Figure S1: Taxonomic distribution of bacterial reads from leaves of *Pavetta indica*. Estimation based on blastn searches against the NCBI nucleotide database on a subset of 1M reads. Only bacterial reads (representing an estimated 7% of total reads) are shown here.

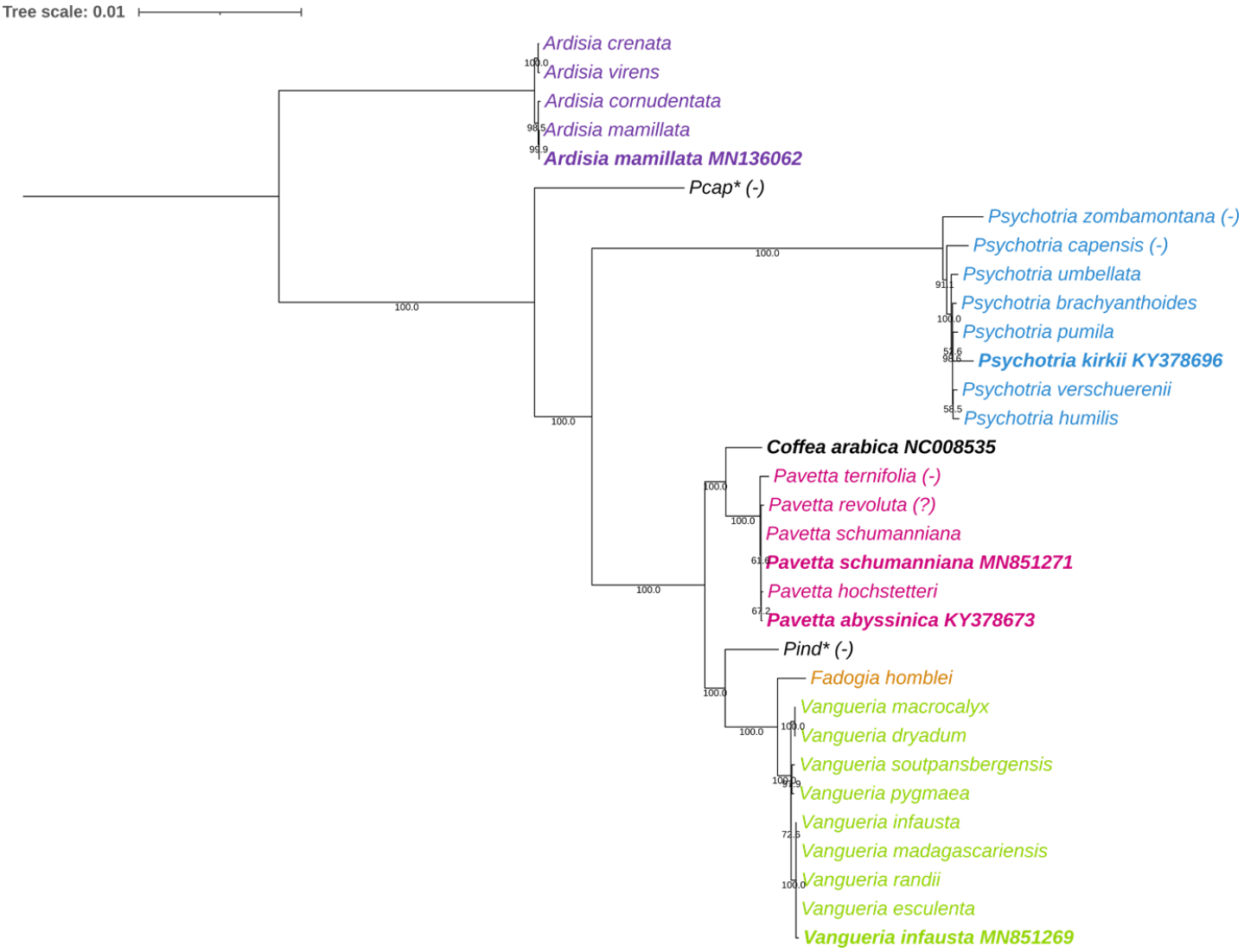

Figure S2: SNP-based chloroplast phylogeny of investigated Primulaceae and Rubiaceae plant species. The *Ardisia* clade was set as outgroup. Numbers on the branches represent bootstrap support values based on 1000 replications. Branches with <50% bootstrap support are collapsed. Names in boldface represent included reference chloroplast sequences. Colours represent the different plant genera. Species marked with * are *Pavetta* species likely misidentified (see text). (-) indicates no *Burkholderia s.l.* symbiont was detected. (?) indicates presence of a *Burkholderia s.l.* symbiont is uncertain (see text). Only one sample per investigated plant species is included in the phylogenetic tree. *Pcap: Pavetta capensis*; *Pind*: *Pavetta indica*. Samples are colour-coded based on the host genus: Purple – *Ardisia*; Blue – *Psychotria*; Pink – *Pavetta*; Green – *Vangueria*; Orange – *Fadogia.* The whole genome alignment used to construct this tree included 73857 positions.

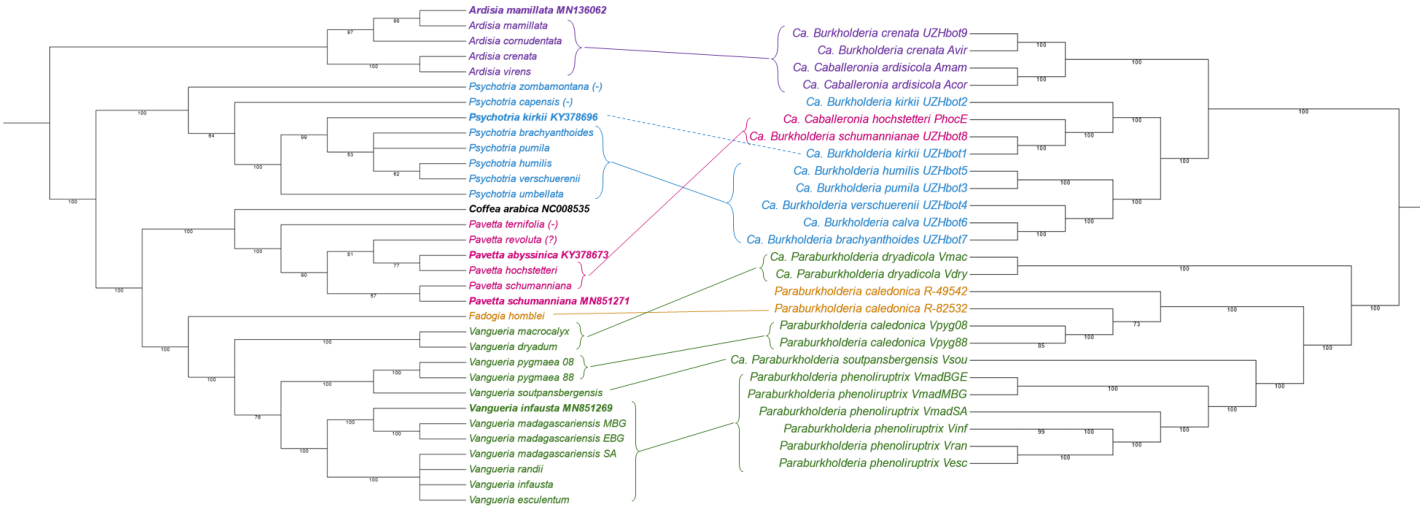

Figure S3: Cophylogenetic patterns between leaf endophytes and their hosts. Left: Chloroplast phylogeny of host plants as in Figure S2. Right: Core-genome phylogeny of leaf endophytes (as in Figure 1). Branch lengths are not representative. Numbers on branches represent bootstrap support values based on 100 bootstrap replications. Connections were drawn between representative groups of endophytes and their host plants. The dotted line for *Psychotria kirkii* indicates that the endophyte and host plant are not derived from the same voucher. Samples are colour-coded based on the host genus: Purple – *Ardisia*; Blue – *Psychotria*; Pink – *Pavetta*; Green – *Vangueria*; Orange – *Fadogia.* Branches with bootstrap support values <50% were collapsed.

