Metagenomics of African *Empogona* and *Tricalysia* (Rubiaceae) reveals the presence of leaf endophytes

Brecht Verstraete [Corresp., Equal first author, 1], **Steven Janssens** [Equal first author, 1, 2], **Petra De Block** [1], **Pieter Asselman** [3], **Gabriela Méndez** [4, 5], **Serigne Ly** [6], **Perla Hamon** [6], **Romain Guyot** [6, 7]

[1] Meise Botanic Garden, Meise, Belgium

[2] Department of Biology, KU Leuven, Leuven, Belgium

[3] Department of Biology, Ghent University, Ghent, Belgium

[4] Grupo de Investigación (BIOARN), Universidad Politécnica Salesiana, Quito, Ecuador

[5] Facultad de ingenieria, Pontificia Universidad Católica del Ecuador, Quito, Ecuador

[6] DIADE, Université de Montpellier, Montpellier, France

[7] Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales, Colombia

Corresponding Author: Brecht Verstraete
Email address: brecht.verstraete@plantentuinmeise.be

Background. Leaf symbiosis is a phenomenon in which host plants of Rubiaceae interact with bacterial endophytes within their leaves. To date, it has been found in around 650 species belonging to 8 genera in 4 tribes; however, the true extent in Rubiaceae remains unknown. Our aim is to investigate the possible occurrence of leaf endophytes in the African plant genera *Empogona* and *Tricalysia* and, if present, to establish their identity.

Methods. Total DNA was extracted from the leaves of four species of the Coffeeae tribe (*Empogona congesta, Tricalysia hensii*, *T. lasiodelphys*, and *T. semidecidua*) and sequenced. Bacterial reads were filtered out and assembled. Phylogenetic analysis of the endophytes was used to reveal their identity and their relationship with known symbionts.

Results. All four species have non-nodulated leaf endophytes, which are identified as *Caballeronia*. The endophytes are distinct from each other but related to other nodulated and non-nodulated endophytes. An apparent phylogenetic or geographic pattern appears to be absent in endophytes or host plants. *Caballeronia* endophytes are present in the leaves of *Empogona* and *Tricalysia*, two genera not previously implicated in leaf symbiosis. This interaction is likely to be more widespread, and future discoveries are inevitable.

Metagenomics of African *Empogona* and *Tricalysia* (Rubiaceae) reveals the presence of leaf endophytes

Brecht Verstraete[1,*], Steven Janssens[1,2,*], Petra De Block[1], Pieter Asselman[3], Gabriela Mendez Silva[4,5], Serigne Ndiawar Ly[6], Perla Hamon[6], Romain Guyot[6,7]

[1] Meise Botanic Garden, Meise, Belgium
[2] Department of Biology, KU Leuven, Leuven, Belgium
[3] Department of Biology, Ghent University, Ghent, Belgium
[4] Grupo de Investigación (BIOARN), Universidad Politécnica Salesiana, Quito, Ecuador
[5] Facultad de ingenieria, Pontificia Universidad Católica del Ecuador, Quito, Ecuador
[6] DIADE, Université de Montpellier IRD, CIRAD, Montpellier, France
[7] Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales, Colombia
[*] These authors contributed equally to this work.

Corresponding Author:
Brecht Verstraete[1]
Nieuwelaan 38, 1860 Meise, Belgium
Email address: brecht.verstraete@meisebotanicgarden.be

Abstract

Background. Leaf symbiosis is a phenomenon in which host plants of Rubiaceae interact with bacterial endophytes within their leaves. To date, it has been found in around 650 species belonging to 8 genera in 4 tribes; however, the true extent in Rubiaceae remains unknown. Our aim is to investigate the possible occurrence of leaf endophytes in the African plant genera *Empogona* and *Tricalysia* and, if present, to establish their identity.
Methods. Total DNA was extracted from the leaves of four species of the Coffeeae tribe (*Empogona congesta*, *Tricalysia hensii*, *T. lasiodelphys*, and *T. semidecidua*) and sequenced. Bacterial reads were filtered out and assembled. Phylogenetic analysis of the endophytes was used to reveal their identity and their relationship with known symbionts.
Results. All four species have non-nodulated leaf endophytes, which are identified as *Caballeronia*. The endophytes are distinct from each other but related to other nodulated and non-nodulated endophytes. An apparent phylogenetic or geographic pattern appears to be absent in endophytes or host plants. *Caballeronia* endophytes are present in the leaves of *Empogona* and *Tricalysia*, two genera not previously implicated in leaf symbiosis. This interaction is likely to be more widespread, and future discoveries are inevitable.

Introduction

Plant-bacteria interactions are considered ubiquitous and a common phenomenon in angiosperms (*Orozco-Mosqueda & Santoyo, 2021*). Many studies have shown the beneficial impact of such interactions; the most widely known example is the nitrogen-fixing *Rhizobia* bacteria that occur in the rhizosphere of several members of the Fabaceae family (*Poole, Ramachandran & Terpolilli, 2018*).

An example of plant endophytes in the phyllosphere is bacterial leaf nodule symbiosis, which is found in a number of taxa in eudicots (Primulaceae and Rubiaceae) and monocots (Dioscoreaceae) (*Miller, 1990*). When leaf nodules are present, it is usually easy to recognize the presence of this particular plant-bacteria interaction. Moreover, these distinctive structures can sometimes be used for the taxonomic characterisation of certain plant lineages in Rubiaceae (e.g. *Van Oevelen et al., 2001*; *Razafimandimbison et al., 2017*). However, leaf nodules are not necessarily always present, and therefore putative leaf endophytes often remain undetected (*Verstraete et al., 2011*; *Lemaire et al., 2012b*). With the development of modern molecular methods and especially with the rapid increase of different high-throughput sequencing techniques, an array of tools became available to detect and study bacterial leaf endophytes (e.g. *Carlier et al., 2016*; *Carlier et al., 2017*; *Danneels et al., 2021*; *Schindler et al., 2021*; *Danneels et al., 2023*).

The Rubiaceae family currently has the highest recorded number of species that are characterized by bacterial leaf nodulation. The presence of “thickened, hard warts” in *Pavetta indica* L. was already noted almost 130 years ago (*Trimen, 1894*) and it was later discovered that these leaf nodules contain endophytic bacteria (*Zimmermann, 1902*). This symbiosis between Rubiaceae and leaf bacteria was then more elaborately described by *von Faber (1912)*. Currently, leaf nodules in Rubiaceae have been observed in the genera *Pavetta* L. (ca 350 spp in the Pavetteae tribe), *Psychotria* L. (ca 80 spp in the Psychotrieae tribe), and *Sericanthe* Robbr. (ca 12 spp in the Coffeeae tribe) (*Lersten & Horner, 1976*; *Miller, 1990*; *Lemaire et al., 2011b*; *Lemaire et al., 2012a*). Because the leaf bacteria cannot survive outside the nodules, culture-independent methods were necessary to establish the identity of these nodulated endophytes: they belong to the genus *Burkholderia* s.l. (e.g. *Van Oevelen et al., 2002*; *Lemaire et al., 2011a*; *Lemaire et al., 2012a*; *Pinto-Carbó et al. 2018*). Furthermore, most plant species seem to harbour unique bacterial lineages. Since the discovery of leaf endophytes, the *Burkholderia* s.l. genus has undergone several taxonomic changes and therefore names such as *Paraburkholderia* and *Caballeronia* can also be encountered in the literature (*Bach et al., 2022*).

The Rubiaceae family also contains species with leaf endophytes that are not housed in conspicuous nodules (*Van Wyk et al., 1990*). Instead, this second phenotype is characterised by endophytes occurring in the intercellular space between the leaf mesophyll cells (*Van Wyk et al., 1990*; *Lemaire et al., 2012b*; *Verstraete et al., 2013a*). This non-nodulating phenotype has been observed in the genus *Psychotria* (22 spp in the Psychotrieae tribe; *Lemaire et al., 2012b*) as well as in the genera *Fadogia* Schweinf., *Fadogiella* Robyns, *Globulostylis* Wernham, *Rytigynia* Blume, and *Vangueria* Juss. (ca 191 spp in the Vanguerieae tribe; *Verstraete et al., 2011*; *Verstraete et al., 2013a*; *Verstraete et al., 2013b*). These non-nodulated endophytes have also

been identified as *Burkholderia* s.l. but they are not necessarily specific to a single host plant species (*Lemaire et al., 2012b*; *Verstraete et al., 2011*; *Verstraete et al., 2013a*; *Verstraete et al., 2013b*). Because the same bacterial species can be found in several host species or even in the soil, this interaction is believed to be less specialised (*Verstraete et al., 2013a*; *Verstraete et al., 2014*).

