



Comparative genomics of *Sphingomonas* from the cupuassu (*Theobroma grandiflorum*) phyllosphere reveals signatures of host adaptation and beneficial interactions

Ana Karla da Silva² · Camila Henriques de Paula² · Jéssica Pereira de Matos² · Dilson Fagundes Ribeiro² · Isabella Ferreira Cordeiro² · Camila Gracyelle de Carvalho Lemes² · Angélica Bianchini Sanchez² · Camila Carrião Machado Garcia^{1,2} · Vinicius A. C. de Abreu³ · Rafael Moysés Alves⁴ · Mauro de Medeiros Oliveira⁵ · João Victor dos Anjos Almeida⁵ · Nalvo F. Almeida⁶ · Alessandro M. Varani⁵ · Leandro Marcio Moreira^{1,2}

Received: 18 November 2025 / Accepted: 23 February 2026 / Published online: 26 March 2026
© The Author(s) 2026

Abstract

Members of the genus *Sphingomonas* are ubiquitous phyllosphere inhabitants recognized for their metabolic versatility and intimate associations with plants. Here, we describe and compare eight high-quality metagenome-assembled genomes (MAGs) of *Sphingomonas* recovered from leaves of the witches'-broom-resistant cupuassu (*Theobroma grandiflorum*) genotype C174. These MAGs represent four known species — *S. adhaesiva* (bin09, bin18, bin28), *S. fennica* (bin11), *S. phyllosphaerae* (bin08, bin10), and *S. taxi* (bin15, bin20) — and one novel lineage proposed as *CandidatusSphingomonascupuassunensis*. Comparative analyses with 15 reference *Sphingomonas* genomes revealed a conserved core of 623 orthologous protein families and extensive genomic divergence, reflecting broad adaptive plasticity within the genus. Functional annotation uncovered substantial variability in xenobiotic degradation, nitrogen metabolism, secretion systems, and carbohydrate-active enzymes. Notably, genes encoding effector-related proteins, including *avrXca*, were identified in *S. phyllosphaerae* (bin08, bin10) and *S. adhaesiva* (bin09, bin28), suggesting potential immunomodulatory interactions with the host plant. Together, these findings highlight the genomic and functional diversity of *Sphingomonas* inhabiting the cupuassu phyllosphere and provide an evolutionary and ecological framework to explore their roles in tropical plant–microbe symbioses and host disease resilience.

✉ Alessandro M. Varani
alessandro.varani@unesp.br

✉ Leandro Marcio Moreira
lmorei@gmail.com

¹ Department of Biological Science, Institute of Exact and Biological Science, Federal University of Ouro Preto, R. Diogo de Vasconcelos, 122. Pilar, Ouro Preto 35400-000, Minas Gerais, Brazil

² Núcleo de Pesquisas em Ciências Biológicas, Universidade Federal de Ouro Preto, Ouro Preto 35400-000, MG, Brazil

³ Faculdade de Computação (FACOMP), Universidade Federal do Pará, Belém, PA, Brazil

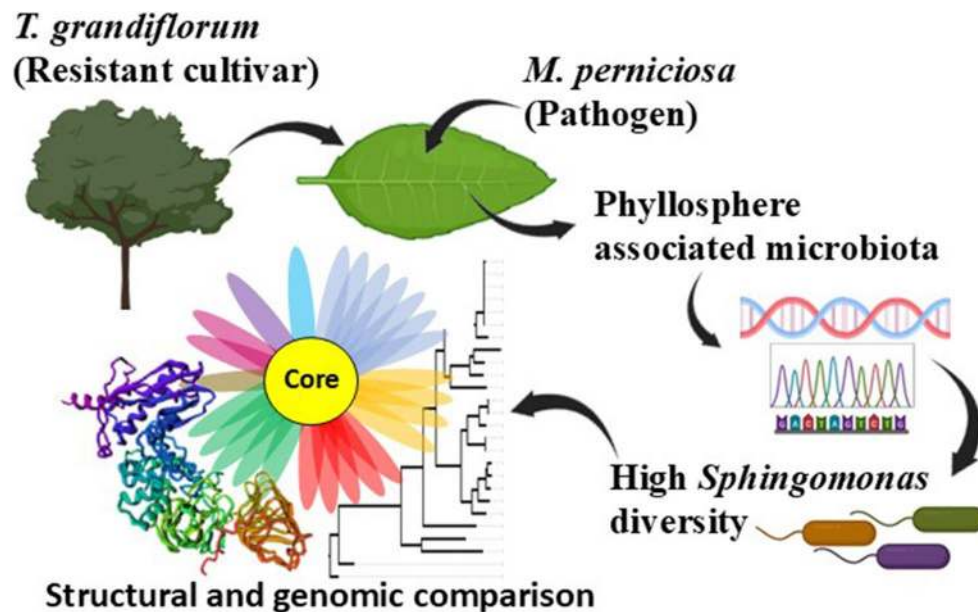
⁴ Embrapa Amazônia Oriental Belém, Belém, PA, Brazil

⁵ Department of Agricultural and Environmental Biotechnology, Faculdade de Ciências Agrárias e Veterinárias, São Paulo State University (UNESP) - College of Agricultural and Veterinary Sciences, Jaboticabal 14884-900, São Paulo, Brazil

⁶ Faculdade de Computação, Universidade Federal de Mato Grosso do Sul, Campo Grande, MS, Brazil

Graphical Abstract

Summary of the steps and methodologies used to analyze the diversity and potential of MAGs identified as belonging to the *Sphingomonas* genus



Keywords *Sphingomonas* · Phyllosphere microbiome · Comparative genomics · Host-associated adaptation · Tropical plant–microbe interactions · *Theobroma grandiflorum* (cupuassu)

1 Introduction

Cupuassu (*Theobroma grandiflorum*, Willd. ex Spreng.) Schum. is a tropical fruit tree native to the Brazilian Amazon, whose pulp is widely used in regional food production and represents an important socioeconomic resource for Amazonian communities (Pereira et al. 2018). However, the crop is severely affected by witches' broom disease (WBD), caused by the basidiomycete *Moniliophthora perniciososa*. Infection leads to hypertrophy, hyperplasia, and loss of apical dominance, resulting in substantial yield reduction and significant management challenges due to the pathogen's high dispersal potential (Meinhardt et al. 2008; Teixeira et al. 2014; Costa et al. 2021).

Recent genomic studies have provided new insights into *T. grandiflorum* biology and disease resistance (Mournet et al. 2020; Falcão et al. 2022; Alves et al. 2024; Abreu et al. 2025). Alves et al. (2024) identified resistant and susceptible clones (C174 and C1074, cultivar BRS Coari) and generated chromosome-scale genome assemblies to investigate the genetic basis of resistance and susceptibility to WBD (Falcão et al. 2022; Abreu et al. 2025). In that study, high-molecular-weight DNA was extracted from young leaf tissue of the resistant genotype C174 for long-read sequencing. Because leaf material inherently contains microbial

epiphytes and endophytes, DNA from the associated phyllosphere microbiota was co-purified and sequenced together with the host genome. Metagenomic assembly of these non-plant reads led to the recovery of 28 high-quality metagenome-assembled genomes (MAGs) (BioProject PRJNA922445). Among these, two groups were particularly abundant: 11 MAGs affiliated with the phylum *Actinomycetota* (Matos et al. 2024) and eight belonging to the genus *Sphingomonas*.

A previous comparative study focusing on *Actinomycetota* MAGs revealed gene clusters potentially involved in secondary metabolite biosynthesis and other functions that could be ecologically relevant to plant–microbe interactions (Matos et al. 2024). These findings suggested that certain phyllosphere bacteria might influence the ecological balance within the cupuassu–pathogen system (Matos et al. 2024). Yet, despite the well-recognized diversity and ecological significance of *Sphingomonas*, no genomic characterization of its representatives associated with cupuassu leaves has been performed to date.

Sphingomonas species are Gram-negative, strictly aerobic Alphaproteobacteria within the family Sphingomonadaceae, ubiquitously distributed across soil, water, and plant environments (White et al. 1996). Members of this genus are notable for their metabolic versatility, resilience to

environmental stress, and capacity to degrade a wide range of xenobiotics (White et al. 1996; Stolz et al. 2000; Basta et al. 2004). Importantly, *Sphingomonas* is among the most prevalent genera inhabiting the plant phyllosphere, where it may establish commensal or mutualistic associations with the host (Li et al. 2022; Lundberg et al. 2022; Peng et al. 2024). Therefore, it is plausible that *Sphingomonas* populations inhabiting cupuassu leaves have developed genomic and metabolic traits that facilitate adaptation to this tropical phyllosphere environment. Exploring these traits may provide insights into the ecological roles and evolutionary strategies of *Sphingomonas* in association with disease-resistant plants.

Building on the genomic resources previously generated for *T. grandiflorum*, we delved into these questions through a comparative genomic analysis of eight *Sphingomonas* MAGs associated with the phyllosphere of resistant cupuassu plants. Our goal was to characterize their taxonomic diversity, genomic composition, and functional potential, and to identify genetic traits that may underlie adaptation to the leaf environment. We specifically aimed to determine whether these genomes harbor metabolic or structural features that could be ecologically relevant to the host–microbiota relationship. Understanding these associations may provide insights into how the phyllosphere microbiota interacts with the host immune landscape, offering a broader ecological context for the resilience of *T. grandiflorum* to witches' broom disease and serving as a genomic framework for future functional studies on tropical phyllosphere symbionts.

2 Materials and methods

2.1 Plant material and DNA extraction

Fresh and young leaves of *Theobroma grandiflorum* (clone C174, cultivar BRS Coari) were collected from mature trees maintained at Embrapa Amazônia Oriental (Pará, Brazil). The samples were immediately frozen in liquid nitrogen and stored at -80°C in a super freezer until they were sent for sequencing (Arizona Genomics Institute - Tucson, AZ, USA). High-molecular-weight DNA was extracted from approximately 5 g of leaf tissue using a modified cetyltrimethylammonium bromide (CTAB) protocol optimized for long-read sequencing. DNA quality was assessed using a Qubit dsDNA High-Sensitivity assay (Thermo Fisher Scientific, USA) and a NanoDrop ND-1000 spectrophotometer. DNA integrity and fragment size distribution were verified using a Femto Pulse system (Agilent Technologies) and pulsed-field gel electrophoresis, while purity was confirmed by restriction enzyme digestion.

2.2 Metagenomic sequencing, assembly and binning

The extracted DNA was sheared and purified with PB Beads (PacBio, Menlo Park, CA, USA). SMRTbell libraries were prepared using the SMRTbell Express Template Prep Kit v2.0 and sequenced on a PacBio Sequel IIe platform (one SMRT Cell, CCS mode) to generate high-fidelity (HiFi) reads. To distinguish plant-derived reads from microbial sequences, raw HiFi reads were classified with Kraken2 (v2.1.5) (Wood et al. 2019) using the PlusPFP standard RefSeq index (version 20222009; <https://benlangmead.github.io/aws-indexes/k2>). Bacterial reads were extracted with KrakenTools v1.2.0 (<https://ccb.jhu.edu/software/krakentools/>). Assembly of the non-plant reads was performed with Hifiasm-meta (v0.19) (Feng et al. 2022), and metagenome-assembled genomes (MAGs) were binned using MetaBAT2 (v2.15) (Kang et al. 2019). MAG quality was evaluated with CheckM (v1.2.4) (Parks et al. 2015) and BUSCO (v6.0.0) (Simão et al. 2015) using the alphaproteobacteria_odb12 and sphingomonas_odb12 lineages dataset.

2.3 Genome annotation and taxonomic assignment

MAG annotation was performed using the Prokaryotic Genome Annotation Pipeline (PGAP, NCBI) (Tatusova et al. 2016), and taxonomic classification was cross-validated using PGAP and Kraken2 with the PlusPFP RefSeq index, and the Genome Taxonomy Database Toolkit (GTDB-Tk (v2.5.2), dataset release r226) (Chaumeil et al. 2022). GTDB-Tk classification was conducted using both the classify_wf and de_novo_wf workflows, which employ 120 single-copy bacterial marker genes (bac120 set) for placement of genomes in the GTDB reference tree and reconstruction of a de novo phylogeny including all *Sphingomonas* MAGs and reference genomes. Average nucleotide identity (ANI) and relative evolutionary divergence (RED) values were obtained from GTDB-Tk outputs and used to delimit species according to established thresholds ($\text{ANI} \geq 95\%$, $\text{AF} \geq 0.65$). Only MAGs classified within the genus *Sphingomonas* were retained for comparative analysis. Pairwise ANI among genomes was calculated using two complementary approaches. First, ANI values reported in NCBI assembly metadata were retrieved when available. Second, pairwise ANI between genomes was calculated de novo using OrthoANIu (<http://www.ezbiocloud.net/tools/orthoaniu>) implemented in the EzBioCloud ANI Calculator (<https://www.ezbiocloud.net/tools/ani>). Species boundaries were interpreted according to the 95–96% ANI threshold, in agreement with the GTDB-Tk framework.