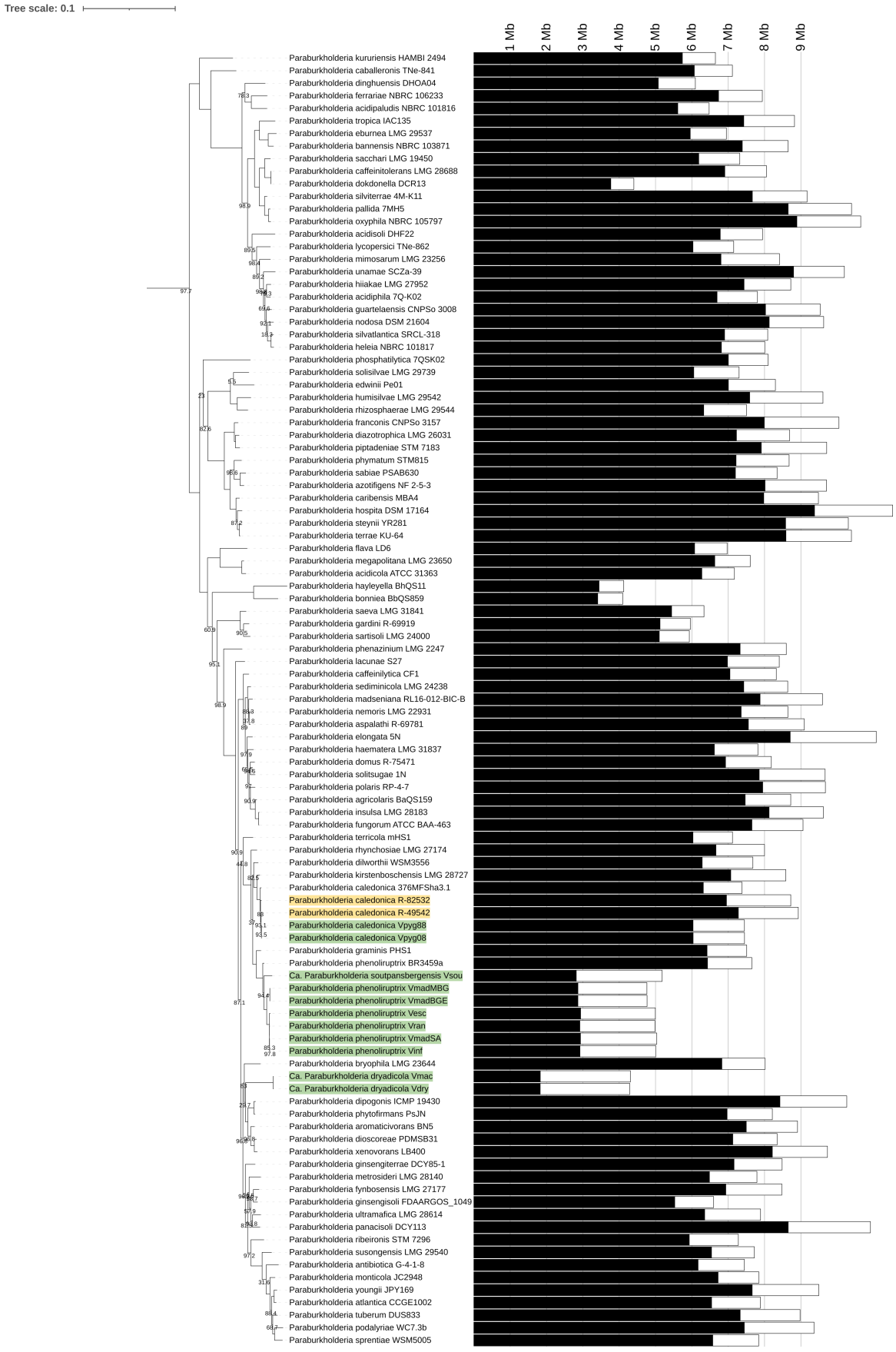

Figure S4. Genome size of *Paraburkholderia sp.*, including leaf endophytes. Representative genomes of *Paraburkholderia sp.* Were downloaded from the NCBI Refseq database. Protein sequences of 40 core marker genes were extracted with FetchMG[4] and aligned using MAFFT v7.4751. A maximum-likelihood phylogeny was constructred with IQ-TREE v.2.0.3[2] with model "LG+F+R9". Branch support values < 100% (SH-aLRT) are displayed on the branches, support values = 100% are omitted for clarity. Samples are colour-coded based on the host genus: Green – *Vangueria*; Orange – *Fadogia*; Black bars represent the coding capacity of the genome (the proportion of the genome coding for functional proteins).

