Unveiling Genetic Diversity: Construction and Analysis of the *Moniliophthora roreri* Pangenome

Isabella Gallego Rendón

Tesis

Diego Mauricio Riaño Pachón

UNIVERSIDAD EAFIT
ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA
BIOLOGÍA
MEDELLÍN
2023

RESUMEN

Antecedentes:
El patógeno del cacao Moniliophthora roreri representa una amenaza significativa para la producción mundial de cacao. Esto se debe a su notable capacidad de dispersión, su propensión a infectar una amplia gama de cultivares de cacao, su adaptabilidad a diferentes nichos ecológicos y su alta patogenicidad. Para desarrollar estrategias de control efectivas, es crucial una comprensión más profunda de su diversidad genética y capacidades funcionales. El objetivo de este estudio fue realizar un análisis pan-genómico de 22 genomas de M. roreri para obtener información sobre su composición genética y sus propiedades funcionales potenciales.

Resultados:
Se identificaron un total de 456,309 genes codificadores de proteínas a partir de los genomas ensamblados, de los cuales el 97.5% fueron asignados a ortogrupos. El pan-genoma se categorizó en categorías de núcleo duro, núcleo blando, accesorio y exclusivo. El análisis derivado proporcionó una perspectiva sobre los patrones de expansión o contracción del grupo de genes en respuesta a la integración de genomas adicionales. La anotación funcional y el análisis de enriquecimiento de términos GO revelaron genes asociados con varios procesos biológicos, describiéndose estos procesos al subrayar los genes potencialmente asociados a mecanismos patogénicos y adaptativos de M. roreri.

Conclusiones:
Este estudio pan-genómico integral proporciona una comprensión fundamental de la composición genética y funcional de M. roreri. Los hallazgos elucidan posibles grupos de genes que podrían ser de interés para futuras investigaciones en el campo de las interacciones planta-patógeno y para el desarrollo de intervenciones dirigidas al control de enfermedades del cacao.

Palabras clave: Pangenoma, Moniliasis, Moniliophthora roreri, Cacao

ABSTRACT

Background:

The cacao pathogen Moniliophthora roreri poses a significant threat to global cacao production. This is due to its distinct dispersal ability, its propensity to infect a wide range of cacao cultivars, its adaptability to different ecological niches, and its high pathogenicity. To develop effective control strategies, a deeper understanding of its genetic diversity and functional capabilities is critical. The objective of this study was to perform a pangenomic analysis of 22 M. roreri genomes to gain insights into its genetic composition and potential functional properties.

Results:

A total of 456,309 protein-coding genes were identified from the assembled genomes, of which 97.5% were assigned to orthogroups. The pangenome was categorized into hard-core, soft-core, accessory, and exclusive categories. Derivative analysis provided a perspective on the gene pool expansion or contraction patterns in response to the integration of additional genomes. Functional annotation and GO terms enrichment analysis revealed genes associated with various biological processes, these processes were described underscoring the potential genes associated to pathogenic and adaptive mechanisms of *M. roreri*.

Conclusions:

This comprehensive pangenomic study provides a fundamental understanding of the genetic and functional makeup of *M. roreri*. The insights elucidate potential gene clusters that might be of interest for future research in the field of plant-pathogen interactions and targeting interventions for cacao disease control.

Keywords: Pangenome, Frosty Pod Rot, Moniliophthora roreri, Cacao

INTRODUCTION

- **Cacao**

Cacao (*Theobroma cacao*) is a crop native to the American tropics found at altitudes between 0 and 1,200 meters above sea level. It is cultivated in 61 tropical countries from Africa, Asia, and Latin America (Jiménez et al., 2022; Nieves-Orduña et al., 2023a). The cultivation of cacao is of high economic importance due to the commercial potential of its seeds, which are the primary source of chocolate production (de Souza et al., 2018). Colombia has positioned itself as the fourth leading cacao producer in Latin America, with an annual production of about 69,040 tons and exports of up to 11,689 tons per year (Fedecacao, 2022).

The cultivated cacao has been traditionally divided into three main cultivars: Forastero, Trinitario, and Criollo. Each has differences in their origins, fruit size, and taste. Another significant difference is the susceptibility of the cultivars to pests and diseases, with the Forastero cultivar being the most common globally but less resistant to pathogens (Jiménez, Alvarez & Mosquera, 2021). Colombian cacao crops are highly variable at the genetic level, which can be explained by the crosses that occur between the Forastero and Trinitario varieties. The environmental conditions in Colombia promote seed dispersion and favor the production potential of cacao ecotypes with different characteristics (Tirado-Gallego et al., 2016). However, this classification does not capture the genetic diversity of cacao populations in the Amazon, and studies based on molecular markers recognize up to 10 different genetic clusters, with three of them corresponding to the traditional cultivars (Nieves-Orduña et al., 2023b).

Currently, one of the main challenges for cacao producers is phytosanitary issues such as witches' broom disease and moniliasis, also known as frosty pod rot disease, caused by *Moniliophthora perniciosa* and *M. roreri*, respectively. These two fungi are phylogenetically close and attack crops of the genera *Theobroma* and *Herrania*. However, they show significant differences related to their host ranges and infection cycles (Meinhardt et al., 2014).

In recent years, the phytosanitary problem has increased due to poor crop management, environmental changes, and a lack of pathogen control strategies in the crops (Aikpokpodion, 2019). Frosty pod rot can cause yield losses of up to 80%, forcing land use changes (Nieves-Orduña et al., 2023a).

- ***M. roreri*: A Fungal Threat to Cacao**

As mentioned earlier, *M. roreri* is the fungus responsible for cacao moniliasis. This fungus is a basidiomycete with a hemibiotrophic life cycle endemic to the Magdalena region of

Colombia (Cubillos, 2017; Phillips-Mora & Wilkinson, 2007). Its host range includes cacao species from the *Herrania* and *Theobroma* genera, and its distribution ranges from altitudes of 0 to 1520 meters above sea level. It thrives in areas with an average annual precipitation ranging from 780 to 5500 mm and temperatures between 18° and 28°C (Jiménez et al., 2022).

Several characteristics contribute to the devastating effect of this pathogen on crops: its high rate of dispersion, the variety of cacao it infects, its ability to adapt to different environments, its high level of pathogenicity, and increased temperatures and precipitation due to global warming (Ortega Andrade et al., 2017). Another important factor when assessing the impact of the fungus on crops is its high rate of genetic diversification (Jaimes et al., 2016).

As previously detailed, the variety of cacao used in cultivation plays a decisive role in the development and infection cycle of the pathogen, as some varieties are more susceptible than others (Jiménez et al., 2022). In Colombia, studies have been carried out on the infection potential of *M. roreri* in different cacao clones. The results indicate that some clones exhibit less severe infections (Ali et al., 2015). These types of studies are especially relevant because they lead to genetic improvement strategies aiming to mitigate and control the consequences of the pathogen on the crops.

Symptoms of moniliasis start about 40 to 60 days after infection, with the appearance of chlorotic spots on the pod surface. As the infection progresses, these spots darken, turning oily with chlorotic borders. Eventually, the mycelium and spores of *M. roreri* become visible on the pod surface, and the characteristic white and powdery spots of the disease appear (Jiménez et al., 2022). The germinated spores penetrate the pod through the stomata and colonize the internal tissues of the cacao fruit. The outcome of the infection can vary depending on the development stage of the fruit when infected. The fruit's susceptibility range decreases as it matures, with pods younger than one month being the most susceptible (Bailey et al., 2018).

To date, the reproductive biology of *M. roreri* has not been definitively identified. However, literature suggests that the fungus reproduces clonally, and its spores reproduce asexually (Ali et al., 2015; Díaz-Valderrama & Aime, 2016).

Currently, the *M. roreri* population has been classified into five genetic groups based on ISSR and AFLP profiling of 94 isolates from Latin America. In this classification, the Bolívar and Co-West groups show the broadest geographic range (Phillips-Mora et al., 2007).

According to some studies on the pathogen's population distribution, it is believed that the spread of *M. roreri* originated in wild *Theobroma* and *Herrania* fruits and later colonized commercial crops, likely facilitated by human activity (Ali et al., 2015). On the other hand,

studies on *M. roreri* have mainly focused on cultural control and fungicides. Still, there isn't enough information on the fungus's genomic aspect, highlighting the need for more in-depth genetic and evolutionary analyses (Tiburcio et al., 2010).

- **Genomics**

As mentioned earlier, *M. roreri* is a fungus endemic to the middle Magdalena region of Colombia, where the highest percentage of genetic diversity is concentrated (Cubillos, 2017). From this origin, the fungus has ventured to other regions, enabling the emergence of different genetic groups. The spread of *M. roreri* across various terrains in Colombia and South America, combined with its significant genetic diversity, highlights its strong adaptive abilities (Espinoza-Lozano et al., 2022).

Notably, *M. roreri*'s genetic composition has been of great interest for scientific exploration. Research incorporating AFLP and ISSR markers revealed several genetic clusters, some unique to certain territories and others with a more widespread presence. Later research using SNPs identified two main genetic groups, highlighting the significant role of northeastern Colombian populations (Ali et al., 2015; Daboussi & Capy, 2003; Melo et al., 2014; Suárez-Contreras, 2017).

There are two genomes of *M. roreri* deposited in the NIH GenBank genetic sequence database under the accession numbers GCA_000488995.1 (Meinhardt et al., 2014) and GCA_001466705.2 (Díaz-Valderrama et al., 2023). The first genome for the pathogen has a size of 52.2 Mb and was published by Meinhardt et al., in 2014. The sample was isolated from a state in Ecuador and the assembly of the genome offered insights into the mechanisms underlying the biotrophic and necrotrophic phases of the pathogen. However, it was then replaced as the reference genome for the species by the genome published by Díaz-Valderrama et al., in 2023, which was isolated from the region of Chiapas, Mexico. This genome has a size of 59.5 Mb and is reportedly the most complete and continuous genome produced for the genus Moniliophthora to date.

A recent publication aimed to elucidate the underlying mechanisms governing the hemibiotrophic life cycle and clonal reproductive biology of *M. roreri* (Minio et al., 2023). In this study, 22 new genomes were sequenced, sourced from isolates across different geographical regions in Central and South America. Additionally, the research identified three previously unreported alleles for the A locus (designated AX3, AX4, and AX5) and a new allele for the B locus (designated BX3). These genomes have sizes ranging from 56.5 Mb to 63.2 Mb, with an average size of 57.9 Mb. The genomes were submitted to a BUSCO analysis (Seppey et al., 2019) and the results portrayed an average completeness of 94.82%, with complete BUSCOs values ranging between 94.5% and 95.0%.