2.4 Comparative genomics of *Sphingomonas*

Genomes of *Sphingomonas* species closely related to the cupuaçu MAGs were retrieved from public repositories NCBI (<https://www.ncbi.nlm.nih.gov/home/genomes/>) and PATRIC (<https://www.bv-brc.org/>) (Gillespie et al. 2011), accessed in July 2023. Candidate reference genomes were first screened for phylogenetic proximity using GTDB-Tk (v2.5.2) (dataset release R226), and those clustering within or adjacent to the MAGs' taxonomic lineages were retained. To ensure data quality and consistency, only assemblies with completeness $\geq 95\%$ and contamination $\leq 5\%$, as assessed by CheckM (v1.2.4), were considered. For most species, all available high-quality genomes were included. In the case of *S. paucimobilis*, where numerous assemblies were available, six complete genomes representing the phylogenetic and ecological diversity of the species were selected (Table S1). In total, 15 *Sphingomonas* reference genomes and one outgroup — *Sphingobium herbicidovorans* strain MH (ATCC 700291/NBRC 16415) — were used in comparative analyses. The outgroup was selected because it belongs to the same family (Sphingomonadaceae) and has a comparable genome size (~4.1 Mb), allowing for balanced phylogenomic inference. Orthologous protein families among all genomes were identified using OrthologSorter (<https://git.facom.ufms.br/bioinfo/orthologsorter>), which employs BLASTp (Camacho et al. 2009) and OrthoMCL (Li et al. 2003) to cluster proteins into (i) unique singletons, (ii) shared families (flexible genome), and (iii) core orthologous groups.

2.5 Phylogenomics and genomic-distance analysis

Phylogenomic reconstruction was based on concatenated alignments of core orthologous proteins. Poorly aligned and divergent regions were trimmed using Gblocks (v0.91b) (Castresana 2000). A maximum-likelihood tree was inferred with RAxML (v8.2.12) (Stamatakis 2014) under the PROTCATJTT substitution model with 1,000 rapid bootstrap replicates. Intraspecies variation and genome-wide sequence conservation were further examined using MUMmer (Kurtz et al. 2004) and MUMi (Deloger et al. 2009) to provide complementary nucleotide-level distance metrics.

2.6 Functional annotation and pathway reconstruction

Functional annotation of metabolic pathways and cellular processes was performed using BlastKOALA v3.0 on the KEGG platform (Kanehisa et al. 2016; Kanehisa and Goto

2000). KEGG Orthology (KO) identifiers were assigned to annotated genes to reconstruct metabolic networks and infer pathway completeness across the genomes. For specialized functional comparisons, genes encoding carbohydrate-active enzymes were identified using the CAZy database (Cantarel et al. 2009) using dbCAN3 (Jinfang et al. 2023), with a focus on glycoside hydrolases, glycosyltransferases, polysaccharide lyases, carbohydrate esterases, carbohydrate-binding modules, and auxiliary activity enzymes. Defense and secretion mechanisms were also characterized: CRISPR arrays and Cas genes were identified and visualized using Proksee (Grant et al. 2023), which automatically detected CRISPR loci, predicted Cas gene composition, and generated genomic context maps for comparative analysis of CRISPR–Cas system types and organization among *Sphingomonas* MAGs. Additionally, presence–absence profiles of secretion system genes (types I–VI, Sec, and Tat) were compiled from the PGAP annotations and confirmed manually.

2.7 Protein structure prediction and superimposition analysis

The three-dimensional structures of beta-lactamase proteins E AvrXca were predicted using ROBETTA (Kim et al. 2004), which integrates homology modeling and ab initio approaches to generate reliable structural models. Structural comparisons were performed through superimpositions using FATCAT 2.0 (Li et al. 2020), a tool that identifies similarities and differences between protein structures by flexible and rigid alignment of .pdb files. Flagellin proteins were modeled with AlphaFold (Jumper et al., 2021), which applies deep learning algorithms to predict high-accuracy structures from amino acid sequences. All predicted structures were evaluated using the confidence scores provided by the respective tools, ensuring their suitability for subsequent comparative analyses.

2.8 Data availability

All *Sphingomonas* metagenome-assembled genomes and raw PacBio HiFi reads have been deposited in DDBJ/ENA/GenBank under BioProject accession PRJNA922445. Individual genome accessions are: JAQIJT0000000000 (Bin28), JAQIJJ0000000000 (Bin18), JAQIJB0000000000 (Bin09), JAQIJD0000000000 (Bin11), JAQIJL0000000000 (Bin20), JAQIJC0000000000 (Bin10), JAQIJA0000000000 (Bin08), and JAQIJG0000000000 (Bin15). OrthologSorter outputs and comparative genomic results are publicly available at <http://jau.facom.ufms.br/Sphingomonas>.

3 Results

3.1 Genomic features and taxonomic assignment of *Sphingomonas* MAGs recovered from the cupuassu phyllosphere

A total of 59,826 reads - individual short sequences of nucleotides - (3.71%) of the cupuassu leaf metagenome were assigned by Kraken2 to the order *Sphingomonadales*, of which 59,440 reads (3.69%) belonged to the family Sphingomonadaceae and 55,340 reads (3.43%) to the genus *Sphingomonas*. These reads were assembled into eight high-quality metagenome-assembled genomes (MAGs) affiliated with *Sphingomonas*.

The assembled MAGs exhibited mean read lengths ranging from 1 kb to 30.3 kb, assembly N50 values of 9.8 kb, and an average coverage depth of approximately 20×. All *Sphingomonas* MAGs showed CheckM completeness above 80% and contamination below 8%, with six MAGs classified as high-quality (completeness > 95%, contamination < 5%), meeting the Minimum Information about Metagenome-Assembled Genomes (MIMAG) criteria.

Taxonomic assignment using Kraken2, and PGAP identified four previously described *Sphingomonas* species: *S. adhaesiva* (Bin28, Bin09, and Bin18), *S. fennica* (Bin11), *S. phyllosphaerae* (Bin10 and Bin08), and *S. taxi* (Bin15 and Bin20). However, phylogenomic placement using GTDB-Tk, which integrates taxonomic topology and ANI, reclassified Bin15. It shares only 86.29% ANI with its closest type strain, *Sphingomonas liriopsis*, and has a RED value of 0.98, confirming it represents a novel species (Table S1).

3.2 Comparative genomics reveals extensive genome plasticity among *Sphingomonas* species

To validate the taxonomic classification, we conducted a comparative genomic analysis of *Sphingomonas* MAGs associated with cupuassu against a curated dataset of 17 publicly available *Sphingomonas* genomes (Table S1). Additionally, *Sphingobium herbicidovorans* strain MH was included as an outgroup for phylogenomic comparison.

A total of 623 orthologous protein families were shared across the 26 genomes analyzed (Fig. 1A). When *S. herbicidovorans* was excluded, the soft-core genome - genes present in the vast majority of strains - ($\geq 96.2\%$ presence) was reduced to only 10 orthologous protein families, underscoring the exceptionally high proportion of flexible genes in *Sphingomonas* genomes. The flexible genome (set of genes present in some, but not all, strains of a species) accounted for approximately 76–87% of all protein-coding sequences across the analyzed genomes, reflecting the metabolic and ecological diversity of this genus.

Further supporting this high degree of genomic plasticity, only a limited number of gene families were shared across all internal nodes of the phylogeny, except for the clade defining the *S. paucimobilis* group, which contained up to 210 protein families exclusive to this species (Fig. 1B).

The number of unique single-copy genes (singletons) varied markedly among the analyzed genomes, from 12 in *S. paucimobilis* strain FDAARGOS_908 to 810 in strain K101 and 801 in strain 20006FA (Fig. 1A). These two strains also showed the highest numbers of unique multi-copy genes (100 and 143, respectively), together representing nearly 20% of their total gene content. Among the cupuassu-associated *Sphingomonas* MAGs, the number of unique single-copy genes (singletons) was substantial, ranging from 100 in Bin28 to 305 in Bin11. The number of unique multi-copy gene families was also variable, with Bin11 again having the highest count (57).

3.3 Phylogenomic relationships and evidence for novel *Sphingomonas* lineages associated with the cupuassu phyllosphere

Phylogenomic analysis based on the concatenated core set of 623 orthologous proteins revealed a clear separation of most taxa into well-supported, species-specific clades (Fig. 1B). Bootstrap values were maximal across nearly all nodes, confirming the robustness of the inferred topology. Complementary MUMi distance analyses indicated substantial genomic dissimilarity among the examined genomes, except for *S. paucimobilis* strains and *S. adhaesiva* DSM 7418 and NBRC 15,099, which exhibited strong genomic conservation (Fig. 1C–D).

Despite this overall resolution, several genomes displayed branching patterns indicative of taxonomic inconsistencies. The *S. paucimobilis* clade was the only group consistently composed of genomes belonging to the same species, showing ANI > 99% relative to the reference strain FDAARGOS_908 (GCA_016027095.1). Within this clade, strain LK11 — originally isolated from *Tephrosia purpurea* leaves — clustered separately and showed 97.03% ANI to the *S. yabuuchiae* type strain (Fig. 1B; Table S1).

Strains previously deposited as *S. taxi* were distributed across distinct clades. The type strain ATCC 55,669, isolated from *Taxus canadensis*, has been reclassified as *S. kyungheensis* and shared only 87.25% ANI with its reference genome. Strains R3119 and R122, both recovered from hospital sinks, are now classified as *S. albertensis* (93.96% ANI) and *S. endophytica* (85.06% ANI), respectively.

Among the cupuassu-associated genomes, Bin 15 — initially annotated as *S. taxi* — displayed 86.29% ANI to *S. liriopsis* and a RED value of 0.97896 in GTDB-Tk, supporting its placement as a distinct and previously undescribed

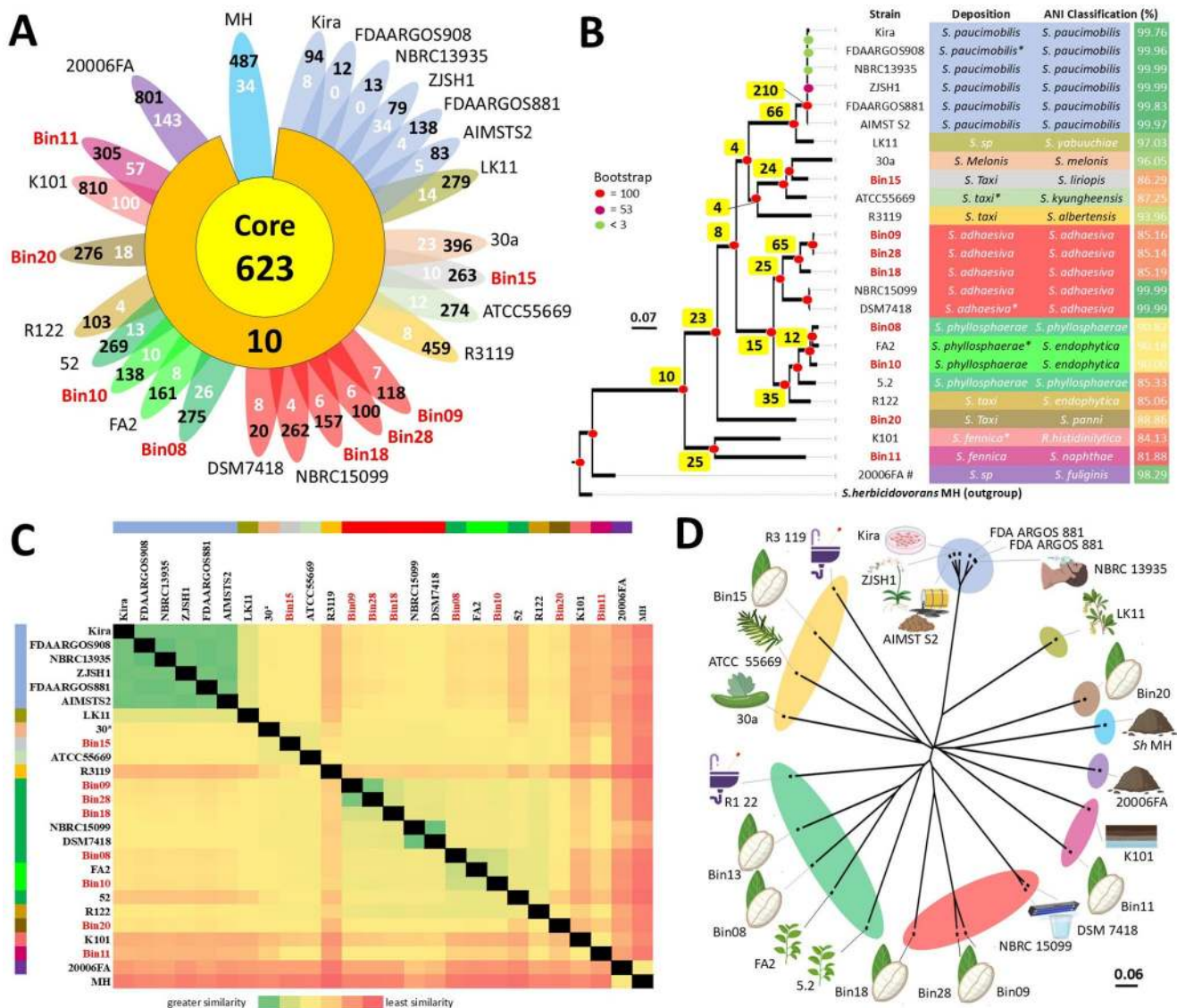


Fig. 1 Comparative genomics and phylogenomic relationships among *Spingomonas* strains. **(A)** Flower plot illustrating shared and unique gene families across 26 *Spingomonas* genomes. The central core represents the 623 orthologous protein families shared by all strains. Each petal corresponds to a single genome, with black numbers indicating unique single-copy genes (singletons) and white numbers representing unique multi-copy gene families. **(B)** Maximum-likelihood phylogenetic tree constructed from the core protein families. Node circles are color-coded by bootstrap support, and yellow highlights indicate

the number of orthologous families at each internal node. Taxon labels show the original and updated NCBI classifications, along with average nucleotide identity (ANI) values relative to reference strains. **(C)** Pairwise genome similarity matrix generated with MUMmer, showing sequence conservation patterns across the analyzed genomes. **(D)** Phylogenetic distribution derived from the MUMmer-based dissimilarity matrix. Background colors correspond to species assignments in panel B, and icons denote the isolation source of each strain (see Table S1)

lineage within *Spingomonas* (Table S1). Among strains originally annotated as *S. phyllosphaerae*, strain 5.2 remained within this species but exhibited only 85.33% ANI to its reference genome (GCA_000427645.1). Strain FA2 is now designated as *S. endophytica*, although it shared merely 90.18% ANI with its reference, highlighting misclassifications in public repositories.