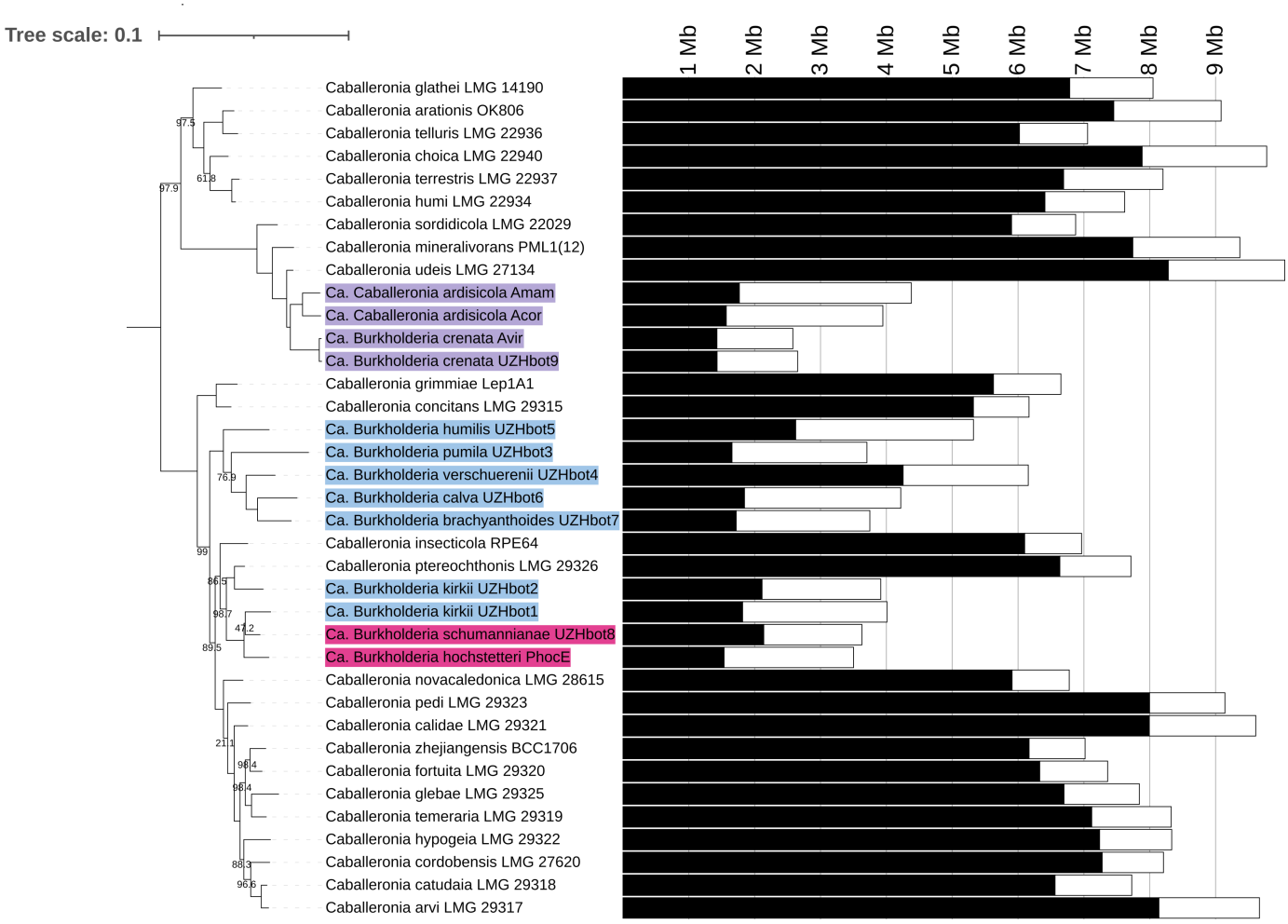

Figure S5. Genome size of *Caballeronia sp.*, including leaf endophytes. Representative genomes of *Caballeronia sp.* Were downloaded from the NCBI Refseq database. Protein sequences of 40 core marker genes were extracted with Fetch-MG[4] and aligned using MAFFT v7.4751. A maximum-likelihood phylogeny was constructred with IQ-TREE v.2.0.3[2] with model "LG+F+R9". Branch support values < 100% (SH-aLRT) are displayed on the branches, support values = 100% are omitted for clarity. Samples are colour-coded based on the host genus: Purple – *Ardisia*; Blue – *Psychotria*; Pink – *Pavetta*; Black bars represent the coding capacity of the genome (the proportion of the genome coding for functional proteins).

