Received: 29 August 2025 | Accepted: 24 December 2025

DOI: 10.1002/tpg2.70196

The Plant Genome

RESOURCES

Integrative chromosome-scale genome analysis of cupuassu provides insights into witches' broom disease resistance and expands genomic resources for *Theobroma*

Vinicius A. C. de Abreu[1] | **Rafael Moysés Alves**[2] | **Mauro de Medeiros Oliveira**[3] | **Vitor Trinca**[3] | **Loeni Ludke Falcão**[4] | **Lucilia Helena Marcellino**[4] | **Antonio Figueira**[5] | **Douglas S. Domingues**[6] | **Alessandro M. Varani**[3]

[1]Laboratório de Bioinformática e Computação de Alto Desempenho (LaBioCad), Faculdade de Computação (FACOMP), Universidade Federal do Pará (UFPA), Belém, Pará, Brazil

[2]Embrapa Amazônia Oriental, Belém, Pará, Brazil

[3]Departamento de Biotecnologia Agropecuária e Ambiental, Faculdade de Ciências Agrárias e Veterinárias (FCAV), Universidade Estadual Paulista (UNESP), Jaboticabal, São Paulo, Brazil

[4]Embrapa Recursos Genéticos e Biotecnologia, Brasília, Brazil

[5]Centro de Energia Nuclear na Agricultura (CENA), Universidade de São Paulo (USP), Piracicaba, São Paulo, Brazil

[6]Departamento de Genética, Escola Superior de Agricultura Luiz de Queiroz (ESALQ), Universidade de São Paulo (USP), Piracicaba, São Paulo, Brazil

Correspondence
Alessandro M. Varani, Departamento de Biotecnologia Agropecuária e Ambiental, Faculdade de Ciências Agrárias e Veterinárias (FCAV), Universidade Estadual Paulista (UNESP), Jaboticabal, São Paulo, Brazil. Email: alessandro.varani@unesp.br

Assigned to Associate Editor Katrien Devos.

Funding information
Conselho Nacional de Desenvolvimento Científico e Tecnológico, Grant/Award Numbers: 304367/2022-2, 310645/2021, 313174/2022; Fundação Amazônia de Amparo a Estudos e Pesquisas, Grant/Award

Abstract

Cupuassu (*Theobroma grandiflorum*) is a fruit tree native to the Brazilian Amazon and increasingly relevant to regional bioeconomies. Its cultivation is severely affected by witches' broom disease (WBD), caused by *Moniliophthora perniciosa*. While a chromosome-scale genome of the susceptible genotype C1074 is available, the lack of a resistant reference has limited investigation into the genomic basis of resistance. Here, we present the first chromosome-scale assembly of the WBD-resistant genotype C174 (415.8 Mb) and a comparative analysis with C1074 integrating structural variant detection, gene-duplication profiling, transposable-element (TE) annotation, and time-resolved host–pathogen transcriptomics. C174 exhibits distinctive tandem and dispersed duplications, genotype-specific TE insertions, and coordinated defense-gene expression, together with higher heterozygosity

Abbreviations: BUSCO, benchmarking universal single-copy orthologs; DEGs, differentially expressed genes; DUF, domain of unknown function; ETI, effector-triggered immunity; LAI, LTR assembly index; LARD, large retrotransposon derivatives; LTR, long terminal repeat; MAPK, mitogen-activated protein kinase; MEME, mixed effects model of evolution; NLR, nucleotide-binding domain leucine-rich repeat receptor; PK, protein kinases; PR, pathogenesis-related; PRR, pattern-recognition receptors; PTI, pattern-triggered immunity; QTLs, quantitative trait loci; RLKs, receptor-like kinases; ROS, reactive oxygen species; SNP, single nucleotide polymorphism; SV, structural variation; TE, transposable element; TF, transcription factor; WBD, witches' broom disease.

Vinicius A. C. de Abreu and Rafael Moysés Alves contributed equally to this work.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2026 The Author(s). *The Plant Genome* published by Wiley Periodicals LLC on behalf of Crop Science Society of America.

Number: #075/2020; Fundação de Amparo à Pesquisa do Estado de São Paulo, Grant/Award Numbers: #2019/25176-0, #2023/04372-1, #2023/10314-4

indicative of broader allelic diversity. Genome-wide, TEs show differential expression and spatial proximity to immune loci, suggesting potential regulatory importance. Although C1074 encodes more nucleotide-binding domain leucine-rich repeat receptors (541 vs. 434), most remain transcriptionally inactive, whereas C174 shows sequential activation of pattern-recognition receptors, mitogen-activated protein kinase components, transcription factors, and pathogenesis-related proteins. Within the Chromosome 6 previously identified resistance quantitative trait locus, two duplicated *DUF4220/DUF594* genes (where DUF is domain of unknown function) unique to C174—orthologous to a maize gene implicated in fungal response before—are infection-induced and display signatures of episodic positive selection. Together, these results establish a high-quality genomic framework for exploring the molecular architecture of WBD response in *T. grandiflorum*. The datasets generated here—including the C174 and C1074 reference genomes, immune-related variant catalog, and prioritized defense-gene lists—constitute a comprehensive open resource for evolutionary, functional, and breeding research of *Theobroma* species.

Plain Language Summary

Witches' broom disease (WBD) severely affects cupuassu (*Theobroma grandiflorum*), a fruit tree essential to Amazonian livelihoods. Here, we present the first chromosome-scale genome of a WBD-resistant cupuassu genotype and compare it with that of a susceptible genotype. The analysis integrates genomic structure, transposable elements, and infection-time transcriptomes, revealing differences associated with defense-related genes and pathways. These results provide a solid framework for understanding the genomic basis of disease response in *T. grandiflorum* and deliver an open, high-quality resource to support future research, molecular breeding, and genome-editing efforts in *Theobroma* crops.

1 | INTRODUCTION

The Amazon region harbors extraordinary plant diversity and several crops of rising economic importance. *Theobroma grandiflorum* (Willd. ex Spreng.) Schum., commonly known as cupuassu, is valued for its flavor-rich fruit pulp and seed butter, supporting regional food and cosmetics industries while thriving in mixed agroforestry systems that promote sustainable livelihoods (Costa et al., 2020; Pugliese et al., 2013; Rosa et al., 2024). Cupuassu cultivation is severely constrained by witches' broom disease (WBD), caused by the hemibiotrophic fungus *Moniliophthora perniciosa*, which can devastate yields and threaten grower income (Alves & Resende, 2008; Bailey et al., 2018).

M. perniciosa infects meristematic tissues and deploys a suite of effectors to subvert host immunity (Barbosa et al., 2018). In response, plants deploy pattern-recognition receptors (PRRs), including receptor-like kinases (RLKs), to detect pathogen-associated molecular patterns, triggering pattern-triggered immunity (PTI) Jones & Dangl, 2006). PTI activates mitogen-activated protein kinase (MAPK) cascades, production of reactive oxygen species (ROS), and modulates hormonal signaling (Meng & Zhang, 2013; Robert-Seilaniantz et al., 2011). If pathogen effectors suppress PTI, plants engage effector-triggered immunity (ETI) via intracellular nucleotide-binding domain leucine-rich repeat receptors (NLRs), which further amplify defense signaling and promote the expression of downstream pathogenesis-related (PR) proteins and other defense responses (dos Santos & Franco, 2023; Jones & Dangl, 2006; X. Li et al., 2023). These layers of defense are tightly regulated by key transcription factor (TF) families responsive to biotic and abiotic stresses, such as AP2/ERF (ethylene-responsive and abiotic stress signaling), bHLH (growth and hormonal regulation), MYB (secondary metabolism and defense), NAC (programmed cell death and abiotic/biotic responses), WRKY (immune gene activation

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and pathogen signaling), and bZIP (hormone signaling and environmental stress adaptation) (Seo et al., 2015).