These new approaches aim to expand the existing understanding of the reproductive biology of the fungus and further explore the pathogen-host interactions surrounding cacao

species, contributing to the development of innovative solutions to address challenges associated with cacao cultivation.

- **Pangenome**

In modern genomics, to fully understand genetic diversity within a species, relying solely on a single reference genome is inadequate (Garcia et al., 2023). Single genome representations naturally miss the depth of variations within a species. This understanding has led to the idea of 'pan-genomes', which encompass all genes seen across different strains of a species (Amir et al., 2020).

The pan-genome consists of two main parts: the 'core genome', which has universally present genes, and the 'accessory' genome, which has genes specific to certain strains that can influence distinct traits (Badet & Croll, 2020). Some species, based on their environment, have broader pan-genomes than others. While the idea of pan-genomes started with studies of prokaryotes, it's now also applied to complex eukaryotes , revealing broad genetic diversity across them (Amir et al., 2020).

The hard-core pangenome, as referred to before, is the set of genes present in 100% of the strains. The soft-core pangenome is defined as the set of genes present in >90% of the strains (Agarwal et al., 2023). The accessory classification are those genes present in more than 1 but less than 100% of the strains (Tettelin et al., 2005); finally, the exclusive category is given to the set of genes unique to each strain (Gong et al., 2023).

While we understand how simple organism pan-genomes evolve, mainly through sharing genes, complex organism pan-genomes change due to a range of genetic activities, like changes in gene organization, evolutionary adaptation, variations in the number of gene sets, and gene sharing between species (McCarthy & Fitzpatrick, 2019).

- **Why now?**

Theobroma cacao is more than just the foundation of a lucrative global industry. It holds a central position in socio-economic systems, particularly in neotropical cultivation areas. Both small-scale farmers and larger regional economies rely heavily on the well-being and sustainability of cacao crops. Any disease affecting these plants can lead to immediate economic losses, disrupt food chains, threaten community stability, and impact regions with a long history of cacao farming (González-Orozco et al., 2020)

The rapid spread of frosty pod rot, caused by the fungus *M. roreri*, has been both sudden and damaging. As previously mentioned, this disease results not only in significant crop losses but also influences farmers to switch to less profitable cacao varieties or abandon cacao farming altogether (Tirado-Gallego et al., 2016). This not only affects the economy

but also erodes long-standing cacao farming traditions and the cultural significance attached to them.

Understanding the biology, genetics, and spread patterns of *M. roreri* is more than just academic curiosity; it's essential. While currently, this fungus is mainly found in the Americas (Phillips-Mora & Wilkinson, 2007), the interconnectedness of today's world and continuous trade in agriculture heighten the risk of it reaching other key cacao-producing areas, like Africa and Asia. Such a development would have major consequences, especially given the vast cacao regions in these continents and their limited readiness to deal with this particular disease.

For the construction and exploration of the *M.roreri* pan-genome, this research uses 22 publicly available genomes collected from various geographic regions (Minio et al., 2023), and aims to uncover the genetic diversity of this fungus at a deeper level, enhancing our understanding of eukaryotic pan-genomes and giving some insights into the genetic contents of the pathogen for future research regarding its adaptability and pathogenicity.

Methods

0. Computational resources

All computational analyses were conducted on the bioinformatics cluster of the Center of Nuclear Energy for Agriculture (CENA) at the University of São Paulo. This infrastructure provided the requisite computational power and storage capabilities to handle large-scale genomic datasets and facilitated the efficient execution of the pangenomic analyses. The use of the CENA cluster ensured reproducibility and reliability in the results, offering high-performance computing resources designed for bioinformatics research. The specific configurations and software versions used for each analysis are provided throughout the methods section of this study.

1. Data collection

The data used in this study was retrieved from a previously published work by Andrea Minio et al., (Minio et al., 2023), the authors generated genomic *M. roreri* data from samples of diverse geographic locations. This data set was chosen for its relevance and completeness in relation to our research objectives. Briefly, the fungal samples were originally obtained from a collection held by the USDA. gDNA was extracted and submitted to whole genome sequencing with Illumina technology. The genomes were then assembled and subjected for quality control. Finally, the coding sequences were extracted and the proteins underwent an annotation process and were mapped against the RefSeq database. More information about the original data collection procedures can be found in (Minio et al., 2023). A brief description of the isolates included in the study is shown in table 1. Further assembly statistics can be found on the Supplemental Material section of the original article.
From the data published only the FASTA files and GFF3 files were taken into account for the present study. The data is available at Zenodo (https://zenodo.org/records/7872498) (Minio & Cantu, 2024). The GFF3 files were processed using the gffreads utility, as part of the Cufflinks suite (Pertea & Pertea, 2020) to extract the coding sequences (CDS) and deduce protein sequences.

Isolate	Country	Region/state	Year of isolation	Genome size
MrC26	Costa Rica	Guayabo Abajo, Cartago	1999	59.6 Mb
MrB3	Bolivia	Porvenir, Caranavi, La Paz	2012	57.5 Mb
MrP5	Peru	Castillo Papayal, Huánuco	1997	57.7 Mb
MrCo52	Colombia	Lebrija, Santander	2006	56.5 Mb
MrE7	Ecuador	Mocache, Los Ríos	1999	57.4 Mb
MrCo44	Colombia	San Jerónimo, Antioquia	2006	57.8 Mb
MrCo8	Colombia	San Jerónimo, Antioquia	1999	58.2 Mb
MrCo29	Colombia	La Suiza, Copoica, Santander	2004	57.5 Mb

MrCo82	Colombia	Lebrija, Santander	2006	56.9 Mb
MrCo84	Colombia	Tumaco, Nariño	2006	56.8 Mb
MrCo25	Colombia	Rio Negro, Santander	2004	57.8 Mb
MrCo74	Colombia	El Carmen de Chucurí, Santander	2006	57.6 Mb
MrCo12	Colombia	Palestina, Caldas	1999	57.5 Mb
MrCo43	Colombia	Tarazá, Antioquia	2006	58.2 Mb
MrCo45	Colombia	Valdivia, Antioquia	2006	57.2 Mb
MrCo58	Colombia	Caparrapi, Cundinamarca	2006	56.5 Mb
MrCo65	Colombia	Bucarasica, Norte de Santander	2006	57.6 Mb
MrCo67	Colombia	Marsella, Risaralda	2006	57.5 Mb
Mr017	Ecuador	Cerecita, Guayas	2019	57.6 Mb
Mr020	Ecuador	Buena Fe, Los Rios	2019	57.8 Mb
Mr030	Ecuador	Mocache, Los Rios	2019	57.4 Mb
MrA4295	Costa Rica	Unknown	1978	57.5 Mb

Table 1. List of the 22 *Moniliophthora roreri* isolates used in the study. Details include the designated isolate code, country of origin, specific region or state of isolation, the year they were isolated and genome size.

2. Identification of groups of orthologous genes

Groups of orthologous genes were identified with OrthoFinder. The OrthoFinder algorithm incorporates all-vs-all similarity searches, MCL clustering, and a species-overlap algorithm to predict orthogroups (Emms & Kelly, 2015). The deduced proteomes for each isolate underwent analysis using OrthoFinder in order to sort the genes into orthogroups.

Similarity searches were computed with Diamond v0.9.24 (Buchfink et al., 2021) with the following parameters “diamond_ultra_sens2, --query-cover 80 --subject-cover 80 --id 80”, that only keeps hits with at least 80% identity over 80% of the sequence length.

Assigning genes into orthogroups heavily depends on a parameter of the MCL algorithm, referred to as Inflation. In order to select the appropriate inflation value we tried the following set of values: {1.2, 1.6, 2.0, 2.4, 2.8, 3.2, 3.6, 4.0, 4.4, 4.8, 5.2, 5.6, 6.0}. The inflation value producing the clustering maximizing the similarity of protein domain architecture of its constituent members was chosen for further analysis using cogeqc in R (Almeida-Silva & Peer, 2023).

With the selected proper inflation value, we identified orthogroups present in all isolates, and we refer to them as hard-core orthogroups. In order to deal with potentially incomplete genome assemblies we also defined a set of soft-core orthogroups, those that are present in at least 90% of the isolates. Furthermore, "accessory orthogroups" were delineated as those appearing in more than one isolate but in fewer than 90% of them. Lastly, "exclusive orthogroups" were classified as those found in a single isolate.

We used the script computeorthogroupstats.py developed by the LabBCES research group within the 'Center for Nuclear Energy in Agriculture' (CENA) at the University of São

Paulo, to compute the distribution of the average size of the identified orthogroups (Riaño-Pachón & Vaz, 2023). This step provided insights into the distribution and variance in orthogroup sizes within the studied datasets. The input for the script was the Orthogroups_I4.4.tsv table generated by the OrthoFinder program.
We utilized the script extractpantranscriptomegroups.py to extract distinct pangenome groups (Riaño-Pachón & Vaz, 2023). For this step the data used was the gene count file generated by the OrthoFinder step. These groups were categorized into hard-core, soft-core, exclusive, and accessory pangenomes. The output of this script was used for crafting a table which was then used for plotting.

For the visualization of gene counts across orthogroups, we employed a log2 scaling to address the wide variance typically observed in biological datasets. This approach compresses the spectrum of larger gene counts while amplifying the smaller ones, enhancing discernibility of patterns and subtle differences. In order to retain the significance of orthogroups with a gene count of zero—a crucial aspect of our dataset definition—we pre processed the gene counts data using the following line of code “sns.heatmap(np.log2(orthogroups_data + 1), …”, which adds a value of one to each count before the logarithmic transformation. By doing this we ensured that the zero values remained zero after the log(2) scaling, thus maintaining the significance of a zero count in the graphic representation.