Strain K101, originally deposited as *S. fennica*, has since been reassigned to *Edaphospingomonas fennica*

and displayed 84.13% ANI to *Rhizorhabdus histidinilytica* (GCA_042433025.1), consistent with its phylogenetic placement outside the main *Spingomonas* clade. Similarly, strain 20006FA — originally classified as *Spingomonas* sp. — formed a basal branch relative to all other *Spingomonas* genomes analyzed. This genome showed 98.29% ANI to *Spingobium fuliginis* (GCA_004152845.1) and had the largest genome in the dataset (5.41 Mb), nearly 1 Mb larger than K101. Notably, the NCBI RefSeq record for 20006FA

has been flagged for potential contamination, which may explain its atypical phylogenetic position.

Analysis of the eight *Sphingomonas* MAGs recovered from the cupuassu phyllosphere confirmed that all displayed ANI values below the established 95–96% species-level threshold when compared with their closest reference strains, indicating the presence of divergent or novel lineages (Fig. 1B; Table S1). Bins 09, 18 and 28 — assigned to *S. adhaesiva* by NCBI — showed ANI < 86% relative to the reference genome and clustered together in an isolated branch adjacent to the *S. adhaesiva* group, suggesting a deeply diverged lineage. Bins 08 and 10, classified as *S. phyllosphaerae* and *S. endophytica*, respectively, exhibited ANI ≤ 91% to their references, supporting their designation as distinct variants within the *phyllosphaerae*–*endophytica* complex.

Bin 11, previously identified as *S. fennica* and now listed as *S. naphthae* in NCBI, showed only 81.88% ANI to its reference genome. Although it clustered near K101, the ANI between them did not exceed 76%, indicating distinct genera—consistent with the smaller genome size of Bin 11 (~1.4 Mb less than K101). In contrast, Bin 15 formed a highly supported, topologically distinct branch separate from all other *Sphingomonas* lineages. Based on its phylogenomic position, genomic distinctiveness (ANI = 86.29%), RED value (0.97896), and unique ecological origin (cupuassu phyllosphere), we propose Bin 15 as a novel species, here designated Candidatus *Sphingomonas cupuassunensis*.

3.4 Functional diversity and metabolic specialization among *Sphingomonas* genomes

Based on KEGG functional annotation, a comparative analysis of metabolic pathways and cellular processes was performed across all *Sphingomonas* genomes (Fig. 2A). Among the cupuassu-associated MAGs, Bin 15 - Candidatus *Sphingomonas cupuassunensis* - and Bin 11 exhibited the lowest representation of annotated genes across most functional categories, consistent with their comparatively smaller genome sizes. In contrast, strain 20006FA displayed the highest number of annotated genes in nearly all categories, in line with its larger genome size relative to the remaining *Sphingomonas* strains analyzed.

To explore these trends further, four functional categories were examined in greater detail to evaluate how metabolic variability might influence strain-specific capacities (Fig. 2B). Notably, the category “xenobiotics biodegradation and metabolism” showed pronounced variability in gene composition across genomes. Interestingly, although Bin 15 and Bin 11 had the lowest total gene counts overall, they were not the strains with the fewest genes within this

specific category. This pattern underscores substantial metabolic heterogeneity related to xenobiotic processing, possibly reflecting adaptations to distinct ecological conditions from which each strain originated.

Closer examination of the subcategories within this functional class revealed extensive enzymatic diversity and modular pathway architecture among *Sphingomonas* genomes (Fig. 3A–B). For instance, strain NBRC15099 was the only genome encoding the multicomponent enzyme complex *dmpKLMNOP*, responsible for phenol-to-catechol conversion, while strain 20006FA uniquely carried genes encoding enzymes involved in the transformation of 2-aminobenzoate to catechol (*andABCD* and *antABC* operons).

This metabolic divergence was also evident among the cupuassu-associated MAGs. Specifically, Bin 11 and Bin 20 encoded enzymes mediating the conversion of 3,4-dihydroxybenzoate to oxaloacetate and pyruvate via the *ligABCIJKgalBD* operon, whereas Bins 15, 09, 28, and 18 appeared to metabolize the same substrate through an alternative pathway leading to succinyl-CoA and acetyl-CoA, catalyzed by the *pobABCGH* enzymes (Fig. 3A–B).

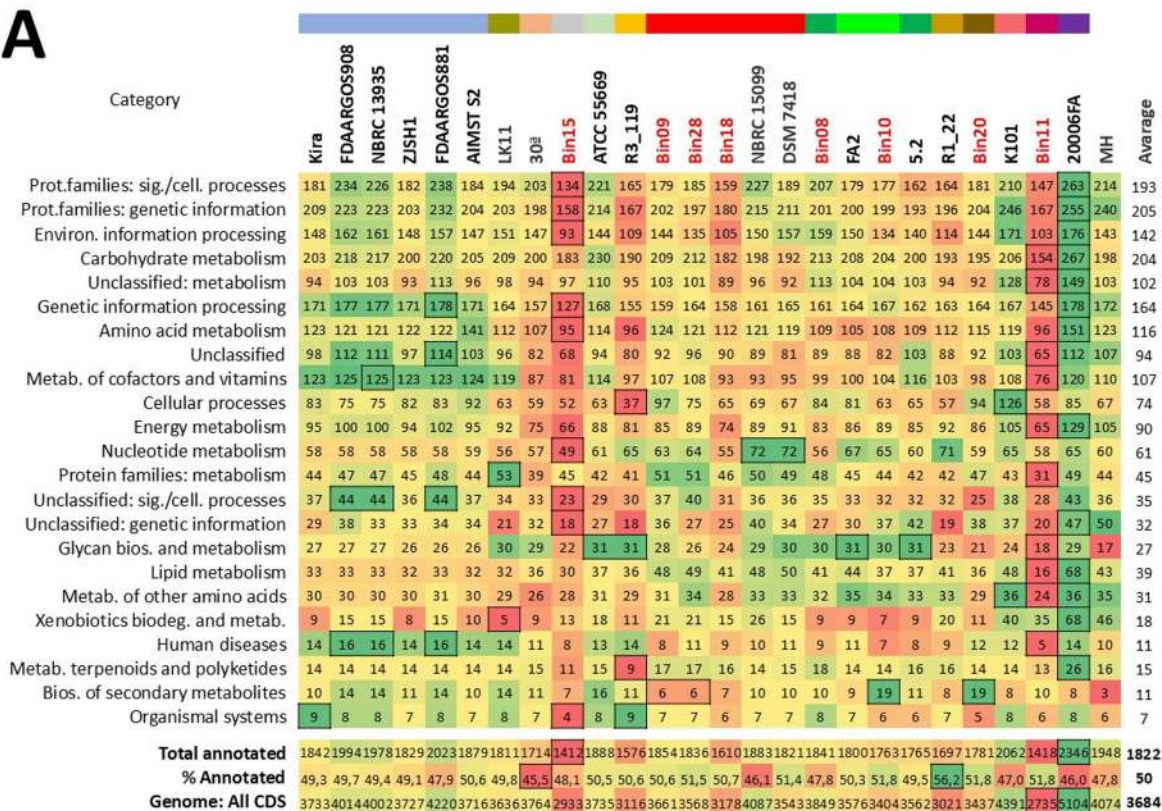
3.5 Nitrogen metabolism reveals limited denitrification capacity and lineage-specific gene losses

Nitrogen metabolism exhibited substantial divergence among the analyzed *Sphingomonas* genomes, despite their taxonomic proximity (Fig. 4A–B). Among the cupuassu-associated MAGs, Bins 15, 09, 28, and 11 lacked genes encoding the nitrite/nitrate transporter (*nrt2*) as well as key enzymes involved in the conversion of nitrite to nitrate (*nasB*, *nasAB*, *narGHI*) and in nitrite reduction to ammonia (*NIT-6*). The absence of these genes suggests a restricted capacity for assimilatory and dissimilatory nitrate reduction within these lineages.

In contrast, only strains K101 and R122 possessed the *hcp* gene, encoding hydroxylamine reductase, the enzyme responsible for the conversion of hydroxylamine to ammonia. Both genomes also carried *nirS*, indicating a potential ability to reduce nitrite to nitric oxide. However, only K101 harbored *norB* and *norC*, which mediate the conversion of nitric oxide (NO) to nitrous oxide (N₂O), suggesting that complete partial denitrification may be limited to this strain.

None of the *Sphingomonas* genomes examined contained gene sets related to the oxidation of ammonia to nitrite, the conversion of ammonia to molecular nitrogen (N₂), or the subsequent reduction of nitric oxide to diazene (N₂H₂) and nitrogen gas. Collectively, these results indicate that nitrogen metabolism within *Sphingomonas* is highly variable and largely incomplete, with most genomes retaining only partial or truncated nitrogen transformation modules.

A



B

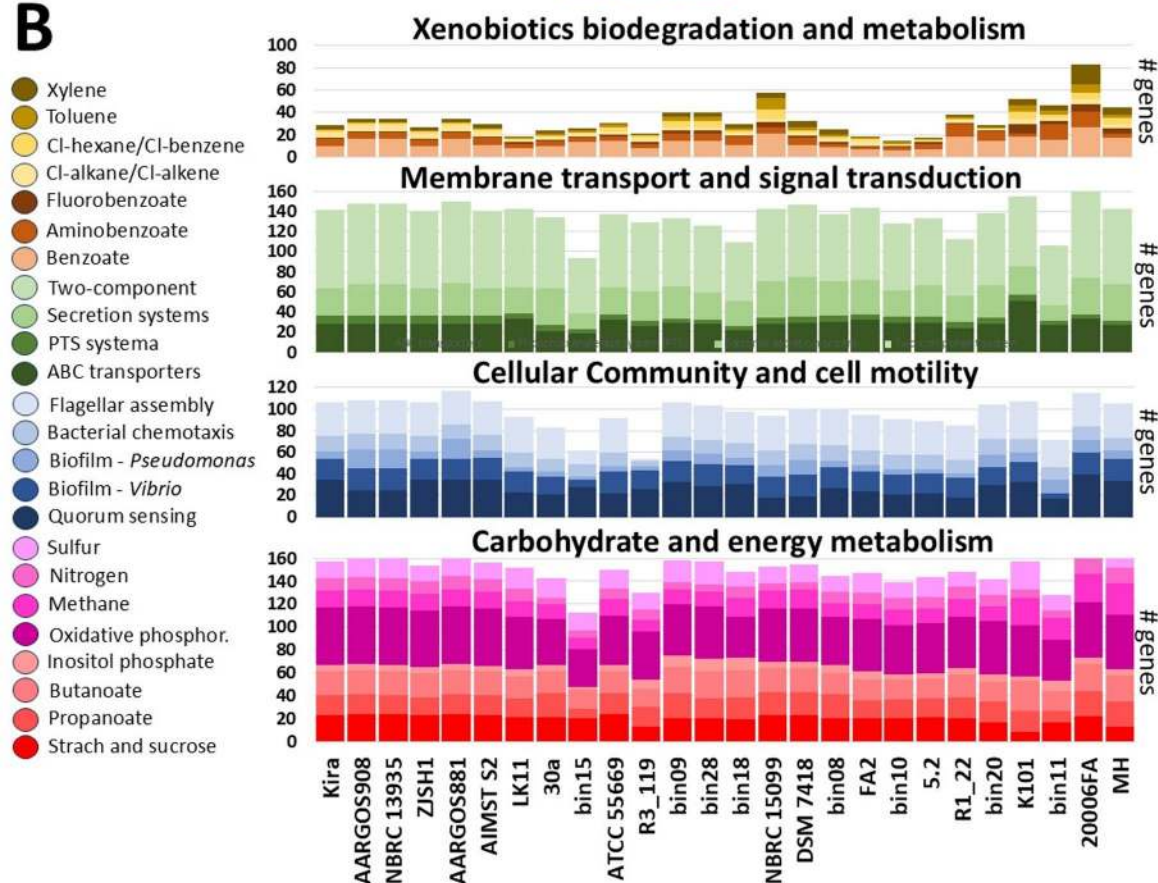


Fig. 2 Comparative analysis of global metabolic profiles across *Sphingomonas* genomes. **(A)** Heatmap summarizing metabolic categories inferred from genome re-annotation using BlastKoala. The last three rows display the total number of genes assigned to each category. Cells outlined in black denote the lowest (red background) and highest (green background) gene counts within each functional category, illustrating the marked functional variability among strains. **(B)** Gene distribution across four representative metabolic pathways, highlighting extensive heterogeneity in genes related to xenobiotic degradation and metabolism

3.6 Carbohydrate metabolism and antibiotic resistance modules reveal extensive functional diversification

Marked compositional variability was observed in genes related to complex carbohydrate metabolism, particularly those encoding glycoside hydrolases (GHs) (Fig. 5A). Strains isolated from hospital environments (R3119 and R122) harbored the lowest numbers of GH genes (≤ 17), whereas *S. paucimobilis* and *S. phyllosphaerae* genomes encoded approximately 100 GHs each. Notably, the cupuassu-associated MAG Bin 08 contained 105 GH genes, of which 58 were predicted to carry N-terminal signal peptides, indicating potential secretion of their encoded enzymes into the extracellular environment.