Although systematic breeding of cupuassu remains in its infancy (≈50 years), on-farm clonal selection has led to the identification of two reference genotypes: "C174," a naturally WBD-resistant clone, and "C1074," an agronomically superior but susceptible clone. These genotypes were independently selected at different sites cultivated by smallholder farmers, vegetatively propagated, and kept as clones at the germplasm collection of Embrapa Amazônia Oriental (Brazilian Agricultural Research Corporation, Belém, Pará, Brazil). Thus, these genotypes are clonally maintained but not clonally related, representing distinct selections whose divergence predates breeding and reflects standing natural variation within Amazonian populations. Both genotypes served as parents of a mapping population used to identify quantitative trait locus (QTL) for WBD resistance (Alves & Chaves, 2023; Mournet et al., 2020). Genetic mapping of the C174 × C1074 biparental population identified a QTL associated with WBD-resistance in the distal region of Chromosome 6, spanning approximately 13.9 cM (≈3.7 Mb), and encompassing defense-related genes such as *TgPR3* and adjacent *NLRs*, *PRs*, and *PRRs* (Alves et al., 2024; Mournet et al., 2020). Complementary transcriptomic analyses revealed that C174 and C1074 display distinct activation of PRR, ROS, terpene-, and TF-mediated defense pathways, indicating contrasting regulatory dynamics rather than single-gene effects (Falcão et al., 2022). Additional studies have also highlighted that host-associated microbial communities may contribute to WBD resistance (de Matos et al., 2024).

The recent release of the chromosome-scale genome of the susceptible clone C1074 provided the first complete view of the cupuassu genome, establishing a foundation for detailed comparative analysis of this resistance locus (Alves et al., 2024). However, the structural and regulatory genomic determinants of resistance in C174 remained largely unexplored. The availability of the C1074 reference genome and public transcriptomic datasets now enables direct comparison between resistant and susceptible genetic backgrounds. Rather than inferring causality, such comparative analyses allow hypothesizing on how variation in gene content, structure, and regulation may underlie resistance-associated traits. In particular, transposable elements (TEs)—major drivers of genomic structural and regulatory novelty in plants (Bourque et al., 2018)—are here examined for their positional and expression patterns relative to defense loci, providing an evidence-based framework for exploring genome organization and evolutionary dynamics in *T. grandiflorum*.

Here, we present the first complete chromosome-scale genome assembly of the WBD-resistant genotype C174, together with a data-driven comparative framework with C1074. We conducted integrative analyses encompassing structural variation (SV), resistome composition (including PRs, PRRs, NLRs, protein kinases [PKs], and TFs), TE dynamics, and transcriptomic reanalysis using genotype-resolved references. This unified approach enables precise characterization of the Chromosome 6 resistance QTL and of genome-wide structural and regulatory contrasts between the two genotypes. The resulting genome–transcriptome resource serves as a foundation for community-level studies, marker development, and functional validation in cupuassu and related *Theobroma* species, supporting the long-term goal of sustainable genetic improvement of these crop species.

Core Ideas

- First chromosome-scale genome of the witches' broom-resistant cupuassu C174 expands resources for *Theobroma*.
- Comparative analysis with the susceptible C1074 identifies transposable element (TE)-associated variation near immune gene neighborhoods.
- Time-course transcriptome reveals phased activation of defense pathways in C174, with limited nucleotide-binding domain leucine-rich repeat receptor (NLR) induction in C1074.
- Duplicated DUF*4220*/DUF*594* genes under positive selection within the resistance quantitative trait locus (QTL) represent candidates for functional analysis.
- Comprehensive genomic and transcriptomic datasets are provided as open resources for evolutionary and breeding studies.

2 | MATERIALS AND METHODS

2.1 | Sampling, DNA/RNA extraction, and sequencing of the resistant genotype C174

Leaves from a tree of the WBD-resistant cupuassu genotype C174 were collected at the Embrapa Amazônia Oriental clonal repository in Belém, Pará, Brazil (1.4359° S, 48.4495° W), and deposited in the Herbarium JABU (http://jabu.jbrj.gov.br/v2) at the Universidade Estadual Paulista, Jaboticabal *campus* (Voucher JABU1369). Both C174 and C1074 are maintained as clonal accessions at Embrapa germplasm collection. Genomic DNA was extracted following the same protocols used for the susceptible genotype C1074 (Alves et al., 2024), ensuring methodological consistency between datasets. Full details of sequencing platforms, read depth, and quality metrics are provided in Table S1 and Supporting Information 1.

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.2 | Genome and transcriptome assembly, annotation, and quality assessment of C174

The genome and annotation for the C1074 were obtained from Alves et al. (2024) (BioProject PRJNA691024). RNA-seq data for both C174 and C1074 across infection time points were retrieved from Falcão et al. (2022) (BioProject PRJNA898247). The genome and transcriptome of C174 were assembled and annotated using the same procedures as for the susceptible genotype C1074 (Alves et al., 2024), thereby ensuring complete methodological consistency between the two *T. grandiflorum* genotypes (Supporting Information 1).

Briefly, one PacBio Sequel IIa SMRT Cell was generated for the genome assembly. Hi-C libraries were prepared using the Phase Genomics Proximo Hi-C Plant kit and sequenced on the Illumina NovaSeq platform (Table S1). For Iso-Seq, half of a PacBio SMRT Cell produced 4.6 million reads (8.1 Gb) derived from a mixture of C174 and C1074 RNA samples (Table S1).

The PacBio HiFi reads were assembled using Hifiasm v0.19.3 (Cheng et al., 2021) under default parameters. Contaminant sequences were removed with Kraken2 (Wood et al., 2019) and the extract_kraken_reads.py utility v1.2 (Lu et al., 2022), employing the PlusPFP index database (May 17, 2021). The primary assembly was indexed with BWA v0.7.17 (H. Li & Durbin, 2009). Restriction maps (*Dpn*II) were generated with Juicer v1.6 (Durand, Shamim, et al., 2016) for scaffolding through 3D-DNA v180419 (Dudchenko et al., 2017), followed by manual curation in Juicebox Assembly Tools v3.1.4 (Durand, Robinson, et al., 2016). Residual gaps were polished with the *run-ASM-pipeline-post-review.sh* script from 3D-DNA and the *close_scaffold_gaps.sh* routine from MaSuRCA v4.1.0 (Zimin et al., 2017). Chromosome numbering followed that of *Theobroma cacao*. Assembly quality and completeness were assessed using Merqury v1.3 (Rhie et al., 2020), Inspector v1.2 (Chen et al., 2021), long terminal repeat (LTR) assembly index (LAI) (Ou et al., 2018), and BUSCO v5.4.5 (Benchmarking Universal Single-Copy Orthologs; Manni et al., 2021) against the *embryophyta_odb10* database (Kriventseva et al., 2018).

Iso-Seq transcripts were processed in SMRT Link 12.0 (PacBio) with default settings. The short-read (retrieved from Falcão et al. [2022] [BioProject PRJNA898247]) and HiFi RNA-seq data were assembled de novo using Trinity v2.14.0 (Haas et al., 2013). Genome-guided assemblies were generated by aligning reads with HISAT2 v2.2.1 (Kim et al., 2019) and minimap2 v2.24 (H. Li, 2021), followed by transcript reconstruction using StringTie2 v2.2.1 (Kovaka et al., 2019). These datasets were integrated with PASA v2.5.3 (Haas et al., 2003) and TransDecoder v5.7.0 (https://github.com/TransDecoder/TransDecoder) to produce a comprehensive transcriptome database for downstream annotation (identification of untranslated regions, isoforms, and intron–exon boundaries corrections).