We currently know that leaf symbiosis (both nodulated and non-nodulated) is present in 8 genera in 4 tribes of Rubiaceae but the true extent remains unknown to date. However, it is likely that leaf symbiosis is more widespread and could occur in other genera as well. *Tilney & Van Wyk (2009)* made histological sections of leaves of *Keetia gueinzii* (Sond.) Bridson (Vanguerieae tribe) and saw "intercellular, non-nodulating, slime-producing bacteria", but this has not been confirmed with molecular data yet. Several *Burkholderia* species have been found to be associated with the roots of coffee plants (i.e. *Coffea arabica* L. and *C. canephora* Pierre ex A.Froehner in *Caballero-Mellado et al. (2004)*, and *C. liberica* W.Bull in *Duong et al. (2021)*) or with the seeds (see review of *Vaughan et al. (2015)*). *Burkholderia* bacteria have also been found to be associated with the leaves of *C. arabica*, although only as epiphytes and not as endophytes (*Vega et al., 2005*).

The genus *Coffea* belongs to the Coffeeae tribe together with the genus *Sericanthe*, for which leaf nodules have been reported (*Lemaire et al., 2011a*). So far, there have been no reports of other taxa in this tribe that contain *Burkholderia* s.l. endophytes, not in leaf nodules nor free between the mesophyll cells. The genera within the Coffeeae tribe that are most closely related to *Sericanthe* are *Diplospora* DC., *Empogona* Hook.f., and *Tricalysia* A.Rich ex DC. (*Arriola et al., 2018*). Since none of these genera have visible nodules in their leaves, if leaf endophytes were to be present, it would have to be non-nodulated ones. Non-nodulated leaf endophytes in Rubiaceae have so far only been found in taxa occurring in Africa (*Lemaire et al., 2012b*; *Verstraete et al., 2013b*). The genus *Diplospora* occurs in (sub)tropical Asia, while the other three genera (*Empogona*, *Sericanthe*, and *Tricalysia*) are found in continental Africa and Madagascar (*POWO, 2023*). It is plausible that the highest likelihood of finding additional taxa with leaf endophytes are taxa closely related to *Sericanthe* and occurring in Africa, and we therefore focus our efforts on *Empogona* and *Tricalysia*.

Our specific aims are 1) to investigate the possible occurrence of non-nodulated leaf endophytes in *Empogona* and *Tricalysia*, 2) if present, to establish their identity and to explore their phylogenetic relationships with other nodulated and non-nodulated endophytes, and 3) to look for patterns in the endophytes and host plants.

Materials & Methods

Plant material, DNA isolation, and sequencing

This study investigates four species of the Coffeeae tribe: *Empogona congesta* (Oliv.) J.E.Burrows, *Tricalysia hensii* De Wild., *T. lasiodelphys* (K.Schum. & K.Krause) A.Chev., and *T. semidecidua* Bridson (Table 1). These species were included in a previous study about chloroplast genome evolution in Rubiaceae (*Ly et al., 2020*). The plant material was obtained

from the collection of Meise Botanic Garden, Belgium. Total DNA isolation from the leaves and DNA sequencing was done as described in *Ly et al. (2020)*. Raw Illumina reads (BGI-seq 500 platform, 2x100 bp paired-end) are available under the BioProject PRJNA880288 (Table 1). While processing the raw sequencing reads, *Ly et al. (2020)* encountered "contamination" (i.e. non-chloroplast reads) and removed those reads to be able to reconstruct the chloroplast genomes of these four species. However, in this study, we are interested in this "contamination" in the raw reads because it contains information on possible leaf endophytes.

Read filtering and assembly

The "contaminants" in the Illumina reads were explored using metagenomic analysis tools. First, Kaiju v.1.9 (*Menzel, Ng & Krogh, 2016*) was used to identify and classify raw reads not belonging to the plant genome with the NCBI non-redundant RefSeq protein database (NCBI nr_euk). When a significant number of "contaminants" was present, it was examined in more detail in the Kaiju output. Second, the presence of "contaminants" was confirmed by exploring the distribution of GC count per sample using FASTQC: several GC count peaks in a single sample might suggest the presence of different organisms in the reads. Finally, reads from "contaminants" showing different GC count were filtered out using KAT v.2.3.4 (k-mer Analysis Tool; *Mapleson et al., 2017*). The reads were analysed with KAT gcp to create a matrix of the number of k-mers found, given k-mer frequency (27-mer) with GC count for each distinct k-mer to explore GC bias. The matrix was displayed via a density plot of the k-mer coverage versus GC count. KAT filter tools were used to filter out reads according to the GC bias (k-mer coverage of 100 to 500X and GC count of 10 to 22%). The reads left after filtering (without trimming, cleaning, or error correction) were subsequently assembled using MaSuRCA v.3.2.6 (*Zimin et al., 2013*) into scaffolds (Table 2) with the default parameters. The draft genome assemblies of the four endophytes are available on GenBank (Table 3) and on Zenodo (*Verstraete et al., 2022a*). Additionally, the reads were assembled using metaSPAdes v.3.15.5 (*Nurk et al., 2017*) and those draft assemblies are also available on Zenodo (*Verstraete et al., 2023*).

Analysis of the assembled bacterial draft genomes

The scaffolds obtained after assembly were compared to a *Burkholderia* s.l. sequence database of 2,288 genomes (representing 22 Gb) downloaded from NCBI (https://www.ncbi.nlm.nih.gov/assembly/?term=burkholderia) as available in September 2019. BLASTn v.2.2.26 (NCBI BLAST) was used for the comparison. Scaffolds with an e-value < 10 e^{-20} were kept. Assembly completeness was assessed using BUSCO v.5.4.3 (*Seppey, Manni & Zdobnov, 2019*) with the proteobacteria_odb10 database downloaded from https://busco.ezlab.org. The assembled draft genomes were annotated using Prokka v.1.14.5 (*Seemann 2014*) and the annotations are available on Zenodo (*Verstraete et al., 2023*).

Assembly of the 16S rRNA genes

A custom pipeline was developed to assemble targeted genes from the raw Illumina reads (Mendez Silva et al., unpublished). In short, raw Illumina reads were mapped to a 16S rRNA reference gene (CP000010.1: c2677815–2676328 *Burkholderia mallei*) using Bowtie2 v.2.4.4 (*Langmead & Salzberg 2012*). Mapped reads were subsequently filtered out and assembled with ABySS v.2.2.1 (*Jackman et al., 2017*).

Phylogenetic analysis of the endophytes

Three genes (16 rRNA, *gyrB*, and *recA*) were identified and used for phylogenetic analysis. A FASTA file with these sequences is available on Zenodo (*Verstraete et al., 2022b*). The 16S rRNA sequences were obtained from the raw Illumina reads, while the *gyrB* and *recA* sequences were obtained from the assembled scaffolds using the BLASTn tool. The *gyrB* nucleotide sequence (NC_006348.1: 3081–5549 *Burkholderia mallei*) and the *recA* nucleotide sequence (NC_006348.1: 290127–291197 *Burkholderia mallei*) were used as queries. Sequences were extracted using the extractseq function as implemented in EMBOSS (*Rice, Longden & Bleasby, 2000*).

The obtained sequences were combined with previously published datasets (*Lemaire et al., 2011b*; *Lemaire et al., 2012b*; *Verstraete et al., 2013b*; *Danneels et al. 2023*) to assess the phylogenetic position of the detected endophytes (Table S1). Automatic sequence alignment was performed with MAFFT v.7.490 (*Katoh et al., 2002*), followed by manual optimisation in Geneious R11. Possible incongruence among the different datasets was tested using a partition homogeneity test (implemented in PAUP*4.0b10a; *Swofford, 2003*). Due to sensitivity issues with the latter test (*Barker & Lutzoni, 2002*), resolution and support values of the different topologies were visually examined (hard versus soft incongruence; *Johnson & Soltis, 1998*). The best-fit nucleotide substitution model for each gene marker was selected using the Akaike information criterion in jModelTest v.2.1.10 (*Posada, 2008*). The model selection test showed that the GTR+I+G model is the most optimal model for 16S rRNA, and that the GTR+G model is the best for *gyrB* and *recA*. Bayesian inference analyses were performed with MrBayes v.3.2.7 (*Huelsenbeck & Ronquist, 2001*) on three individual data partitions and a combined data matrix under a mixed-model approach. Ten million generations were run, and parameters and trees were sampled every 1,000th generation. Chain convergence and ESS parameters were checked with Tracer v.1.7.2 (*Rambaut et al., 2018*). Bayesian posterior probability values above or equal to 0.95 are regarded as statistically supported (*Alfaro, Zoller & Lutzoni, 2003*).