A similar pattern of compositional variability was detected among genes associated with antibiotic biosynthesis and degradation (Fig. 5B–C). Unique to several cupuassu-associated genomes were two β -lactamase genes (families 15 and 16) and two monooxygenase genes (families 9 and 13), the latter potentially linked to aromatic compound metabolism or antibiotic-modifying reactions. Within the β -lactamase group, family 15, present exclusively in Bins 09 and 28, showed high protein sequence identity (99.75%) and conserved genomic synteny, reflecting the overall genomic similarity between these two strains. In contrast, family 16 β -lactamases, found only in Bins 18 and 20, exhibited lower protein identity (58.72%) and substantial divergence in flanking gene organization, consistent with the phylogenetic distance separating these genomes (Fig. 5D).

Two notable observations emerged from this analysis. First, TonB-like protein-encoding genes flanked both β -lactamase loci. Although TonB itself is not directly involved in antibiotic degradation, its established role in siderophore-mediated transport raises the possibility of indirect interactions with siderophore-mimicking β -lactam antibiotics such as cefiderocol, which utilize TonB-dependent receptors for cellular entry. Second, despite pronounced sequence divergence, predicted tertiary structures of the β -lactamases showed remarkable structural conservation with orthologs from other taxa — namely 8RLK (class C β -lactamase from *Pseudomonas aeruginosa*) and 7U1C

(β -lactamase from *Caulobacter vibrioides*) — as well as among each other, supporting potential functional conservation despite evolutionary divergence (Fig. 5D).

3.7 Secretion systems and CRISPR-Cas architecture reveal functional differentiation among *Sphingomonas* lineages

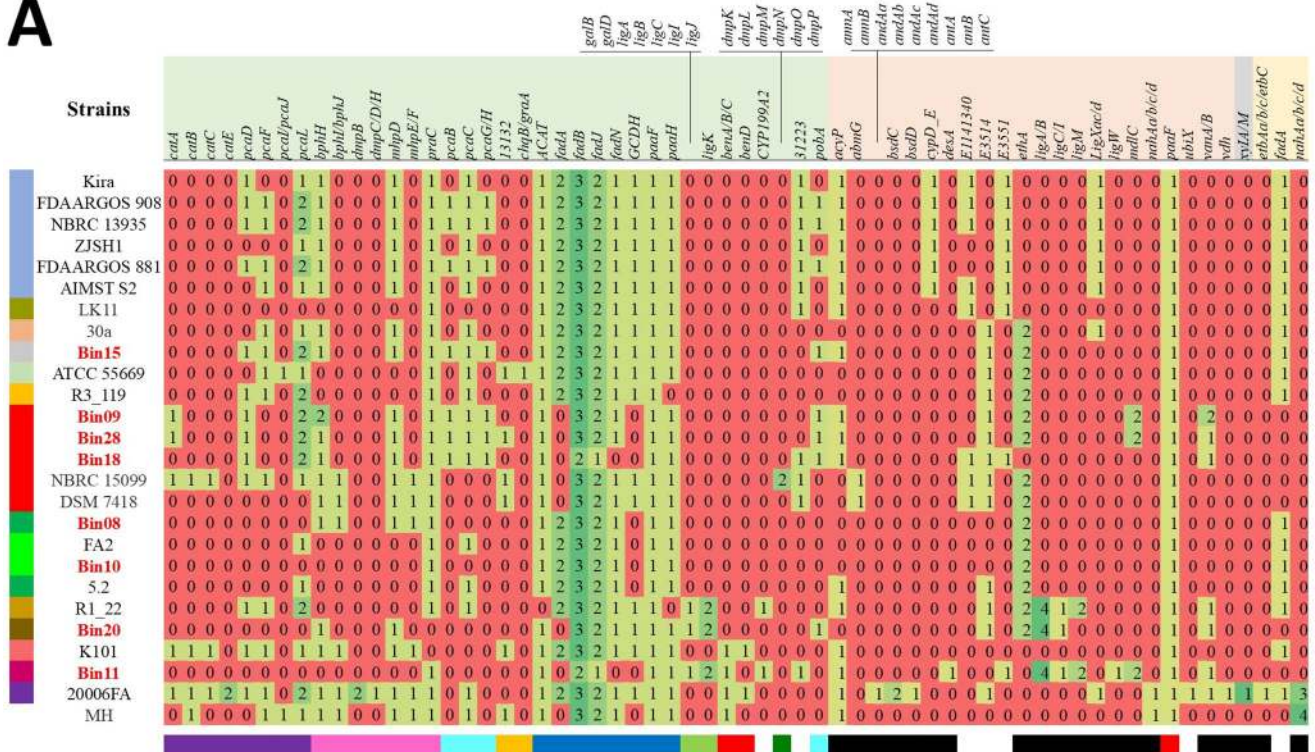
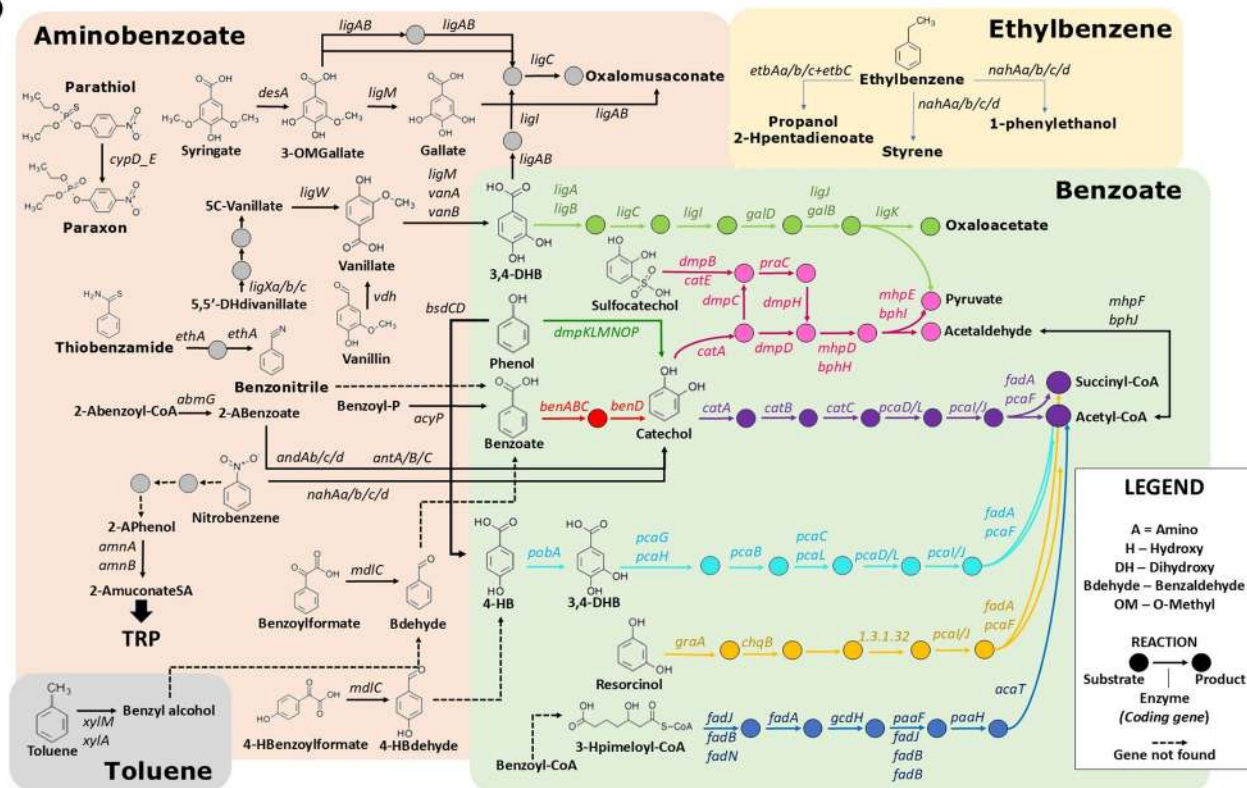
Comparative analysis of the secretion machinery revealed consistent differences among *Sphingomonas* genomes (Fig. 6A). Genes encoding the Sec and Tat systems were universally conserved across all strains, as expected for the genus. In contrast, Type V and Type VII secretion systems were absent in all genomes examined.

Type I secretion system (T1SS) genes were detected only in R3119, Bin 15, and Bin 20, all of which are associated with cupuassu. The Type II system (T2SS) was missing only in Bin 15 and Bin 11, whereas Type IV secretion system (T4SS) genes were identified in 30a, Bin 09, 5.2, 20006FA, and MH. The Type VI system (T6SS) was present exclusively in FDAARGOS, NBRC15099, DSM7418, Bin 20, and Bin 11, while Type III secretion system (T3SS) components were observed in four genomes — NBRC15099, DSM7418, FA2, and the cupuassu-associated Bin 08 (Fig. 6B).

Flagellar assembly genes were detected in all genomes except R3119, which also exhibited a markedly reduced repertoire of chemotaxis-related genes (only eight). This reduction is consistent with its overall smaller complement of genes associated with cellular motility and signaling, supporting the loss of structural flagellar components in this strain.

Regarding defense systems against exogenous DNA, three distinct CRISPR-Cas clusters were identified across the dataset (Fig. 6C): (a) The first cluster (Type II) was found exclusively in *S. paucimobilis*, Bin 15 - Candidatus *Sphingomonas cupuassunensis*; R122, K101, and Bin 11; (b) The second cluster (Type I) occurred only in *S. adhaesiva* genomes (NBRC15099 and DSM7418); (c) The third cluster, also Type I, was uniquely detected in cupuassu-associated MAGs (Bin 09, Bin 28, Bin 18, and Bin 20).

Analysis of the composition and synteny of this third CRISPR locus revealed that Cas genes are located downstream of the repeat region and adjacent to a set of dsm (Defense-Related Modules) genes, which are frequently implicated in bacterial defense against mobile genetic elements (Fig. 6D). Between Bin 09 and Bin 28, the main structural difference was the fragmentation of the CRISPR gene in Bin 28, caused by the insertion of a transposable element. This event reduced the number of CRISPR repeats and spacers from 51 in Bin 09 to only 20 in Bin 28.

A**B**

Bin 20 exhibited overall synteny similar to Bin 09, with-out CRISPR gene fragmentation, but with a truncated *dsmM* gene. Although the Cas gene arrangement was conserved,

Bin 20 retained a smaller number of spacers (28) and pre-served *dsmA* and *dsmD*, but lacked the helicase gene typically associated with the *dsm* defense module. Collectively,

Fig. 3 Diversity of phenolic compound degradation pathways in *Sphingomonas* genomes. **(A)** Gene presence–absence matrix illustrating the distribution of enzymes involved in phenolic compound degradation across the analyzed genomes. Numbers indicate gene copy counts. The heatmap, based on gene abundance, enables visual comparison among strains. The colored bar in the last row corresponds to the specific metabolic pathway detailed in panel B. **(B)** Integrated metabolic map summarizing phenolic compound degradation pathways derived from the matrix in panel A. Arrows and metabolite nodes are color-coded according to the associated pathways: toluene (purple), aminobenzoate (salmon), ethylbenzene (orange), and benzoate (green). Within benzoate metabolism, the subpathways include benzoate (red), catechol (purple, meta-cleavage; pink, ortho-cleavage), phenol (dark green), 3,4-dihydroxybenzoate—3,4-DHB—(light green, meta-cleavage; aquamarine, ortho-cleavage), resorcinol (orange), and 3-hydroxypimeloyl-CoA (blue). Enzymes encoded by the corresponding genes are annotated above or beside each reaction step

these findings indicate that while CRISPR-Cas systems are broadly conserved across *Sphingomonas*, their architecture and association with auxiliary defense genes display substantial lineage-specific variation, reflecting differential evolutionary pressures across environmental and host-associated contexts.

3.8 PTI-modulating effector analysis reveals conserved AvrXca homologs in cupuassu-associated *Sphingomonas*

Because bacterial effector proteins secreted through specialized secretion systems can modulate or attenuate plant immune responses, we examined the genomes of all *Sphingomonas* strains for genes encoding such effectors. Only one effector candidate was identified: the *avrXca* gene, detected exclusively in four of the eight cupuassu-associated MAGs (Bin 08, Bin 09, Bin 10, and Bin 28), with Bin 08 harboring two non-identical copies (Fig. 7A). A vestigial copy of *avrXca* was also found in R122, suggesting ancestral retention or partial pseudogenization.

The two *avrXca* copies in Bin 08 shared 46% amino acid identity and 80% similarity with each other, while the copies in the other strains were nearly identical (up to 99% identity). When compared against 100 sequences retrieved from NCBI, *avrXca* homologs displayed 54–76% identity and 81–90% coverage, confirming their divergence from previously characterized effectors.

Predicted protein structure analyses revealed three major domains: (i) an N-terminal lipoprotein signal peptide (SEC/SPII, confidence score 0.9999; residues 1–25), (ii) a central peptidase domain, and (iii) a C-terminal acetyltransferase domain (Fig. 7B). Structural prediction of the Bin 09 AvrXca sequence confirmed this architecture, with increased uncertainty localized to the N-terminal region—likely reflecting a signal peptide cleaved prior to secretion.

To assess evolutionary relationships, BLASTp searches were performed using the Bin 09 AvrXca sequence as

query, and the top 100 homologous proteins were used to reconstruct a phylogenetic profile (Fig. 7C). All close homologs were found in plant-associated bacteria, including the phytopathogenic genera *Dickeya* and *Xanthomonas* (Gammaproteobacteria) as well as the symbiotic Alphaproteobacteria *Asticcacaulis*, *Sphingomonas*, *Sphingopyxis*, *Novosphingobium*, and *Sphingobium*. These groups formed well-supported clades, suggesting a conserved functional role across plant-associated taxa.