Genome annotation was conducted in two main phases following current best practices for plant genomes (Vuruputoor et al., 2023). In the first phase, TEs and other repetitive sequences were identified with EDTA-GUI (Extensive de novo TE Annotator - Graphical User Interface; https://github.com/Marcos-Fernando/EDTA-GUI), also available through the AnnoTEP-DB platform (https://plantgenomics.ncc.unesp.br/AnnoTEP-DB/). In the second phase, the soft-masked assembly was subsequently annotated by integrating multiple gene predictors, including BRAKER v3.0.4 (Lars et al., 2023), EVidenceModeler v2.1.0 (Haas et al., 2008), and PASA (Haas et al., 2003). For functional annotation Blast2GO v6.0 (Conesa et al., 2005) was employed. Telomeric and centromeric repeat regions were identified using quarTeT (commit e1a2f72) (Lin et al., 2023) and the Centromics pipeline (https://github.com/ShuaiNIEgithub/Centromics), respectively. Comprehensive details of the annotation procedures are provided in Supporting Information 1.

2.3 | LTR age estimation and TE comparative analyses

Insertion times of intact LTR retrotransposons were estimated using LTR_retriever (Ou & Jiang, 2017), integrated in the EDTA-GUI pipeline. Calculations assumed a neutral mutation rate (μ) of 1.5×10^{-9} substitutions per site per year, consistent with estimates for long-lived perennial trees (Buschiazzo et al., 2012) and reflecting the generation time of cupuassu (~10–15 years). These values provide model-based approximations intended to indicate relative timing rather than absolute chronological events, as ancient insertions may be partially eroded or nested, resulting in divergence-derived ages that are inherently approximate.

TE libraries generated with the EDTA-GUI pipeline for C174 and C1074 were compared using cd-hit-est-2d (Fu et al., 2012) with parameters: -c 0.80 -n 10 -aS 0.8 -aL 0.8 -G 1. To support expression analyses, a pan-genome TE library was also constructed by merging the genotype-specific libraries using the panEDTA workflow (Ou et al., 2024), producing a nonredundant and lineage-aware reference set for downstream quantification (see Section 2.5).

RepeatMasker v4.1.7-p1 (http://www.repeatmasker.org) was used in sensitive mode (-div 20 -cutoff 250) to map TEs in both genomes. TE–gene associations were inferred using the TE_Density tool (Teresi et al., 2022), which quantifies the spatial co-occurrence of annotated TEs and gene features.

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2.4 | Resistome and TFs annotation

PRRs and NLRs were identified using Resistify v1.1.5 (M. Smith et al., 2025). Transcriptional regulators (TRs), TFs, and PKs were annotated using iTAK v2.0.2 (Zheng et al., 2016). PR proteins were identified via hmmscan (Eddy, 2011) against the Pfam database (Mistry et al., 2020), using an *E*-value cutoff of 1e-5.

2.5 | Transcriptional reprogramming in response to pathogen inoculation

Short-read RNA-seq reads were analyzed in three contrasts: (i) inoculated versus mock at 24 h after inoculation (hai), (ii) inoculated versus mock at 48 hai, and (iii) 48 hai versus 24 hai within the inoculated samples, to identify genes whose expression changes during this time interval. There were three biological replicates for each genotype, treatment, and time point (Falcão et al., 2022). The RNA-seq reads from C174 and C1074 were aligned to their respective genomes using STAR v2.7.1a (Dobin et al., 2012), with GTFs (Gene Transfer Format files) derived from GFF3 gene models via gffread (Pertea & Pertea, 2020). Splice sites were extracted using *hisat2_extract_splice_sites.py*. STAR was run in two-pass mode (–twopassMode Basic), generating coordinate-sorted BAM files. Splice junction annotations (–sjdbFileChrStartEnd) and noncanonical intron filtering (–outFilterIntronMotifs RemoveNoncanonical) were applied. For TE quantification, options –outFilterMultimapNmax 100 and –winAnchorMultimapNmax 100 were used.

Differential expression of genes and TEs was performed using DESeq2 (Love et al., 2014) via TEtranscripts (Jin et al., 2015). Raw *p*-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) correction, and genes or TEs with an adjusted *p*-value (padj) ≤ 0.05 were considered statistically significant. A threshold of absolute $\log_2$fold change ≥ 1 or ≤ -1 was applied to identify genes and TEs with meaningful expression changes. To be classified as differentially expressed, genes or TEs had to meet both criteria simultaneously.

2.6 | Comparative genomic analyses between C1074 and C174

Comparative genome alignments were performed using D-GENIES (Cabanettes & Klopp, 2018) for dot plot visualization and DupGen_finder (Qiao et al., 2019) to classify homologous genes into syntenic and collinear categories. Following Tang et al. (2008), *synteny* was defined as the conservation of genes on homologous chromosomes between C174 and C1074, while *collinearity* denotes the subset of those syntenic regions retaining conserved gene order.

SVs were identified using MUM&Co v3.8 (O'Donnell & Fischer, 2020), which detects and classifies insertion, deletion, and rearrangement events—including TE-associated categories such as *insertion_mobile* and *deletion_mobile*—through BLAST-based comparative analysis. Telomeric and potential centromeric regions were detected with quarTeT v1.2.1 (Lin et al., 2023). Chromosome-scale ideograms were generated using jcvi plotting tools (https://github.com/tanghaibao/jcvi/wiki/Miscellaneous-plotting).

Orthogroups were inferred using OrthoFinder v2.5.5 (Emms & Kelly, 2019), which integrates normalized all-versus-all sequence similarity searches and graph-based clustering. Pairwise comparisons were performed using DIAMOND v2.1.8 (Buchfink et al., 2021) with the –ultra-sensitive option to ensure high-confidence detection of distant homologs and minimize false negatives in orthogroup assignment. Only the longest protein isoform per gene was retained.

Following OrthoFinder's standard output, genes were classified into three categories: (i) shared orthogroups, containing homologous genes present in both genotypes; (ii) orthogroup-specific genes, belonging to gene families found in only one genotype (C174 or C1074), representing genotype-specific expansions; and (iii) exclusive genes, which were not assigned to any orthogroup and therefore lacked detectable orthologs in the other genotype.

In parallel, intra-genomic duplicated pairs were identified using DupGen_finder, which classifies duplicates as whole-genome, tandem, proximal, transposed, or dispersed. Consequently, certain genes categorized as exclusive or orthogroup-specific could still possess intragenomic paralogs and were therefore included in Ka/Ks analysis (Section 2.10). Genes lacking detectable paralogs were excluded from Ka/Ks estimation.

Functional enrichment analyses of gene ontology (GO) terms were performed using GOATOOLS (Klopfenstein et al., 2018), applying Benjamini–Hochberg FDR correction (FDR ≤ 0.05).

2.7 | Single nucleotide polymorphism density analysis

To quantify sequence polymorphism between the two genotypes, C174 and C1074 assemblies were aligned in both directions with Winnowmap2 (distinct = 0.9998). Alignments were converted to sorted BAM files, and variants were called using *bcftools* (mpileup and call) (Danecek et al., 2021; Jain et al., 2022) (k = 19; asm10 preset) after masking high-copy *k*-mers identified by *meryl* (Rhie et al., 2020) (top 0.02% most frequent) with a minimum mapping quality of 20 and base quality 20. Variant calls were normalized against the corresponding reference and filtered to retain only biallelic single nucleotide polymorphisms (SNPs);

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

indels and low-quality sites were discarded. SNP counts were summarized in nonoverlapping 150-kb windows using *bedtools* (Quinlan & Hall, 2010) and normalized to SNPs per kilobase (variants/[window_size/1000]). Genome-wide and per-chromosome statistics (mean, median, P95, P99) were computed with custom AWK/Python scripts. For the Chromosome 6 resistance QTL, windows overlapping the interval were compared to the chromosomal background using a two-sided Mann–Whitney *U* test to assess enrichment of SNP density. All analyses used only uniquely aligned primary segments (–secondary = no) and were repeated in both alignment directions to confirm robustness.

