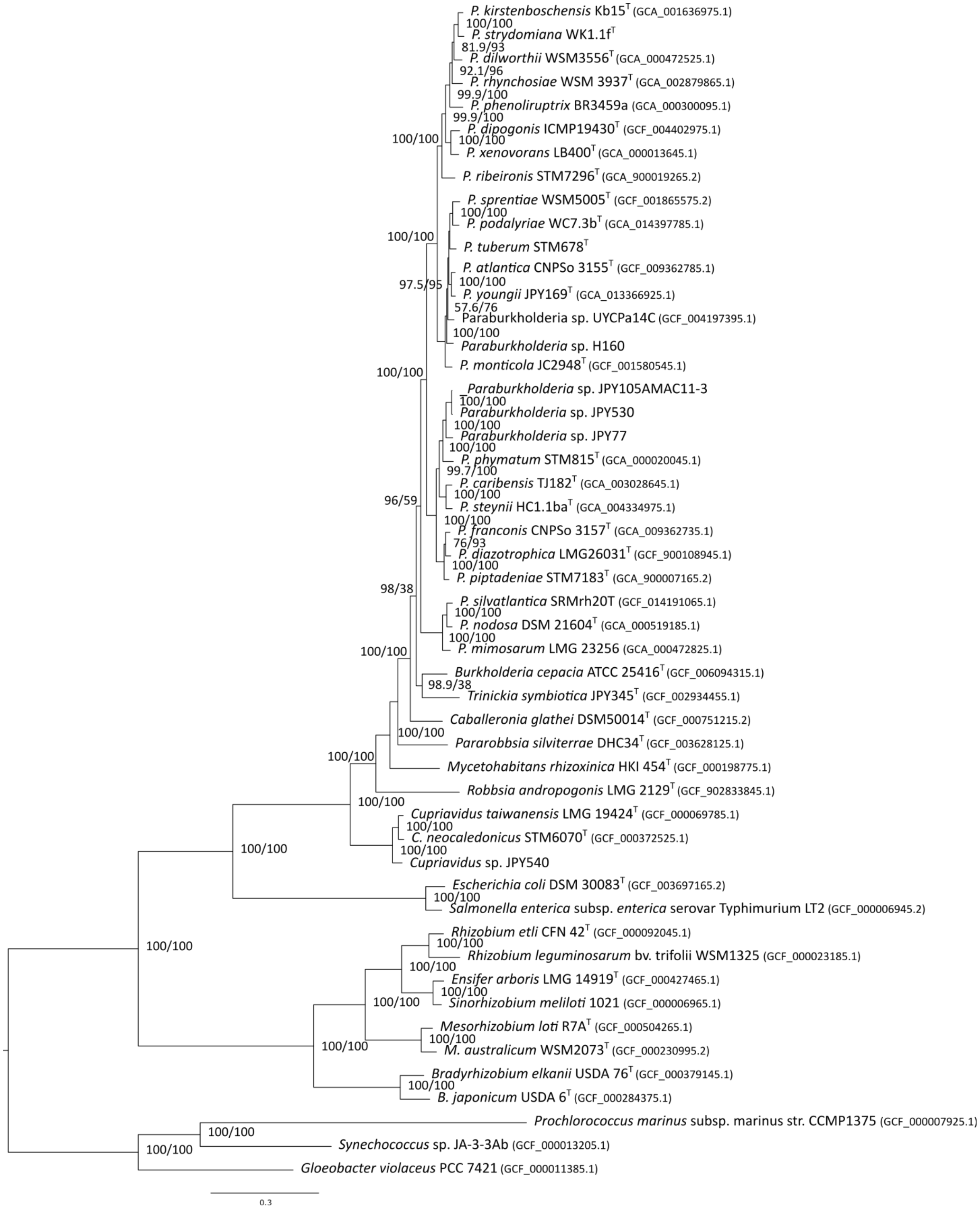
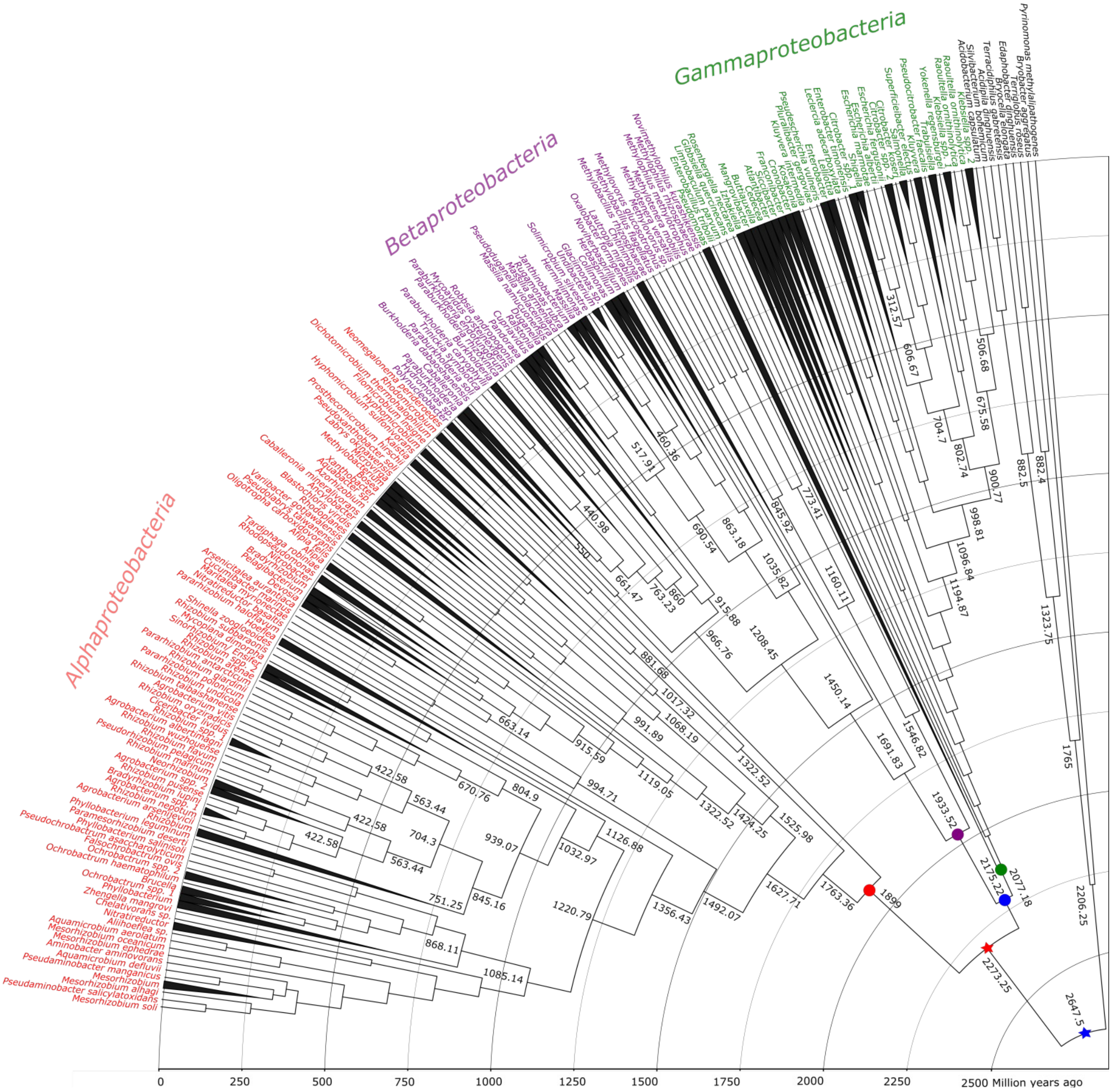
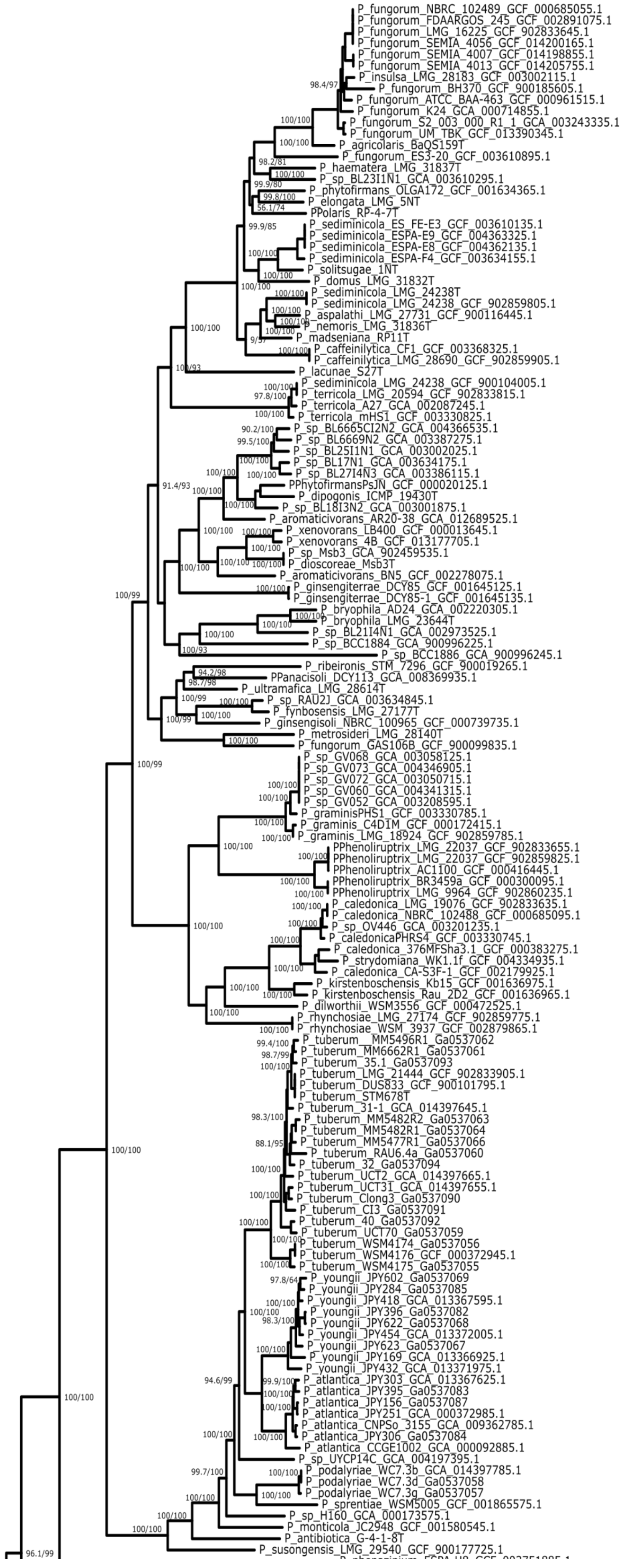
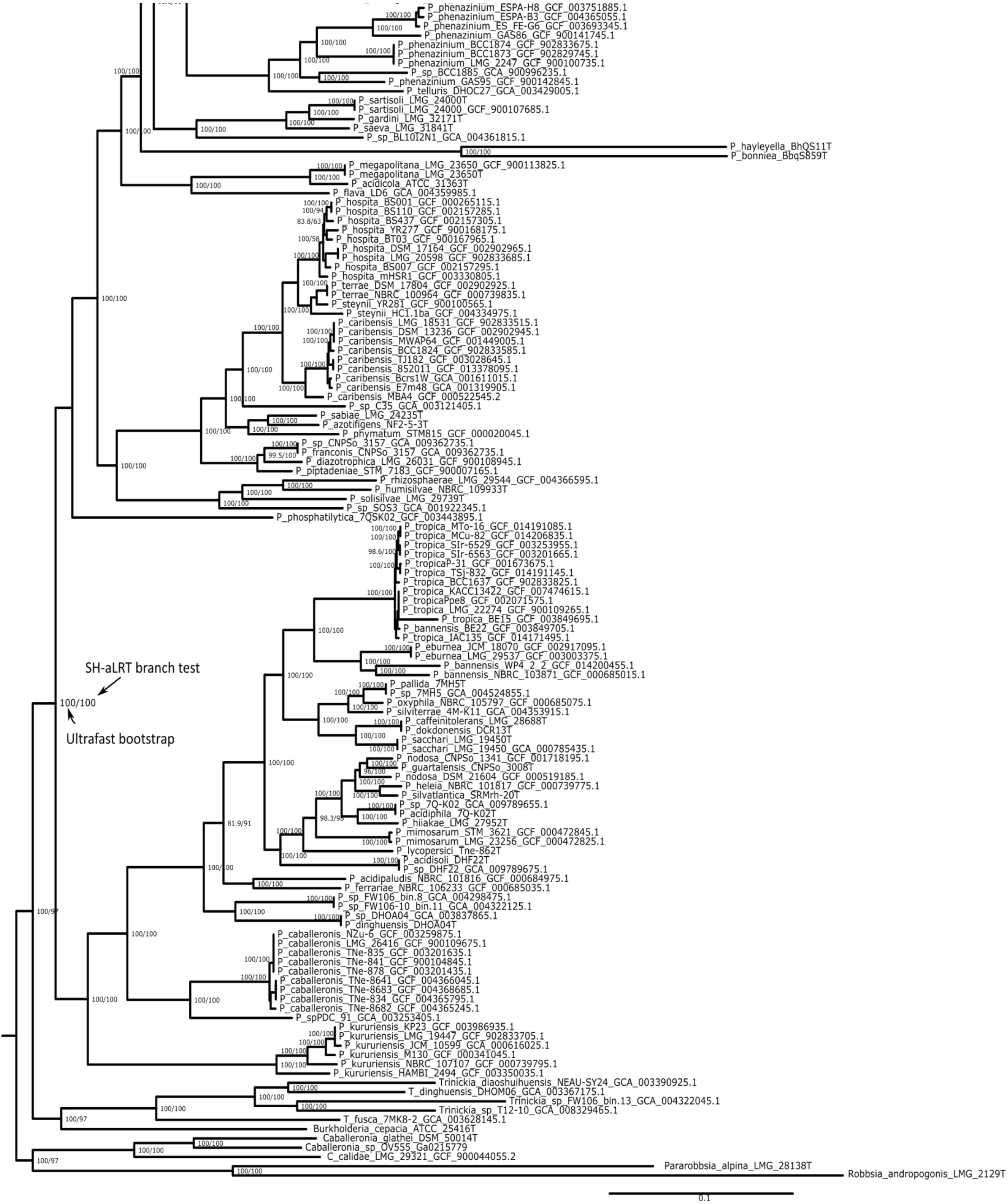


Supplementary Fig. S1. A ML phylogeny of 50 taxa used to fix the tree topology in the BEAST analysis of rhizobial *Paraburkholderia*. The phylogeny was inferred with IQ-TREE [Nguyen et al 2015] using amino acid sequences of 63 conserved core loci. The individual gene sequence alignments were generated from the whole genome sequences using the UBCG pipeline [Na et al., 2018], manually selected based on the completeness of the sequences, and concatenated using FASconCAT-G [Kück and Meusemann, 2010]. Accession numbers for genome sequences appear in parenthesis. The phylogeny included beta-rhizobia, enterobacteria, alpha-rhizobia and cyanobacteria strains for either tree-calibration or outgroup purposes. Branch support was inferred from 1,000 UFBoot [Hoang et al., 2018] and SH-aLRT tests [Guindon et al., 2010]. The UFBoot values appear on the left of each pair of the support values. The scale bar indicates the number of amino acid changes per site.