Results

Read filtering and assembly

Several species of Rubiaceae were recently sequenced using a whole genome sequencing approach with DNA extracted from leaves to establish their phylogenetic relationships (*Ly et al., 2020*). During the quality analysis steps, two types of raw reads with very different GC count per read (one peak at 39% and a second at 63%) were found, suggesting the probable presence of

multiple organisms in the raw reads. To understand the origin of the different GC count peaks, a metagenomic approach was applied on the raw reads from *Empogona* and *Tricalysia* species.

For *E. congesta*, 37% of all reads could be assigned to a taxon name based on the NCBI non-redundant RefSeq protein database. About 30% of all reads or 80% of the named reads is assigned to Burkholderiaceae (Fig. S1A). For *T. semidecidua*, 30% of all reads is assigned to a taxon name and about 25% of all reads or 83% of the named reads is assigned to Burkholderiaceae (Fig. S1C). For *T. lasiodelphys*, 23% of all reads is assigned to a taxon name and about 18% of all reads or 75% of the named reads is assigned to Burkholderiaceae (Fig. S1E). For *T. hensii*, 20% of all reads is assigned to a taxon name and about 14% of all reads or 73% of the named reads is assigned to Burkholderiaceae (Fig. S1G). The Kaiju output can also be found in Supplemental Data S1.

To confirm the presence of raw reads belonging to bacteria, the levels of k-mer coverage and GC count per distinct k-mer were analysed. The density plots of the k-mer coverage and the GC count indicate a low to medium k-mer coverage with a wide spread of the GC count (Figs. S1B, S1D, S1F, S1H, left part of the plots) suggesting the presence of reads with sequencing error and a medium coverage genome (i.e. the plant genome). However, the plots also show high coverage (approximately 700X) with GC counts of 15 to 20% (Figs. S1B, S1D, S1F, S1H, upper right of the plots). These unexpected reads with high coverage and high GC count were extracted from the set of raw reads using KAT.

The filtered bacterial reads from *E. congesta* were assembled into 632 scaffolds with an assembly size of 3.8 Mb, while the bacterial reads from *T. hensii, T. lasiodelphys*, and *T. semidecidua* were assembled into 769, 736, and 1,578 scaffolds with assembly sizes of 5.7, 7.8, and 4 Mb, respectively. The assembly N50 ranged from 9.8 to 48 Kb (Table 2).

Analysis of the assembled bacterial draft genomes

For the host species *E. congesta,* 612 of the 632 scaffolds (97%) resulted in a strong hit against the *Burkholderia* s.l. database (e-value cut-off 10 e^{-20}). For the draft genomes of the host species *T. lasiodelphys, T. hensii*, and *T. semidecidua*, there was a lower proportion of hits (89%, 50%, and 31%, respectively) (Table 3). Removing scaffolds without a strong hit against the *Burkholderia* s.l. genome database increased the N50 of the assemblies. However, this did not result in a large decrease in assembly sizes, indicating that only small-size scaffolds were discarded. BUSCO analysis indicated high completeness of the assemblies (> 90%), except for the draft genome of the host species *T. lasiodelphys* (75.3%). In contrast to the other species, the large number of scaffolds and low N50 value for *T. lasiodelphys* suggest a fragmented and incomplete assembly. The genome assemblies should be considered as rough drafts, since bias might have been introduced when filtering the sequences by k-mer/GC count and BLAST. However, it is unlikely that host plant sequences remain present in the assemblies.

Phylogenetic analysis of leaf endophytes in *Empogona* and *Tricalysia*

The phylogenetic position of the endophytes found in *Empogona congesta*, *Tricalysia hensii*, *T. lasiodelphys*, and *T. semidecidua* were inferred from the 16S rRNA, *gyrB*, and *recA* sequences. The combined dataset demonstrated that all four non-nodulating leaf endophytes of *Empogona* and *Tricalysia* belong to *Burkholderia* s.l., more specifically, to the genus *Caballeronia* (Fig. 1). The non-nodulated endophyte of *T. lasiodelphys* is related to the nodulated endophytes of *Sericanthe andongensis* (Hiern) Robbr. and the non-nodulated endophytes of *Psychotria psychotrioides* (DC.) Roberty (BPP: 0.63). The non-nodulated endophyte of *E. congesta* falls within a highly supported clade of nodulated endophytes of several *Pavetta* species and non-nodulated endophytes of several *Globulostylis* species (BPP: 1.00). The non-nodulated endophyte of *T. hensii* is found as sister to the nodulated *Candidatus* Burkholderia kikwitensis (BPP: 1.00), nested within a clade of several other nodulated endophytes of *Psychotria* species. The non-nodulated endophyte of *T. semidecidua* falls in a clade with *Caballeronia fortuita* and *C. novacaledonica* (BPP: 0.97).

Discussion

Detection of non-nodulated leaf endophytes in *Empogona* and *Tricalysia*

The majority of the studies on phylogenetic relationships within the Rubiaceae family to date has relied on phylogenetic approaches using individual nuclear or plastid DNA markers, or a combination of both (*Wikström, Bremer & Rydin, 2020*). However, phylogenomic approaches using more comprehensive amounts of genetic data are becoming more and more common, e.g. mitochondrial genomic data (*Rydin, Wikström & Bremer, 2017*), plastid genomes (*Ly et al., 2020*; *Wikström, Bremer & Rydin, 2020*), or a combination of hundreds of nuclear genes (*Antonelli et al., 2021*; *Thureborn et al., 2022*). The onset of the high-throughput sequencing methodology provides a novel tool to also detect leaf endophytes in Rubiaceae. High-throughput sequencing allows for the sequencing of total DNA, which is subsequently cleaned using bioinformatic filtering to only retain desired sequences, i.e. plant DNA sequences in most cases (e.g. *Charr et al., 2020*). For example, in the study of *Ly et al. (2020)*, the objective was to obtain whole chloroplast genomes from 27 species in the Rubiaceae family. Before chloroplast genome assembly, the raw data was “checked in order to detect potential contamination”, with the unwanted reads being removed from further analysis. However, this “contamination” could be valuable on its own and possibly contain information on the presence of leaf endophytes. In fact, previous studies that used a DNA sequencing approach to detect leaf endophytes in Rubiaceae (e.g. *Van Oevelen et al., 2001*; *Lemaire et al., 2011b*; *Verstraete et al., 2013a*) also extracted total DNA but then eliminated the plant DNA by targeting bacterial DNA.

Our study is based on the unpublished raw data of *Ly et al. (2020)* (but made available here) and looks for evidence of leaf endophytes in the read “contamination”, specifically focussing on the genera *Empogona* and *Tricalysia*. These two genera belong to the Coffeeae tribe and are closely related to *Sericanthe* (*Arriola et al., 2018*), a genus known for its leaf nodulated symbiosis (*Lemaire et al., 2011a*). Unlike *Sericanthe*, *Empogona* and *Tricalysia* do not have visible nodules in their leaves; the detected leaf endophytes are therefore non-nodulated

endophytes. In fact, by examining total DNA, we detected bacterial reads in *E. congesta*, *T. hensii, T. lasiodelphys*, and *T. semidecidua* (Table 3), indicating the presence of leaf endophytes.

Our reassessment of the original reads of the study of *Ly et al. (2020)* shows that we are dealing with leaf endophytes that are present in large proportions. Epiphytic contamination is unlikely because the leaves were cleaned with sterile water before the extraction of the total DNA. We have found that for each of the four investigated species, a large proportion of the reads is assigned to a limited taxonomic group. In *T. semidecidua,* this even reaches 83% of the named reads. This finding is in line with all previous studies about Rubiaceae endophytes. Although at this point, we cannot rule out that there might be more complex communities within the leaves, the fact remains that the particular *Burkholderia*-Rubiaceae interaction has been demonstrated for this new group of plants.

As such, the fact that leaf endophytes are detected in plants previously not implicated in leaf symbiosis is not unexpected. The wider occurrence of plant-bacteria interactions in Rubiaceae has been suggested before (*Lemaire et al., 2012b*; *Verstraete et al., 2013a*). It is therefore likely that additional hidden plant-bacteria interactions will be found when a systematic survey of leaf symbiosis in Rubiaceae is done.