The pangenome categories were based on the prevalence of gene groups among the isolates., as follows: the pan-genome itself consists of the entire set of genes along the 22 strains; the hard-core groups are those genes in 100% of the strains; the soft-core groups were defined as those present in at least 90% of the strains; the accessory category as those present in more than one but less than 90% of the strains, and finally, the exclusive category was delimited as the gene groups appearing in a single strain.
In order to determine the stabilization point of the pangenome trajectory, the custom R script derivadas.r (Gallego, 2023) was used. The dataset used was the orthogroup gene count obtained from the OrthoFinder analysis. The data was subsequently filtered to retain only rows corresponding to the "Pan-Genome" classification. For this subset, the data was grouped by the number of genotypes, and the mean values for the number of genes and groups were computed. Derivatives between consecutive elements in the `NumberGenotypes` and `NumberGroups` columns were calculated to assess the rate of change. A non-parametric loess regression model was fitted to these derivatives against the number of genotypes, providing a smooth curve to capture the underlying trend in the scatter plot.

3. Identification of protein domains

Protein domains were identified with the tool hmmscan from the HMMER suite v3.2.1 (Eddy, 2011) using default parameters and the PFAM database v34.0. These results were exploited by cogeqc to select the appropriate inflation value for MCL (see above).

4. Functional Annotation and enrichment analysis

We attributed Gene Ontology terms to the proteins of the deduced proteomes of the 22 isolates, using the tool PANNZER2 (Protein ANNotation with Z-scoRE)(Törönen & Holm, 2022). From the results, the columns containing the GO terms ID and gene ID were extracted into a new file, which was then used as the input for an R script (LabBCES, 2024) that uses the GOdb R package v3.17 (Carlson, 2019), which consists on a set of annotation maps that describe the Gene Ontology assembled from the GO data obtained previously.
We identified enriched GO terms associated with biological processes for the orthogroups classified as hard-core, soft-core, accessory and exclusive. For the analysis, the data was submitted to statistical tests, with the Classic Fisher's exact test , and the p-value was corrected for multiple testing using the BH method (Benjamini & Hochberg, 1995). These statistical parameters were implemented on the GO terms testing package topGO (Alexa & Rahnenfuhrer, 2023). The overrepresented Ontology terms sorted by adjusted p-values were taken into account, and the most representative biological processes were plotted.
REVIGO web server (Supek et al., 2011) was used to summarize the overrepresented biological processes derived from the enrichment analysis. This platform simplifies extensive GO term lists by selecting a representative subset through a clustering approach based on semantic similarity measures.

Results

1. Identification of Orthologous Groups

The analysis of domain architecture concordance for the orthogroups generated with different inflation values, realized with cogeqc, highlighted 4.4 as the proper inflation for this dataset (Figure 1). Briefly, the `assess_orthogroups` function of the cogeqc package was utilized to cross-examine the domain annotations of genes within each orthogroup against the set of domain annotations derived from individual genomes. For each specified inflation value, ranging from 1.2 to 6.0 with increments of 0.4, the orthogroup data was assessed, and a mean consistency score was computed.Summary statistics for the orthogroups with different inflation values are shown in Table 2.

Inflation Value	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0	4.4	4.8	5.2	5.6	6.0
# genes in OGs	446590	446590	446590	446572	446414	445996	445623	445250	444887	444548	444215	443849	443517
# unassigned genes	9719	9719	9719	9737	9895	10313	10686	11059	11422	11761	12094	12460	12792
% genes in OGs	97.9	97.9	97.9	97.9	97.8	97.7	97.7	97.6	97.5	97.4	97.3	97.3	97.2
% unassigned genes	2.1	2.1	2.1	2.1	2.2	2.3	2.3	2.4	2.5	2.6	2.7	2.7	2.8
# OGs	26715	25206	27143	27406	27607	27798	27939	28095	28215	28330	28403	28487	28557
# species-specific OGs	39	30	65	81	118	171	184	226	249	268	286	301	306

# genes in species-specific OGs	179	60	539	617	697	786	775	812	822	811	811	825	829
% genes in species-specific OGs	0.0	0.0	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Mean OGs size	16.7	17.7	16.5	16.3	16.2	16.0	15.9	15.8	15.8	15.7	15.6	15.6	15.5
Median OGs size	22.0	22.0	22.0	22.0	22.0	22.0	22.0	22.0	22.0	21.0	21.0	21.0	21.0
# OGs with all species present	14127	13965	14061	14027	13992	13964	13950	13929	13908	13886	13876	13859	13848
# of single-copy OGs	13766	13354	13775	13779	13783	13782	13781	13781	13778	13766	13763	13751	13745

Table 2. Set of values given for the parameter “inflation” of the MCL clustering algorithm on OrthoFinder. The highlighted column was selected as the optimal inflation value by the cogeqc R tool and was used for further analysis.

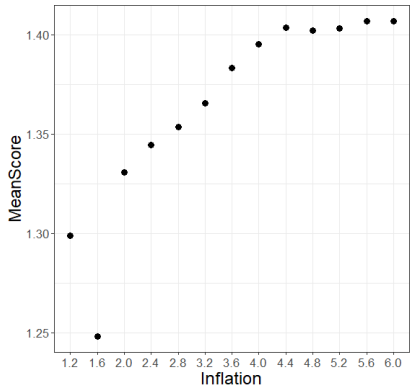


Figure 1. Mean Consistency Score Across Varying Inflation Values. The graph displays the relationship between the established inflation values (x-axis) and the corresponding mean consistency scores (y-axis). Each point on the plot represents the mean score for a specific inflation value.

We had 456,309 protein-coding genes for the 22 isolates of *M. roreri* and according to the OrthoFinder analysis (Table 2 and 3), 97.5% of these could be assigned to orthogroups, i.e., 444,887. With that, the hard-core genome of *M. roreri* comprises 307,081 protein-coding genes in 13,908 groups (69,02% of the total protein-coding genes). These genes, as defined earlier, are accounted for on 100% of the isolates, suggesting that they are essential for the basic biological processes of the pathogen (see enrichment analysis below). The size of the soft-core (i.e., number of groups present in at least 90% of the isolates) is 15,145, with 333,880 protein-coding genes. The accessory group is made up of 110,185 protein-coding genes distributed in 12,821 groups, i.e., groups present in more than one strain but less than 90% of the strains. Finally, 249 exclusive groups were found, with 822 protein-coding genes in total. These genes are strain-specific and can be related to the isolate's most recent adaptations, conferring specific advantages or playing roles in unique interactions with the host or environment.

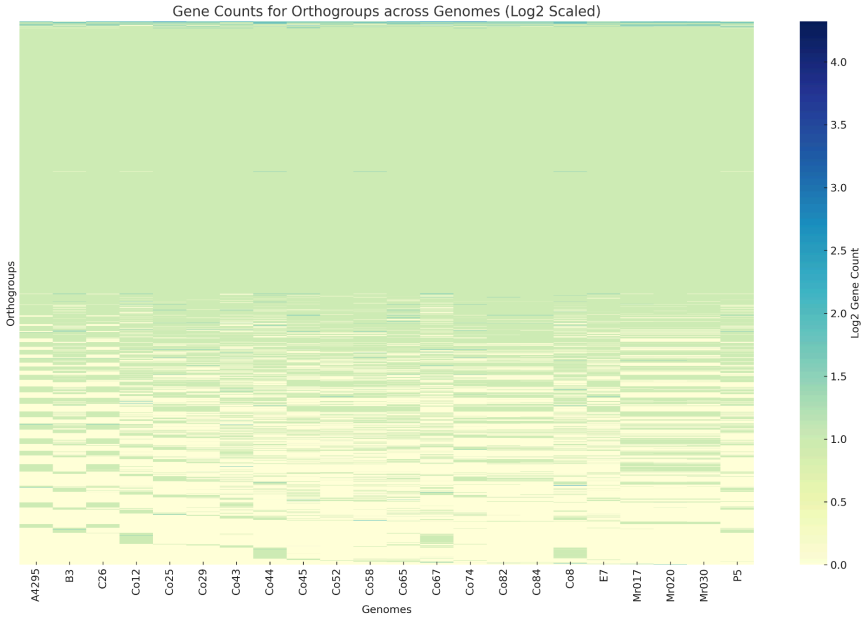


Figure 2. Heatmap of Gene Counts in Orthogroups Across Genomes. This visualization presents the distribution of gene counts for various orthogroups (Y-axis) across multiple genomes (X-axis). To provide a clearer representation and accommodate the range of gene counts, a logarithmic (Log2) scale was employed, with the color intensity corresponding to the Log2-transformed gene counts. The color scale bar ranges from 0.0 to 4.0, indicating the variation in gene counts. Darker shades denote higher gene counts, offering insights into the distribution and abundance of genes across the orthogroups for different genomes.

Moniliophthora roreri genomic content statistics evaluated in 22 isolates	
Statistic	**Value**
Number of genes present in pangenome	456309
Number of orthogroups	28215
Number of genes present in orthogroups	444887
Number of core groups	13908
Number of proteins present in core groups	307081
Number of groups present in 90% of the species (19.8)	15145
Number of proteins present in soft-core groups	333880
Number of accessory groups	12821
Number of proteins present in accessory groups	110185
Number of exclusive groups	249
Number of proteins present in exclusive groups	822

Table 3. General statistics of the pan-genome of *M. roreri*. The table contains the number of orthogroups for each class (pan, hard-core, soft-core, accessory and exclusive) and the number of protein-coding genes assigned in each group.

One question that often arises is whether the sample size is enough to capture the genomic diversity of the organism under study, or whether there is not enough sampling effort that could capture the genomic diversity of the organism. In order to assess that, the number of groups obtained from the orthogroups extraction step (inflation 4.4) was plotted against the number of genotypes (Figure 3), this allowed us to analyze the fluctuation of the number of groups detected as a function of the increasing number of genotypes. These trajectories were calculated for each pangenome classification (hard-core, soft-core, accessory and exclusive) and for the entire pangenome.

Our analysis of the pangenome trajectory curve revealed key insights into the pangenome's dynamics. The slopes of the pangenome curve, calculated using the first derivative for each point on the x-axis representing the number of genotypes, provided a perspective on the rate of change in the number of orthogroups (Figure 4). The point where the derivative approached zero was of particular interest. This point corresponds to the number of genotypes/strains capturing the majority of the species' genomic diversity, suggesting that the inclusion of additional genomes does not introduce new orthogroups, thus not significantly enhancing the overall diversity.