Supplementary Fig. S2. A dated ML phylogeny generated using the ‘chronopl’ function (with λ = 0) from the Ape package using R (4). The phylogeny represents the phylum *Pseudomonadota* and was inferred from the nucleotide sequences of 92 genes extracted from the genome sequences of 809 taxa largely distributed across genera in the Alpha- and Betaproteobacteria.The phylogeny was rooted with a few taxa from the phylum *Acidobacteria*. The same dataset was used by Rahimlou et al (2021). The blue and red stars indicate the nodes whose dates were fixed for calibration purposes, i.e., the divergence of the phyla *Pseudomonadota* (Hydrobacteria) and *Acidobacteria* (Terrabacteria), and the divergence of the lineage *Alphaproteobacteria* and lineage *Betaproteobacteria-Gammaproteobacteria,* dated at 3540–2830 Ma (blue star) and 2928–2154 Ma (red star), respectively [Supplementary Table S1]. The nodes marked with solid circles indicate major taxonomic groups which were supported by high ultrafast bootstrap values in IQ-TREE [Nguyen et al., 2015]. Red, purple and green solid circles indicate the nodes of the classes *Alphaproteobacteria*, *Betapreoteobacteria* and *Gammaproteobacteria*, respectively. The colour of the taxon names of these three classes corresponds to that of the mentioned circles. The blue circle indicates the divergence of *Betaproteobacteria* and *Gammaproteobacteria*. Genome accession numbers for the taxa in this phylogenetic tree are provided at https://figshare.com/s/bda511ccb0470957d3ed.

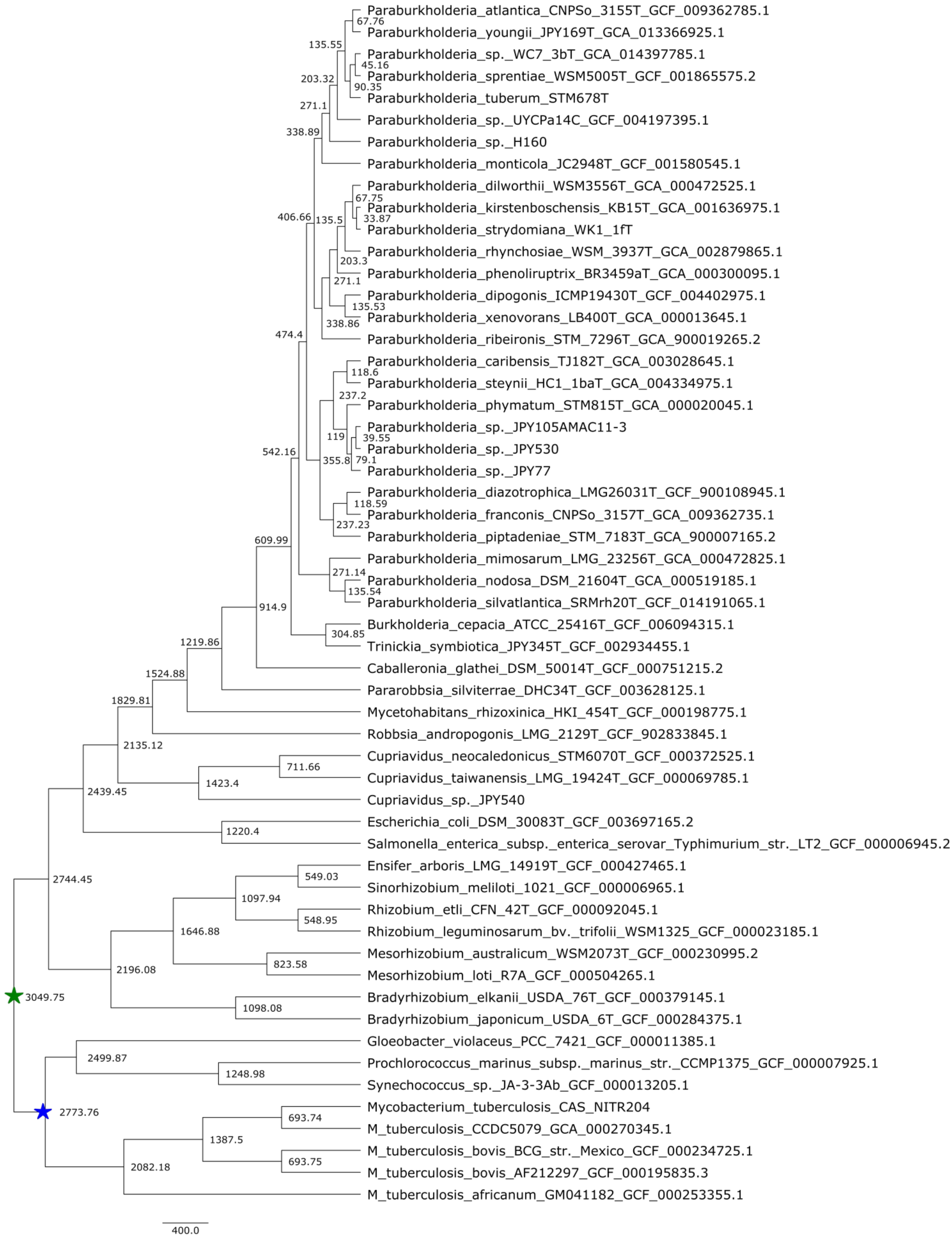

P_fungorum_NBRC_102489_GCF_000685055.1
P_fungorum_FDAARGOS_245_GCF_002891075.1
P_fungorum_LMG_16225_GCF_902833645.1
P_fungorum_SEMIA_4056_GCF_014200165.1
P_fungorum_SEMIA_4007_GCF_014198855.1
P_fungorum_SEMIA_4013_GCF_014205755.1
P_insulsa_LMG_28183_GCF_003002115.1
98.4/97
P_fungorum_BH370_GCF_900185605.1
P_fungorum_ATCC_BAA-463_GCF_000961515.1
P_fungorum_K24_GCA_000714855.1
100/100
P_fungorum_S2_003_000_R1_1_GCA_003243335.1
P_fungorum_UM_TBK_GCF_013390345.1
100/100
P_agricolaris_BaQS159T
P_fungorum_ES3-20_GCF_003610895.1
98.2/81
P_haematera_LMG_31837T
100/100
P_sp_BL23I1N1_GCA_003610295.1
99.9/80
P_phytofirmans_OLGA172_GCF_001634365.1
99.8/100
P_elongata_LMG_5NT
56.1/74
PPolaris_RP-4-7T
99.9/85
P_sediminicola_ES_FE-E3_GCF_003610135.1
P_sediminicola_ESPA-E9_GCF_004363325.1
P_sediminicola_ESPA-E8_GCF_004362135.1
100/100
P_sediminicola_ESPA-F4_GCF_003634155.1
P_solitsugae_1NT
100/100
P_domus_LMG_31832T
100/100
100/100
P_sediminicola_LMG_24238T