The identity of leaf endophytes and their phylogenetic relationships

Our metagenomic analysis revealed that the bacterial leaf endophytes in *Empogona congesta*, *Tricalysia hensii, T. lasiodelphys*, and *T. semidecidua* belong to the family Burkholderiaceae. This is fully expected as so far, all leaf endophytes in Rubiaceae host plants have been identified as *Burkholderia* s.l. (e.g. *Lemaire et al., 2012b*; *Verstraete, Janssens & Rønsted, 2017*; *Pinto-Carbó et al., 2018; Sinnesael, 2020; Georgiou et al., 2021*).

After finding out the preliminary identity of the leaf endophytes, three genetic markers were extracted in order to achieve a more accurate identification, as well as to include the newly discovered leaf endophytes in a phylogenetic framework. Analysis of 16S rRNA, *gyrB*, and *recA* has already been extensively used to delineate species within *Burkholderia* s.l., as well as to unravel phylogenetic relationships at the generic level (*Verstraete et al., 2011*; *Verstraete et al., 2013a*; *Lemaire et al., 2011a*; *Lemaire et al., 2011b*). The use of these three markers particularly allows for comparison with other leaf endophytes and *Caballeronia* type strains. Even though the use of genomic data would be preferable and is common in recent literature about free-living *Burkholderia* s.l. (*Bach et al., 2022*), genomic information about leaf endophytes is often lacking. Such genomic data is usually extracted from pure cultures, but this is not possible for nodulated leaf endophytes, as they cannot be cultivated (*Sinnesael et al., 2019*). The study of the genomes of non-nodulated endophytes shows more promise but this has only just begun (*Danneels et al., 2023*).

The endophytes of *Empogona* and *Tricalysia* are identified as members of the genus *Caballeronia* (Fig. 1). The non-nodulated endophyte of *T. lasiodelphys* is most closely related to nodulated endophytes of *Sericanthe* and non-nodulated endophytes of *Psychotria*, while the endophyte of *E. congesta* is related to nodulated endophytes of *Pavetta* and non-nodulated

endophytes of *Globulostylis*. The endophyte of *T. hensii* is most closely related to nodulated and non-nodulated endophytes of *Psychotria*. The endophyte of *T. semidecidua* falls in a clade with *Caballeronia fortuita* and *C. novacaledonica*. The type strain of *C. fortuita* was isolated from *Fadogia homblei* (Rubiaceae) rhizosphere soil in South Africa (*Verstraete et al., 2014*; *Peeters et al., 2016*), while the type strain of *C. novacaledonica* was isolated from *Costularia* (Cyperaceae) rhizosphere soil in New Caledonia (*Guentas et al., 2016*). Besides the fact that all these endophytes belong to the genus *Caballeronia*, there does not seem to be much of a phylogenetic pattern.

The study of *Van Oevelen et al. (2001)*, which was the first to identify leaf endophytes in Rubiaceae host plants (i.e. in a few *Psychotria* species), found that the 16S rRNA sequences of the endophytes were similar to that of *Burkholderia glathei*. As a result, they assigned the Rubiaceae leaf endophytes to the genus *Burkholderia* (*Van Oevelen et al., 2001*). However, since then, several changes have been made to the taxonomy of this genus. First, all leaf endophytes were transferred to *Paraburkholderia*, when *Burkholderia* s.l. was split into a pathogenic group (*Burkholderia* s.s.) and a lineage of environmental bacteria (*Paraburkholderia*) (*Sawana, Adeolu & Gupta, 2014*). Later, *Paraburkholderia* was further subdivided and a new genus was created, *Caballeronia* (*Dobritsa & Samadpour, 2016*), which holds all nodulated endophytes as well as the non-nodulated endophytes of *Globulostylis* and *Psychotria*. The present study also designates the newly discovered non-nodulated endophytes of *Empogona* and *Tricalysia* as species of the genus *Caballeronia* (Fig. 1). The non-nodulated endophytes of the Vanguerieae genera (*Fadogia*, *Fadogiella*, *Rytigynia*, and *Vangueria*) remain in *Paraburkholderia*, except for those of *Globulostylis* (Supplemental Data S2). This is in agreement with what was previously known (*Verstraete et al., 2013b*).

None of the investigated host plants has conspicuous bacterial leaf nodules in their leaves, and the endophytes are therefore non-nodulated endophytes. When looking at the phylogenetic tree of the leaf endophytes (Fig. 1), there is no apparent phylogenetic pattern for nodulation. The non-nodulated endophytes in *Empogona* and *Tricalysia* are not clustered, although they all belong to the genus *Caballeronia*. The newly found endophytes are related to other nodulated or non-nodulated leaf endophytes. Also, when analysing the results in a larger framework, no phylogenetic pattern is apparent for all non-nodulated endophytes in Rubiaceae since the majority of the Vanguerieae endophytes are situated within the genus *Paraburkholderia* (Supplemental Data S2). However, we hypothesize that leaf nodulation should be considered as a character of the host plants, rather than of the leaf endophytes (see also *Lemaire et al., 2012b*).

Patterns in the host plants

In this study, we found *Burkholderia* s.l. endophytes in two genera of Rubiaceae that were previously not known to take part in leaf symbiosis. This brings the total number of genera in Rubiaceae for which leaf symbiosis is (molecularly) confirmed to ten: *Psychotria* (Psychotrieae tribe), *Pavetta* (Pavetteae tribe), *Fadogia*, *Fadogiella*, *Globulostylis*, *Rytigynia*, *Vangueria* (Vanguerieae tribe), and *Empogona*, *Sericanthe*, *Tricalysia* (Coffeeae tribe).

Finding phylogenetic patterns is however challenging. For the five Vanguerieae genera, the presence of *Burkholderia* s.l. endophytes is consistent at the genus level (*Verstraete et al., 2013a*) and the plants with leaf symbiosis only occur in Africa and Madagascar. For the genus *Pavetta*, it is generally assumed that most of the species have leaf nodules (*Lersten & Horner, 1976*; *Miller, 1990*; *Lemaire et al., 2011b*) and these are found in the Paleotropics (*POWO, 2023*). The presence or absence of nodules, as well as their form, has been used in the past to classify subgeneric taxa (e.g. *P.* series *Enodulosae*; *Bremekamp, 1939*). However, within *Pavetta*, the phylogenetic distribution of species without nodules is irregular (*Bremekamp, 1934*). Within the pantropical genus *Psychotria*, the situation is even more complex. The number of (known) species with nodules (ca 80 spp; *Lemaire et al., 2012b*) is rather limited compared to the total number of species in the genus (ca 1645; *POWO, 2023*), meaning that the nodulated form of leaf symbiosis is not a frequent character in *Psychotria*. Also, nodulating *Psychotria* plants are restricted to Africa and Madagascar. Unfortunately, a detailed list with the presence and absence of nodules is missing, so it remains uncertain to date to what extent leaf nodulation is present in *Psychotria*. The few nodulating *Psychotria* species that were included in molecular studies were not recovered as a monophyletic group (*Razafimandimbison et al., 2014*). Besides nodulating species, there are also some species without nodules but with non-nodulated endophytes (*Lemaire et al., 2012b*). Although these non-nodulating *Psychotria* plants were found in a clade (clade III in *Lemaire et al. (2012b)*) separate from the nodulating species (clade II in *Lemaire et al. (2012b)*), they also do not form a monophyletic group. Finally, the species *Psychotria lucens* Hiern was used in the past as "negative control" (*Van Oevelen et al., 2001*) and later several other species without bacterial endophytes were found (*Lemaire et al., 2012b*). This means that all three conditions occur in *Psychotria*. However, the taxonomic delimitation of *Psychotria* has changed many times (*Razafimandimbison et al., 2014*) and finding evolutionary patterns within this megagenus remains challenging.

