood values computed assuming a Γ law model (through the PUZZLE program with eight rate classes), for which it is not possible to achieve the calculation for the 375 trees. As it can be seen on Fig S4, the PCA obtained with 50 topologies was quite similar to the one with 375 topologies (Fig. 2), suggesting that our tree sampling is correct. Interestingly, the impact of the Γ law model was less important for proteins than for rRNA (compare the distance between P1 and P1-Γ and between R and R-Γ), as expected from the values of the shape parameter α (0.71 for P1 and 0.32 for R).

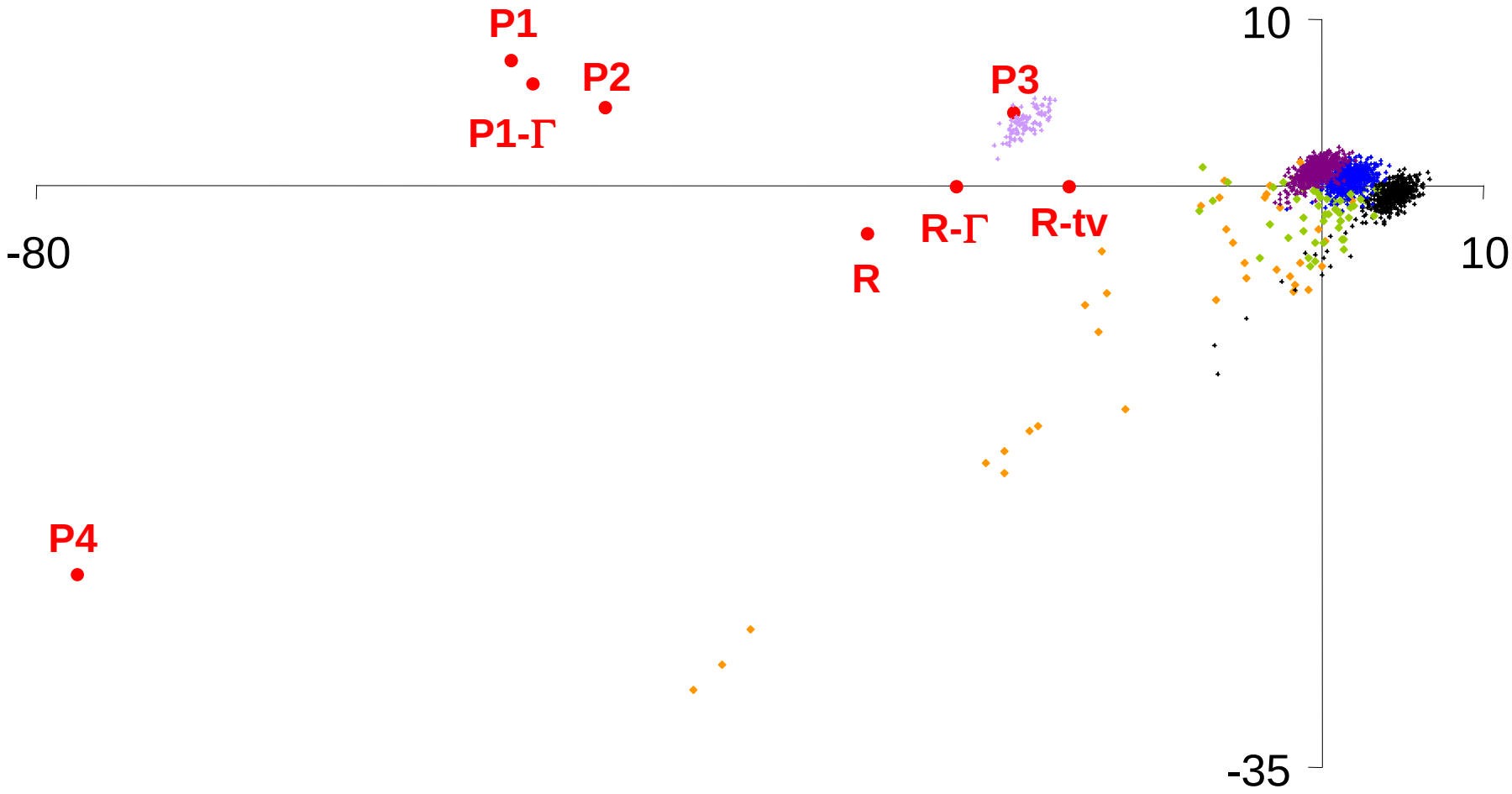


Fig. S4: Principal component analysis based on the likelihood of the first 50 topologies. This figure is very similar to the one obtained with 375 topologies (Fig. 2), indicating that our sampling of topologies have small influence on our conclusions. With this reduced sample, it is possible to compute the likelihood of fusion P1 with a Γ law model. Interestingly, the new point (P1-Γ) is very close to P1.