P_sediminicola_LMG_24238_GCF_902859805.1
100/100
P_aspalathi_LMG_27731_GCF_900116445.1
100/100
P_nemoris_LMG_31836T
100/100
100/100
P_madseniana_RP11T
9/57
100/100
P_caffeinilytica_CF1_GCF_003368325.1
P_caffeinilytica_LMG_28690_GCF_902859905.1
100/93
P_lacunae_S27T
100/100
P_sediminicola_LMG_24238_GCF_900104005.1
97.8/100
P_terricola_LMG_20594_GCF_902833815.1
P_terricola_A27_GCA_002087245.1
100/100
P_terricola_mHS1_GCF_003330825.1
90.2/100
P_sp_BL6665CI2N2_GCA_004366535.1
99.5/100
P_sp_BL6669N2_GCA_003387275.1
P_sp_BL25I1N1_GCA_003002025.1
100/100
P_sp_BL17N1_GCA_003634175.1
100/100
100/100
P_sp_BL27I4N3_GCA_003386115.1
91.4/93
100/100
PPhytofirmansPsJN_GCF_000020125.1
P_dipogonis_ICMP_19430T
100/100
P_sp_BL18I3N2_GCA_003001875.1
P_aromaticivorans_AR20-38_GCA_012689525.1
100/100
100/100
P_xenovorans_LB400_GCF_000013645.1
P_xenovorans_4B_GCF_013177705.1
100/100
P_sp_Msb3_GCA_902459535.1
100/100
100/100
P_dioscoreae_Msb3T
100/100
P_aromaticivorans_BN5_GCF_002278075.1
100/100
P_ginsengiterrae_DCY85_GCF_001645125.1
100/99
P_ginsengiterrae_DCY85-1_GCF_001645135.1
P_bryophila_AD24_GCA_002220305.1
100/100
P_bryophila_LMG_23644T
100/100
100/100
P_sp_BL21I4N1_GCA_002973525.1
P_sp_BCC1884_GCA_900996225.1
100/93
P_sp_BCC1886_GCA_900996245.1
P_ribeironis_STM_7296_GCF_900019265.1
94.2/98
PPanacisoli_DCY113_GCA_008369935.1
98.7/98
P_ultramafica_LMG_28614T
100/99
100/100
P_sp_RAU2J_GCA_003634845.1
100/99
100/100
P_fynbosensis_LMG_27177T
P_ginsengisoli_NBRC_100965_GCF_000739735.1
P_metrosideri_LMG_28140T
100/100
P_fungorum_GAS106B_GCF_900099835.1
100/99
P_sp_GV068_GCA_003058125.1
P_sp_GV073_GCA_004346905.1
P_sp_GV072_GCA_003050715.1
100/100
P_sp_GV060_GCA_004341315.1
100/100
P_sp_GV052_GCA_003208595.1
P_graminisPHS1_GCF_003330785.1
100/100
P_graminis_C4D1M_GCF_000172415.1
100/100
P_graminis_LMG_18924_GCF_902859785.1
100/100
PPhenoliruptrix_LMG_22037_GCF_902833655.1
100/100
PPhenoliruptrix_LMG_22037_GCF_902859825.1
PPhenoliruptrix_AC1100_GCF_000416445.1
100/100
PPhenoliruptrix_BR3459a_GCF_000300095.1
100/100
PPhenoliruptrix_LMG_9964_GCF_902860235.1
100/100
P_caledonica_LMG_19076_GCF_902833635.1
100/100
P_caledonica_NBRC_102488_GCF_000685095.1
100/100
P_sp_OV446_GCA_003201235.1
100/100
P_caledonicaPHRS4_GCF_003330745.1
P_caledonica_376MFSha3.1_GCF_000383275.1
P_strydomiana_WK1.1f_GCF_004334935.1
100/100
P_caledonica_CA-S3F-1_GCF_002179925.1
100/100
P_kirstenboschensis_Kb15_GCF_001636975.1
100/100
P_kirstenboschensis_Rau_2D2_GCF_001636965.1
P_dilworthii_WSM3556_GCF_000472525.1