The genera *Empogona*, *Sericanthe*, and *Tricalysia* are closely related (*Arriola et al., 2018*), and the species of *Empogona* (*Tosh et al., 2009*) and *Sericanthe* (*Robbrecht, 1978*) were once included in *Tricalysia*. Perhaps it is therefore not so surprising to find leaf endophytes in these genera. The difference between *Sericanthe* on the one hand and *Empogona* and *Tricalysia* on the other, is that the former has leaf nodules, while the latter do not. A next step would be to investigate additional species of *Empogona* and *Tricalysia* to elucidate the true extent of leaf symbiosis in these two genera and to find out whether leaf symbiosis has any phylogenetic signal. Another observation worth investigating is that all tree genera are related to *Diplospora* (*Arriola et al., 2018*) and *Discospermum* Dalzell (*Tosh et al., 2009*). However, the major difference is that these two genera are found in (sub)tropical Asia, while the other three genera (*Empogona*, *Sericanthe*, and *Tricalysia*) are exclusively found in continental Africa and Madagascar (*POWO, 2023*). So far, the nodulated symbiosis is predominantly found in Africa (except for a few noduled *Pavetta* species in (sub)tropical Asia) and non-nodulated symbiosis is even restricted to that area. A broader screening of the Coffeeae tribe would therefore be useful to demonstrate the presence or absence of a geographic pattern in leaf symbiosis. However, for

this, a new phylogenetic framework of the Coffeeae is needed, which shows the relationships between the different genera and onto which the character "leaf symbiosis" can then be plotted.

Conclusions

Metagenomic analysis revealed that bacterial endophytes are present in the leaves of *Empogona* and *Tricalysia*, two genera not previously implicated in leaf symbiosis. This result is another step towards discovering the true extent of leaf symbiosis (both nodulated and non-nodulated) in the Rubiaceae family. The endophytes belong to the genus *Caballeronia* and are not housed in leaf nodules. No phylogenetic signals have been found in the endophytes, nor does there appear to be a phylogenetic or geographical pattern in the host species. However, leaf symbiosis is predominantly found in Africa (as are both *Empogona* and *Tricalysia*), so additional plant-bacteria interactions are likely to be found on this continent.

Acknowledgements

The authors thank Dr Mathilde Dupeyron for her input on the manuscript.

References

Alfaro ME, Zoller S, Lutzoni F. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution* 20(2): 255–266. DOI: 10.1093/molbev/msg028

Antonelli A, Clarkson JJ, Kainulainen K, Maurin O, Brewer GE, Davis AP, Epitawalage N, Goyder DJ, Livshultz T, Persson C, Pokorny L, Straub SCK, Struwe L, Zuntini AR, Forest F, Baker WJ. 2021. Settling a family feud: a high-level phylogenomic framework for the Gentianales based on 353 nuclear genes and partial plastomes. *American Journal of Botany* 108(7): 1143–1165. DOI: 10.1002/ajb2.1697

Arriola AH, Davis AP, Davies NM, Meve U, Liede-Schumann S, Alejandro GJD. 2018. Using multiple plastid DNA regions to construct the first phylogenetic tree for Asian genera of Coffeeae (Ixoroideae, Rubiaceae). *Botanical Journal of the Linnean Society* 188: 132–143. DOI: 10.1093/botlinnean/boy059

Bach E, Pereira Passaglia LM, Jiao J, Gross H. 2022. *Burkholderia* in the genomic era: from taxonomy to the discovery of new antimicrobial secondary metabolites. *Critical Reviews in Microbiology* 48(2): 121–160. DOI: 10.1080/1040841X.2021.1946009

Barker FK, Lutzoni FM. 2002. The utility of the incongruence length difference test. *Systematic Biology* 51(4): 625–637. DOI: 10.1080/10635150290102302

Bremekamp CEB. 1934. A monograph of the genus *Pavetta* L. *Repertorium novarum specierum regni vegetabilis* 37: 1–61. DOI: 10.1002/fedr.19340370102

Bremekamp CEB. 1939. A monograph of the genus *Pavetta* L.: additions and emendations. *Repertorium novarum specierum regni vegetabilis* 47: 12–28. DOI: 10.1002/fedr.19390470106

Caballero-Mellado J, Martínez-Aguilar L, Paredes-Valdez G, Estrada-de los Santos P. 2004. *Burkholderia unamae* sp. nov., an N_2-fixing rhizospheric and endophytic species.

International Journal of Systematic and Evolutionary Microbiology 54: 1165–1172. DOI: 10.1099/ijs.0.02951-0

Carlier A, Cnockaert M, Fehr L, Vandamme P, Eberl L. 2017. Draft genome and description of *Orrella dioscoreae* gen. nov. sp. nov., a new species of Alcaligenaceae isolated from leaf acumens of *Dioscorea sansibarensis*. *Systematic and Applied Microbiology* 40: 11–21. DOI: 10.1016/j.syapm.2016.10.002

Carlier A, Fehr L, Pinto-Carbó M, Schäberle T, Reher R, Dessein S, König G, Eberl L. 2016. The genome analysis of *Candidatus* Burkholderia crenata reveals that secondary metabolism may be a key function of the *Ardisia crenata* leaf nodule symbiosis. *Environmental Microbiology* 18: 2507–2522. DOI: 10.1111/1462-2920.13184

Charr J-C, Garavito A, Guyeux C, Crouzillat D, Descombes P, Fournier C, Ly SN, Raharimalala EN, Rakotomalala J-J, Stoffelen P, Janssens S, Hamon P, Guyot R. 2020. Complex evolutionary history of coffees revealed by full plastid genomes and 28,800 nuclear SNP analyses, with particular emphasis on *Coffea canephora* (Robusta coffee). *Molecular Phylogenetics and Evolution* 151: 106906. DOI: 10.1016/j.ympev.2020.106906

Danneels B, Blignaut M, Marti G, Sieber S, Vandamme P, Meyer M, Carlier A. 2023. Cyclitol metabolism is a central feature of *Burkholderia* leaf symbionts. *Environmental Microbiology* 25(2): 454–472. DOI: 10.1111/1462-2920.16292

Danneels B, Viruel J, Mcgrath K, Janssens S, Wales N, Wilkin P, Carlier A. 2021. Patterns of transmission and horizontal gene transfer in the *Dioscorea sansibarensis* leaf symbiosis revealed by whole-genome sequencing. *Current Biology* 31(12): 2666–2673.e4. DOI: 10.1016/j.cub.2021.03.049

Dobritsa AP, Samadpour M. 2016. Transfer of eleven species of the genus *Burkholderia* to the genus *Paraburkholderia* and proposal of *Caballeronia* gen. nov. to accommodate twelve species of the genera *Burkholderia* and *Paraburkholderia*. *International Journal of Systematic and Evolutionary Microbiology* 66: 2836–2846. DOI: 10.1099/ijsem.0.001065

Duong B, Nguyen HX, Phan HV, Colella S, Trinh PQ, Hoang GT, Nguyen TT, Marraccini P, Lebrun M, Duponnois R. 2021. Identification and characterization of Vietnamese coffee bacterial endophytes displaying in vitro antifungal and nematicidal activities. *Microbiological Research* 242: 126613. DOI: 10.1016/j.micres.2020.126613

Georgiou A, Sieber S, Hsiao CC, Grayfer T, Gorenflos López JL, Gademann K, Eberl L, Bailly A. 2021. Leaf nodule endosymbiotic *Burkholderia* confer targeted allelopathy to their *Psychotria* hosts. Scientific Reports 11: 22465. DOI: 10.1038/s41598-021-01867-2

Guentas L, Gensous S, Cavaloc Y, Ducousso M, Amir H, De Georges de Ledenon B, Moulin L, Jourand P. 2016. *Burkholderia novacaledonica* sp. nov. and *B. ultramafica* sp. nov. isolated from roots of *Costularia* spp. pioneer plants of ultramafic soils in New Caledonia. Systematic and Applied Microbiology 39(2): 151–159. DOI: 10.1016/j.syapm.2016.03.008

Huelsenbeck JP, Ronquist F. 2001 MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8): 754–755. DOI: 10.1093/bioinformatics/17.8.754

Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Research* 27(5): 768–777. DOI: 10.1101/gr.214346.116

Johnson LA, Soltis DE. 1998. Assessing congruence: empirical examples from molecular data. In: Soltis DE, Soltis PS, Doyle JJ, eds. *Molecular Systematics of Plants 2: DNA Sequencing*. Boston: Kluwer, 297–348.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30(14): 3059–3066. DOI: 10.1093/nar/gkf436

Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9: 357–359. DOI: 10.1038/nmeth.1923

Lemaire B, Lachenaud O, Persson C, Smets E, Dessein S. 2012b. Screening for leaf-associated endophytes in the genus *Psychotria*. *FEMS Microbiology Ecology* 81(2): 364–372. DOI: 10.1111/j.1574-6941.2012.01356.x

Lemaire B, Robbrecht E, Van Wyk B, Van Oevelen S, Verstraete B, Prinsen E, Smets E, Dessein S. 2011a. Identification, origin, and evolution of leaf nodulating symbionts of *Sericanthe* (Rubiaceae). *Journal of Microbiology* 49(6): 935–941. DOI: 10.1007/s12275-011-1163-5

Lemaire B, Vandamme P, Merckx V, Smets E, Dessein S. 2011b. Bacterial leaf symbiosis in angiosperms: host specificity without co-speciation. *PLoS ONE* 6: e24430. DOI: 10.1371/journal.pone.0024430

Lemaire B, Van Oevelen S, De Block P, Verstraete B, Smets E, Prinsen E, Dessein S. 2012a. Identification of the bacterial endosymbionts in leaf nodules of *Pavetta* (Rubiaceae). *International Journal of Systematic and Evolutionary Microbiology* 62(1): 202–209. DOI: 10.1099/ijs.0.028019-0

Lersten NR, Horner HT. 1976. Bacterial leaf nodule symbiosis in angiosperms with emphasis on Rubiaceae and Myrsinaceae. *The Botanical Review* 42(2): 145–214.