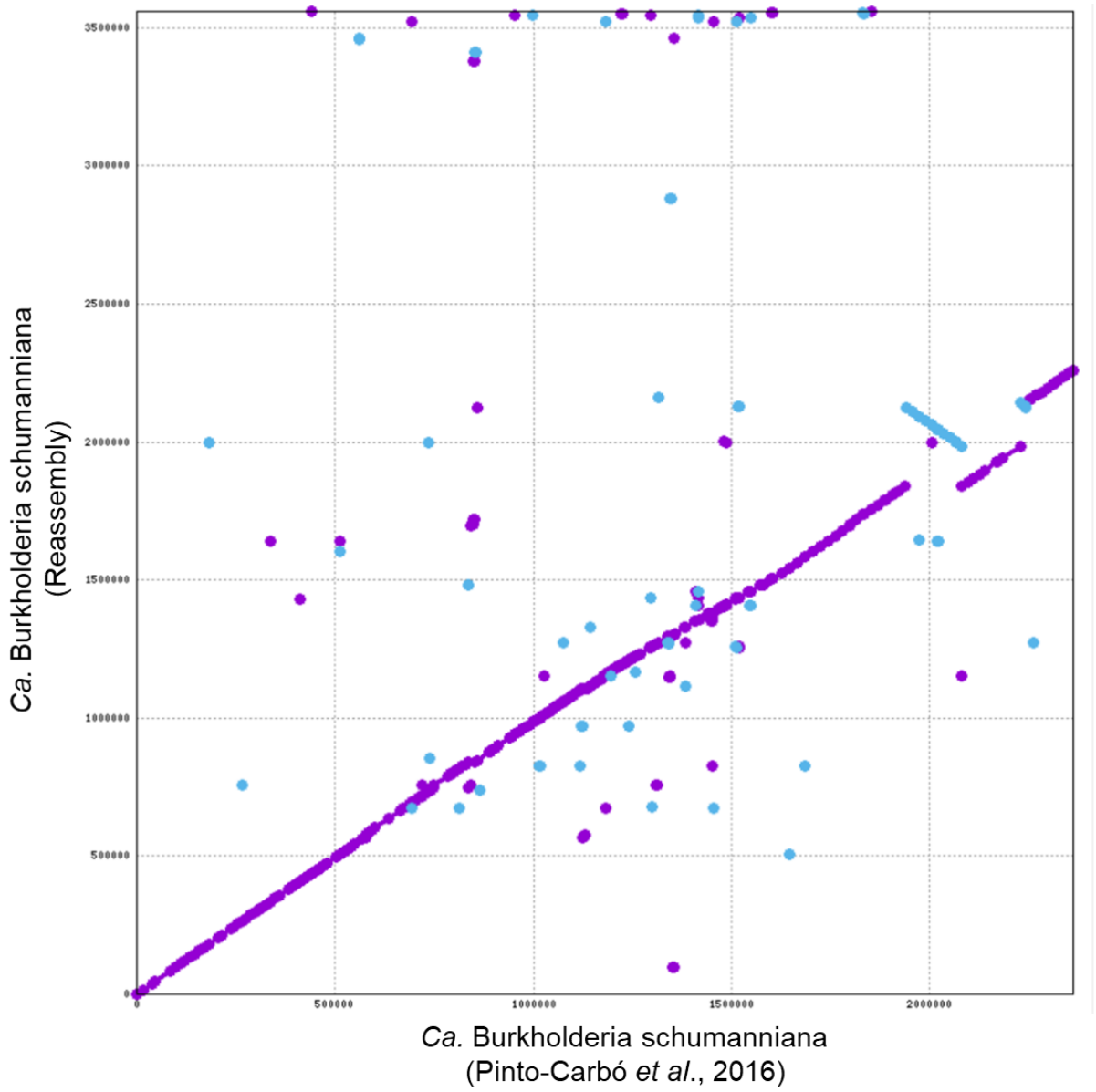

Figure S6: Dotplot of the reassembly of *Ca.* Burkholderia schumanniana UZHbot8 versus the published genome sequence. Numbers on X- and Y-axis represent the position in the assembly, in bases. Dots represent aligned subsequences between both assemblies. Purple dots represent sequences aligned in the same sense, while blue dots represent sequences aligned in opposite sense.