100/100
P_rhynchosiae_LMG_27174_GCF_902859775.1
100/100
P_rhynchosiae_WSM_3937_GCF_002879865.1
99.4/100
P_tuberum__MM5496R1_Ga0537062
98.7/99
P_tuberum_MM6662R1_Ga0537061
100/100
P_tuberum_35.1_Ga0537093
P_tuberum_LMG_21444_GCF_902833905.1
P_tuberum_DUS833_GCF_900101795.1
P_tuberum_STM678T
P_tuberum_31-1_GCA_014397645.1
98.3/100
P_tuberum_MM5482R2_Ga0537063
P_tuberum_MM5482R1_Ga0537064
P_tuberum_MM5477R1_Ga0537066
88.1/95
P_tuberum_RAU6.4a_Ga0537060
100/100
P_tuberum_32_Ga0537094
100/100
P_tuberum_UCT2_GCA_014397665.1
P_tuberum_UCT31_GCA_014397655.1
100/100
P_tuberum_Clong3_Ga0537090
P_tuberum_CI3_Ga0537091
100/100
P_tuberum_40_Ga0537092
P_tuberum_UCT70_Ga0537059
100/100
P_tuberum_WSM4174_Ga0537056
P_tuberum_WSM4176_GCF_000372945.1
100/100
P_tuberum_WSM4175_Ga0537055
97.8/64
P_youngii_JPY602_Ga0537069
P_youngii_JPY284_Ga0537085
100/100
P_youngii_JPY418_GCA_013367595.1
100/100
P_youngii_JPY396_Ga0537082
98.3/100
P_youngii_JPY622_Ga0537068
P_youngii_JPY454_GCA_013372005.1
P_youngii_JPY623_Ga0537067
100/100
P_youngii_JPY169_GCA_013366925.1
P_youngii_JPY432_GCA_013371975.1
94.6/99
99.9/100
P_atlantica_JPY303_GCA_013367625.1
100/100
P_atlantica_JPY395_Ga0537083
100/100
P_atlantica_JPY156_Ga0537087
100/100
P_atlantica_JPY251_GCA_000372985.1
P_atlantica_CNPSo_3155_GCA_009362785.1
100/100
P_atlantica_JPY306_Ga0537084
100/100
P_atlantica_CCGE1002_GCA_000092885.1
P_sp_UYCP14C_GCA_004197395.1
99.7/100
100/100
P_podalyriae_WC7.3b_GCA_014397785.1
P_podalyriae_WC7.3d_Ga0537058
100/100
P_podalyriae_WC7.3g_Ga0537057
100/100
P_sprentiae_WSM5005_GCF_001865575.1
100/100
P_sp_H160_GCA_000173575.1
100/100
P_monticola_JC2948_GCF_001580545.1
P_antibiotica_G-4-1-8T
96.1/99
P_susongensis_LMG_29540_GCF_900177725.1



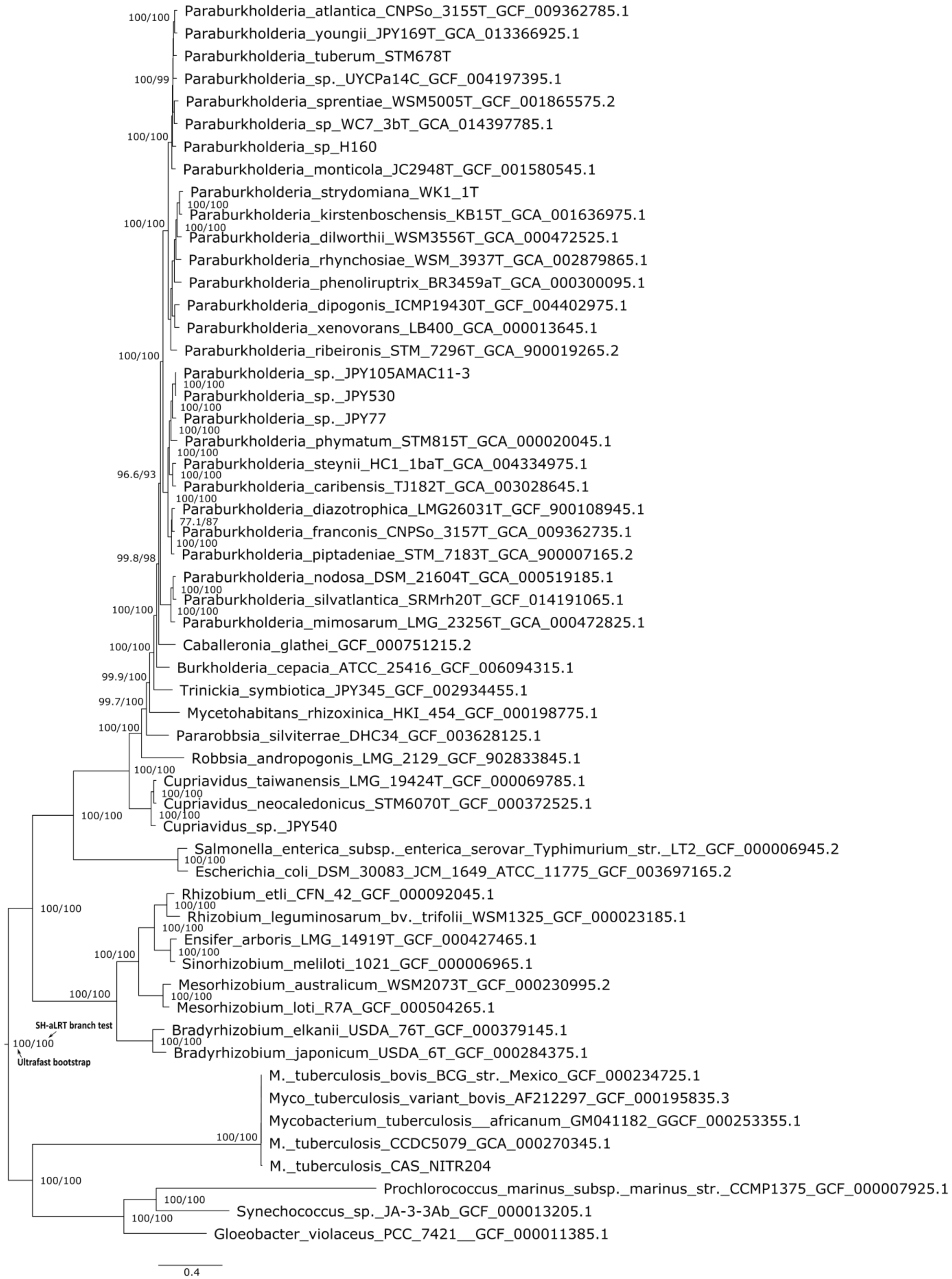
Supplementary Fig. S3. A ML phylogeny constructed using the All *Paraburkholderia* dataset with 263 taxa. The phylogeny was inferred from an amino acid sequence alignment of 92 conserved core loci extracted from the genome sequences using the UBCG pipeline [Na et al., 2018], and their sequences concatenated using FASconCAT-G [Kück and Meusemann, 2010]. Genome accession numbers are included in the titles of the isolates. This phylogeny was also used for the ASR analysis. Branch support was inferred from 1,000 UFBoot [Hoang et al., 2018] and SH-aLRT tests [Guindon et al., 2010]. The phylogeny was rooted with the lineage containing *Caballeronia*, *Pararobbsia* and *Robbsia* species. The scale bar indicates the number of amino acid changes per site.