Ly SN, Garavito A, De Block P, Asselman P, Guyeux C, Charr JC, Janssens S, Mouly A, Hamon P, Guyot R. 2020. Chloroplast genomes of Rubiaceae: comparative genomics and molecular phylogeny in subfamily Ixoroideae. *PLoS ONE* 15(4): e0232295. DOI: 10.1371/journal.pone.0232295

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 3(4): 574–576. DOI: 10.1093/bioinformatics/btw663

Menzel P, Ng KL, Krogh A. 2016. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications* 7: 11257. DOI: 10.1038/ncomms11257

Miller IM. 1990. Bacterial leaf nodule symbiosis. In: Callow JA, ed. *Advances in Botanical Research 17*. San Diego: Academic Press, 163–234.

Nurk S, Meleshko D, Korobeynikov A, Pevzner PA. 2017. metaSPAdes: a new versatile metagenomic assembler. *Genome Research* 27(5): 824–834. DOI: 10.1101/gr.213959.116

Orozco-Mosqueda MaC, Santoyo G. 2021. Plant-microbial endophytes interactions: scrutinizing their beneficial mechanisms from genomic explorations. *Current Plant Biology* 25: 100189. DOI: 10.1016/j.cpb.2020.100189

Peeters C, Meier-Kolthoff JP, Verheyde B, De Brandt E, Cooper VS, Vandamme P. 2016. Phylogenomic study of *Burkholderia glathei*-like organisms, proposal of 13 novel *Burkholderia* species and emended descriptions of *Burkholderia sordidicola*, *Burkholderia zhejiangensis*, and *Burkholderia grimmiae*. Frontiers in Microbiology 7: 877. DOI: 10.3389/fmicb.2016.00877

Pinto-Carbó M, Gademann K, Eberl L, Carlier A. 2018. Leaf nodule symbiosis: function and transmission of obligate bacterial endophytes. *Current Opinion in Plant Biology* 44: 23–31. DOI: 10.1016/j.pbi.2018.01.001

Poole P, Ramachandran V, Terpolilli J. 2018. *Rhizobia*: from saprophytes to endosymbionts. *Nature Reviews Microbiology* 16: 291–303. DOI: 10.1038/nrmicro.2017.171

Posada D. 2008. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* 25(7): 1253–1256. DOI: 10.1093/molbev/msn083.
POWO 2023. Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Available at http://www.plantsoftheworldonline.org/ (accessed 31 March 2023)
Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic Biology* 67(5): 901–904. DOI: 10.1093/sysbio/syy032
Razafimandimbison SG, Kainulainen K, Wikström N, Bremer B. 2017. Historical biogeography and phylogeny of the pantropical Psychotrieae alliance (Rubiaceae), with particular emphasis on the Western Indian Ocean Region. *American Journal of Botany* 104: 1407–1423. DOI: 10.3732/ajb.1700116
Razafimandimbison SG, Taylor CM, Wikström N, Pailler T, Khodabandeh A, Bremer B. 2014. Phylogeny and generic limits in the sister tribes Psychotrieae and Palicoureeae (Rubiaceae): evolution of schizocarps in *Psychotria* and origins of bacterial leaf nodules of the Malagasy species. *American Journal of Botany* 101: 1102–1126. DOI: 10.3732/ajb.1400076
Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. *Trends in Genetics* 16(6): 276–277. DOI: 10.1016/S0168-9525(00)02024-2
Robbrecht E. 1978. *Sericanthe*, a new African genus of Rubiaceae (Coffeeae). *Bulletin du Jardin botanique national de Belgique / Bulletin van de National Plantentuin van België* 48: 3–78. DOI: 10.2307/3667918
Rydin C, Wikström N, Bremer B. 2017. Conflicting results from mitochondrial genomic data challenge current views of Rubiaceae phylogeny. *American Journal of Botany* 104(10): 1522–1532. DOI: 10.3732/ajb.1700255
Sawana A, Adeolu M, Gupta RS. 2014. Molecular signatures and phylogenomic analysis of the genus *Burkholderia*: proposal for division of this genus into the emended genus *Burkholderia* containing pathogenic organisms and a new genus *Paraburkholderia* gen. nov. harboring environmental species. *Frontiers in Genetics* 5: 429. DOI: 10.3389/fgene.2014.00429
Schindler F, Fragner L, Herpell JB, Berger A, Brenner M, Tischler S, Bellaire A, Schönenberger J, Li W, Sun X, Schinnerl J, Brecker L, Weckwerth W. 2021. Dissecting metabolism of leaf nodules in *Ardisia crenata* and *Psychotria punctata*. *Frontiers in Molecular Biosciences* 8: 1–23. DOI: 10.3389/fmolb.2021.683671
Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30(14): 2068–2069. DOI: 10.1093/bioinformatics/btu153
Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. In: Kollmar M, ed. *Gene Prediction. Methods in Molecular Biology, vol 1962*. New York: Humana, 227–245. DOI: 10.1007/978-1-4939-9173-0_14
Sinnesael A. 2020. Bacterial leaf symbiosis – Origin, function, evolutionary gain, and transmission mode of endophytes in bacteriophilous Rubiaceae. D. Phil. Thesis, KU Leuven.
Sinnesael A, Leroux O, Janssens SB, Smets E, Panis B, Verstraete B. 2019. Is the bacterial leaf nodule symbiosis obligate for *Psychotria umbellata*? The development of a *Burkholderia*-free host plant. *PLoS ONE* 14: e0219863. DOI: 10.1371/journal.pone.0219863
Swofford DL. 2003. PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
Tilney PM, Van Wyk AE. 2009. Taxonomy of the genus *Keetia* (Rubiaceae-subfam. Ixoroideae-tribe Vanguerieae) in southern Africa, with notes on bacterial symbiosis as well as the

structure of colleters and the ‘stylar head’ complex. *Bothalia* 39: 165–175. DOI: 10.4102/abc.v39i2.242

Tosh J, Davis AP, Dessein S, De Block P, Huysmans S, Fay MF, Smets E, Robbrecht E. 2009. Phylogeny of *Tricalysia* (Rubiaceae) and its relationships with allied genera based on plastid DNA data: resurrection of the genus *Empogona*. *Annals of the Missouri Botanical Garden* 96: 194–213. DOI: 10.3417/2006202

Trimen H. 1894. *A hand book of the flora of Ceylon. Part II Connaraceae–Rubiaceae*. London: Dulau & Co.

Thureborn O, Razafimandimbison SG, Wikström N, Rydin C. 2022. Target capture data resolve recalcitrant relationships in the coffee family (Rubioideae, Rubiaceae). *Frontiers in Plant Science* 13: 967456. DOI: 10.3389/fpls.2022.967456

Van Oevelen S, De Wachter R, Vandamme P, Robbrecht E, Prinsen E. 2002. Identification of the bacterial endosymbionts in leaf galls of *Psychotria* (Rubiaceae, angiosperms) and proposal of “*Candidatus* Burkholderia kirkii” sp. nov. *International Journal of Systematic and Evolutionary Microbiology* 52: 2023–2027. DOI: 10.1099/00207713-52-6-2023

Van Oevelen S, Prinsen E, De Wachter R, Robbrecht E. 2001. The taxonomic value of bacterial symbiont identification in African *Psychotria* (Rubiaceae). *Systematics and Geography of Plants* 71: 557–563. DOI: 10.2307/3668700

Van Wyk AE, Kok PDF, van Bers NL, van der Merwe CF. 1990. Non-pathological bacterial symbiosis in *Pachystigma* and *Fadogia* (Rubiaceae): its evolutionary significance and possible involvement in the aetiology of gousiekte in domestic ruminants. *South African Journal of Science* 86: 93–96.