2.8 | In silico heterozygosity estimation

PacBio HiFi reads from each genotype were converted to 21-mer and 31-mer frequency histograms with *meryl* (v1.4.1) (Rhie et al., 2020). Histograms were modeled using GenomeScope2 (Ranallo-Benavidez et al., 2020). The heterozygosity interval (min–max) reported by GenomeScope2 was recorded, and the midpoint was taken as the point estimate of genome-wide heterozygosity.

To obtain an orthogonal and assembly-based validation, HiFi reads from each genotype were aligned back to their respective genome assemblies using minimap2 v2.26 (H. Li, 2021). The resulting BAM files were sorted and indexed with SAMtools v1.21, and variants were called with bcftools v1.17 (Danecek et al., 2021). Only high-confidence biallelic SNPs with genotype 0/1, mapping quality ≥ 20, and read depth ≥ 10 were retained. The proportion of heterozygous sites was then calculated as (number of heterozygous SNPs ÷ assembly size in bp) × 100, providing an independent estimate of allelic variation.

2.9 | Comparative QTL analysis

The Chromosome 6 resistance-associated QTL previously mapped by Mournet et al. (2020) in the *C174* × *C1074* biparental population was used as the genomic reference interval for comparative analysis. The corresponding region was identified and aligned between both parental genomes using MCScan (Python version) (https://github.com/tanghaibao/jcvi/wiki/MCscan-(Python-version), allowing synteny-based comparison and visualization of the homologous intervals.

Cross-species comparisons were also performed against the *T. cacao* genome v2 (Argout et al., 2017) to assess structural conservation and rearrangements in this chromosomal region. All figures were refined for consistency and graphical clarity using Inkscape (https://inkscape.org/).

2.10 | Ka/Ks and positive selection analysis

Intragenomic duplicated gene pairs identified by DupGen_finder (Section 2.7) were used to estimate nonsynonymous (Ka) and synonymous (Ks) substitution rates using the *calculate_Ka_Ks_pipeline* workflow described by Qiao et al. (2019) (https://github.com/qiao-xin/Scripts_for_GB/tree/master/calculate_Ka_Ks_pipeline). Briefly, the protein sequences of each duplicate pair were aligned using MAFFT v7.490 (Katoh & Standley, 2013) with the L-INS-i algorithm, and codon alignments were generated using PAL2NAL v14 (Suyama et al., 2006). Ka and Ks values were calculated using *KaKs_Calculator* 2.0 (Wang et al., 2010), adopting the γ-MYN method (a modified version of the Yang–Nielsen model) under the Tamura–Nei substitution framework (Tamura & Nei, 1993).

For gene families containing more than two paralogs, Ka/Ks was computed only for the high-scoring pair—the pair showing the highest sequence similarity and alignment score—to avoid redundancy and artificial inflation of substitution rates. This procedure reflects the implementation of *DupGen_finder* (Qiao et al., 2019) and follows the best-practice recommendations of *KaKs_Calculator* for reliable pairwise substitution estimation (Wang et al., 2010).

All Ka/Ks estimates were accompanied by Fisher's exact-test *p*-values, and only pairs with $p \leq 0.05$ were retained for downstream interpretation. To ensure robustness, gene pairs with Ks > 5.0 (to avoid substitution saturation) were excluded. Pairs with Ka/Ks > 1 were considered indicative of potential positive selection, whereas those with Ka/Ks < 1 reflected purifying selection.

2.11 | Branch-site tests for episodic selection

To refine the evidence of adaptive evolution within the resistance QTL, we focused on two C174-specific duplicates *DUF4220/DUF594* (*TgrandC174G00000023870* and *TgrandC174G00000023871*) (where DUF is domain of unknown function), which (i) are physically located inside the WBD-resistance QTL on Chromosome 6, (ii) are significantly induced during infection (24–48 hai and 48 hai contrasts), and (iii) exhibit Ka/Ks > 1.

To contextualize their evolutionary divergence, we retrieved all annotated DUF4220/DUF594 homologs from *T. grandiflorum* (C174 and C1074), *T. cacao* (Criollo v2.1), and *Herrania umbratica*, representing closely related genomes. As external references, we included the functionally characterized *Zea mays* ortholog (LOC103634932), previously reported as a fungal-responsive gene (Miranda et al., 2017; Zhu et al., 2017), and *Arabidopsis thaliana* AT5G45460, which was included as an outgroup to orient the phylogeny.