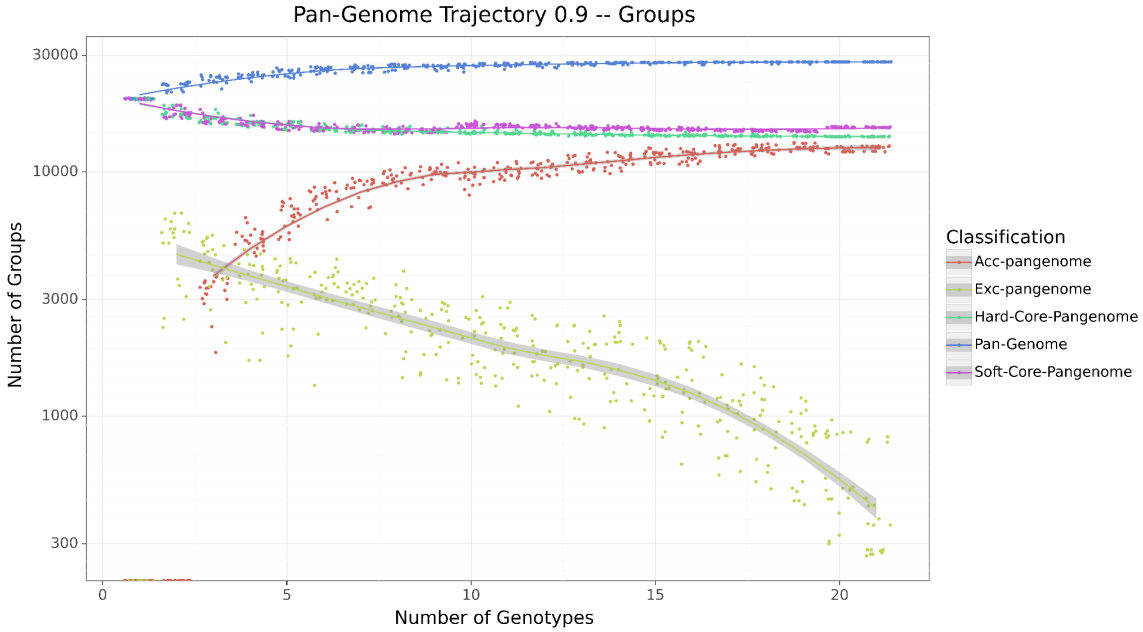


Figure 3. The figure visualizes the trajectory of pangenome groups of *M.roreri* as the number of analyzed genotypes increases. The x-axis represents the number of genotypes, with a total of 22 genotypes studied. The y-axis denotes the number of orthogroups identified by OrthoFinder. Five distinct lines on the graph correspond to the different classifications of the genome: hard-core

(cyan), soft-core (purple), accessory (red), exclusive (green), and the pangenome itself (blue). As we move along the x-axis, each increment signifies the inclusion of a new genotype in the analysis, and the corresponding y-values represent the fluctuation in the number of groups for each classification. Notably, for each x-value (each genotype), there are 20 data points on the graph. This structure arises because the analysis considers a pool of 20 genotypes at each step, quantifying the number of groups present within that pool. This process is iteratively performed 22 times to encompass all 22 genotypes in the study.

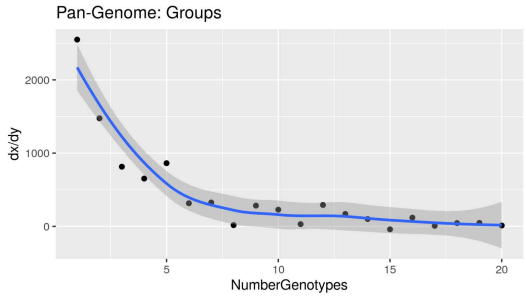


Figure 4. Derivative Analysis of the Pangenome Trajectory. The plot visualizes the first derivative values of the pangenome trajectory concerning the cumulative number of genotypes included in the pangenome analysis. On the x-axis, we have the "NumberGenotypes", representing the sequential addition of genomes, while the y-axis displays the corresponding "DerivativeValues", indicating the rate of change in the number of gene groups. Each point on the plot corresponds to a genome and its impact on the expansion or contraction of the pan-genome. A higher derivative value suggests that the addition of that particular genome contributed to a rapid increase in unique gene groups, whereas a lower or negative value indicates a relative saturation or decrease in the unique gene groups. The smooth line represents a loess regression, providing a trend of how the pan-genome evolves as more genomes are incorporated.

Number of Genotypes	1st Derivative Values	Number of Genotypes	1st Derivative Values
1	2565.00	11	129.50
2	1407.8	12	215.15
3	917.50	13	1.15
4	744.40	14	221.10
5	593.85	15	-19.35
6	393.45	16	62.65
7	258.40	17	60.80
8	187.40	18	85.20
9	282.15	19	12.50
10	147.05	20	15.05

Table 4. Derivative values of the pangenome curve. The table presents two columns: "Number of Genotypes" and "Derivative Values". The "Number of Genotypes" column represents the cumulative number of genotypes added to the pangenome analysis. The "Derivative Values" column provides the calculated first derivative values of the pan-genome groups concerning the number of genotypes, indicating the rate of change in the number of gene groups as more genomes are

incorporated into the analysis. Higher derivative values suggest a rapid increase in the number of gene groups with the addition of genomes, while lower or negative values denote a deceleration or decline, respectively.

2. Functional Annotation and enrichment analysis

- Functional Annotation

The table 5 was crafted using a custom python script (Gallego, 2023). The table contains information regarding the functional annotation of each of the 22 genomes used for the pangenome construction.

Genome Name	Total Annotated Genes	Genes with Pfam Domains	Distinct Pfam Domains	Genes with GO terms	Distinct GO terms
Mr017	20384	14217	12487	12157	5128
B3	4408	3100	4468	2643	3216
Co12	20413	14159	12490	12094	5127
Mr020	20404	14223	12490	12157	5131
Co29	20431	14221	12465	12139	5141
Co82	20221	14075	12460	12001	5141
Co74	20445	14225	12498	12116	5148
Co44	20703	14359	12522	12263	5118
Co67	20546	14254	12486	12190	5123
Co43	20780	14472	12496	12352	5130
Co84	20287	14126	12503	12042	5146
Mr030	20393	14230	12493	12163	5128
Co58	3768	2532	3246	2157	2620
P5	2497	1759	2473	1490	2364
A4295	20501	14264	12473	12186	5147
Co52	20165	14081	12474	12044	5122
E7	21815	15614	15424	13488	6105
C26	20471	14262	12471	12177	5147
Co65	20440	14142	12476	12083	5145
Co8	21200	14644	12531	12494	5119
Co25	20451	14210	12499	12113	5148
Co45	20288	14092	12472	12028	5133

Table 5. Annotation metrics for each genome utilized in the study. The table enumerates several key metrics for each genome, including the genome name or identifier, the total number of annotated genes, the number of genes associated with Pfam protein domains, the unique count of Pfam protein domains identified, the count of genes annotated with Gene Ontology (GO) terms, and

the unique count of GO terms associated with the genes.

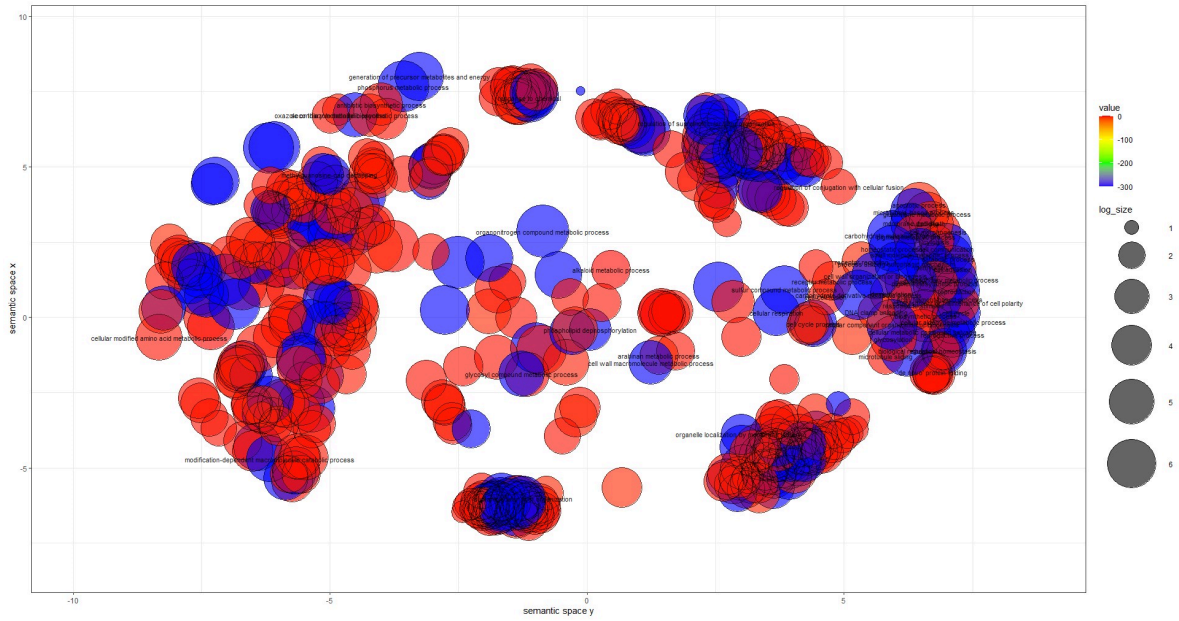
- **Enrichment analysis**

To provide a comprehensive understanding of the specialized roles of different gene sets within the *M. roreri* pangenome, a functional enrichment analysis was conducted for each gene classification—hard-core, soft-core, accessory, and exclusive. This analysis elucidated the biological processes or functions that are significantly overrepresented within each category.