Fig. S5: Phylogeny based on the Ala-tRS sequences (548 positions). The evolutionary distances were computed using the PUZZLE software with a JTT+Γ model. The tree was computed by the Neighbor Joining method.

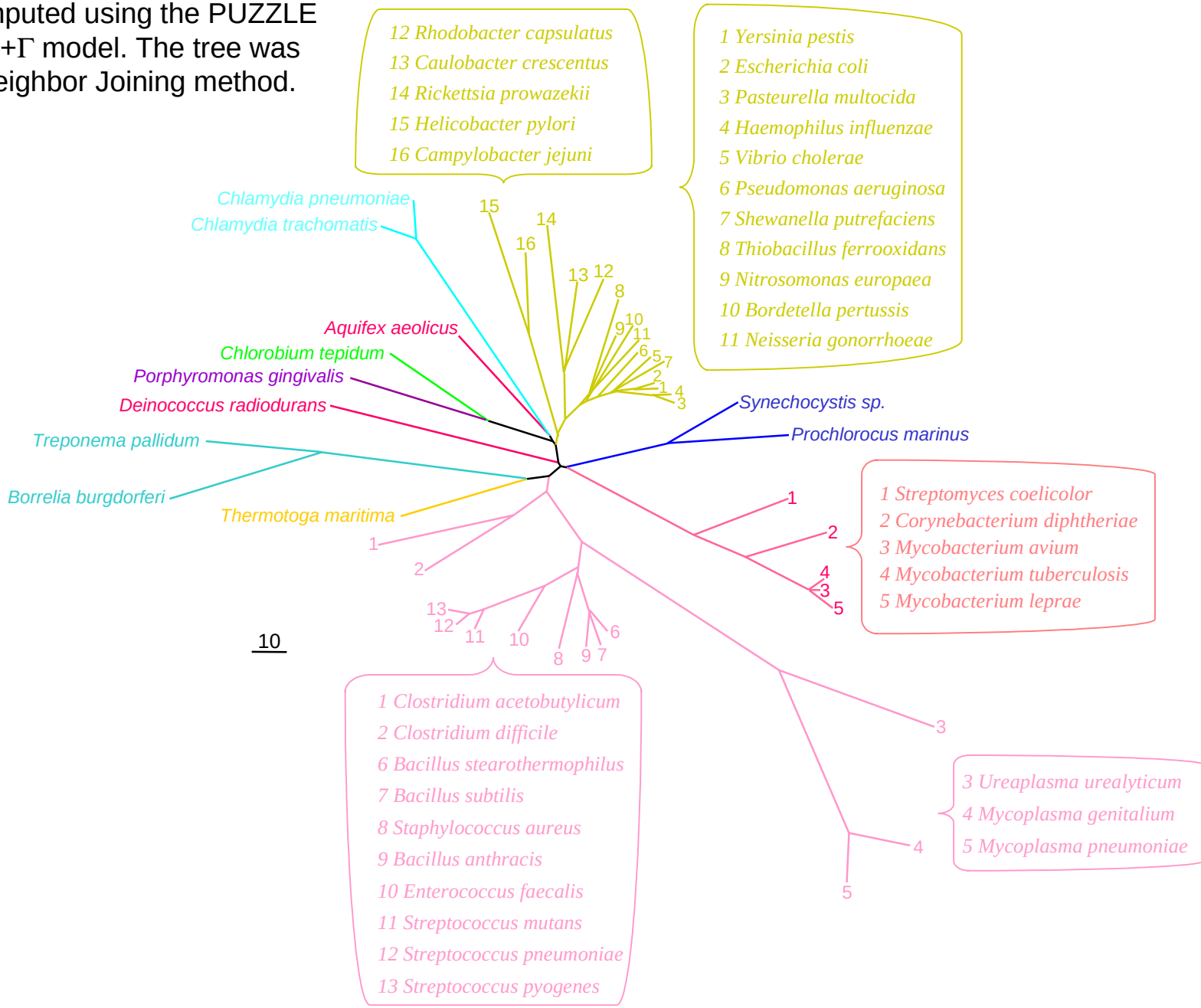