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

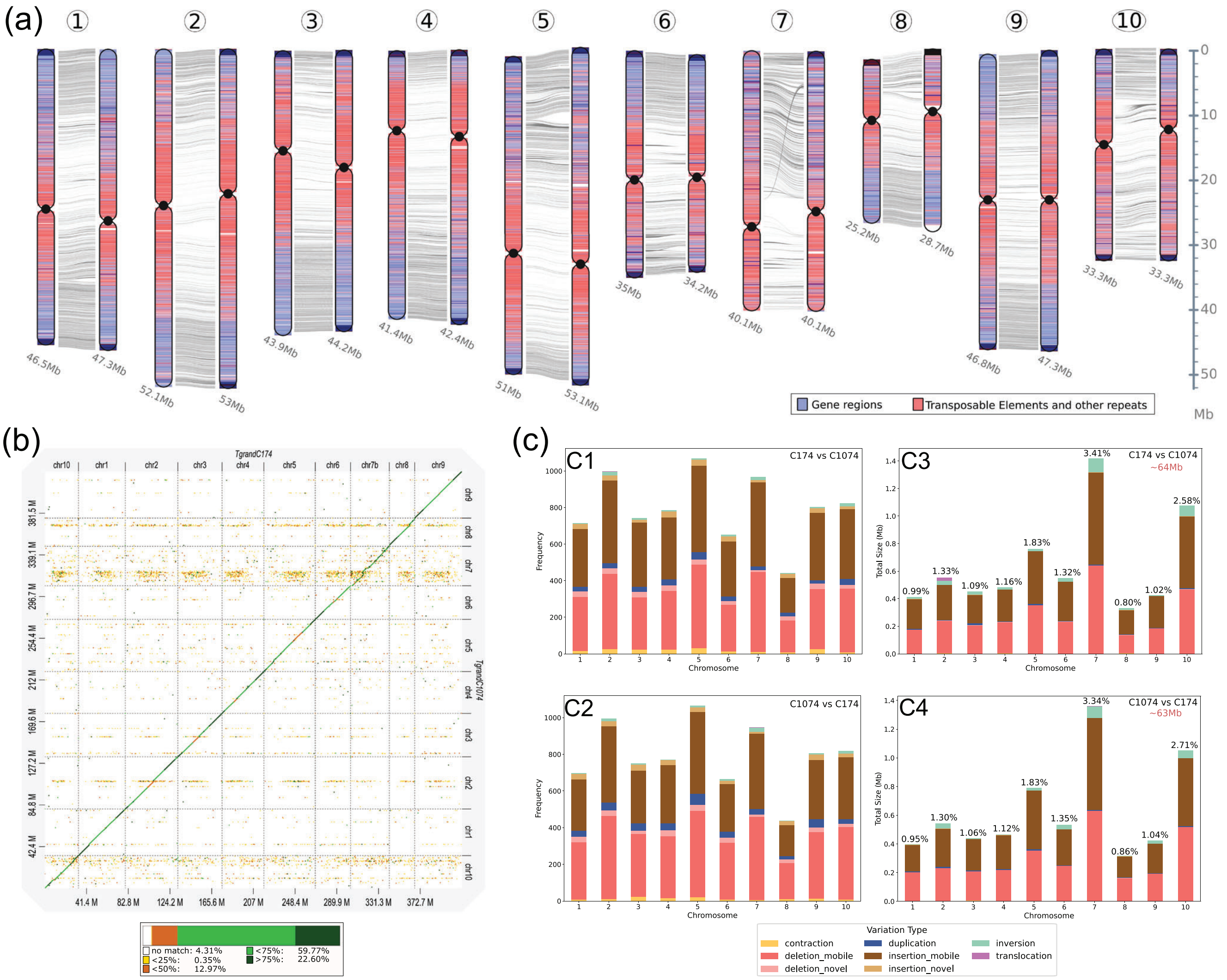

FIGURE 1 Genome-wide comparisons of gene or transposable element (TE) regions between the witches' broom disease resistance C174 and susceptible C1074 genotypes. (a) Representation of the 10 chromosomes. Gene-rich regions are shown in blue, and TE-rich regions are shown in red. Predicted centromeric regions are depicted as black circles. Telomeric repeats are shown as black squares at the edges of the chromosomes. (b) D-GENIES alignment dot plot. (c) Two-way structural variation (SV) analysis. C1 and C2 frequency of SV types by chromosome. C174 versus C1074 (C1) and C1074 versus C174 (C2). C3 and C4 total size of SVs by chromosome. C174 versus C1074 (C3) and C1074 versus C174 (C4). The values shown above each bar in panels C3 and C4 correspond to the proportion of SV identified on each chromosome. The SVs include contraction, deletion mobile (potentially related to TEs), deletion novel (not related to TEs), duplication, insertion mobile (potentially related to TEs), insertion novel (not related to TEs), inversion, and translocation.

All branch-site tests were performed independently of the tree outgroup to avoid potential bias.

Protein sequences were aligned using MAFFT v7.490 (L-INS-i algorithm), and codon alignments were generated with PAL2NAL v14 to preserve the reading frame. Gaps and in-frame stop codons were trimmed using trimAl v1.4 (strict mode) (Capella-Gutierrez et al., 2009). A maximum-likelihood phylogeny was inferred with IQ-TREE3 (Wong et al., 2025) under the GTR+F+R6 model with 1000 ultrafast bootstraps.

Episodic selection was tested using three complementary approaches implemented in HyPhy v2.5.64: (i) aBSREL (M. D. Smith et al., 2015), which evaluates each branch for episodic $\omega > 1$; (ii) Mixed effects model of evolution (MEME) (Murrell et al., 2012), which detects site-specific positive selection ($p \leq 0.05$, posterior probability ≥ 0.80); and (iii) RELAX (Wertheim et al., 2014), which measures global intensification ($K > 1$) or relaxation ($K < 1$) of selection pressure. All p-values were corrected for multiple testing using the Holm–Bonferroni method.

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3 | RESULTS

3.1 | Assembly metrics of the WBD-resistant C174 genome and elevated *k*-mer heterozygosity relative to the susceptible C1074

The chromosome-scale assembly of the resistant genotype C174 genome spans 415.8 Mb across 10 chromosomes (Figure 1A), approximately 8 Mb shorter than the susceptible C1074 genome (Table S1). Using 27× PacBio HiFi coverage, the assembly achieved a base-level error rate <0.07%, BUSCO completeness of 98.8%, and an LAI (Ou et al., 2018) of 14.79, all consistent with a high-quality reference genome.

Merqury analysis yielded *k*-mer completeness values of 88% for C1074 and 86% for C174, while Inspector reported mapping rates of 95.99% and 99.87% and base-level error rates of 0.0016% and 0.0668%, respectively (Table S1). Because both assemblies were generated in haploid mode, heterozygous *k*-mers from alternative haplotypes are partially absent, resulting in completeness values below 100%. Such values are typical of high-quality haploid plant assemblies and are fully supported by BUSCO, LAI, and Inspector metrics, confirming the chromosome-scale accuracy of both genomes (Table S1).

All centromeres and 12 telomeric regions were successfully assembled, with only 29 residual gaps (mean size = 101 bp; range = 100–140 bp; total = 2.94 kb, <0.001% of the genome). These small, noncoding gaps are located outside annotated gene regions and are thus considered minor (Table S1).

GenomeScope2 (k = 21 and 31) estimated per-base heterozygosity at 1.07% for C174 and 0.59% for C1074 (Table S2). Independent self-mapping SNP calls yielded concordant estimates (0.90% and 0.62%, respectively; Table S2), indicating higher allelic diversity in the resistant genotype C174.

Chromosome-scale dot plots and DupGen_finder analyses revealed extensive gene conservation between C174 and C1074 (97.0% synteny), but reduced collinearity (76.3%), consistent with small-scale rearrangements and transpositions that disrupted local gene order since the divergence of the two genotypes. Approximately 4.3% of each genome (~18 Mb) is genotype-specific (Figure 1B).

Together, these results confirm the high contiguity and accuracy of the C174 assembly while highlighting clear structural divergence relative to C1074.

3.2 | Protein-coding is highly shared with limited genotype-specific expansions

The C174 genome encodes 30,157 protein-coding genes, 1,224 fewer than C1074, which has 31,381 genes (Table 1). Both genomes exhibit an average of five introns and six exons per gene, and a 3.4 kb average gene length. Gene models account for approximately 25% of the total genome size in both genomes (Table S3). The BUSCO assessment of the annotated gene sets returned a completeness score of 99.4% for C174, indicating that the predicted proteome encompasses nearly the entire conserved gene repertoire of *T. grandiflorum* and provides a solid basis for downstream comparative and functional analyses (Table S3).

Orthogroup clustering revealed that 95.9% of the genes in C174 are shared with C1074. Genes not assigned to any orthogroup were considered exclusive, totaling 702 in C174 and 590 in C1074. Additionally, 888 genes in C174 and 1089 in C1074 were classified as orthogroup-specific, representing genotype-restricted expansion or contraction (Table 1).

3.3 | TEs predominantly drive SV between C174 and C1074

The C174 and C1074 genomes show extensive SV differences. An intersection of the SV genomic coordinates against the annotated TE coordinates (≥80% coverage and ≥80% identity to annotated TEs) showed that 77% of C174 SVs and 80% of C1074 SVs overlap TEs (Table 1; Table S4; Figure 1C).

The most frequent SVs were insertions_mobile and deletions_mobile, distributed genome-wide but particularly enriched on chromosomes 5, 6, 7, and 10. Chromosome 7 contained the largest number and cumulative size of SVs (~14% of all events, spanning ≈14 Mb, equivalent to ~34% of its length and ~3.3% of the total genome). Overall, only two translocations were detected: a 213 kb event on Chromosome 2 in C174 and a 20 kb event on Chromosome 7 in C1074 (Table S4). Inversions accounted for ~0.6% of each genome, while other SV types—contractions, duplications, novel deletions, and insertions—collectively represented <0.5%.

Altogether, SVs encompassed ~64 Mb in C174 and ~63 Mb in C1074, each representing ~15% of their respective genome sizes, highlighting the prevalence of TE-associated variation in the genomic differentiation (Table S4).

3.4 | Single-nucleotide divergence and coupled SNP and SV hotspots on Chromosomes 7 and 10

The genome-wide median density was 6.2 SNP kb^{-1} (~0.62% of aligned bases), but values varied markedly among chromosomes (Table S5). Chromosomes 1 and 9 were the most conserved (<5 SNP kb^{-1}). In contrast, Chromosome 7 reached median values >11 SNP kb^{-1} and displayed an

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Comparative genomic, structural, and transcriptomic differences between the resistant cupuassu genotype (C174) and the susceptible genotype (C1074).