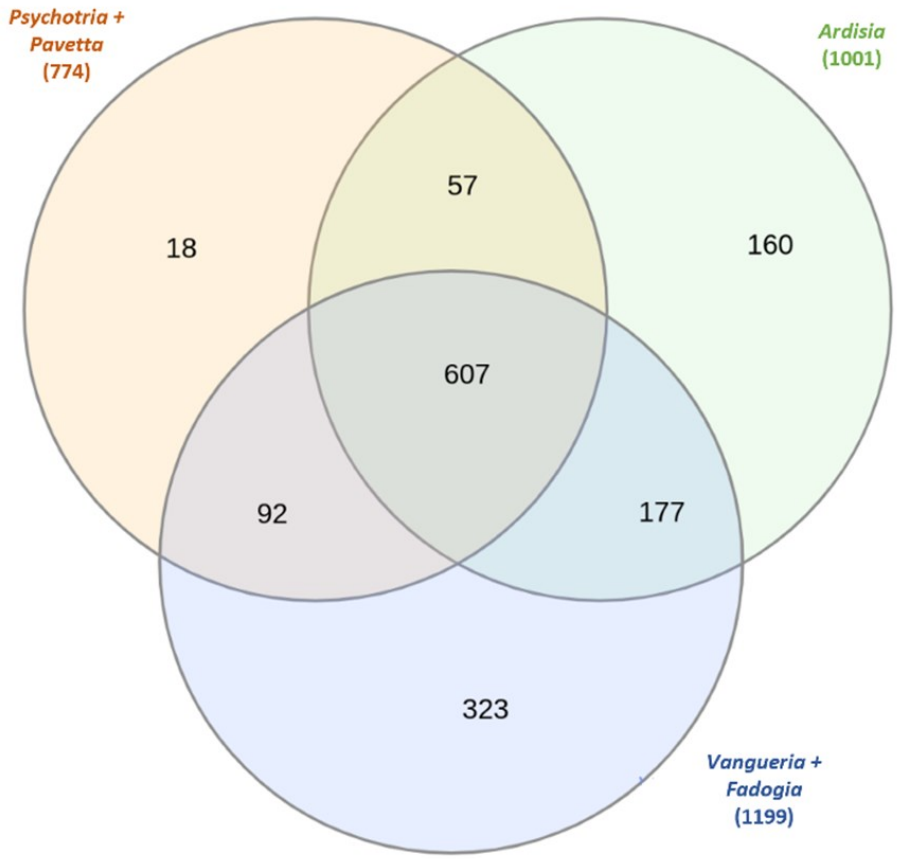

Figure S7: Core-genome overlap between leaf endophytes of different plant genera. Venn diagram showing the overlap of the core genome between endophytes of different plant hosts. Numbers between brackets represent the total core genome size of a certain clade.