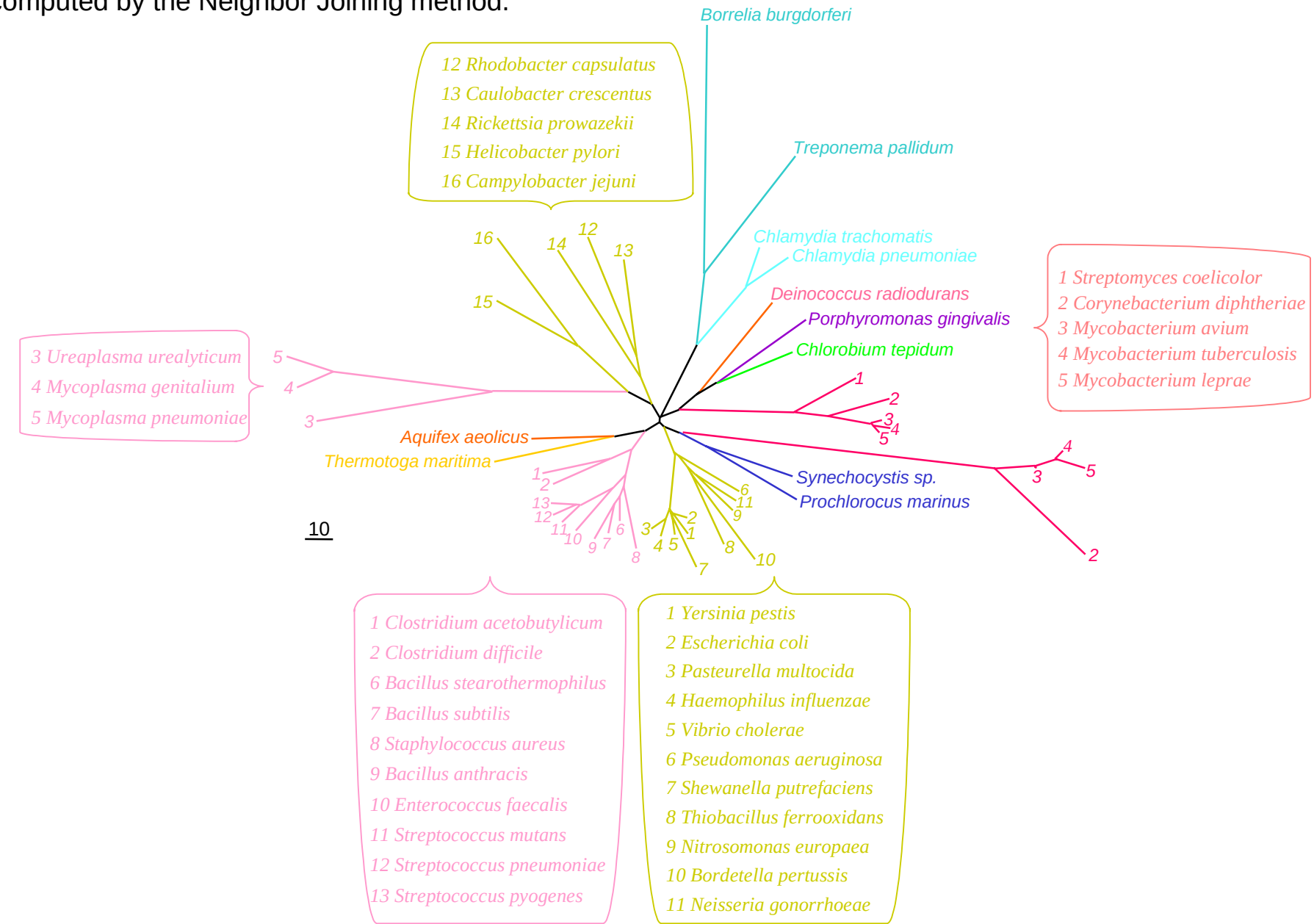


Fig. S6: Phylogeny based on the Cys-tRS sequences (316 positions). The evolutionary distances were computed using the PUZZLE software with a JTT+Γ model. The tree was computed by the Neighbor Joining method.