	C174	C1074
Number of genes	30,157	31,381
Number of genes in orthogroups	29,455 (97.7%)	30,791 (98.1%)
Positive selection	90	148
Neutral selection	-	-
Purifying selection	19,372	19,712
Number of unassigned genes (exclusive)	702 (2.3%)	590 (1.9%)
Positive selection	10	3
Neutral selection	-	-
Purifying selection	200	173
Number of genes in species-specific orthogroups	888 (2.9%)	1089 (3.5%)
Positive selection	3	10
Neutral selection	80	-
Purifying selection	372	471
Structural variation	7988 (77% TEs)	7.952 (80% TEs)
Deletions (mobile/novel)	3566 (84% TEs)	3885 (92% TEs)
Mobile	3322 (86% TEs)	3638 (95% TEs)
Novel	244 (58% TEs)	247 (39% TEs)
Insertions (mobile/novel)	3871 (70% TEs)	3489 (68% TEs)
Mobile	3627 (72% TEs)	3248 (71% TEs)
Novel	244 (49% TEs)	241 (29% TEs)
Duplications	278 (73% TEs)	364 (74% TEs)
Contractions	185 (60% TEs)	126 (44% TEs)
Inversions	96 (98% TEs)	86 (100% TEs)
Translocation	2 (100% TEs)	2 (100% TEs)
Unique and intact transposable elements (TEs)	1668	1665
Shared and repeated TEs in both genomes (271 clusters)	280	286
Length occupied	124 Mb (29%)	125 Mb (29%)
Exclusive TEs	1388	1379
Length occupied	181 Mb (43%)	174 Mb (41%)
Recent LTR elements	56	90
Length occupied	47 Mb (11%)	45 Mb (10%)
Differentially expressed genes (DEGs)	1169	1127
Unassigned genes (exclusive)	32	20
Downregulated	17	9
Upregulated	15	12
Species-specific orthogroups	38	5
Downregulated	6	2
Upregulated	32	3
Shared orthogroups	1099	1102
Downregulated	713	542
Upregulated	386	560

Abbreviation: LTR, long terminal repeat.

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

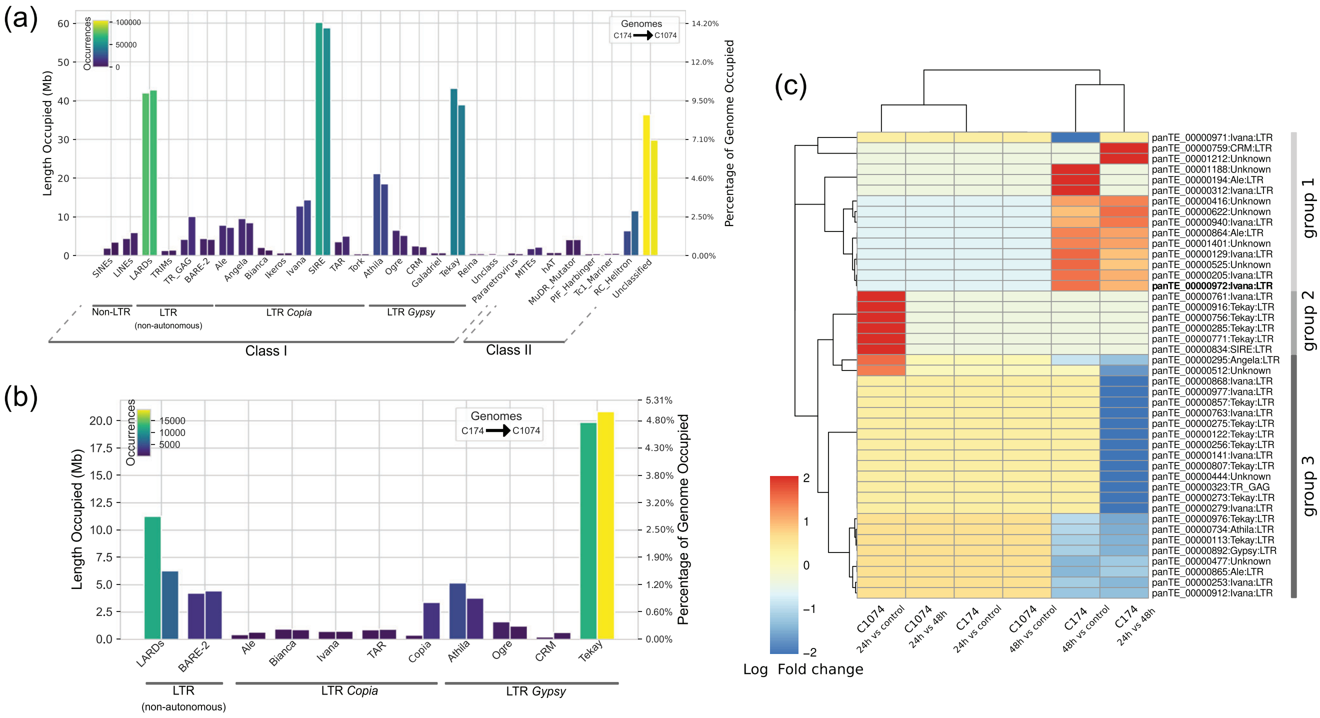

FIGURE 2 Comparative genomic analyses of transposable element (TE) content between the C174 and C1074 genotypes. (a) Distribution of the genomic regions occupied (length or percentage) by Class I and Class II TE lineages. (b) Distribution of recent and exclusive long-terminal repeat-retrotransposons (LTR-RT) insertion events. (c) Heatmap of TE differential expression analysis at 24 h after inoculation (hai) with *Moniliophthora. perniciosa*, in the transition 24 hai versus 48 hai, and at 48 hai.

extended upper tail up to 35 SNP kb^{-1} (P99) (Table S5), indicative of large, highly differentiated haplotype blocks (Figure S1). Elevated SNP density was also observed on Chromosome 10 (∼9 SNP kb^{-1}). Notably, these are the same two chromosomes that concentrate the highest load of SVs (Figure 1C).

3.5 | Conserved TE content with genotype-specific sequence and expression differences

Approximately 280 Mb (67%) of the C174 and C1074 genomes were masked for TE-related sequences. Despite SVs presumably resulting from TE activity, comparative analysis revealed broadly conserved patterns of TE occurrence and size distribution among genotypes (Figure 2A; Table S6). The most abundant TE lineages are LTR *Copia* SIRE, LTR *Gypsy* Tekay, and non-autonomous large retrotransposon derivative (LARD) elements, which together span ∼140 Mb (∼35%) of each genome. Unknown repeats (elements lacking homology to reference databases) vary slightly: 36 Mb in C174 (8.7%) versus 29 Mb in C1074 (7%). Estimation of insertion time suggested two bursts of LTR *Copia* expansion at ∼2.8 and ∼17–19 million years ago (Ma) and one major LTR *Gypsy* expansion at ∼2.1–2.7 Ma (Figure S2).

Pairwise comparison of TE sets using cd-hit-est-2d, identified 271 shared TE clusters, along with 1388 and 1379 unique TE sequences in C174 and C1074, respectively (Table 1). Despite similar overall content, substantial sequence-level variation suggests genotype-specific or recent mobilization events not homogenized across both genotypes.

Analysis of LTR retrotransposons with identical LTR pairs recovered 146 putatively recent insertions (56 in C174; 90 in C1074) (Figure 2B). These correspond to ∼11% and ∼10% of total genome length, respectively, and are dominated by LTR-*Gypsy* (Tekay, Athila), nonautonomous LARDs, and *Copia* BARE-2 families (Figure 2B).