Enrichment of GO term - Hard-core genes

Hard-core gene GO term overrepresentation analysis indicated that these genes, universally present across all examined *M. roreri* samples, showed significant enrichment in a variety of biological processes. These processes span from RNA modifications and cellular localization mechanisms to metabolic activities related to cell wall synthesis, chemical responses, structural organization, and secondary metabolite production (Figure 5).

a)



b)

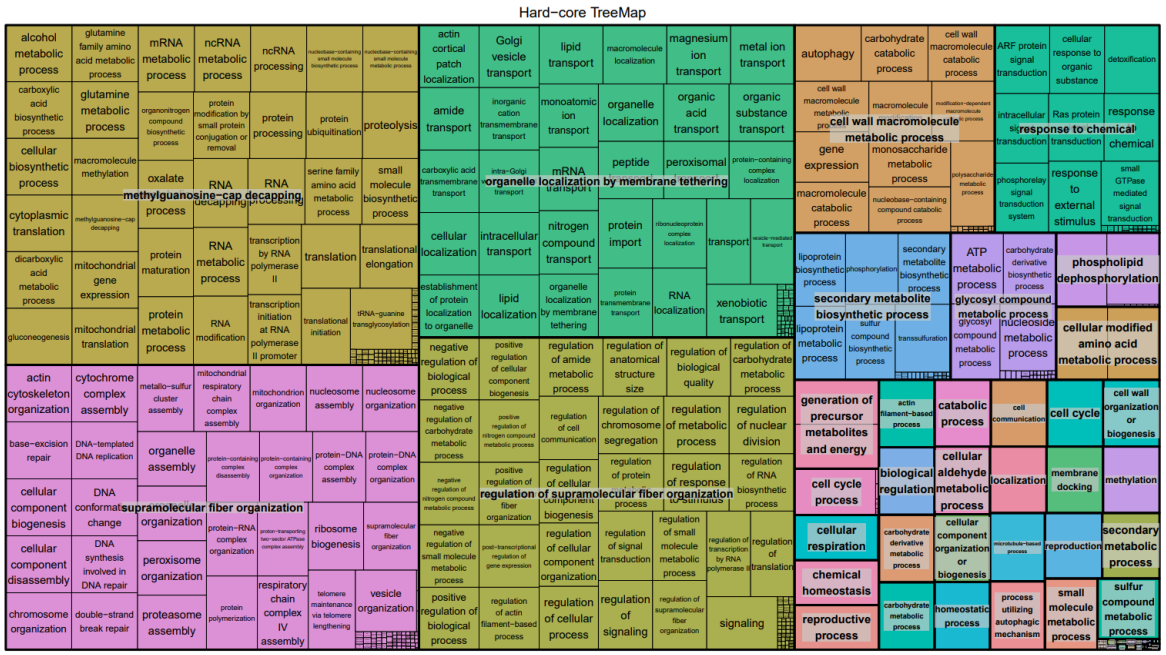
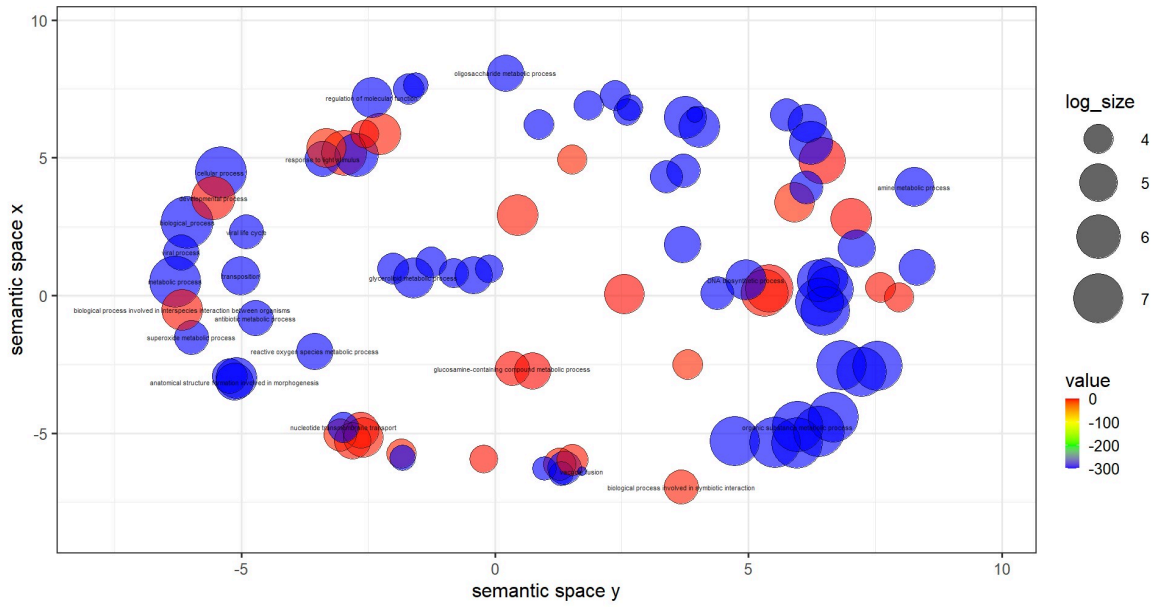


Figure 5. Overrepresented biological processes for the hard-core group. These figures showcase the enrichment analysis results for the hard-core category, composed of two distinct visual representations labeled as (a) and (b). **a)** Scatterplot to visually map out cluster representatives in a two-dimensional space, derived from the application of multidimensional scaling to a matrix showcasing the semantic similarities of GO terms. The axes are non informative. The color variation represents the p-value (>=0.05) assigned to each GO term on the enrichment analysis. The size of the bubbles in the plot is given by the frequency of the GO term in the Gene Ontology Annotation database. Larger bubbles represent a general GO term, while smaller ones indicate more specific or specialized terms. **b)** Tree Map visualization depicting the hierarchy of the enriched GO terms of the dataset. Each square in this plot signifies the high-level groups according to their structure and abundance.

Enrichment of GO term - Accessory genes

Accessory gene enrichment analysis revealed that these genes, present in certain *M. roreri* strains but not universally across all 22 examined strains, demonstrated significant enrichment predominantly in various metabolic processes, cellular activities, and DNA-related mechanisms. Notably, these enriched categories spanned both primary and specialized metabolic pathways, cellular functionalities, and potentially genomic adaptability, as illustrated in Figure 6.

a)



b)

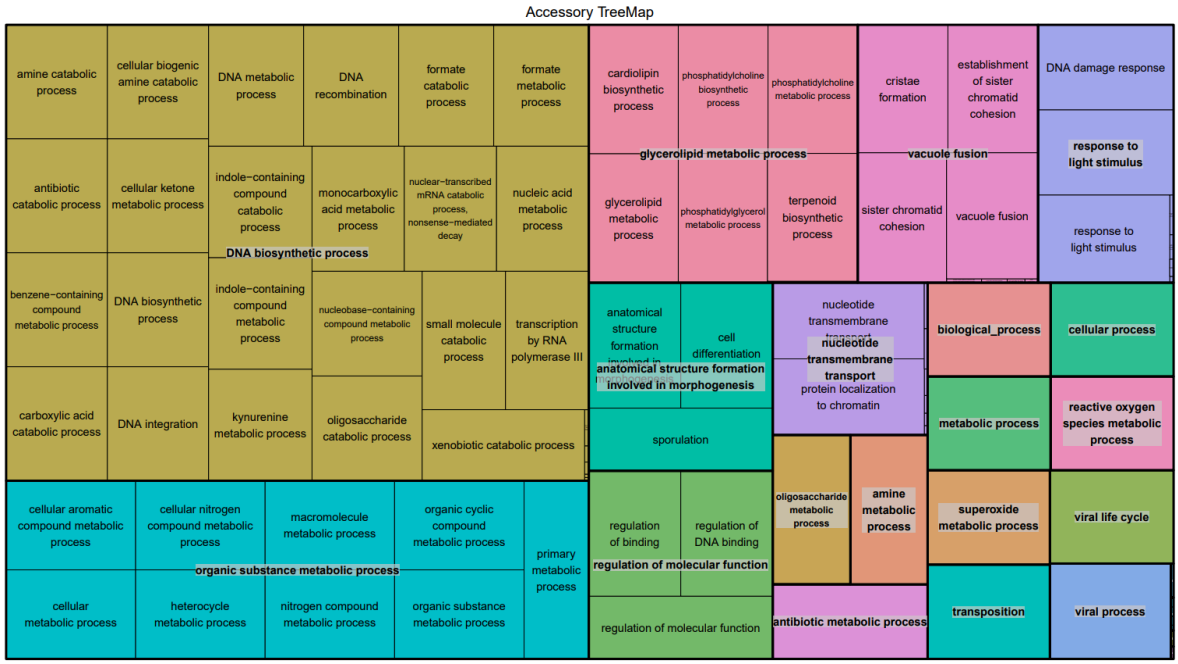


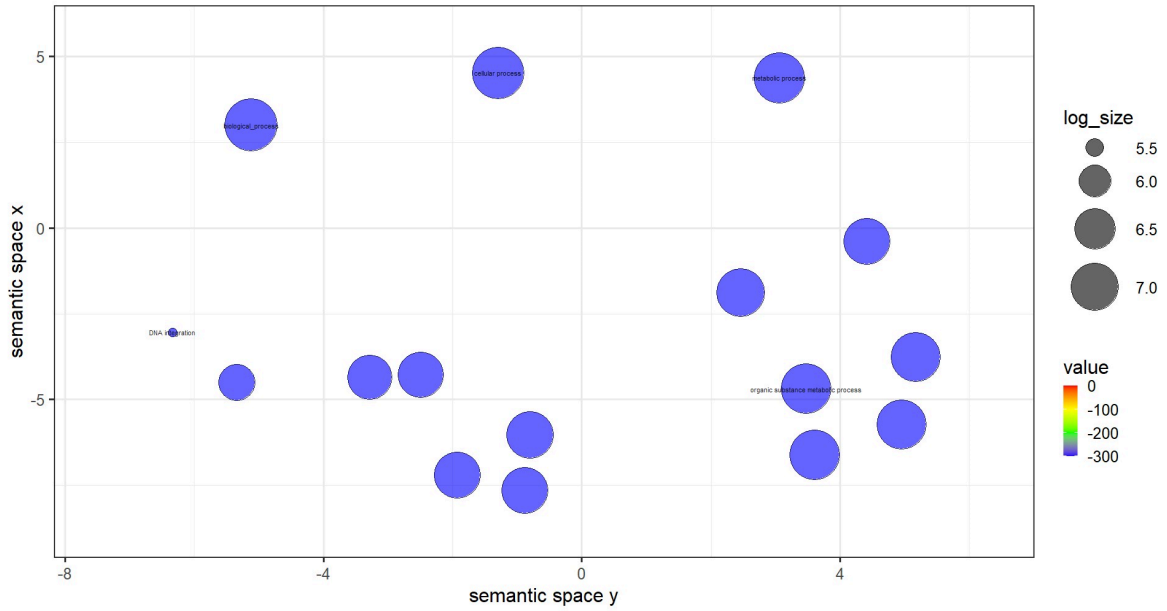
Figure 6. Overrepresented biological processes for the accessory group. These figures showcase the enrichment analysis results for the accessory category, composed of two distinct visual representations labeled as (a) and (b). **a)** Scatterplot to visually map out cluster representatives in a two-dimensional space, derived from the application of multidimensional scaling to a matrix showcasing the semantic similarities of GO terms. The axes are non informative. The color variation represents the p-value (>=0.05) assigned to each GO term on the enrichment

analysis. The size of the bubbles in the plot is given by the frequency of the GO term in the Gene Ontology Annotation database. Larger bubbles represent a general GO term, while smaller ones indicate more specific or specialized terms. **b)** Tree Map visualization depicting the hierarchy of the enriched GO terms of the dataset. Each square in this plot signifies the high-level groups according to their structure and abundance.