Comparative structural modeling of AvrXca from Bin 09 and representative orthologs from each genus revealed high conservation of tertiary structure, despite considerable sequence divergence. The catalytic motifs and domain organization were preserved, and structural superposition metrics indicated strong three-dimensional similarity to known effector homologs (Fig. 7D). Collectively, these findings highlight the presence of conserved PTI-modulating effector candidates within specific *Sphingomonas* lineages associated with cupuassu, warranting further experimental validation.

3.9 Flagellar structure and domain diversification across *Sphingomonas* lineages

The bacterial flagellum is a filamentous organelle primarily responsible for motility, but it also plays an important role in host interactions by eliciting pattern-triggered immunity (PTI) in both animals and plants. Given its dual structural and immunological significance, we conducted a detailed examination of the genes and proteins associated with flagellar assembly, with particular emphasis on flagellin, the principal structural subunit of the filament.

All flagellin-encoding sequences identified in the analyzed genomes were aligned and compared to infer their evolutionary relationships. The resulting phylogenetic tree resolved the sequences into four major groups (Fig. 8A). Group I, comprised the two flagellin-like copies found exclusively in strain K101, which were entirely divergent from all other sequences; Group II, formed a clade restricted to the seven *S. paucimobilis* strains, with the LK11 copy representing the most divergent branch—mirroring the phylogenetic pattern observed in the core-genome tree; Group III, the largest and most heterogeneous cluster, contained 34 sequences in total, including two independent subclades formed by *S. paucimobilis* flagellins. A similar arrangement was observed for the cupuassu-associated strains (Bins 18, 09, and 28), which grouped closely with copies from *S. adhaesiva* (NBRC15099 and DSM7418), and Group IV, including copies from Bin 08, Bin 10, 5.2, NBRC15099, FA2, and R122. Although the flagellin-like sequence from Bin 11 associated with this clade, it showed pronounced divergence, as did the copy from Bin 20, which occupied an isolated position between Groups II and III.

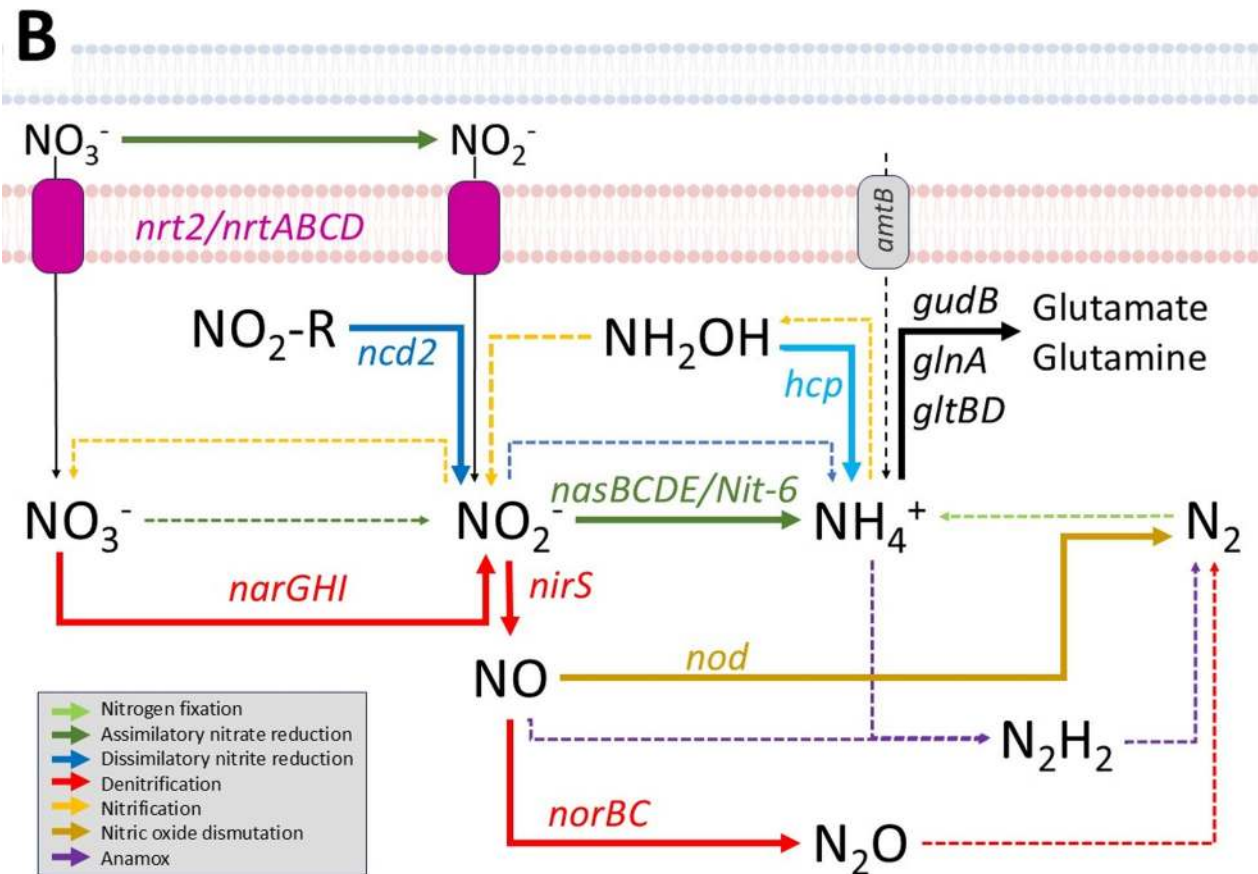
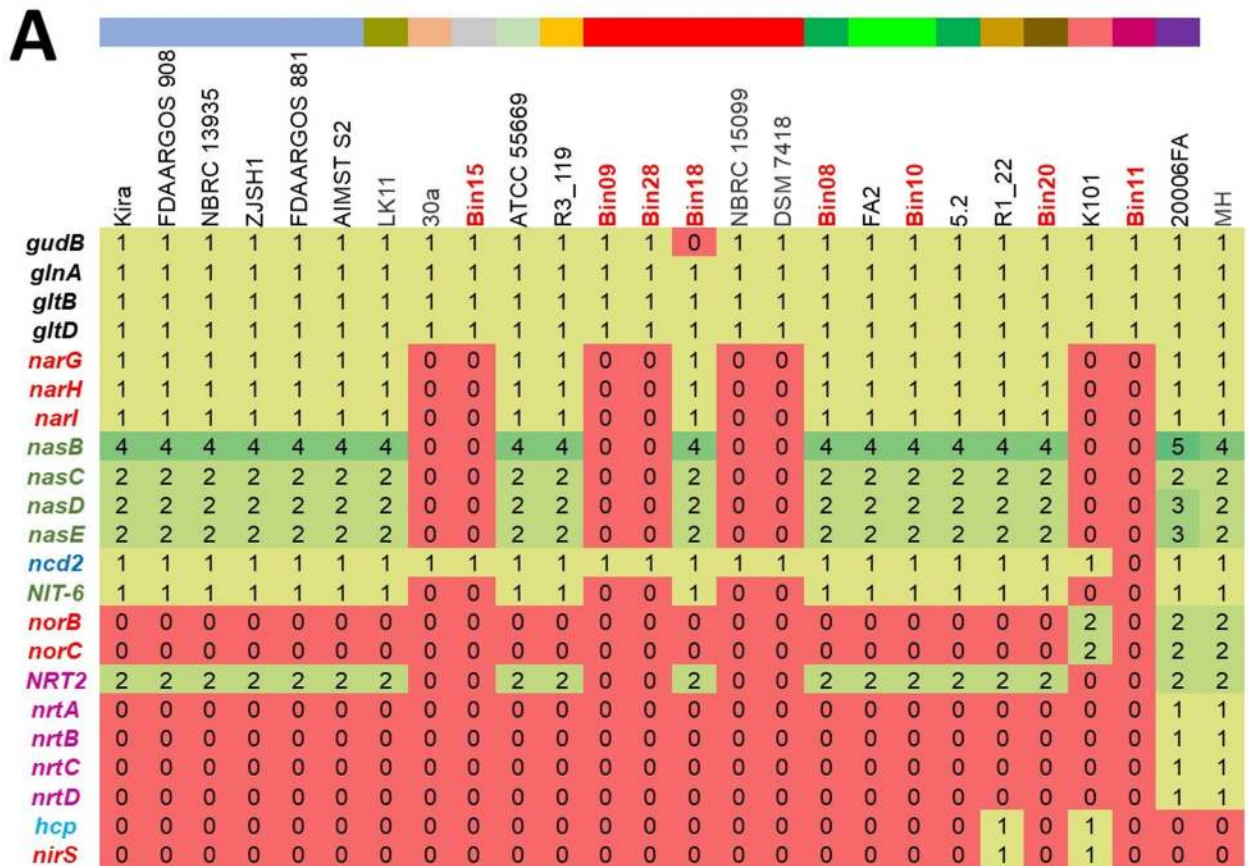


Fig. 4 Comparative analysis of nitrogen metabolism across *Sphingomonas* genomes. (A) Presence–absence matrix of genes involved in nitrogen metabolism across the analyzed genomes. Numbers indicate gene copy counts, and the heatmap facilitates comparison among strains based on functional gene abundance. The color of each gene name in the first column corresponds to the metabolic pathway shown in panel B. (B) Integrated nitrogen metabolism map derived from the gene matrix in panel A. Arrows and gene names are color-coded according to their corresponding pathways. Dashed arrows denote reactions for which no homologous genes were detected in any *Sphingomonas* genome. Abbreviations: NO_3^- – nitrate; NO_2^- – nitrite; NH_4^+ – ammonium; N_2 – nitrogen gas; NO – nitric oxide; N_2O – nitrous oxide; N_2H_2 – diazene; NH_2OH – hydroxylamine; $\text{NO}_2\text{-R}$ – organic nitrite

The observed phylogenetic divergence was consistent with marked compositional hypervariability within the structural domains of flagellin (Fig. 8B). The N- and C-terminal regions — corresponding to the D0 and D1 domains — were highly conserved, whereas the central D2 and D3 domains exhibited extensive variability. These patterns are functionally consistent with their structural roles: D0 and D1 form the inner filament core, essential for polymerization and stability, while D2 and D3 are surface-exposed and mediate environmental interactions, including host recognition and adhesion (Fig. 8C).

To further investigate this variability, we performed secondary and tertiary structure topology analyses of the flagellin proteins (Fig. 8D). Structural modeling revealed that alterations in secondary structural elements directly contribute to the hypervariable topology of the D2 and D3 domains, particularly in flagellins encoded by the cupuassu-associated MAGs. These include the divergent copies from Bin 20 and one of the paralogs in Bin 10 (Bin10_c2-1), which display distinct surface-exposed conformations that may reflect adaptation to the phyllosphere environment.

4 Discussion

The chromosome-scale genome sequencing of the resistant clone 174 (cultivar BRS Coari) of *Theobroma grandiflorum* represents a major milestone for understanding the genomic architecture and host–microbe interactions in this tropical species (Abreu et al. 2025). Beyond enabling studies on genome evolution, domestication, and defense-related gene repertoires, it also provides a unique opportunity to investigate the composition and potential functional roles of its phyllosphere microbiota.

In a recent study, we demonstrated that several *Actinomyces* strains associated with the phyllosphere of this resistant cultivar harbor biosynthetic gene clusters potentially involved in the production of bioactive metabolites, suggesting that certain bacterial members of the cupuassu microbiome may contribute to pathogen suppression,

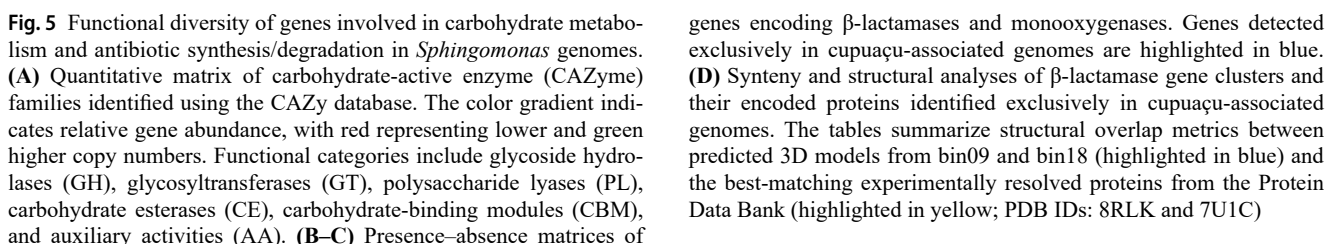
particularly against *Moniliophthora perniciosa*, the causal agent of the WBD (Matos et al. 2024). Building upon this framework, the present study focused on the *Sphingomonas* lineages recovered from the same genomic dataset. These MAGs, co-recovered during high-molecular-weight DNA sequencing of leaf tissue, represent 28.5% of all reconstructed MAGs and constitute the most abundant bacterial genus identified in the cupuassu phyllosphere.

Members of the genus *Sphingomonas* are widely recognized for their ecological versatility and symbiotic relationships with plants. Across different hosts, they have been implicated in pesticide degradation in rice (Farhan et al. 2021), drought tolerance in soybean (Asaf et al. 2017), induced systemic resistance in *Brassica* and strawberry (Olimi et al. 2022; Mazoyon et al. 2023), and growth promotion in tomato and spinach (Khan et al. 2014; Sultana et al. 2024). Despite this growing body of evidence, the present work provides the first genomic-level characterization of *Sphingomonas* strains associated with cupuassu and raises the possibility that some lineages may play a role in shaping host defense dynamics.