MLSA: 92 genes (UBCG)

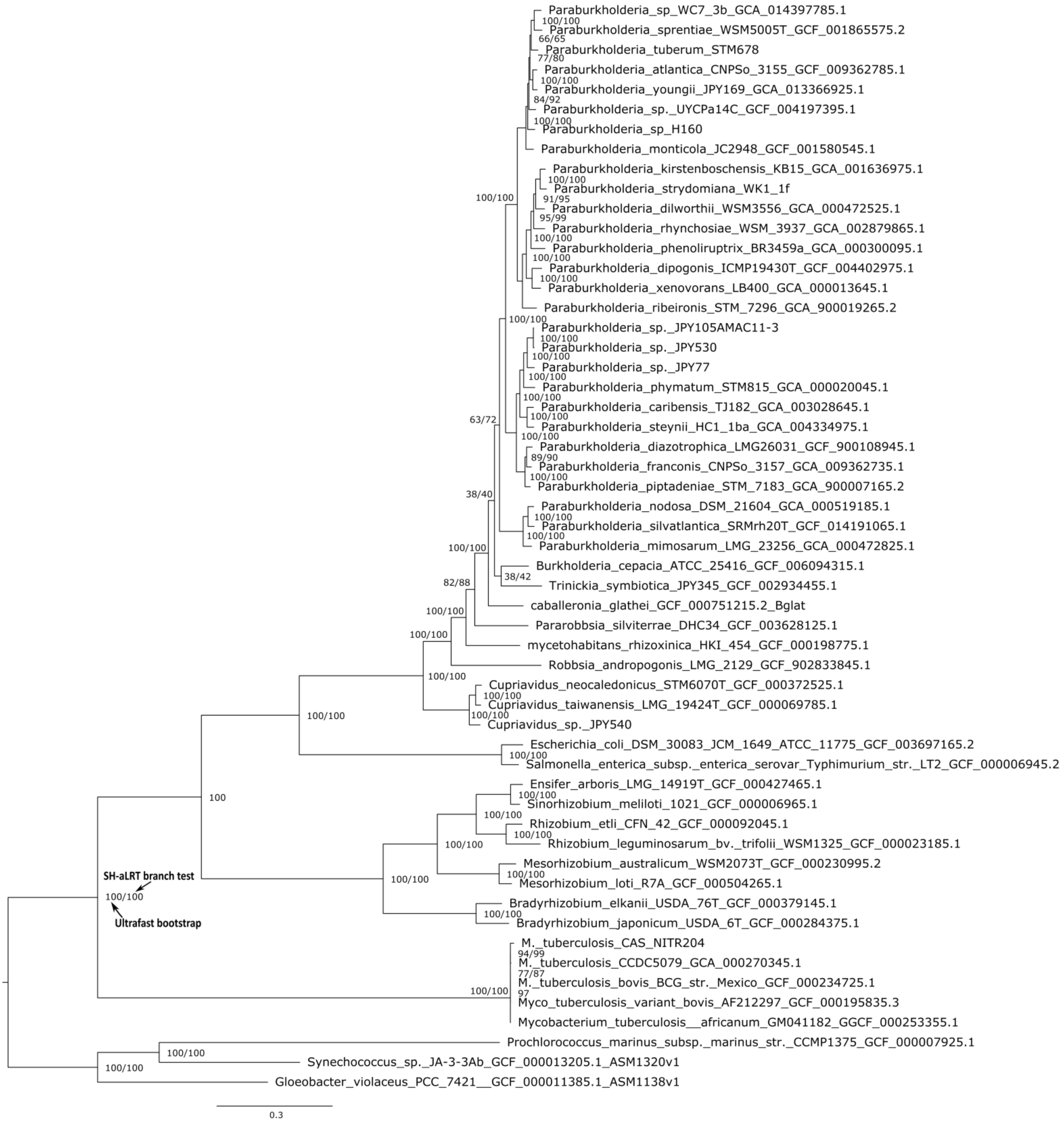
Supplementary Fig. S4. A ML phylogeny inferred from the *Paraburkholderia* species dataset with 96 taxa. The dataset consisted of nucleotide sequences of the 92UBCG-loci. The (type) strains of type species of other genera of *Burkholderia sensu lato* (i.e., *Caballeronia glathei* DSM 50014^{T} [Dobritsa et al., 2016], *Burkholderia cepacia* ATCC 25416^{T} [Yabuuchi et al., 1992], *Trinickia symbiotica* JPY345^{T} [Estrada-de los Santos et al., 2018], *Mycetohabitans rhizoxinica* HKI 454^{T} [Estrada-de los Santos et al., 2018], *Pararobbsia alpina* LMG 28138^{T} [Lin et al., 2020; Weber and King, 2017] and *Robbsia andropogonis* LMG 2129^{T}) [Lopes-Santos et al., 2017] were used for outgroup purposes, with the phylogeny rooted with *R. andropogonis* LMG 2129^{T}. Branch support was inferred from 1,000 UFBoot [Hoang et al., 2018] and SH-aLRT tests [Guindon et al., 2010]. UFBoot values appear on the left of each pair of support values. Only support higher than 70% is indicated. The scale bar indicates the number of amino acid changes per site. Genome accession numbers for the taxa in this phylogeny are mentioned in other phylogenies (Supplementary Figs S3 and S5-S7).



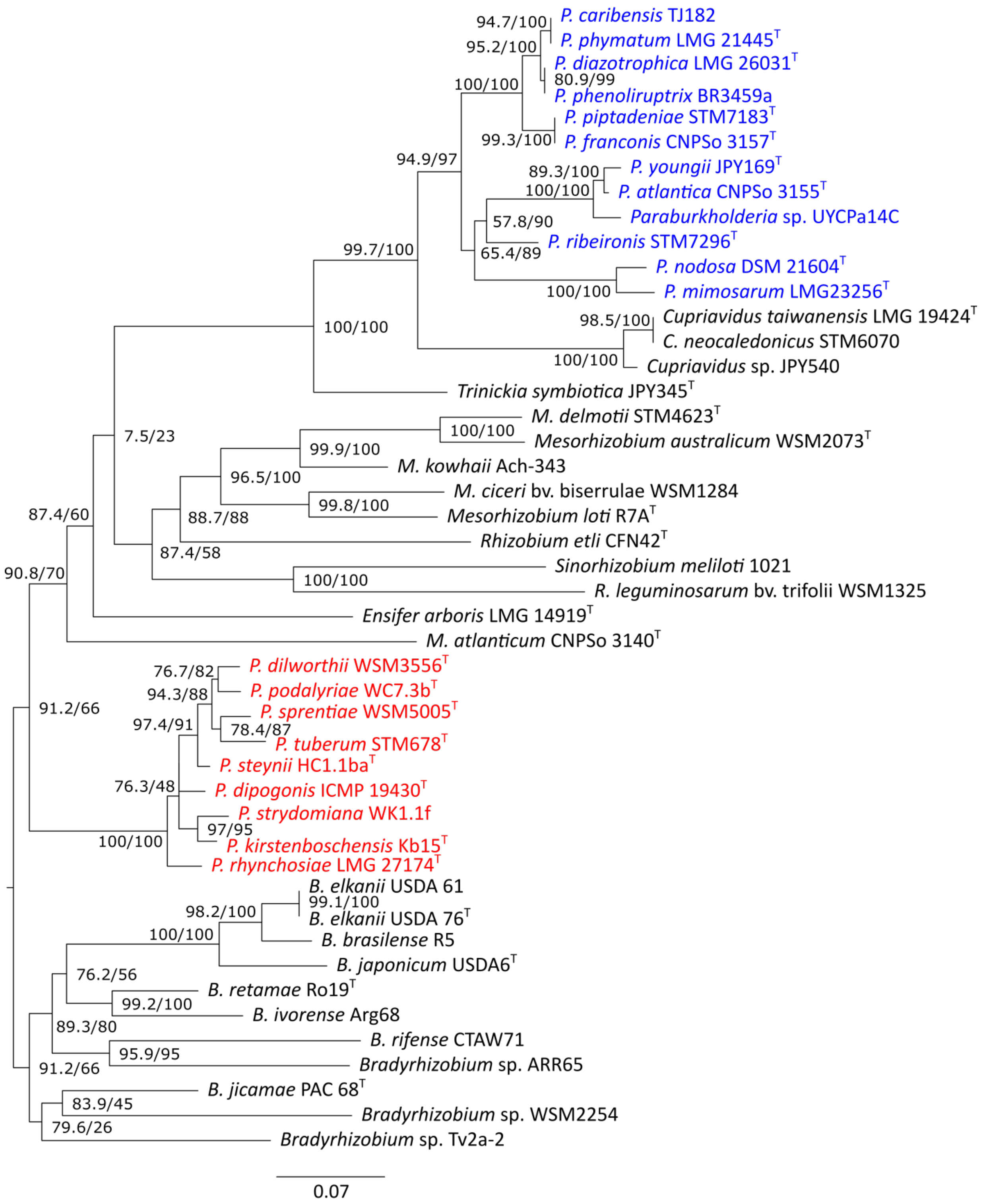