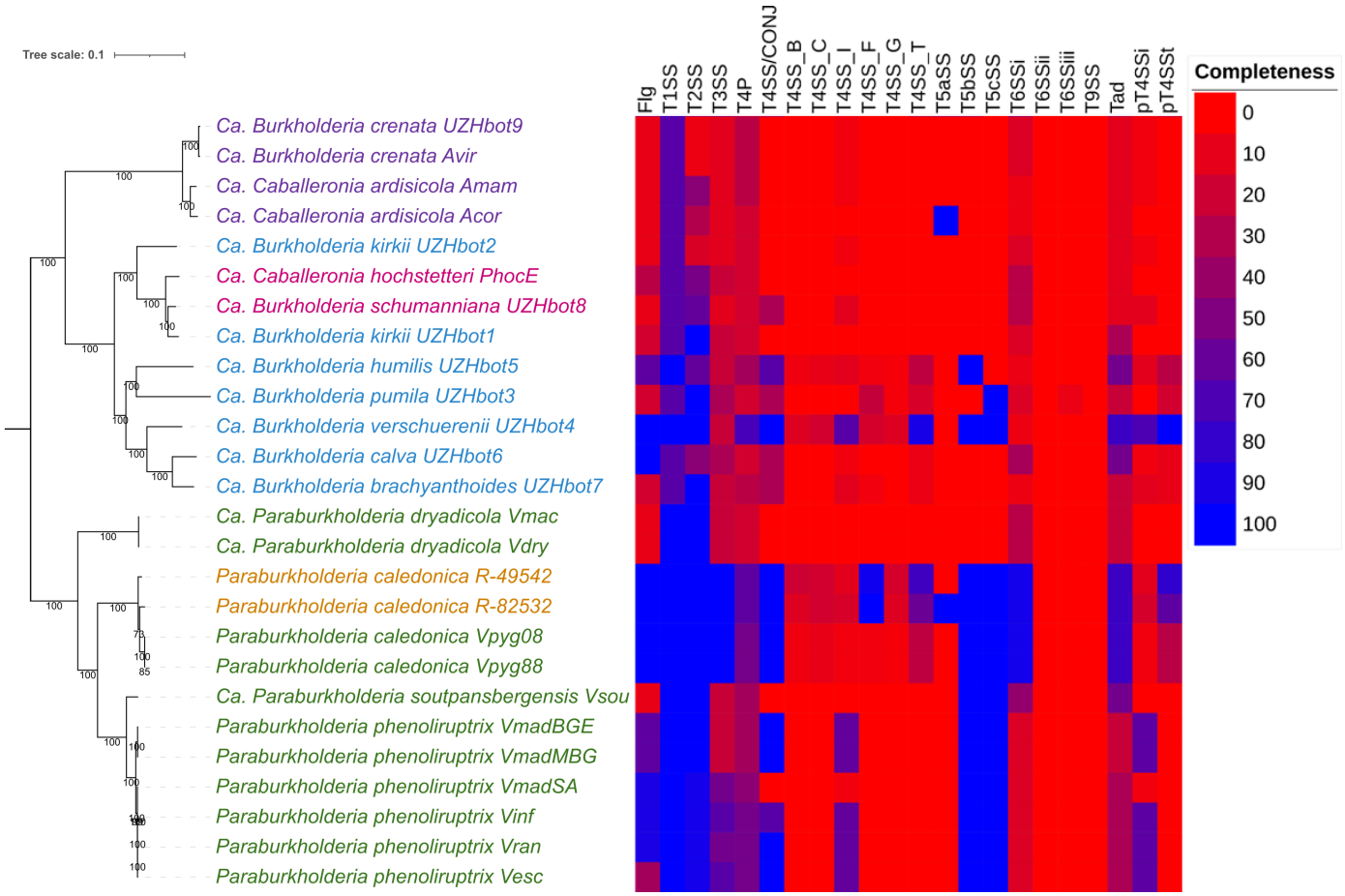

Figure S8: Presence and completeness of flagellar apparatus and secretion systems in leaf endophytes. Completeness refers to the proportion of genes in a certain cluster that are found present in a certain genome. Coloured names represent the host species of the endophytes: Purple – *Ardisia* spp.; Blue – *Psychotria* spp.; Pink – *Pavetta* spp.; Green – *Vangueria* spp.; Orange – *Fadogia* spp. Abbreviations: SA – South Africa; 08/88 last two digits of *V. pygmaea* voucher number; BGE – Botanic Garden Edinburgh; MBG – Meise Botanic Garden; Flg – Flagellar apparatus; TXSS – Type X Secretion System; T4P – Type IV Pilus; Tad – Tight Adherence pilus.