TE RNA-seq analysis using the TEtranscripts approach revealed 44 LTR retrotransposons (∼13 Mb per genome) showing differential expression during *M. perniciosa* infection (Figure 2C). Three general temporal patterns were observed as follows: (i) early induction (upregulation in C1074 at 24 hai), (ii) late induction (upregulation in C174 at 48 hai), and (iii) late repression (downregulation in C174 at 48 hai). Most differentially expressed LTR retrotransposons have estimated insertion ages >0.1 Ma, indicating that transcriptional activation is not restricted to the youngest elements. One *Copia*-Ivana element on Chromosome 7 of C174 shows no detectable LTR divergence and is strongly upregulated at 48 hai, whereas its relative copy on Chromosome 1 of C1074 (∼0.25 Ma) remains transcriptionally silent.

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Collectively, these results reveal that although both genotypes share a general conserved TE landscapes, subtle differences in sequence composition and transcriptional activation might contribute to genotype-specific regulatory plasticity, rather than differences in total TE abundance or age.

3.6 | TE proximity to genes reveals distinct insertion patterns near immune-related loci

Across both genotypes, the general landscape of TEs near genes is relatively uniform (Figure 3A). Considering ± 10 kb up- and downstream, as well as intragenic regions of genes, the TE environment is consistently dominated by LARDs, *Copia*-SIRE/Ivana retrotransposons, and *Gypsy* elements such as Tekay and Athila. Non-autonomous TR_GAGs contribute only marginally. This background pattern is broadly conserved between C174 and C1074.

Defense-related gene families, however, show a different pattern (Figures 3B,C and 4). At NLR loci, C174 exhibits a higher frequency of intragenic LARD and *Gypsy*-Athila insertions, whereas C1074 shows enrichment for *Gypsy*-Tekay and *Copia*-Ivana elements (Figure 3B). PR genes present the opposite: intragenic regions in C174 are often flanked by *Gypsy*-Tekay, whereas in C1074 LARDs predominate (Figure 4A). Similarly, PK and TF loci show genotype-specific TE associations both up- and downstream of coding regions (Figure 4B,C). In contrast, PRR genes maintain comparable TE compositions in both genomes, mirroring the genome-wide profile (Figure 3C).

Together, these observations indicate that defense-related loci in *T. grandiflorum* differ in their local TE composition between the resistant and susceptible genotypes, revealing distinct genotype-specific insertion histories and local sequence contexts around defense genes.

3.7 | Genotypes diverge for global expression dynamics of exclusive and orthogroup-specific genes

RNA-seq analysis comparing inoculated samples with controls revealed 1169 and 1127 differentially expressed genes (DEGs) in C174 and C1074, respectively (Table 1). While both genotypes share many orthogroups, their expression dynamics diverge substantially, particularly for exclusive and orthogroup-specific transcripts (Figure 5).

Despite similar mean Log_2FC values in shared orthogroups at 24 hai (0.032 ± 0.769 for C1074; 0.044 ± 0.712 for C174) (Figure 5A), C1074 exhibited an early increase in DEG number with 475 upregulated and 158 downregulated genes. In contrast, C174 displayed a minimal response, with only 15 upregulated and nine downregulated genes. This pattern reversed when comparing 48 hai versus 24 hai within inoculated samples (temporal transition; Figure 5A): C174 induced 149 and repressed 311 genes, while C1074 exhibited a reduced response with 34 upregulated and 345 downregulated genes. The shift in expression magnitude was significant (C174: -0.065 ± 0.923; C1074: 0.009 ± 0.783; $p = 4.28\text{e-}20$). By 48 hai, C174 showed 222 upregulated and 393 downregulated genes, whereas C1074 displayed only 51 upregulated and 39 downregulated genes (Figure 5C). Mean Log_2FC remained significantly different (C174: -0.030 ± 0.945; C1074: 0.051 ± 0.765; $p = 1.86\text{e-}23$). Exclusive and orthogroup-specific genes followed similar trends but with lower DEG counts.

3.8 | Distinctive evolutionary selection of exclusive and orthogroup-specific genes

Ka/Ks analysis revealed distinct selective pressures acting in both genotypes (Table 1; Figure 5D). A small subset of genes (103 in C174 and 161 in C1074) was under positive selection (Ka/Ks > 1, p-value < 0.05), suggesting possible adaptive evolution. C1074 contained more positively selected orthogroup genes than C174 (148 vs. 90), and more positively selected species-specific orthogroup genes (10 vs. 3), despite the absence of differential expression in these genes (Table 1). Neutral selection (Ka/Ks ≈ 1) was not observed in any category.

3.9 | Structural remodeling and genotype-specific duplications define the Chromosome 6 resistance QTL in cupuassu

Microsynteny analysis of the previously identified WBD resistance QTL on Chromosome 6 (Mournet et al., 2020) revealed differences in gene content and organization between the C174 and C1074 genotypes when compared to *T. cacao* (Figure 6). Four gene clusters associated with plant–pathogen interaction and resistance mechanisms were identified in *T. grandiflorum*, highlighting specific expansions relative to *T. cacao* (Figure 6; Figure S3; Table 2). These expansions, particularly of NLRs (e.g., 84 NLRs in C1074 vs. 39 in C174; Figure 6B,D), appear driven mainly by tandem duplications, consistent with localized diversification within resistance-related gene clusters. Comprehensive gene annotations for the QTL region revealed ≈420 genes in C174 and ≈460 in C1074 (Table S7A,B). In *T. cacao*, the syntenic region is notably contracted and harbors fewer defense-related genes, suggesting that the more elaborate configuration in cupuassu emerged after the cupuassu–cacao divergence.

Consistent with this structural remodeling, the Chromosome 6 resistance QTL shows a localized enrichment of fixed