Supplementary Fig. S5. A timeline of rhizobial *Paraburkholderia* species and strains generated from the divergence times estimated using the chronopl function in APE. The dated phylogeny was inferred from the modified or second version of the rhizobial *Paraburkholderia* dataset with 55 taxa, using the amino acid sequence alignments of 63 core loci extracted from the whole genome sequences using the UBCG pipeline [Na et al., 2018]. The dated phylogeny also included enterobacteria, alpha-rhizobia, mycobacteria and cyanobacteria for outgroup and calibration purposes. The stars on the nodes indicate the calibration points. The green star indicates the divergence of Hydrobacteria and Terrabacteria [Supplementary Table S1]. The blue star indicates the divergence of *Actinobacteria* and *Cyanobacteria* [Supplementary Table S1].



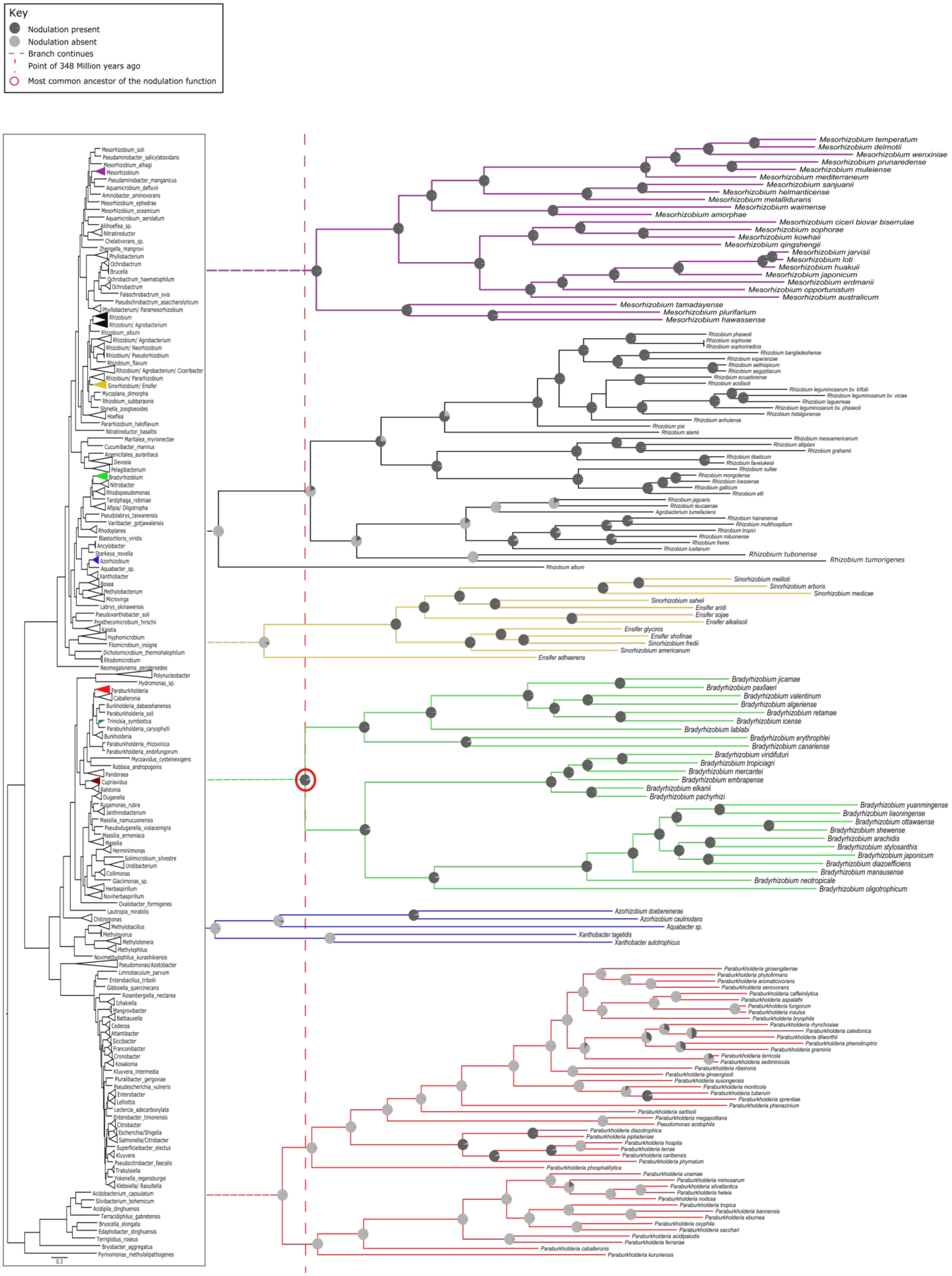
Supplementary Fig. S6. A ML phylogeny of 55 taxa inferred with IQ-TREE [Nguyen et al 2015] using the second version of the Rhizobial *Paraburkholderia* dataset. The dataset contained nucleotide sequence alignments of 63 core loci extracted from the whole genome sequences using the UBCG pipeline [Na et al., 2018]. The 63 locus sequence alignments were manually selected based on the completeness of the sequences, and was concatenated using FASconCAT-G [Kück and Meusemann, 2010]. Branch support was inferred from 1,000 UFBoot [Hoang et al., 2018] and SH-aLRT tests [Guindon et al., 2010]. The phylogeny was rooted with Terrabacteria. The scale bar indicates the number of nucleotide changes per site. This phylogeny was also used for the molecular dating analysis shown in Supplementary Fig. S5.