Enrichment of GO term - Exclusive genes

The exclusive genes, unique to specific strains of the pathogen and not universally present across the analyzed strains, displayed significant enrichment in several metabolic and cellular processes. Specifically, these genes were found to be enriched in categories regarding macromolecule metabolism, DNA integration, organic substances processing mechanisms, among other biological processes (Figure 7).

a)



b)

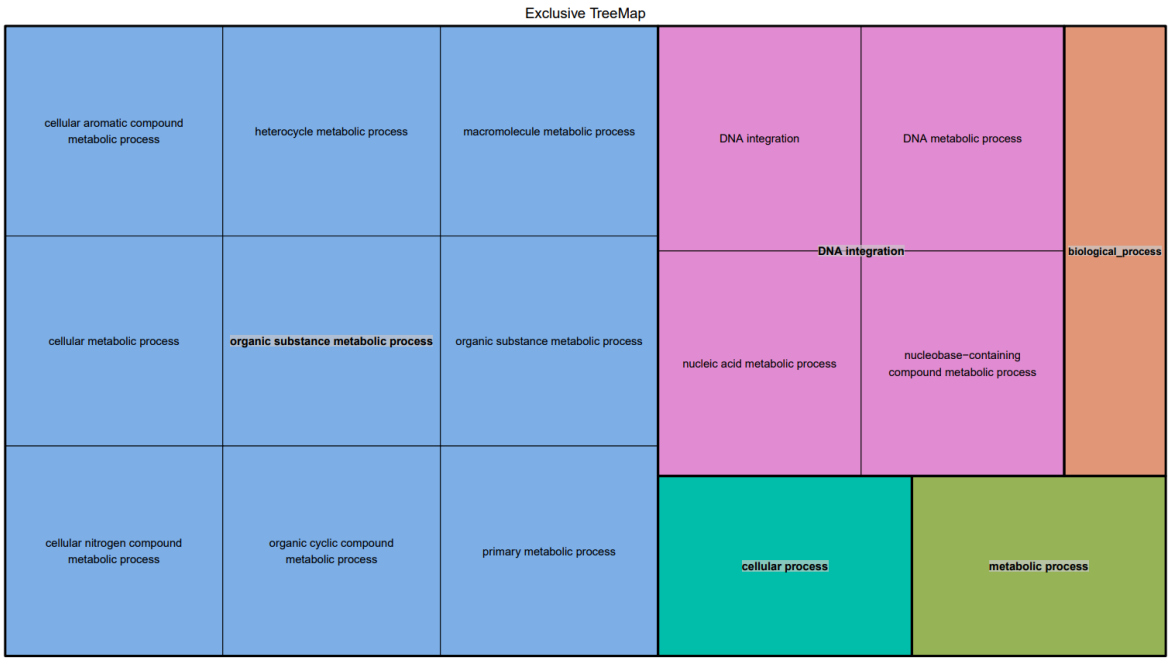


Figure 7. Overrepresented biological processes for the exclusive group. These figures showcase the enrichment analysis results for the exclusive category, composed of two distinct visual representations labeled as (a) and (b). **a)** Scatterplot to visually map out cluster representatives in a two-dimensional space, derived from the application of multidimensional scaling to a matrix showcasing the semantic similarities of GO terms. The axes are non informative. The color variation represents the p-value (>=0.05) assigned to each GO term on the enrichment analysis. The size of the bubbles in the plot is given by the frequency of the GO term in the Gene Ontology Annotation database. Larger bubbles represent a general GO term, while smaller ones indicate more specific or specialized terms. **b)** Tree Map visualization depicting the hierarchy of the enriched GO terms of the dataset. Each square in this plot signifies the high-level groups according to their structure and abundance.

Discussion

We constructed a global pangenome for the cacao pathogen *Moniliophthora roreri* using 22 publicly available genomes. From our analysis, we identified 456,309 protein-coding genes, of which 97.5% were assigned to orthogroups (Figure 2). Further functional annotation associated many of these genes with specific Gene Ontology terms. By categorizing the pangenome into hard-core, soft-core, accessory, and exclusive groups, we observed a range of distinct biological processes that were overrepresented in the GO term enrichment analysis. These findings provide a clearer understanding of the genomic characteristics and potential functionalities of *M. roreri*.

The pangenome of an organism provides an overview of its entire genetic makeup, including genes that might be present in some strains but absent in others. The

categorizations, namely hard-core, soft-core, accessory, and exclusive genes, each provide distinct insights.

Hard-core genes represent the most conserved genetic elements across all strains. These genes might be crucial for the basic biological functions of the fungus (McCarthy & Fitzpatrick, 2019). Soft-core genes, being present in over 90% of strains, might hint at functions that are beneficial but not absolutely essential (Agarwal et al., 2023). In some cases, the implementation of the soft core approach primarily aids in addressing the challenges posed by incomplete genomes, a consequence of technical limitations. These genes might be recent additions to the genome or older genes undergoing a phase of loss in a few strains.

In our analysis of the core genes of *M. roreri,* several GO terms emerged as significantly enriched, shedding light on the central and conserved roles these genes might play a crucial role in the organism's biology. Notably, the enrichment of 'Methylguanosine−cap Decapping' suggests an essential role in mRNA processing and stability, indicative of robust co-transcriptional regulatory mechanisms (Cowling, 2010). The 'Organelle Localization by Membrane Tethering' terms point towards an organelle positioning and dynamics regulating system through membrane contact sites (MCSs) (Bohnert, 2020). This system is essential for many cellular processes, including metabolism, signaling and cell division (Huang et al., 2020).

The enrichment in 'Cell Wall Macromolecule Metabolic Process' GO term reveals an important set of genes involved in the chemical reactions and pathways that form part of the cell wall macromolecules, associated to cell division, signaling and nutrient uptake (*QuickGO::Term GO:0044036*, n.d.). Previous studies have shown that numerous proteins displaying varied expression between the biotrophic and necrotrophic stages are linked to cell surface alterations or to the cell surface itself (Minio et al., 2023; Vasconcelos et al., 2021), the findings of the present study reinforce those results.

Another overrepresented term on the core genome 'Response to Chemical' is related to the mechanisms underlying the pathogen's response to chemical stress signals, including reactive oxygen species (ROS), toxins or antifungal agents on the environment (Simaan et al., 2019). Further diving into these genes could provide insights into the molecular regulation mechanisms of *M. roreri* to chemical stress signals and their role in its pathogenic lifestyle.

The 'Secondary Metabolite Biosynthetic Process' GO term enrichment in *Moniliophthora roreri'*s core genome doesn't reveal any essential mechanism for the pathogens growth or development, however, the genes related to this biological process play important roles in ecological interactions, including pathogenicity (Minio et al., 2023).

Accessory genes (Tettelin et al., 2005) provide insights into the diverse functionalities that different strains might have acquired or lost over evolutionary time. These genes could be responsible for variations in virulence, drug resistance, or ecological adaptability among different strains. In plant pathogens, the study of accessory genes has garnered significant attention within the research community due to their potential to harbor genes that modulate interactions with the host, including crucial elements such as effectors and metabolic gene clusters (Bertazzoni et al., 2018).

Among the overrepresented GO terms associated with the accessory genes, the enrichment of both 'Cellular aromatic compound metabolic process' and 'Heterocycle metabolic process' suggests a potential capability for the breakdown and utilization of complex organic compounds, which could be crucial in specialized ecological niches or during specific stages of host infection. In previous fungal pathogens studies (Shalaby et al., 2012), it has been suggested that the inhibition of the fungal sensory pathways to detect aromatic compounds of the host could serve as a control strategy for plant diseases. Therefore, the identification of these genes in the pathogen in the current study provides a foundational knowledge base that could facilitate the development of such control strategies.

It is recognized that variation among genotypes arises from gene acquisition, facilitated by multiple molecular processes. However, certain pathogens, including some in the eukaryotic realm, exhibit a slower rate of gene gain. For these pathogens, diversity is often shaped by gene losses, leading to distinct attributes. This interplay between gene acquisition and loss shapes the pangenome of a species (Brockhurst et al., 2019). While eukaryotes predominantly do not evolve through horizontal gene transfer (HGT), exceptions exist, especially when the source of genes is bacteria or plant organellar genomes (Wang et al., 2020). The acquisition of genes via horizontal transfer can lead to rapid adaptations, which might include enhanced pathogenicity or adaptability to specific hosts or environments.

In this context, a study by Tiburcio et al., provides an insightful example. They observed that only a few species within the *Moniliophthora* genus exhibit traits related to plant pathogenicity. This suggests that the evolution of these traits in species like *M. roreri* is not solely based on modifications to pre-existing genes. Instead, there's a possibility that such species have incorporated novel genes associated with plant pathogenicity through HGT mechanisms. This could mean that the pathogen has acquired genes (perhaps from other pathogens or environmental microbes) that enhance its ability to infect cacao plants or resist plant defenses.

The presence of ‘DNA integration mechanisms’, as suggested by the enrichment of the corresponding GO terms in both accessory and exclusive genomes (Figures 6 and 7), might indicate an evolutionary strategy employed by *Moniliophthora roreri*. This strategy could involve frequent interactions with other microbes, either in the soil or on the cacao

plant, leading to gene exchange events. Such interactions could be a rich area of investigation, offering insights into the ecological dynamics that drive the evolution of this pathogen.