Vaughan MJ, Mitchell T, McSpadden Gardener BB. 2015. What’s inside that seed we brew? A new approach to mining the coffee microbiome. *Applied and Environmental Microbiology* 81: 6518–6527. DOI: 10.1128/AEM.01933-15

Vega FE, Pava-Ripoll M, Posada F, Buyer JS. 2005. Endophytic bacteria in *Coffea arabica* L. *Journal of Basic Microbiology* 45: 371–380. DOI: 10.1002/jobm.200410551

Verstraete B, Janssens S, De Block P, Asselman P, Mendez Silva G, Ly SN, Hamon P, Guyot R. 2022a. Metagenomics of African *Empogona* and *Tricalysia* (Rubiaceae) reveals the presence of leaf endophytes. DOI: 10.5281/zenodo.6090258

Verstraete B, Janssens S, De Block P, Asselman P, Mendez Silva G, Ly SN, Hamon P, Guyot R. 2022b. Metagenomics of African *Empogona* and *Tricalysia* (Rubiaceae) reveals the presence of leaf endophytes – Fasta files. DOI: 10.5281/zenodo.7333199

Verstraete B, Janssens S, De Block P, Asselman P, Mendez Silva G, Ly SN, Hamon P, Guyot R. 2023. Metagenomics of African *Empogona* and *Tricalysia* (Rubiaceae) reveals the presence of leaf endophytes – Assembly and annotation Files. DOI: 10.5281/zenodo.7787854

Verstraete B, Janssens S, Lemaire B, Smets E, Dessein S. 2013b. Phylogenetic lineages in Vanguerieae (Rubiaceae) associated with *Burkholderia* bacteria in sub-Saharan Africa. *American Journal of Botany* 100: 2380–2387. DOI: 10.3732/ajb.1300303

Verstraete B, Janssens S, Rønsted N. 2017. Non-nodulated bacterial leaf symbiosis promotes the evolutionary success of its host plants in the coffee family (Rubiaceae). *Molecular Phylogenetics and Evolution* 113: 161–168. DOI: 10.1016/j.ympev.2017.05.022

Verstraete B, Janssens S, Smets E, Dessein S. 2013a. Symbiotic ß-proteobacteria beyond legumes: *Burkholderia* in Rubiaceae. *PLoS ONE* 8(1): e55260. DOI: 10.1371/journal.pone.0055260

Verstraete B, Peeters C, Van Wyk B, Smets E, Dessein S, Vandamme P. 2014. Intraspecific variation in *Burkholderia caledonica*: Europe vs. Africa and soil vs. endophytic isolates. *Systematic and Applied Microbiology* 37:194–199. DOI: 10.1016/j.syapm.2013.12.001

Verstraete B, Van Elst D, Steyn H, Van Wyk B, Lemaire B, Smets E, Dessein S. 2011. Endophytic bacteria in toxic South African plants: identification, phylogeny and possible involvement in gousiekte. *PLoS ONE* 6: e19265. DOI: 10.1371/journal.pone.0019265

von Faber 1912. Das erbliche Zusammenleben von Bakterien und tropischen Pflanzen. *Jahrbücher für Wissenschaftliche Botanik* 51: 285–375.

Wikström N, Bremer B, Rydin C. 2020. Conflicting phylogenetic signals in genomic data of the coffee family (Rubiaceae). *Journal of Systematics and Evolution* 58(4): 440–460. DOI: 10.1111/jse.12566

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. 2013. The MaSuRCA genome assembler. *Bioinformatics* 29(21): 2669–2677. DOI: 10.1093/bioinformatics/btt476

Zimmermann A. 1902. Über Bakterienknoten in den Blättern einiger Rubiaceen. *Jahrbücher für wissenschaftliche Botanik* 37: 1–11.

Figure 1

Phylogenetic tree of *Burkholderia* s.l., focussing on *Caballeronia*, based on 16S rRNA, *gyrB*, and *recA* sequences.

The four non-nodulated endophytes in *Empogona* and *Tricalysia* belong to the genus *Caballeronia* and are indicated in bold. Thick lines indicate Bayesian Posterior Probability (BPP) values higher than or equal to 0.95, thin lines indicate BPP support values lower than 0.95.