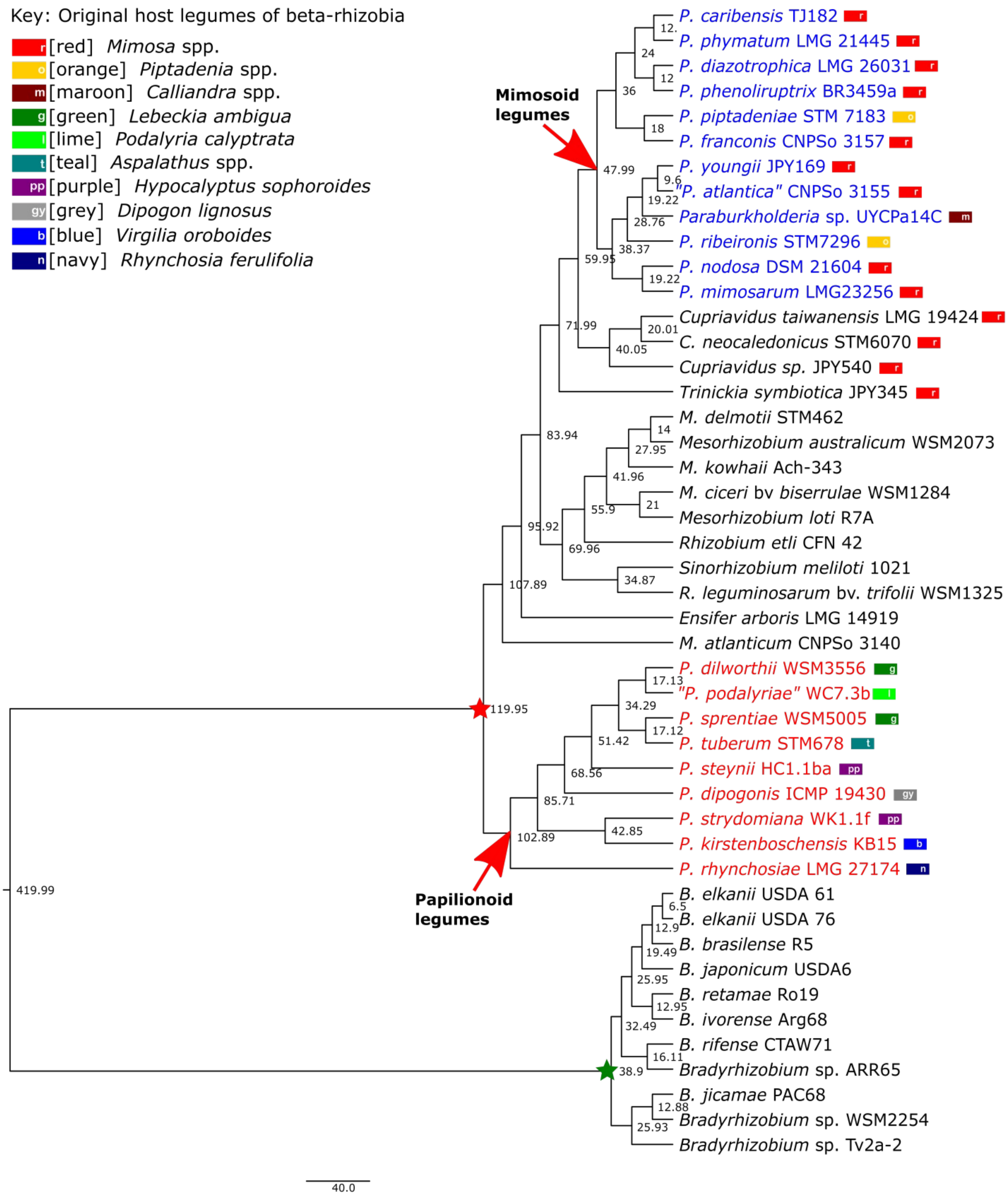
Supplementary Fig. S7. A ML phylogeny of 55 taxa inferred with IQ-TREE [Nguyen et al 2015] using the second version of the Rhizobial *Paraburkholderia* dataset. The dataset contained amino acid sequence alignments of 63 core loci extracted from the whole genome sequences using the UBCG pipeline [Na et al., 2018]. The 63 gene sequence alignments were manually selected based on the completeness of the sequences, and was concatenated using FASconCAT-G [Kück and Meusemann, 2010]. Branch support was inferred from 1,000 UFBoot [Hoang et al., 2018] and SH-aLRT tests [Guindon et al., 2010]. The phylogeny was rooted with Terrabacteria. The scale bar indicates the number of amino acid changes per site. This phylogeny was also used for the molecular dating analysis shown in Supplementary Fig. S5.



Supplementary Fig. S8. A *nodACD* ML phylogeny constructed with IQ-TREE based upon amino acid sequences. The phylogeny largely included beta-rhizobia, as well as representative strains or species for the putative nodulation loci clusters observed in alpha-rhizobia. The species in blue and red font are rhizobial *Paraburkholderia* species native to South/Central America and South Africa, respectively. The phylogeny was rooted at mid-point. Branch support was inferred from 1,000 UFBoot [Hoang et al., 2018] and SH-aLRT tests [Guindon et al., 2010]. UFBoot values are on the left of each pair of support values. The scale bar indicates the number of amino acid changes per site. The type strains are indicated with a superscript T. Genome accession numbers for the taxa in this phylogenetic tree are provided at https://figshare.com/s/bda511ccb0470957d3ed.



Supplementary Fig. S9. A ML phylogeny of the phylum *Pseudomonadota* (Supplementary Fig. S2) showing the lineages which include rhizobia and the ancestral state reconstruction (performed using MrBayes Ancestral States in R) of the genetic predisposition to harbor nodulation loci.. The diagram illustrates that the MRCA of the rhizobial nodulation function emerged within the *Bradyrhizobium* lineage before 348 Ma (date estimated in this study, Table 1), as indicated by the pie in a red circle and the dashed vertical line. The pies at the nodes indicate the estimated probabilities of the binary nodulation states. The coloured solid triangles on the phylogenetic tree represent lineages with the ability to nodulate legumes. Zoomed-in topologies of these lineages are shown in corresponding colours, with the exception of small lineages of *Cupriavidus* and *Trinickia*.



Supplementary Fig. S10. A dated phylogeny of *nodACD* genes showing the original host legumes of mostly rhizobial *Paraburkholderia*. The phylogeny is the same as the one in Fig. 3, with the exception that it shows the original host legumes of all beta-rhizobia included. The species in blue and red font are rhizobial *Paraburkholderia* species native to South/Central America and South Africa, respectively. Whereas species in black font are non-*Paraburkholderia* rhizobial species. Divergence dates are shown at the nodes.