Eight *Sphingomonas* MAGs were identified and analyzed, revealing extensive genomic heterogeneity. Genome sizes ranged from 3.0 to 4.3 Mbp, reflecting contrasting ecological strategies — smaller genomes (e.g., Bins 11, 15 - Candidatus *Sphingomonas cupuassunensis*; and 18) suggest specialization and possible symbiotic dependence, whereas larger genomes (e.g., Bin 08) indicate increased adaptive potential through accessory gene content and mobile elements (Wang et al. 2024). Comparative analyses further demonstrated that these MAGs share less than 91% ANI with any known *Sphingomonas* reference genome, suggesting that several represent previously undescribed species (Konstantinidis and Tiedje 2005; Konstantinidis et al. 2006; Richter and Rosselló-Móra 2009; Parks et al. 2018). This emphasizes the value of genome-resolved metagenomics for uncovering novel microbial diversity in non-model tropical systems (Brown et al. 2015; Hug et al. 2016; Parks et al. 2017).

Functional analyses revealed broad metabolic diversity across these genomes, consistent with the well-documented biochemical plasticity of the genus (White et al. 1996). The most variable pathways included those involved in aromatic compound degradation (Stolz 2009), such as benzoate, catechol, phenol, and 3,4-dihydroxybenzoate catabolism. These substrates are often produced by plants as part of chemical defense responses, and their bacterial catabolism may facilitate persistence within the phyllosphere environment by mitigating potentially toxic secondary metabolites (Delmotte et al. 2009).

Nitrogen metabolism pathways were also heterogeneously distributed among strains. Bins 18, 08, 10, and 20



genes encoding β -lactamases and monooxygenases. Genes detected exclusively in cupuaçu-associated genomes are highlighted in blue. **(D)** Synteny and structural analyses of β -lactamase gene clusters and their encoded proteins identified exclusively in cupuaçu-associated genomes. The tables summarize structural overlap metrics between predicted 3D models from bin09 and bin18 (highlighted in blue) and the best-matching experimentally resolved proteins from the Protein Data Bank (highlighted in yellow; PDB IDs: 8RLK and 7U1C)

Similarly, carbohydrate metabolism genes, particularly those encoding glycoside hydrolases, showed large variation. Bins 08 and 10 were enriched in glycosyl hydrolase families involved in polysaccharide modification and degradation — an attribute likely beneficial in nutrient-scarce leaf environments — while Bin 11 encoded fourfold fewer such genes. These differences mirror patterns reported for

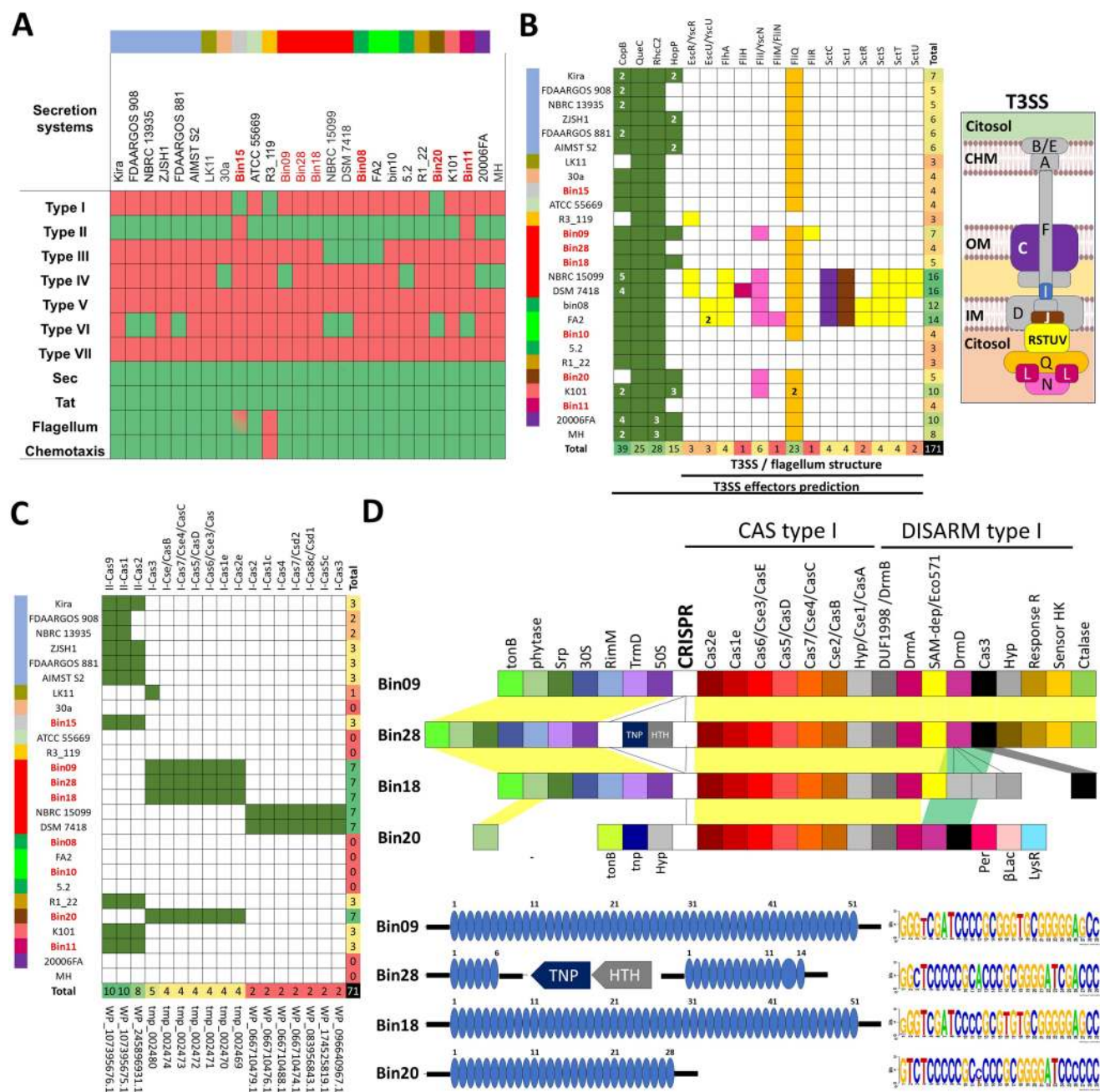


Fig. 6 Genomic landscape of secretion systems and defense mechanisms against invasive genetic elements in *Spingomonas*. **(A)** Presence (green) and absence (red) matrix of genes associated with secretion systems, flagellar biosynthesis, and regulatory components across the analyzed *Spingomonas* strains. Cells containing both colors indicate partial systems, where only subsets of the corresponding genes were detected. **(B)** Presence-absence profile of genes encoding components of the type III secretion system (T3SS, Sct apparatus). The color of each matrix cell corresponds to the specific protein subunit depicted in the schematic representation, while numbers within the cells indicate the total number of paralogous copies identified per genome. Owing

to their evolutionary relationship, several genes identified here may function in both the T3SS and flagellar systems. **(C)** Distribution of CRISPR–Cas–associated genes across *Spingomonas* genomes. Three distinct CRISPR–Cas system types were detected, occurring predominantly in strains associated with cupuassu. **(D)** Synteny and structural organization of type I CRISPR–Cas gene clusters in genomes bin09, bin28, bin18, and bin20. Conserved syntenic blocks are shown in yellow, and lines connect fragmented loci across genomes. Below, the CRISPR repeat composition and consensus sequence are depicted for each genome

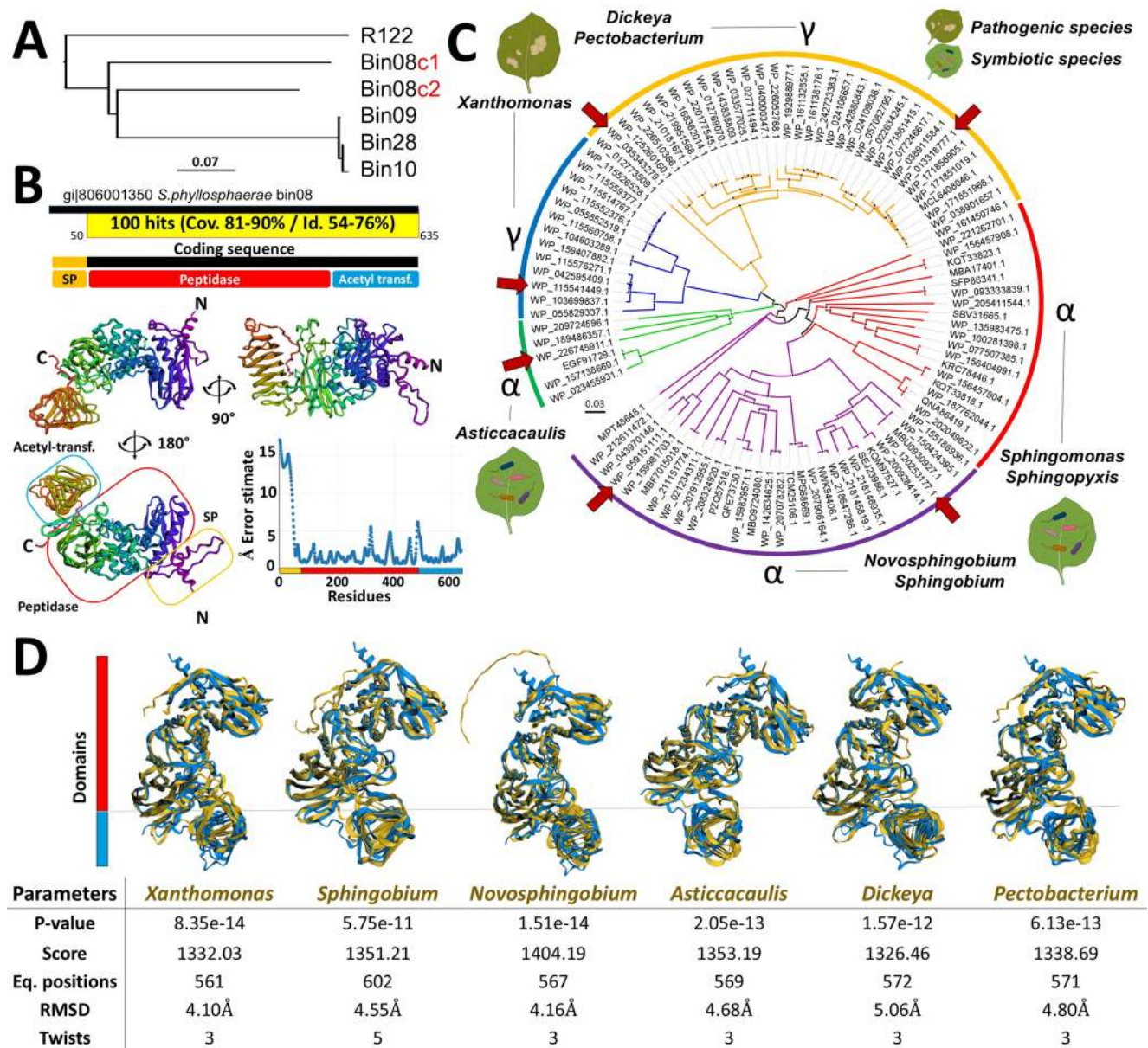


Fig. 7 Comparative and structural characterization of the AvrXca gene and protein identified in *Sphingomonas* genomes associated with cupuassu. **(A)** Phylogenetic tree derived from the multiple sequence alignment of AvrXca protein sequences. A hypothetical gene from *S. taxi* strain R122, showing partial sequence identity to AvrXca, was used as an outgroup. **(B)** Summary of BLASTn results for *avrXca* genes identified in *Sphingomonas*. Cov. – coverage; Id. – identity. Below, domain prediction based on UniProt is shown: SP – signal peptide, Acetyltransferase domain (blue outline), and Peptidase domain (orange outline). The yellow region indicates the signal peptide. The adjacent graph displays the estimated structural error (in Å) during topology prediction of AvrXca from bin09. **(C)** Phylogenetic analysis of 100 orthologous AvrXca copies identified in the NCBI database using the bin09

sequence as query. Distinct and well-supported clades correspond to: purple (*Novosphingobium* and *Sphingobium*), red (*Sphingomonas* and *Sphingopyxis*), and green (*Asticcacaulis*), all within Alphaproteobacteria and primarily symbiotic. Yellow (*Dickeya*, *Pectobacterium*) and blue (*Xanthomonas*) represent Gammaproteobacteria, mostly phytopathogenic. Red arrows indicate proteins selected for comparative structural analysis in panel D. **(D)** Comparative structural analysis of AvrXca from bin09 (blue) with orthologous proteins from *Xanthomonas*, *Sphingobium*, *Novosphingobium*, *Asticcacaulis*, *Dickeya*, and *Pectobacterium* (yellow). Structural superposition metrics (RMSD) are provided below each overlay. Orthologs analyzed correspond to those indicated with red arrows in panel C

other *Sphingomonas* species occupying contrasting ecological niches (Zhao et al. 2017; Sorouri et al. 2023).