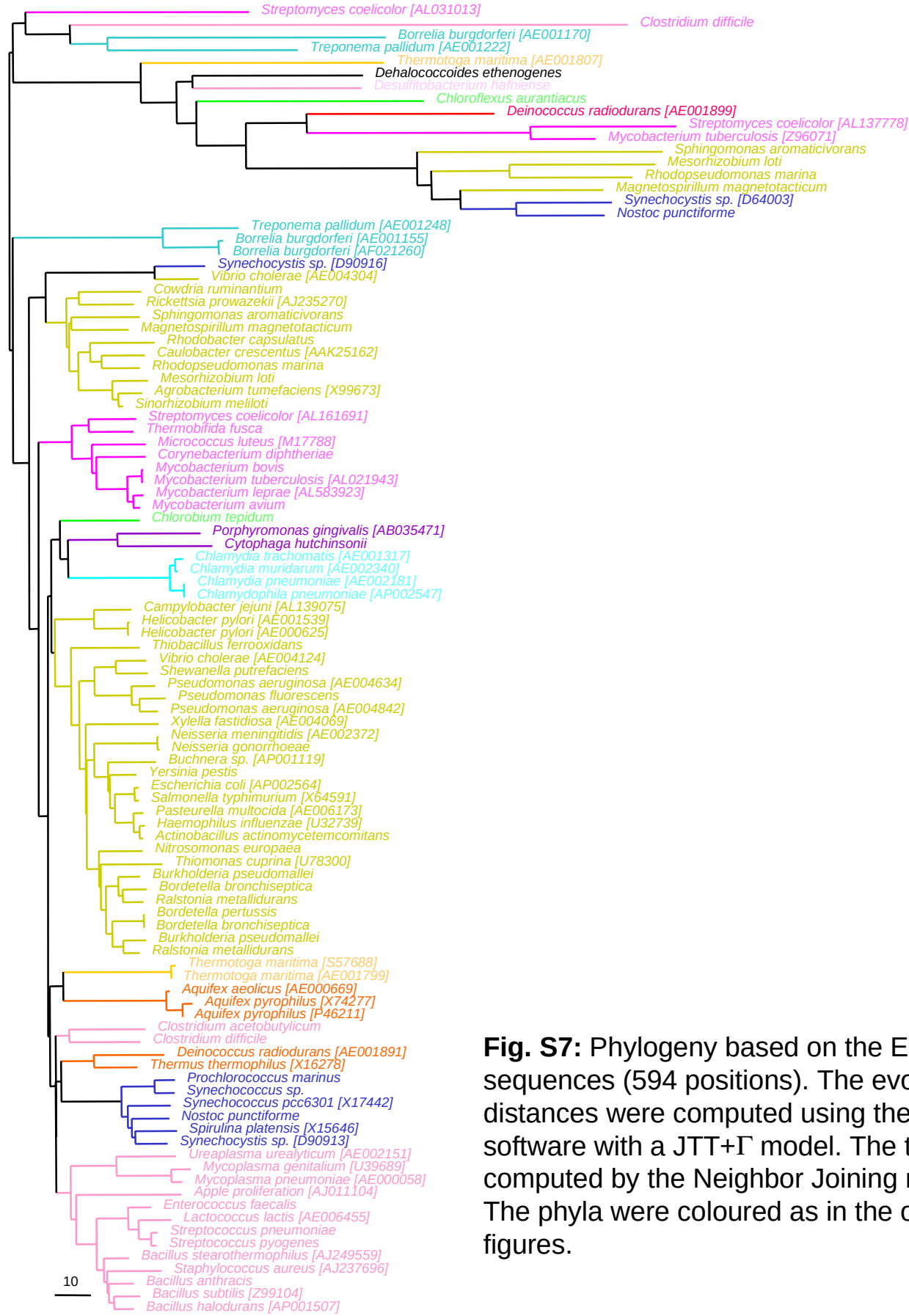


Fig. S7: Phylogeny based on the EF-G sequences (594 positions). The evolutionary distances were computed using the PUZZLE software with a JTT+Γ model. The tree was computed by the Neighbor Joining method. The phyla were coloured as in the other figures.