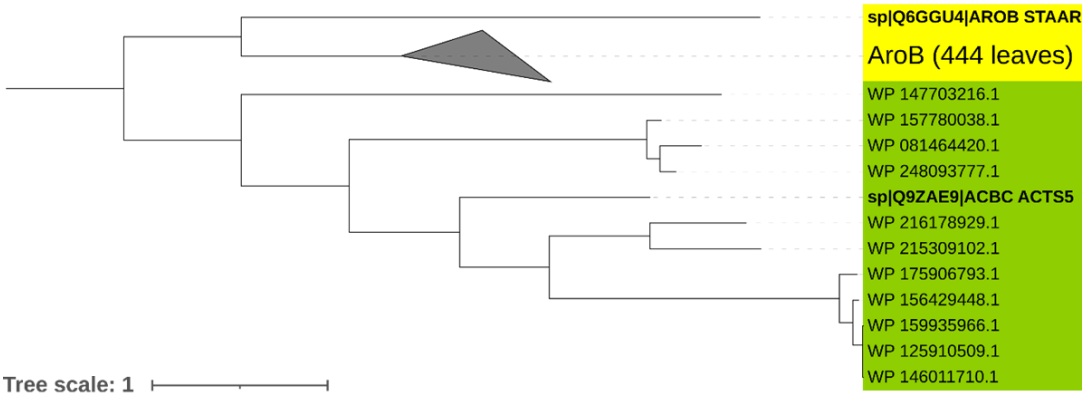

Figure S9: Distribution of putative EEVS in the *Burkholderiaceae* other than leaf symbiotic bacteria. The amino acid sequence of the 2-epi-5-epi-valiolone synthase from *Actinoplanes* sp. ATCC 31044 (accession Q9ZAE9) was retrieved from the UniProt database and used as query in a BlastP search against the NCBI RefSeq database using the NCBI Blast online service with default settings except: the "max target sequences" parameter was increased to 5000; Taxonomic filters were applied to limit hits to the family *Burkholderiaceae* (taxid: 119060); and only hits with e-value < 10^{-3} were considered. The search retrieved 456 matches, which were aligned using MAFFT v7.475[1] in "auto" mode together with the EEVS sequence of *Actinoplanes* sp. ATCC 31044 (UniProt accession Q9ZAE9) and the DHQS AroB sequence of Staphylococcus aureus (UniProt accession Q6GGU4). Sequences corresponding to leaf nodule *Burkholderia* were manually removed from the alignment. Maximum-likelihood phylogenetic analysis was performed using IQ-TREE v.2.0.3[2] with the "JTT+R6" model and visualized in iTOL[3]. Tree labels correspond to NCBI RefSeq or UniProt accession numbers, with labels in bold indicating reference EEVS and AroB sequences. The green-colored range indicates putative EEVS samples, and the yellow-colored range indicated putative DHQS (AroB) sequences.

K-cluster

S-cluster

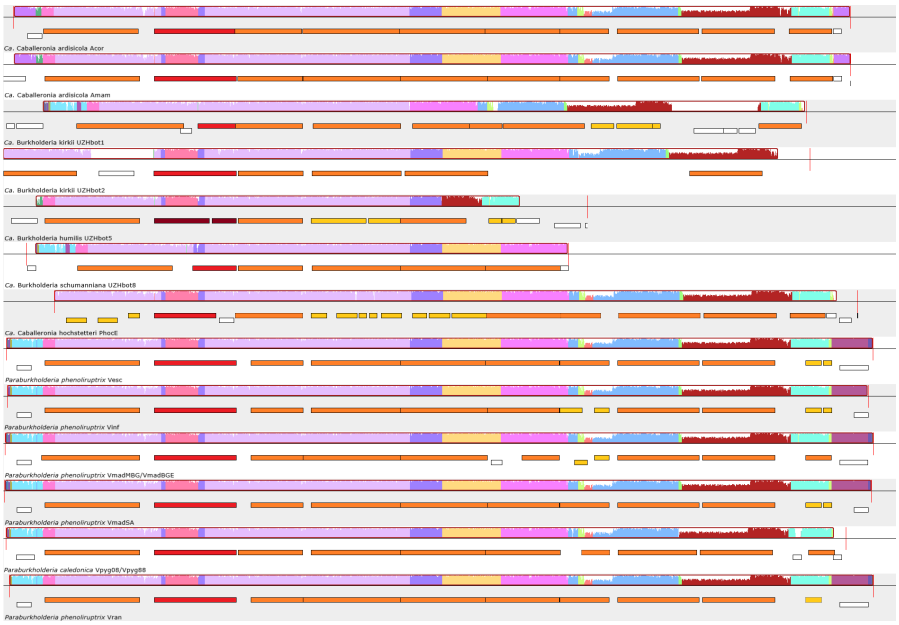

Figure S10. Genetic organization of K-clusters (top) and S-clusters (bottom) in genomes and MAGs of leaf endophytes. Alignments of cluster regions extracted from individual MAGs were created using ProgressiveMauve v2.3.0[5] with default settings. For each genome aligned, the colored boxes on top indicate locally colinear blocks (LCBs). Boxes on the bottom of each panel represent predicted CDSs. The colour codes of the CDS boxes are: Red: EEVS - Dark Red: EEVS pseudogene; Orange: Other cluster gene - Light Orange: Cluster pseudogene.

Supplementary Figure references

1. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–80.

2. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245.

3. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R, Teeling E. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol 37:1530–1534.

4. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, Rasmussen S, Brunak S, Pedersen O, Guarner F, de Vos WM, Wang J, Li J, Doré J, Ehrlich SD, Stamatakis A, Bork P. 2013. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods 10:1196–1199.

5. Darling, A. E., Mau, B., & Perna, N. T. C.-P. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One, 5, e11147 ST-progressiveMauve: multiple genome ali. https://doi.org/10.1371/journal.pone.0011147