From an evolutionary perspective, the discovery of distinct CRISPR-Cas architectures exclusively within the

cupuassu-associated MAGs may reflect adaptation to the dynamic phyllosphere, a habitat characterized by intense microbial competition and frequent exposure to mobile genetic elements and phages (Faure et al. 2019). The

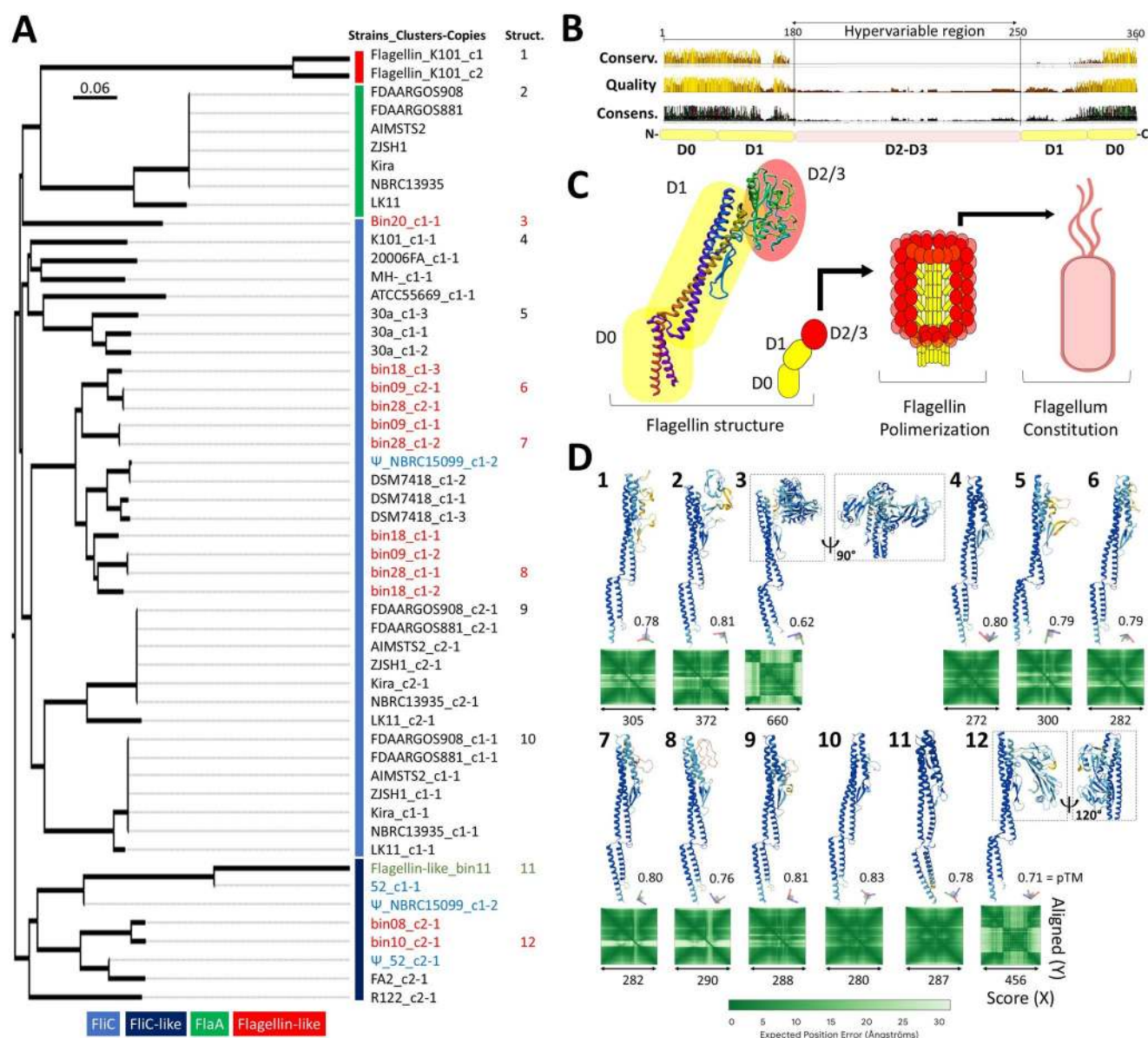


Fig. 8 Structural and evolutionary diversity of flagellin in *Sphingomonas* genomes. **(A)** Phylogenetic reconstruction of flagellin orthologs identified across the analyzed genomes. Red highlights denote sequences derived from cupuassu-associated genomes (Gi-strains). Numbers 1–12 correspond to representative proteins from each phylogenetic clade selected for structural topology comparison in panel D. **(B)** Multiple sequence alignment of flagellin proteins used to infer the phylogeny. The N- and C-terminal regions, corresponding to the

D0 and D1 domains essential for filament polymerization, are highly conserved. In contrast, the central region encoding the D2 and D3 domains—surface-exposed and implicated in host interactions—is highly variable. **(C)** Predicted 3D model of the flagellin protein from bin08. The yellow background highlights the conserved D0 and D1 domains forming the filament core. **(D)** Comparative structural topology of flagellin proteins from 12 representative genomes. Protein identifiers correspond to those in panel A

presence of lineage-specific CRISPR arrays, often coupled with defense-related *dsm* modules, supports the hypothesis that these bacteria have evolved specialized genomic defense mechanisms to maintain stability and ecological fitness within the plant-associated community (Sarver et al. 2025).

Structural and motility-associated genes also exhibited marked variation. The flagellin (FliC) proteins showed

high conservation in core D0/D1 domains but pronounced sequence and structural divergence in the surface-exposed D2/D3 domains. This hypervariability, especially evident in Bin 20 and Bin 08, may reflect local adaptation to host immune recognition or surface adhesion, as these domains are known to act as pathogen-associated molecular patterns (PAMPs) that elicit pattern-triggered immunity (PTI) in plants (Nürnberger et al. 2004; Boller and He 2009).

perniciosa, whose biotrophic phase is sensitive to perturbations in the leaf microenvironment (Meinhardt et al. 2008).

Overall, the detection of effector-like genes, diversified CAZyme and CRISPR repertoires, and structural variants of flagellin expands our understanding of the functional diversity and ecological strategies of *Sphingomonas* in tropical phyllospheres (Berendsen et al. 2012; Sohrabi et al. 2023). These results reinforce the notion that host-associated microbiota are integral components of plant adaptation and defense and highlight the importance of genome-resolved approaches for elucidating plant–microbe coevolution in non-model species.

Together, these genomic and structural insights converge into a unified model that integrates metabolic versatility, defense-related genetic systems, and immunomodulatory interactions between *Sphingomonas* and *T. grandiflorum*. This conceptual framework, illustrated in Fig. 9, summarizes the potential mechanisms through which phyllosphere-associated *Sphingomonas* may contribute to the protection of cupuassu against *Moniliophthora perniciosa*, highlighting the interplay between PTI-, ETI-, and metabolism-driven responses. Furthermore, since cupuassu (*Theobroma grandiflorum*) is endemic to the Amazon, the unique genes found in our bacterial genomes likely reflect a specific geographic signal. Compared to global datasets, the flexible genome of these bacteria appears specialized to the Amazonian environment and the host's chemistry. This suggests a coevolutionary process where the endemic plant selects for microbial strains that are uniquely adapted to its specific ecological niche.

Building on this specialized adaptation, the co-occurrence of *Actinomyces* and *Sphingomonas* within the *Theobroma grandiflorum* microbiome suggests a strategic functional partitioning that enhances host resilience. While *Actinomyces* are typically associated with complex polymer degradation and the production of specialized secondary metabolites for chemical defense, our genomic data indicate that *Sphingomonas* taxa may provide metabolic stability through their flexible genomes, specialized in degrading host-derived aromatic compounds and maintaining membrane integrity via sphingolipids. This potential functional complementarity (where *Actinomyces* act as a putative 'biosynthetic pharmacy' and *Sphingomonas* as 'resilience specialists') could create a synergistic barrier against pathogens, such as *Moniliophthora perniciosa*, by combining niche-specific metabolic versatility with robust antimicrobial protection.

5 Conclusions

This study provides novel insights into the genomic diversity, functional potential, and adaptive features of *Sphingomonas* strains associated with the phyllosphere of a

resistant *T. grandiflorum* cultivar. The recovery of multiple, genomically distinct MAGs — some likely representing new species — alongside the identification of genes involved in aromatic compound degradation, nitrogen and carbohydrate metabolism, and immune modulation underscores the ecological versatility of this genus.

The presence of PTI- and ETI-related signatures, including variable flagellin domains and *avrXca* homologs, suggests potential immunomodulatory capabilities within specific *Sphingomonas* lineages. While these observations do not establish direct causality in plant resistance, they provide a genomic framework for future hypothesis-driven experimentation on microbiome-mediated disease suppression.

Collectively, these findings reinforce the concept of the microbiome as an active component of plant physiology, offering new perspectives for the biotechnological exploitation of phyllosphere bacteria in sustainable crop management and in the understanding of tropical plant–microbe symbioses.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s13199-026-01124-y>.

Acknowledgements We thank the Arizona Genomics Institute (Tucson, AZ, USA) for providing HMW DNA extraction and PacBio Sequel II sequencing support. The Article Processing Charge for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES (ROR identifier: 00x0ma614).

Author contributions AKS, LMM, and AMV designed the project. AMV assembled the genomes. AKS, LMM, NFA, and AMV annotated the genomes. AKS, LMM, and NFA developed the orthologsorter genome comparison platform. AKS, JPM, DFR, CHP, IFC, CGCL, ABS, NFA, CCMG, AMV, and LMM analyzed the genomes and generated the results. AKS, CHP, AMV, and LMM wrote the paper. All authors reviewed the manuscript.

Funding The Article Processing Charge (APC) for the publication of this research was funded by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) (ROR identifier: 00x0ma614). This research was supported by the São Paulo Research Foundation (FAPESP) under grant number 2019/25176-0. Partial funding was also provided by the Minas Gerais State Research Foundation (FAPEMIG) under grant number APQ-02357-17. Research fellowships were awarded to LMM and AMV by the National Council for Scientific and Technological Development (CNPq) under grant numbers 306039/2021-4 and 304367/2022-2, respectively. Additionally, LMM received support from the Federal University of Ouro Preto (UFOP) grants. The funding bodies did not influence the study's design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data availability The data referring to the genetic information generated for this work were deposited at the NCBI.

Declarations

Declaration of generative AI in scientific writing No artificial intelligence tool was used to generate the results, discussion, or write the manuscript.

Conflict of interest The authors have no conflicts of interest to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abreu VAC, Alves RM, Oliveira MM, Trinca V, Falcão LL, Marcelino LH, Figueira A, Domingues DS, Varani A (2025) Integrative chromosome-scale analysis of cupuassu reveals structural and regulatory innovations shaping witches' broom disease resistance and provides a resource for functional studies in *Theobroma*. *Authorea Preprints*. <https://doi.org/10.22541/au.175683923.35704359/v1>
- Alves RM, Abreu VAC, Oliveira RP, Almeida JVDA, Oliveira MM, Silva SR, Paschoal AR, Almeida SS, Souza PAF, Ferro JA, Miranda VFO, Figueira A, Domingues DS, Varani AM (2024) Genomic decoding of *Theobroma grandiflorum* (cupuassu) at chromosomal scale: evolutionary insights for horticultural innovation. *Gigascience* 13:giae027. <https://doi.org/10.1093/gigascience/giae027>
- Asaf S, Khan AL, Khan MA, Imran QM, Yun BW, Lee IJ (2017) Osmoprotective functions conferred to soybean plants via inoculation with *Sphingomonas* sp. LK11 and exogenous trehalose. *Microbiol Res* 205:135–145. <https://doi.org/10.1016/j.micres.2017.08.009>
- Basta T, Keck A, Klein J, Stolz A (2004) Detection and characterization of conjugative degradative plasmids in xenobiotic-degrading *Sphingomonas* strains. *J Bacteriol* 186(12):3862–3872. <https://doi.org/10.1128/JB.186.12.3862-3872.2004>
- Berendsen RL, Pieterse CM, Bakker PA (2012) The rhizosphere microbiome and plant health. *Trends Plant Sci* 17(8):478–486. <https://doi.org/10.1016/j.tplants.2012.04.001>
- Boller T, He SY (2009) Innate immunity in plants: an arms race between pattern recognition receptors in plants and effectors in microbial pathogens. *Science* 324(5928):742–744. <https://doi.org/10.1126/science.1171647>
- Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF (2015) Unusual biology across a group comprising more than 15% of domain bacteria. *Nature* 523(7559):208–211. <https://doi.org/10.1038/nature14486>
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10(1):421. <https://doi.org/10.1186/1471-2105-10-421>
- Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The carbohydrate-active enzymes database (CAZy): an expert resource for glycogenomics. *Nucleic Acids Res* 37(Suppl 1):D233–D238. <https://doi.org/10.1093/nar/gkn663>
- Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17(4):540–552. <https://doi.org/10.1093/oxfordjournals.molbev.a026334>
- Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH (2022) GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics* 38(23):5315–5316. <https://doi.org/10.1093/bioinformatics/btac672>
- Costa JL, Paschoal D, da Silva EM, Silva JS, do Carmo RM, Carrera E, López-Díaz I, Rossi ML, Freschi L, Mieczkowski P, Peres LEP, Teixeira PJPL, Figueira A (2021) *Moniliophthora perniciosa*, the causal agent of witches' broom disease of cacao, interferes with cytokinin metabolism during infection of Micro-Tom tomato and promotes symptom development. *New Phytol* 231(1):365–381. <https://doi.org/10.1111/nph.17386>
- Delmotte N, Knief C, Chaffron S, Innerebner G, Roschitzki B, Schlappbach R, von Mering C, Vorholt JA (2009) Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc Natl Acad Sci USA* 106(38):16428–16433. <https://doi.org/10.1073/pnas.0905240106>
- Deloger M, El Karoui M, Petit MA (2009) A genomic distance based on MUM indicates discontinuity between most bacterial species and genera. *J Bacteriol* 191(1):91–99. <https://doi.org/10.1128/jb.01202-08>
- Falcão LL, Silva-Wernecke JO, Albuquerque PSB, Alves RM, Grynberg P, Togawa RC, Costa MMC, Brigido MM, Marcellino LH (2022) Comparative transcriptomics of cupuassu (*Theobroma grandiflorum*) offers insights into the early defense mechanism to *Moniliophthora perniciosa*, the causal agent of witches' broom disease. *J Plant Interact* 17(1):991–1005. <https://doi.org/10.1080/17429145.2022.2144650>
- Farhan M, Ahmad M, Kanwal A, Butt ZA, Khan QF, Raza SA, Qayyum H, Wahid A (2021) Biodegradation of chlorpyrifos using isolates from contaminated agricultural soil, its kinetic studies. *Sci Rep* 11(1):10320. <https://doi.org/10.1038/s41598-021-88264-x>
- Faure G, Shmakov SA, Yan WX, Cheng DR, Scott DA, Peters JE, Makarova KS, Koonin EV (2019) CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat Rev Microbiol* 17(8):513–525. <https://doi.org/10.1038/s41579-019-0204-7>
- Feng X, Cheng H, Portik D, Li H (2022) Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nat Methods* 19(6):671–674. <https://doi.org/10.1038/s41592-022-01478-3>
- Gillespie JJ, Wattam AR, Cammer SA, Gabbard JL, Shukla MP, Dalay O, Driscoll T, Hix D, Mane SP, Mao C, Nordberg EK, Scott M, Schulman JR, Snyder EE, Sullivan DE, Wang C, Warren A, Williams KP, Xue T, Yoo HS, Zhang C, Zhang Y, Will R, Kenyon RW, Sobral BW (2011) PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infect Immun* 79(11):4286–4298. <https://doi.org/10.1128/iai.00207-11>
- Grant JR, Enns E, Marinier E, Mandal A, Herman EK, Chen C-Y, Graham M, Van Domselaar G, Stothard P (2023) Proksee: in-depth characterization and visualization of bacterial genomes. *Nucleic Acids Res* 51(W1):W484–W492. <https://doi.org/10.1093/nar/gkad326>
- Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hermsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF (2016) A new view of the tree of life. *Nat Microbiol* 1(5):1–6. <https://doi.org/10.1038/nmicrobiol.2016.48>
- Innerebner G, Knief C, Vorholt JA (2011) Protection of *Arabidopsis thaliana* against leaf-pathogenic *Pseudomonas syringae* by *Sphingomonas* strains in a controlled model system. *Appl Environ Microbiol* 77(10):3202–3210. <https://doi.org/10.1128/AEM.00133-11>
- Jinfang Z, Ge Q, Yan Y, Zhang X, Huang L, Yin Y (2023) Dbcan3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Res* 51(W1):115–121. <https://doi.org/10.1093/nar/gkad328>
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589. <https://doi.org/10.1038/s41586-021-03819-2>

- Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28(1):27–30. <https://doi.org/10.1093/nar/28.1.27>
- Kanehisa M, Sato Y, Morishima K (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol* 428(4):726–731. <https://doi.org/10.1016/j.jmb.2015.11.006>
- Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z (2019) MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7:e7359. <https://doi.org/10.7717/peerj.7359>
- Khan AL, Waqas M, Kang S-M, Al-Harrasi A, Hussain J, Al-Rawahi A, Al-Khiziri S, Ullah I, Ali L, Jung H-Y, Lee IJ (2014) Bacterial endophyte *Sphingomonas* sp. LK11 produces gibberellins and IAA and promotes tomato plant growth. *J Microbiol* 52(8):689–695
- Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. *Nucl Acids Res* 32(12):W526–W531. <https://doi.org/10.1093/nar/gkh468>
- Konstantinidis KT, Tiedje JM (2005) Towards a genome-based taxonomy for prokaryotes. *J Bacteriol* 187(18):6258–6264. <https://doi.org/10.1128/jb.187.18.6258-6264.2005>
- Konstantinidis KT, Ramette A, Tiedje JM (2006) The bacterial species definition in the genomic era. *Philos Trans R Soc B* 361(1475):1929–1940. <https://doi.org/10.1098/rstb.2006.1920>
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5(2):R12. <https://doi.org/10.1186/gb-2004-5-2-r12>
- Lebeis SL, Herrera Paredes S, Lundberg DS, Breakfield N, Gehring J, McDonald M, Malfatti S, Glavina del Rio T, Jones CD, Tringe SG, Dangel JL (2015) Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science* 349(6250):860–864. <https://doi.org/10.1126/science.aaa8764>
- Li L, Stoeckert CJ, Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13(9):2178–2189. <https://doi.org/10.1101/gr.1224503>
- Li Z, Jaroszewski L, Iyer M, Sedova M, Godzik A (2020) FATCAT 2.0: towards a better understanding of the structural diversity of proteins. *Nucleic Acids Res* 48(W1):W60–W64. <https://doi.org/10.1093/nar/gkaa443>
- Li PD, Zhu ZR, Zhang Y, Xu J, Wang H, Wang Z, Li H (2022) The phyllosphere microbiome shifts toward combating melanose pathogen. *Microbiome* 10(1):56. <https://doi.org/10.1186/s40168-022-01234-x>
- Lundberg DS, de Pedro Jové R, Pramoj Na Ayutthaya P, Karasov TL, Shalev O, Poersch K, Ding W, Bollmann-Giolai A, Bezrukov I, Weigel D (2022) Contrasting patterns of microbial dominance in the *Arabidopsis thaliana* phyllosphere. *Proc Natl Acad Sci USA* 119(52):e2211881119. <https://doi.org/10.1073/pnas.2211881119>
- Matos JP, Ribeiro DF, Silva AK, Paula CH, Cordeiro IF, Lemes CGC, Sanchez AB, Rocha LCM, Garcia CCM, Almeida NF, Alves RM, Abreu VAC, Varani AM, Moreira LM (2024) Diversity and potential functional role of phyllosphere-associated actinomyces isolated from cupuassu (*Theobroma grandiflorum*) leaves: implications for ecosystem dynamics and plant defense strategies. *Mol Genet Genomics* 299(1):73. <https://doi.org/10.1007/s00438-024-02162-1>
- Mazoyon C, Catterou M, Alahmad A, Mongelard G, Guénin S, Sarazin V, Dubois F, Duclercq J (2023) *Sphingomonas sediminicola* Dae20 is a highly promising beneficial bacteria for crop biostimulation due to its positive effects on plant growth and development. *Microorganisms* 11(8):2061. <https://doi.org/10.3390/microorgani11082061>
- Meinhardt LW, Rincones J, Bailey BA, Aime MC, Griffith GW, Zhang D, Pereira GA (2008) *Moniliophthora perniciosa*, the causal agent of witches' broom disease of cacao: what's new from this old foe? *Mol Plant Pathol* 9(5):577–588. <https://doi.org/10.1111/j.1364-3703.2008.00496.x>
- Mendes R, Kruijt M, de Bruijn I, Dekkers E, van der Voort M, Schneider JHM, Piceno YM, DeSantis TZ, Andersen GL, Bakker PAHM, Raaijmakers JM (2011) Deciphering the rhizosphere microbiome for disease-suppressive bacteria. *Science* 332(6033):1097–1100. <https://doi.org/10.1126/science.1203980>
- Mournet P, de Albuquerque PSB, Alves RM, Silva-Werneck JO, Rivalan R, Marcellino LH, Clément D (2020) A reference high-density genetic map of *Theobroma grandiflorum* (Willd. ex Spreng) and QTL detection for resistance to witches' broom disease (*Moniliophthora perniciosa*). *Tree Genet Genomes* 16(6):89. <https://doi.org/10.1007/s11295-020-01479-3>
- Nürnberg T, Brunner F, Kemmerling B, Piater L (2004) Innate immunity in plants and animals: striking similarities and obvious differences. *Immunol Rev* 198(1):249–266. <https://doi.org/10.1111/j.0105-2896.2004.01119.x>
- Olmi E, Kusstatscher P, Wicaksono WA, Abdelfattah A, Cernava T, Berg G (2022) Insights into the microbiome assembly during different growth stages and storage of strawberry plants. *Environ Microbiome* 17(1):21. <https://doi.org/10.1186/s40793-022-00415-3>
- Parker JE, Barber CE, Fan MJ, Daniels MJ (1993) Interaction of *Xanthomonas campestris* with *Arabidopsis thaliana*: characterization of a gene from *X. c. pv. raphani* that confers avirulence to most *A. thaliana* accessions. *Molecular Plant-Microbe Interactions (MPMI)* 6(2):216–224. <https://doi.org/10.1094/mpmi-6-216>
- Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25(7):1043–1055. <https://doi.org/10.1101/gr.186072.114>
- Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW (2017) Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol* 2(11):1533–1542. <https://doi.org/10.1038/s41564-017-0012-7>
- Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 36(10):996–1004. <https://doi.org/10.1038/nbt.4229>
- Peng X, Wang H, Zhou X (2024) The phyllosphere microbial community structure of three *Camellia* species upon anthracnose. *Forests* 15(12):2080. <https://doi.org/10.3390/f15122080>
- Pereira ALF, Abreu VKG, Rodrigues S (2018) Cupuassu — *Theobroma grandiflorum*. *Exotic Fruits: Ref Guide* 159–162. <https://doi.org/10.1016/B978-0-12-803138-4.00021-6>
- Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* 106(45):19126–19131. <https://doi.org/10.1073/pnas.0906412106>
- Sarver E, González-Moreno KJ, Christensen KG, Lefevers HM, Corbin KR (2025) Phyllosphere synthetic microbial communities: a new frontier in plant protection. *BMC Plant Biol* 25(1):949. <https://doi.org/10.1186/s12870-025-06935-7>
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Sohrabi R, Paasch BC, Liber JA, He SY (2023) Phyllosphere microbiome. *Annu Rev Plant Biol* 74(1):539–568. <https://doi.org/10.1146/annurev-arplant-102820-032704>
- Sorouri B, Rodriguez CI, Gaut BS, Allison SD (2023) Variation in *Sphingomonas* traits across habitats and phylogenetic clades.

- Front Microbiol 14:1146165. <https://doi.org/10.3389/fmicb.2023.1146165>
- Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>
- Stolz A (2009) Molecular characteristics of xenobiotic-degrading sphingomonads. *Appl Microbiol Biotechnol* 81(5):793–811. <https://doi.org/10.1007/s00253-008-1752-3>
- Stolz A, Schmidt-Maag C, Denner EB, Busse HJ, Egli T, Kämpfer P (2000) Description of *Sphingomonas xenophaga* sp. nov. for strains BN6T and N,N which degrade xenobiotic aromatic compounds. *Int J Syst Evol Microbiol* 50(1):35–41. <https://doi.org/10.1099/00207713-50-1-35>
- Sultana R, Islam SMN, Sriti N, Ahmed M, Shuvo SB, Rahman MH, Jashim AII (2024) *Sphingomonas panaciterrae* PB20 increases growth, photosynthetic pigments, antioxidants, and mineral nutrient contents in spinach (*Spinacia oleracea* L.). *Heliyon* 10(3):e25596. <https://doi.org/10.1016/j.heliyon.2024.e25596>
- Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* 44(14):6614–6624. <https://doi.org/10.1093/nar/gkw569>
- Teixeira PJPL, Thomazella DPT, Reis O, Prado PFV, do Rio MCS, Fiorin GL, José J, Costa GGL, Negri VA, Mondego JMC, Mieczkowski P, Pereira GAG (2014) High-resolution transcript profiling of the atypical biotrophic interaction between *Theobroma cacao* and the fungal pathogen *Moniliophthora perniciosa*. *Plant Cell* 26(11):4245–4269. <https://doi.org/10.1105/tpc.114.130807>
- Videira SS, De Araujo JLS, da Silva Rodrigues L, Baldani VLD, Baldani JI (2009) Occurrence and diversity of nitrogen-fixing *Sphingomonas* bacteria associated with rice plants grown in Brazil. *FEMS Microbiol Lett* 293(1):11–19. <https://doi.org/10.1111/j.1574-6968.2008.01475.x>
- Vogel C, Innerebner G, Zingg J, Guder J, Vorholt JA (2012) Forward genetic in planta screen for identification of plant-protective traits of *Sphingomonas* sp. strain Fr1 against *Pseudomonas syringae* DC3000. *Appl Environ Microbiol* 78(16):5529–5535. <https://doi.org/10.1128/aem.00639-12>
- Wang Z, Lajoie G, Jiang Y, Zhang M, Chu C, Chen Y, Fang S, Jin G, Jiang M, Lian J, Li Y, Liu Y, Ma K, Mi X, Qiao X, Wang X, Wang X, Xu H, Ye W, Zhu L, Zhu Y, He F, Kembel SW (2024) Host specificity of plant-associated bacteria is negatively associated with genome size and host abundance along a latitudinal gradient. *Ecol Lett* 27(6):e14447. <https://doi.org/10.1111/ele.14447>
- White DC, Sutton SD, Ringelberg DB (1996) The genus *Sphingomonas*: physiology and ecology. *Curr Opin Biotechnol* 7(3):301–306. [https://doi.org/10.1016/s0958-1669\(96\)80034-6](https://doi.org/10.1016/s0958-1669(96)80034-6)
- Wood DE, Lu J, Langmead B (2019) Improved metagenomic analysis with Kraken 2. *Genome Biol* 20(1):257. <https://doi.org/10.1186/s13059-019-1891-0>
- Zhao Q, Yue S, Bilal M, Hu H, Wang W, Zhang X (2017) Comparative genomic analysis of 26 *Sphingomonas* and *Sphingobium* strains: dissemination of bioremediation capabilities, biodegradation potential and horizontal gene transfer. *Sci Total Environ* 609:1238–1247. <https://doi.org/10.1016/j.scitotenv.2017.07.249>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.