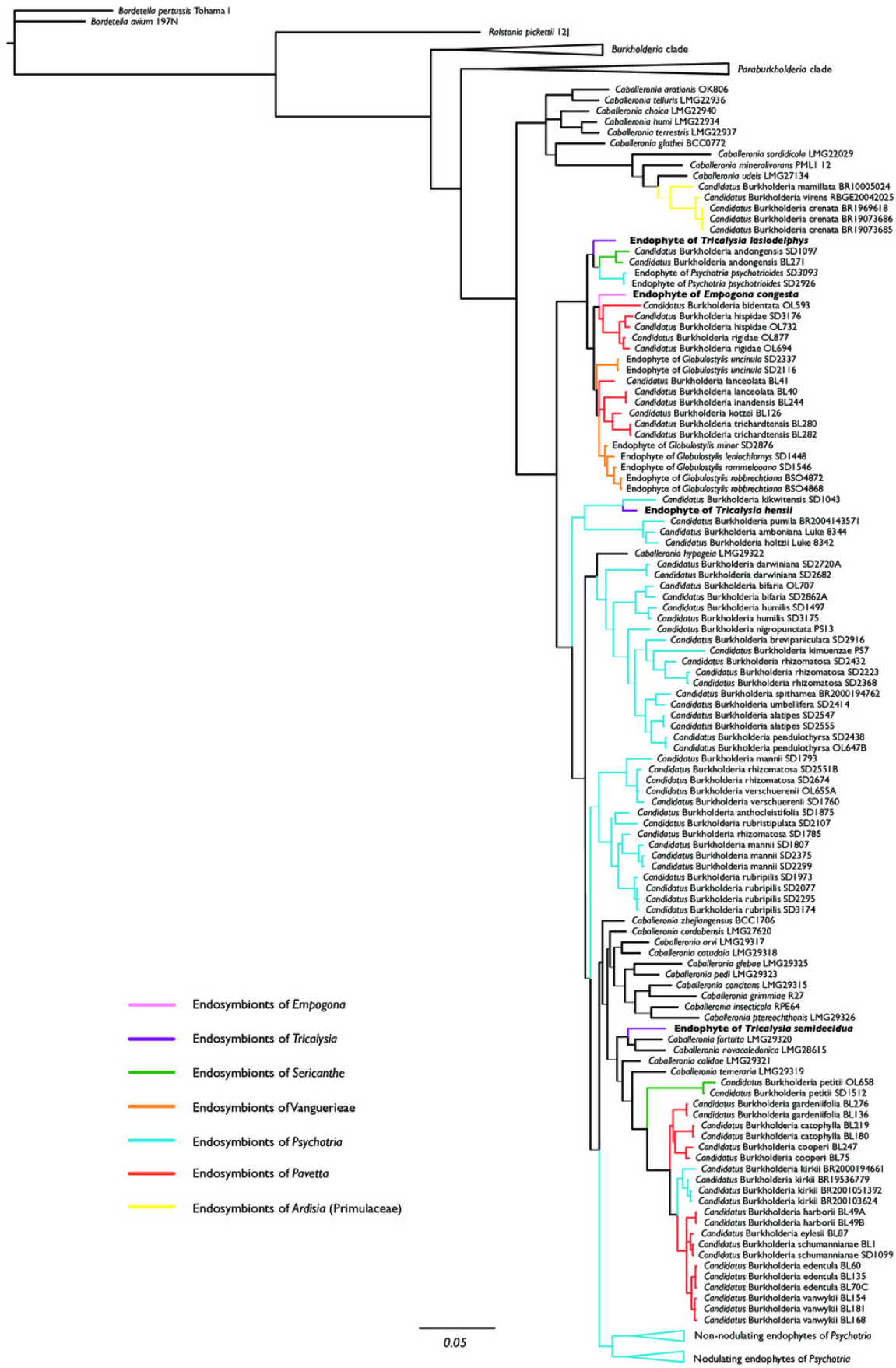
Bordetella pertussis Tohama I
Bordetella avium 197N
Ralstonia pickettii 12J
Burkholderia clade
Paraburkholderia clade
Caballeronia arationis OK806
Caballeronia telluris LMG22936
Caballeronia choica LMG22940
Caballeronia humi LMG22934
Caballeronia terrestris LMG22937
Caballeronia glathei BCC0772
Caballeronia sordidicola LMG22029
Caballeronia mineralivorans PML1 12
Caballeronia udeis LMG27134
Candidatus Burkholderia mamillata BR10005024
Candidatus Burkholderia virens RBGE20042025
Candidatus Burkholderia crenata BR1969618
Candidatus Burkholderia crenata BR19073686
Candidatus Burkholderia crenata BR19073685
Endophyte of Tricalysia lasiodelphys
Candidatus Burkholderia andongensis SD1097
Candidatus Burkholderia andongensis BL271
Endophyte of Psychotria psychotrioides SD3093
Endophyte of Psychotria psychotrioides SD2926
Endophyte of Empogona congesta
Candidatus Burkholderia bidentata OL593
Candidatus Burkholderia hispidae SD3176
Candidatus Burkholderia hispidae OL732
Candidatus Burkholderia rigidae OL877
Candidatus Burkholderia rigidae OL694
Endophyte of Globulostylis uncinula SD2337
Endophyte of Globulostylis uncinula SD2116
Candidatus Burkholderia lanceolata BL41
Candidatus Burkholderia lanceolata BL40
Candidatus Burkholderia inandensis BL244
Candidatus Burkholderia kotzei BL126
Candidatus Burkholderia trichardtensis BL280
Candidatus Burkholderia trichardtensis BL282
Endophyte of Globulostylis minor SD2876
Endophyte of Globulostylis leniochlamys SD1448
Endophyte of Globulostylis rammelooana SD1546
Endophyte of Globulostylis robbrechtiana BSO4872
Endophyte of Globulostylis robbrechtiana BSO4868
Candidatus Burkholderia kikwitensis SD1043
Endophyte of Tricalysia hensii
Candidatus Burkholderia pumila BR2004143571
Candidatus Burkholderia amboniana Luke 8344
Candidatus Burkholderia holtzii Luke 8342
Caballeronia hypogeia LMG29322
Candidatus Burkholderia darwiniana SD2720A
Candidatus Burkholderia darwiniana SD2682
Candidatus Burkholderia bifaria OL707
Candidatus Burkholderia bifaria SD2862A
Candidatus Burkholderia humilis SD1497
Candidatus Burkholderia humilis SD3175
Candidatus Burkholderia nigropunctata PS13
Candidatus Burkholderia brevipaniculata SD2916
Candidatus Burkholderia kimuenzae PS7
Candidatus Burkholderia rhizomatosa SD2432
Candidatus Burkholderia rhizomatosa SD2223
Candidatus Burkholderia rhizomatosa SD2368
Candidatus Burkholderia spithamea BR2000194762
Candidatus Burkholderia umbellifera SD2414
Candidatus Burkholderia alatipes SD2547
Candidatus Burkholderia alatipes SD2555
Candidatus Burkholderia pendulothyrsa SD2438
Candidatus Burkholderia pendulothyrsa OL647B
Candidatus Burkholderia mannii SD1793
Candidatus Burkholderia rhizomatosa SD2551B
Candidatus Burkholderia rhizomatosa SD2674
Candidatus Burkholderia verschuerenii OL655A
Candidatus Burkholderia verschuerenii SD1760
Candidatus Burkholderia anthocleistifolia SD1875
Candidatus Burkholderia rubristipulata SD2107
Candidatus Burkholderia rhizomatosa SD1785
Candidatus Burkholderia mannii SD1807
Candidatus Burkholderia mannii SD2375
Candidatus Burkholderia mannii SD2299
Candidatus Burkholderia rubripilis SD1973
Candidatus Burkholderia rubripilis SD2077
Candidatus Burkholderia rubripilis SD2295
Candidatus Burkholderia rubripilis SD3174
Caballeronia zhejiangensus BCC1706
Caballeronia cordobensis LMG27620
Caballeronia arvi LMG29317
Caballeronia catudaia LMG29318
Caballeronia glebae LMG29325
Caballeronia pedi LMG29323
Caballeronia concitans LMG29315
Caballeronia grimmiae R27
Caballeronia insecticola RPE64
Caballeronia ptereochthonis LMG29326
Endophyte of Tricalysia semidecidua
Caballeronia fortuita LMG29320
Caballeronia novacaledonica LMG28615
Caballeronia calidae LMG29321
Caballeronia temeraria LMG29319
Candidatus Burkholderia petitii OL658
Candidatus Burkholderia petitii SD1512
Candidatus Burkholderia gardeniifolia BL276
Candidatus Burkholderia gardeniifolia BL136
Candidatus Burkholderia catophylla BL219
Candidatus Burkholderia catophylla BL180
Candidatus Burkholderia cooperi BL247
Candidatus Burkholderia cooperi BL75
Candidatus Burkholderia kirkii BR2000194661
Candidatus Burkholderia kirkii BR19536779
Candidatus Burkholderia kirkii BR2001051392
Candidatus Burkholderia kirkii BR200103624
Candidatus Burkholderia harborii BL49A
Candidatus Burkholderia harborii BL49B
Candidatus Burkholderia eylesii BL87
Candidatus Burkholderia schumannianae BL1
Candidatus Burkholderia schumannianae SD1099
Candidatus Burkholderia edentula BL60
Candidatus Burkholderia edentula BL135
Candidatus Burkholderia edentula BL70C
Candidatus Burkholderia vanwykii BL154
Candidatus Burkholderia vanwykii BL181
Candidatus Burkholderia vanwykii BL168
Non-nodulating endophytes of Psychotria
Nodulating endophytes of Psychotria
Endosymbionts of Empogona
Endosymbionts of Tricalysia
Endosymbionts of Sericanthe
Endosymbionts of Vanguerieae
Endosymbionts of Psychotria
Endosymbionts of Pavetta
Endosymbionts of Ardisia (Primulaceae)
0.05

Table 1(on next page)

Provenance of the material of the four investigated plant species, deposited at Meise Botanic Garden (https://www.botanicalcollections.be), and information about the raw sequencing reads obtained from the total DNA isolated.

Table 1:
Provenance of the material of the four investigated plant species, deposited at Meise Botanic Garden (https://www.botanicalcollections.be), and information about the raw sequencing reads obtained from the total DNA isolated.

Species name	Barcode of voucher	Country	Number of reads	Number of nucleotides (Gb)	NCBI accession number
Empogona congesta	BR6202001591004	Zambia	2 x 65,763,918	13.15	SRR21547710
Tricalysia hensii	BR0000012568055	D.R. Congo	2 x 61,592,477	12.31	SRR21547709
Tricalysia lasiodelphys	BR0000009955950	Cameroon	2 x 62,973,455	12.59	SRR21547708
Tricalysia semidecidua	BR6202001590007	Zambia	2 x 65,171,012	13.03	SRR21547707

Table 2(on next page)

Statistics on the filtered and assembled bacterial reads in *Empogona* and *Tricalysia*.

Table 2:
Statistics on the filtered and assembled bacterial reads in *Empogona* and *Tricalysia*.

Host species	Number of filtered reads (100 bp)	Number of scaffolds	Assembly N50 (bp)	Assembly size (Mb)	Average coverage
Empogona congesta	2 x 14,051,241	632	9,820	3.863	727X
Tricalysia hensii	2 x 11,799,331	736	48,029	7.818	301X
Tricalysia lasiodelphys	2 x 9,204,217	2,971	2,095	4.340	424X
Tricalysia semidecidua	2 x 23,251,991	1,578	22,085	4.082	1139X

Table 3(on next page)

Statistics about the scaffolds after BLASTn filtering against *Burkholderia* s.l. genomes.

Table 3:
Statistics about the scaffolds after BLASTn filtering against *Burkholderia* s.l. genomes.

Host species	Number of scaffolds with BLASTn hits against *Burkholderia* s.l. genomes (e-value 10 e^{-20})	Assembly N50 of filtered scaffolds (bp)	Assembly size of filtered scaffolds (Mb)	Complete BUSCO scores	Missing BUSCO scores	NCBI accession number
Empogona congesta	612	10,140	3.762	93.2%	1.8%	JAQFVJ000000000
Tricalysia hensii	369	60,447	6.951	99.6%	0.4%	JAQFVG000000000
Tricalysia lasiodelphys	2,644	2,165	4.145	75.3%	6.4%	JAQFVH000000000
Tricalysia semidecidua	490	24,168	3.919	95.4%	1.4%	JAQFVI000000000