The prominence of broader terms such as 'Macromolecule metabolic process,' 'Cellular metabolic process,' and 'Organic substance metabolic process' underscores the accessory genome's involvement in diverse and dynamic metabolic activities, potentially augmenting the core metabolism under certain conditions (McCarthy & Fitzpatrick, 2019).

The enrichment of both 'Cellular aromatic compound metabolic process' and 'Heterocycle metabolic process' suggests a potential capability for the breakdown and utilization of complex organic compounds, which could be crucial in specialized ecological niches or during specific stages of host infection. Broad terms like 'Biological process,' 'Cellular process,' and 'Metabolic process' further emphasize the diverse biological functions encapsulated within the accessory genes.

According to (Brockhurst et al., 2019), pangenomes can be classified as open or closed, depending on the proportion of core versus accessory gene content. The accessory genome of *Moniliophthora roreri* constitutes approximately 24.13% of its entire pangenome, which would enter into the closed pangenome definition proposed by the mentioned study. This type of pangenomes are usually found among niche specialist pathogens, however, taking into account the broad range of hosts affected by *M.roreri* (Díaz-Valderrama et al., 2022), this doesn't seem to fit into the described pathogenic mechanism of this specific species.

On the other hand, mathematical models have demonstrated that the pangenome measurement is ruled by the Heap’s law, which explains the behavior of the pangenome trajectory curve in reference to the number of new genes discovered as new genome sequences are added to the analysis. According to the function proposed in this definition, a closed pangenome is considered when the pangenome size curve reaches a plateau as more genomes are added, meaning that a significant proportion of the species diversity has been captured. Alternatively, when the trajectory of the curve continues growing with the addition of more genomes, an open pangenome is considered (Richard, 2020).

Upon examining the pangenomic trajectory of *M. roreri* (Figure 3), a stabilization trend is discernible around the tenth genotype. However, when assessing the curve's slope using its first derivative (Figure 4), it is evident that the curve does not reach zero, suggesting an open pangenome. Yet, considering the values obtained for the first derivative (Table 4), one could argue that the curve approaches a stabilization point closely. This implies that the number of genomes utilized in constructing the pangenome captures a significant portion of the species' overall diversity. Nonetheless, this also indicates that additional strains are required to fully encompass the complete diversity.

Further analysis regarding the accessory genome of the species would provide insights into the host range and pathogen adaptability.

Collectively, these findings suggest that *M. roreri*'s accessory genome equips it with supplementary metabolic and cellular capabilities, as well as a range of DNA integration mechanisms potentially conferring advantages in specific environmental contexts or host interactions, in addition to potentially harboring the genes related to the pathogens host range and its ability to infect alternative hosts in order to persist in cacao plantations.

Exclusive genes offer the most direct clues about strain-specific adaptations. For instance, a strain-specific gene could be responsible for increased virulence in a particular environmental condition or resistance against a specific fungicide (Gong et al., 2023).

A deeper examination of the enriched 'Biological Process' GO term among the exclusive genome revealed associated 'child terms' related to the disruption of anatomical structures in other organisms, leading to damage or temporary subversion of that structure. During the infection cycle of *M. roreri*, there is a stage where spores germinate, forming a germ tube that penetrates the epidermis or stomata of the host, initiating the infection cycle (Jiménez et al., 2022). This infection process could be investigated in the context of the mentioned genes, providing pertinent insights into the pathogenic mechanisms of the fungus.

Conclusions

The comprehensive pangenomic analysis of the cacao pathogen, *Moniliophthora roreri*, has unveiled critical insights into its genetic composition and potential functional capabilities. With the identification of 456,309 protein-coding genes across the 22 studied genomes, 97.5% were classified into orthogroups. This high degree of orthogroup assignment underscores the shared genetic heritage among the studied strains.

The subsequent categorization of the pangenome into hard-core, soft-core, accessory, and exclusive genes has paved the way for a better understanding of the organism's genetic variability. Notably, the presence of hard-core genes across all studied strains accentuates their pivotal role in the fundamental biological processes of the fungus. In contrast, the accessory and exclusive genes, while present in a proportion of the strains, suggest functions that are not indispensable, but favor the pathogens infection mechanisms and environmental adaptation systems.

Delving deeper into the functional annotations, the Gene Ontology (GO) term enrichment analysis was instrumental in spotlighting genes associated with different biological processes. Particularly, the identification of genes linked to the pathogenicity and adaptability of *M. roreri* can serve as a foundation for future research.

In conclusion, this study offers a general view of the genetic and functional landscape of *M. roreri*, laying a robust groundwork for subsequent research in plant-pathogen interactions. The insights aim not only enhance our understanding of the pathogen but also highlight potential starting points for devising targeted interventions in cacao disease management. As the global cacao industry struggles with the challenges posed by pathogens like *M. roreri*, such in-depth genomic studies become indispensable in charting a path towards sustainable and resilient cacao production.

References

Agarwal, V., Stubits, R., Nassrullah, Z., & Dillon, M. M. (2023). Pangenome insights into the diversification and disease specificity of worldwide Xanthomonas outbreaks. *Frontiers in Microbiology*, *14*, 1213261. https://doi.org/10.3389/fmicb.2023.1213261

Aikpokpodion, P. (2019). *Theobroma Cacao: Deploying Science for Sustainability of Global Cocoa Economy*. BoD – Books on Demand.

Alexa, A., & Rahnenfuhrer, J. (2023). *topGO: Enrichment Analysis for Gene Ontology* (2.54.0) [Computer software]. Bioconductor version: Release (3.18). https://doi.org/10.18129/B9.bioc.topGO

Ali, S. S., Shao, J., Strem, M. D., Phillips-Mora, W., Zhang, D., Meinhardt, L. W., & Bailey, B. A. (2015). Combination of RNAseq and SNP nanofluidic array reveals the center of genetic diversity of cacao pathogen Moniliophthora roreri in the upper Magdalena Valley of Colombia and its clonality. *Frontiers in Microbiology*, *6*. https://www.frontiersin.org/articles/10.3389/fmicb.2015.00850

Almeida-Silva, F., & Peer, Y. V. de. (2023). *cogeqc: Systematic quality checks on comparative genomics analyses* (1.6.0) [Computer software]. Bioconductor version: Release (3.18). https://doi.org/10.18129/B9.bioc.cogeqc

Amir, R., Sani, Q.-A., Maqsood, W., Munir, F., Fatima, N., Siddiqa, A., & Ahmad, J. (2020). Chapter 6—Pan-genomics of plant pathogens and its applications. In D. Barh, S. Soares, S. Tiwari, & V. Azevedo (Eds.), *Pan-genomics: Applications, Challenges, and Future Prospects* (pp. 121–145). Academic Press. https://doi.org/10.1016/B978-0-12-817076-2.00006-8

Badet, T., & Croll, D. (2020). The rise and fall of genes: Origins and functions of plant pathogen pangenomes. *Current Opinion in Plant Biology*, *56*, 65–73.

https://doi.org/10.1016/j.pbi.2020.04.009

Bailey, B. A., Evans, H. C., Phillips-Mora, W., Ali, S. S., & Meinhardt, L. W. (2018). Moniliophthora roreri, causal agent of cacao frosty pod rot. *Molecular Plant Pathology*, *19*(7), 1580–1594. https://doi.org/10.1111/mpp.12648

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bertazzoni, S., Williams, A. H., Jones, D. A., Syme, R. A., Tan, K.-C., & Hane, J. K. (2018). Accessories Make the Outfit: Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. *Molecular Plant-Microbe Interactions*®, *31*(8), 779–788. https://doi.org/10.1094/MPMI-06-17-0135-FI

Bohnert, M. (2020). Tether Me, Tether Me Not—Dynamic Organelle Contact Sites in Metabolic Rewiring. *Developmental Cell*, *54*(2), 212–225. https://doi.org/10.1016/j.devcel.2020.06.026

Brockhurst, M. A., Harrison, E., Hall, J. P. J., Richards, T., McNally, A., & MacLean, C. (2019). The Ecology and Evolution of Pangenomes. *Current Biology*, *29*(20), R1094–R1103. https://doi.org/10.1016/j.cub.2019.08.012

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), Article 4. https://doi.org/10.1038/s41592-021-01101-x

Carlson, M. (2019). GO.db: A set of annotation maps describing the entire Gene Ontology. *Bioconductor*. http://bioconductor.org/packages/GO.db/

Cowling, V. H. (2010). Regulation of mRNA cap methylation. *Biochemical Journal*, *425*(Pt 2), 295–302. https://doi.org/10.1042/BJ20091352

Cubillos, G. (2017). Frosty Pod Rot, disease that affects the cocoa (Theobroma cacao) crops in Colombia. *Crop Protection*, *96*, 77–82. https://doi.org/10.1016/j.cropro.2017.01.011

Daboussi, M.-J., & Capy, P. (2003). Transposable Elements in Filamentous Fungi. *Annual Review of Microbiology*, *57*(1), 275–299. https://doi.org/10.1146/annurev.micro.57.030502.091029

de Souza, P. A., Moreira, L. F., Sarmento, D. H. A., & da Costa, F. B. (2018). Cacao—Theobroma cacao. In S. Rodrigues, E. de Oliveira Silva, & E. S. de Brito (Eds.), *Exotic Fruits* (pp. 69–76). Academic Press. https://doi.org/10.1016/B978-0-12-803138-4.00010-1

Díaz-Valderrama, J. R., & Aime, M. C. (2016). The cacao pathogen Moniliophthora roreri (Marasmiaceae) possesses biallelic A and B mating loci but reproduces clonally. *Heredity*, *116*(6), Article 6. https://doi.org/10.1038/hdy.2016.5

Díaz-Valderrama, J. R., Rubio-Rojas, K. B., Fernández, N. B., Kijpornyongpan, T., Phillips-Mora, W., Weil, C. F., Gribskov, M., & Aime, M. C. (2023). Genome sequence of the cacao pathogen Moniliophthora roreri belonging to the invasive A1B1 mating type: A resource for genomic efforts in cacao pathology. *PhytoFrontiers™*. https://doi.org/10.1094/PHYTOFR-11-22-0128-A

Díaz-Valderrama, J. R., Zambrano, R., Cedeño-Amador, S., Córdova-Bermejo, U., Casas, G. G., García-Zurita, N., Sánchez-Arévalo, J. A. J., Arévalo-Gardini, E., Dávila, D., Ruiz, J., Pinchi-Dávila, X., Quispe-Chacón, Z. R., Chia-Wong, J. A., Hurtado-Gonzales, O. P., Rodríguez-Callañaupa, C. A., Maldonado-Fuentes, C., Pérez-Callizaya, E., Leiva-Espinoza, S., Oliva-Cruz, M., … Aime, M. C. (2022). Diversity in the invasive cacao pathogen Moniliophthora roreri is shaped by agriculture. *Plant Pathology*, *71*(8), 1721–1734. https://doi.org/10.1111/ppa.13603

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

Espinoza-Lozano, F., Amaya-Márquez, D., Pinto, C. M., Villavicencio-Vásquez, M., Sosa del Castillo, D., & Pérez-Martínez, S. (2022). Multiple Introductions of Moniliophthora roreri from the Amazon to the Pacific Region in Ecuador and Shared High Azoxystrobin Sensitivity. *Agronomy*, *12*(5), Article 5. https://doi.org/10.3390/agronomy12051119

Gallego, I. (2023). *M.roreri pangenome: Construction and analysis* [Computer software]. https://github.com/igallegor97/M.roreri_pangenome

Garcia, J. F., Morales-Cruz, A., Cochetel, N., Minio, A., Figueroa-Balderas, R., Rolshausen, P. E., Baumgartner, K., & Cantu, D. (2023). *Comparative pangenomic insights into the distinct evolution of virulence factors among grapevine trunk pathogens* (p. 2023.09.03.555958). bioRxiv. https://doi.org/10.1101/2023.09.03.555958

Gong, Y., Li, Y., Liu, X., Ma, Y., & Jiang, L. (2023). A review of the pangenome: How it affects our understanding of genomic variation, selection and breeding in domestic animals? *Journal of Animal Science and Biotechnology*, *14*(1), 73. https://doi.org/10.1186/s40104-023-00860-1

González-Orozco, C. E., Galán, A. A. S., Ramos, P. E., & Yockteng, R. (2020). Exploring the diversity and distribution of crop wild relatives of cacao (Theobroma cacao L.) in Colombia. *Genetic Resources and Crop Evolution*, *67*(8), 2071–2085. https://doi.org/10.1007/s10722-020-00960-1

Huang, X., Jiang, C., Yu, L., & Yang, A. (2020). Current and Emerging Approaches for Studying Inter-Organelle Membrane Contact Sites. *Frontiers in Cell and Developmental Biology*, *8*, 195. https://doi.org/10.3389/fcell.2020.00195

Jaimes, Y. Y., Gonzalez, C., Rojas, J., Cornejo, O. E., Mideros, M. F., Restrepo, S., Cilas, C., & Furtado, E. L. (2016). Geographic Differentiation and Population Genetic Structure of Moniliophthora roreri in the Principal Cocoa Production Areas in Colombia. *Plant Disease*, *100*(8), 1548–1558. https://doi.org/10.1094/PDIS-12-15-1498-RE

Jiménez, D. L., Alvarez, J. C., & Mosquera, S. (2022). Frosty pod rot: A major threat to cacao plantations on the move. *Tropical Plant Pathology*, *47*(2), 187–200. https://doi.org/10.1007/s40858-021-00472-y

McCarthy, C. G. P., & Fitzpatrick, D. A. (2019). Pan-genome analyses of model fungal species. *Microbial Genomics*, *5*(2), e000243. https://doi.org/10.1099/mgen.0.000243

Meinhardt, L. W., Costa, G. G. L., Thomazella, D. P., Teixeira, P. J. P., Carazzolle, M. F., Schuster, S. C., Carlson, J. E., Guiltinan, M. J., Mieczkowski, P., Farmer, A., Ramaraj, T., Crozier, J., Davis, R. E., Shao, J., Melnick, R. L., Pereira, G. A., & Bailey, B. A. (2014). Genome and secretome analysis of the hemibiotrophic fungal pathogen, Moniliophthora roreri, which causes frosty pod rot disease of cacao: Mechanisms of the biotrophic and necrotrophic phases. *BMC Genomics*, *15*(1), 164. https://doi.org/10.1186/1471-2164-15-164

Melo, B. L. B., de Souza, J. T., Santos, R. M. F., Rehner, S. A., Solis, K. H., Suarez, C., Hebbar, P. K., Lemos, L. S. L., & Gramacho, K. P. (2014). Development of microsatellites for the cacao frosty pod rot pathogen, Moniliophthora roreri. *Forest Pathology*, *44*(4), 320–324. https://doi.org/10.1111/efp.12103

Minio, A., & Cantu, D. (2024). *Moniliophthora roreri and Moniliophthora perniciosa data* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7872498

Minio, A., Figueroa-Balderas, R., Cohen, S. P., Ali, S. S., Carriel, D., Britto, D., Stack, C., Baruah, I. K., Marelli, J.-P., Cantu, D., & Bailey, B. A. (2023). Clonal reproduction of Moniliophthora roreri and the emergence of unique lineages with distinct genomes during range expansion. *G3 Genes|Genomes|Genetics*, *13*(9), jkad125. https://doi.org/10.1093/g3journal/jkad125

Nieves-Orduña, H. E., Krutovsky, K. V., & Gailing, O. (2023a). Geographic distribution, conservation, and genomic resources of cacao Theobroma cacao L. *Crop Science*, *63*(4), 1750–1778. https://doi.org/10.1002/csc2.20959

Nieves-Orduña, H. E., Krutovsky, K. V., & Gailing, O. (2023b). Geographic distribution, conservation, and genomic resources of cacao Theobroma cacao L. *Crop Science*, *63*(4), 1750–1778. https://doi.org/10.1002/csc2.20959

Ortega Andrade, S., Páez, G. T., Feria, T. P., & Muñoz, J. (2017). Climate change and the risk of spread of the fungus from the high mortality of Theobroma cocoa in Latin America. *Neotropical Biodiversity*, *3*(1), 30–40. https://doi.org/10.1080/23766808.2016.1266072

Pertea, G., & Pertea, M. (2020). *GFF Utilities: GffRead and GffCompare* (9:304). F1000Research. https://doi.org/10.12688/f1000research.23297.2

Phillips-Mora, W., Aime, M. C., & Wilkinson, M. J. (2007). Biodiversity and biogeography of the cacao (Theobroma cacao) pathogen Moniliophthora roreri in tropical America. *Plant Pathology*, *56*(6), 911–922. https://doi.org/10.1111/j.1365-3059.2007.01646.x

Phillips-Mora, W., & Wilkinson, M. J. (2007). Frosty pod of cacao: A disease with a limited geographic range but unlimited potential for damage. *Phytopathology*, *97*(12), 1644–1647. https://doi.org/10.1094/PHYTO-97-12-1644

QuickGO::Term GO:0044036. (n.d.). Retrieved October 28, 2023, from https://www.ebi.ac.uk/QuickGO/term/GO:0044036

Riaño-Pachón, D., & Vaz, F. (2023). *SCPT - SugarCane PanTranscriptome* [Computer software]. https://github.com/labbces/SCPT

Richard, G.-F. (2020). Eukaryotic Pangenomes. In H. Tettelin & D. Medini (Eds.), *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (pp. 253–291). Springer International Publishing. https://doi.org/10.1007/978-3-030-38281-0_12

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. In M. Kollmar (Ed.), *Gene Prediction: Methods and Protocols* (pp. 227–245). Springer. https://doi.org/10.1007/978-1-4939-9173-0_14

Shalaby, S., Horwitz, B. A., & Larkov, O. (2012). Structure–Activity Relationships Delineate How the Maize Pathogen *Cochliobolus heterostrophus* Uses Aromatic Compounds as Signals and Metabolites. *Molecular Plant-Microbe Interactions®*, *25*(7), 931–940. https://doi.org/10.1094/MPMI-01-12-0015-R

Simaan, H., Lev, S., & Horwitz, B. A. (2019). Oxidant-Sensing Pathways in the Responses of Fungal Pathogens to Chemical Stress Signals. *Frontiers in Microbiology*, *10*, 567. https://doi.org/10.3389/fmicb.2019.00567

Suárez-Contreras, L. Y. (2017). Diversidad genética de Moniliophthora roreri mediante Polimorfismo de Longitud de Fragmentos Amplificados (AFLPs). *Revista Colombiana de Ciencias Hortícolas*, *11*(2), 425–434. https://doi.org/10.17584/rcch.2017v11i2.7342

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE*, *6*(7), e21800. https://doi.org/10.1371/journal.pone.0021800

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli,

S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., … Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial “pan-genome.” *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955. https://doi.org/10.1073/pnas.0506758102

Tiburcio, R. A., Costa, G. G. L., Carazzolle, M. F., Mondego, J. M. C., Schuster, S. C., Carlson, J. E., Guiltinan, M. J., Bailey, B. A., Mieczkowski, P., Meinhardt, L. W., & Pereira, G. A. G. (2010). Genes Acquired by Horizontal Transfer Are Potentially Involved in the Evolution of Phytopathogenicity in Moniliophthora perniciosa and Moniliophthora roreri, Two of the Major Pathogens of Cacao. *Journal of Molecular Evolution*, *70*(1), 85–97. https://doi.org/10.1007/s00239-009-9311-9

Tirado-Gallego, P. A., Lopera-Álvarez, A., & Ríos-Osorio, L. A. (2016). *Estrategias de control de Moniliophthora roreri y Moniliophthora perniciosa en Theobroma cacao L.: Revisión sistemática*. http://www.scielo.org.co/scielo.php?pid=s0122-87062016000300009&script=sci_arttext

Törönen, P., & Holm, L. (2022). PANNZER—A practical tool for protein function prediction. *Protein Science*, *31*(1), 118–128. https://doi.org/10.1002/pro.4193

Vasconcelos, A. A., José, J., Tokimatu, P. M., Camargo, A. P., Teixeira, P. J. P. L., Thomazella, D. P. T., do Prado, P. F. V., Fiorin, G. L., Costa, J. L., Figueira, A., Carazzolle, M. F., Pereira, G. A. G., & Baroni, R. M. (2021). Adaptive evolution of Moniliophthora PR-1 proteins towards its pathogenic lifestyle. *BMC Ecology and Evolution*, *21*, 84. https://doi.org/10.1186/s12862-021-01818-5