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

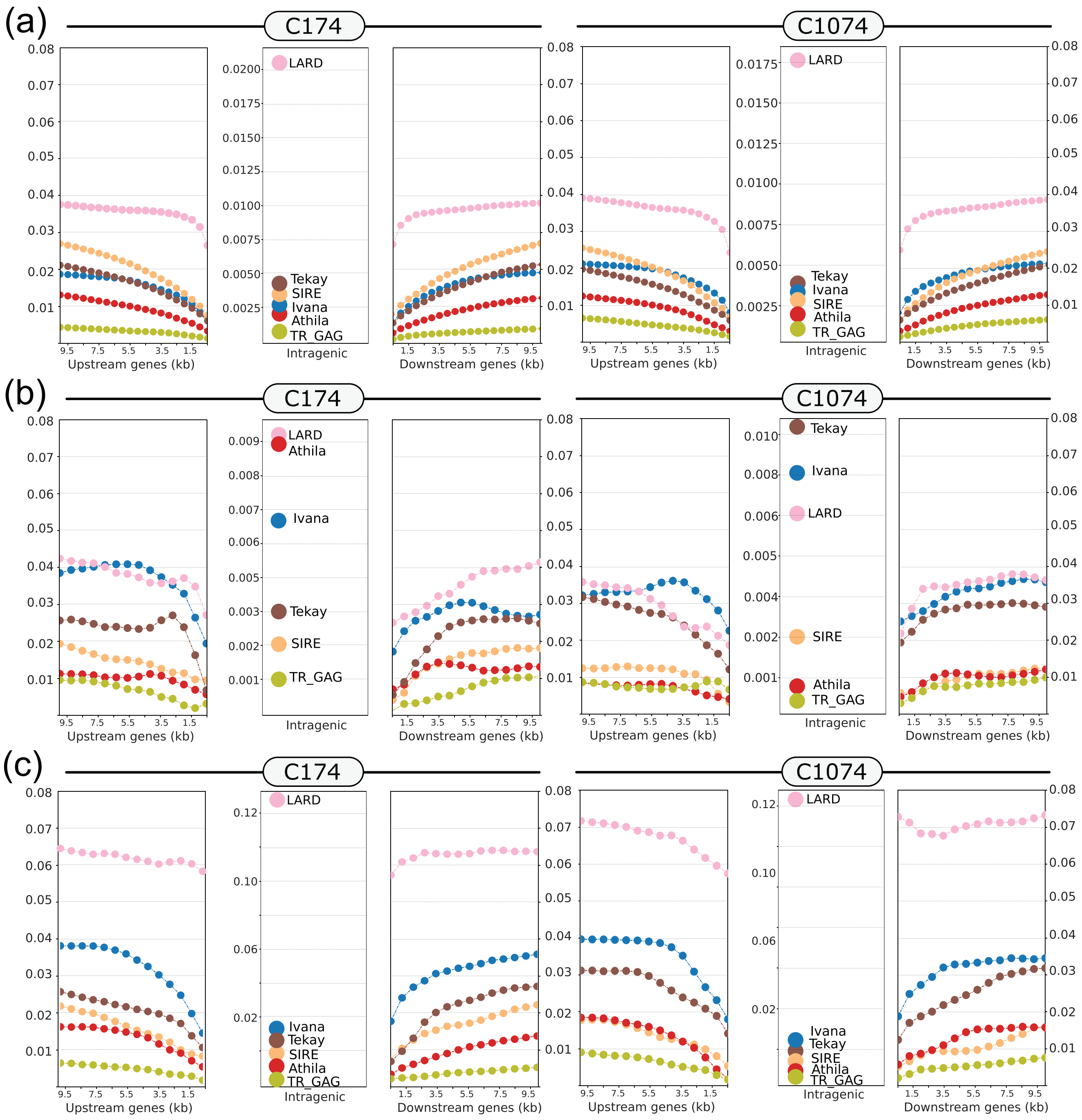

FIGURE 3 Average transposable element (TE) density and location in C174 and C1074 genomes by category of genes. (a) Reference distribution for all annotated genes. (b) Nucleotide-binding leucine-rich-repeat (NLR) immune receptors. (c) Pattern-recognition receptor (PRR) genes. The left panel represents the TE density values upstream of genes, the middle panel represents the intragenic TE density of genes, and the right panel represents the TE density values downstream of genes.

SNPs between the two genotypes (i.e., invariant within each clone but differing between them). Within the QTL interval, the resistant genotype shows a median of 8.7 SNP kb^{-1} versus 6.5 SNP kb^{-1} in the surrounding chromosomal background (+34%). The reciprocal comparison (C1074 on C174) mirrors this pattern, increasing from 5.7 to 8.0 SNP kb^{-1} (+42%) (Figure S4). Both enrichments are significant (Mann–Whitney $U = 3929, p = 0.013$; $U = 4195, p = 0.003$), indicating that C174 carries a differentiated haplotype precisely at the QTL.

Furthermore, transcriptomic profiling revealed divergent gene expression dynamics within the QTL (Table 2; Table S7). Notably, four immune-related genes (three PRRs and one NLR) were upregulated in C174, whereas none were

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

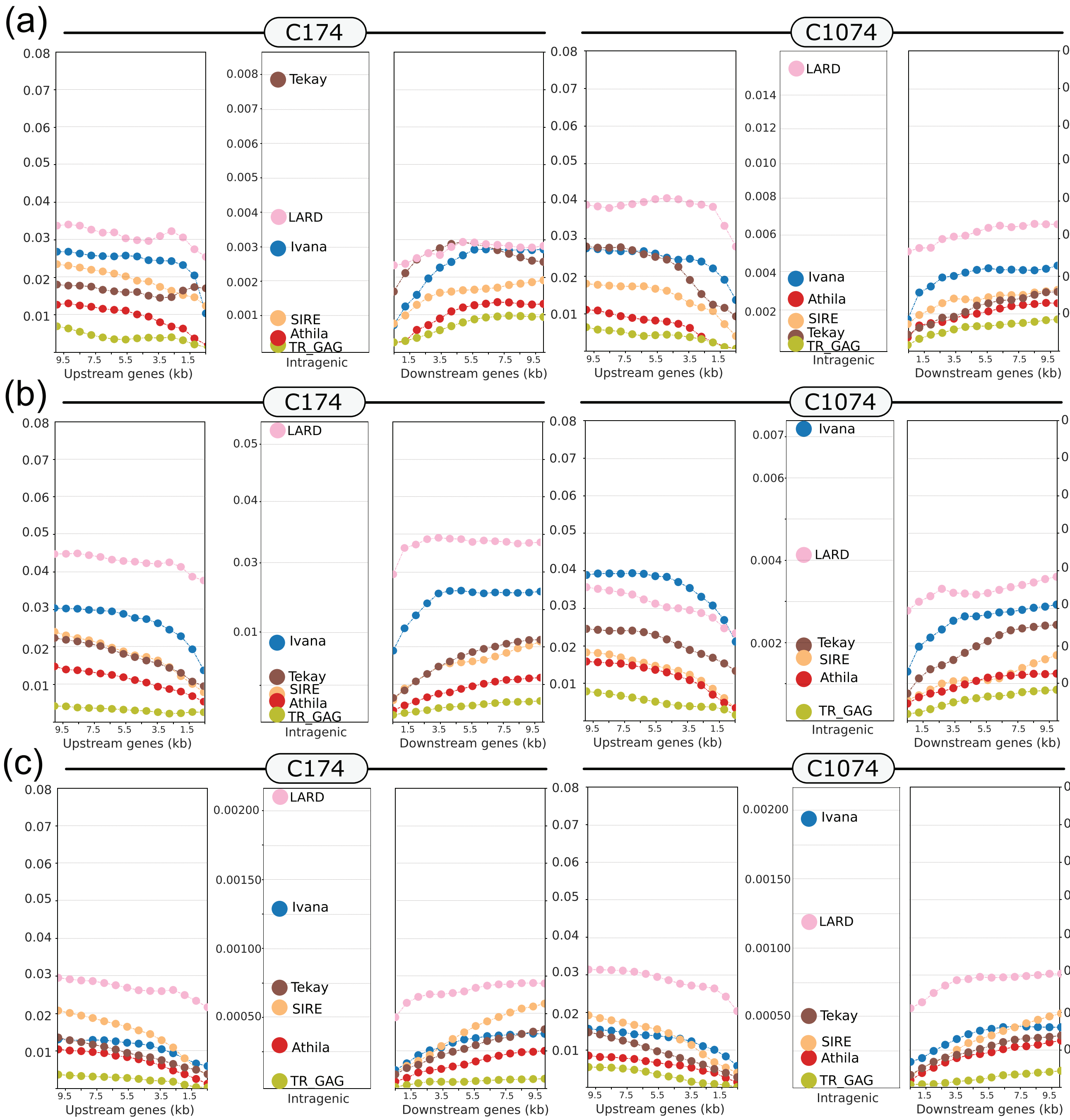

FIGURE 4 Average transposable element (TE) density and location in C174 and C1074 genomes by category of genes. Panel structure as in Figure 3. (a) Pathogenesis-related (PR) genes. (b) Protein kinase (PK) genes. (c) Transcription factor (TF). The left panel represents the TE density values upstream of genes, the middle panel represents the intragenic TE density of genes, and the right panel represents the TE density values downstream of genes.

DEGs in C1074, suggesting a limited immune response in the susceptible genotype (Table S7). The *TgPR3* chitinase, previously described as potentially associated with WBD resistance (Santana Silva et al., 2020), was identified in both genomes (*TgrandC174G00000024019* and *TgrandC1074G00000024418*). These orthologs are under strong purifying selection (Ka/Ks ≈ 0.177), yet neither was differentially expressed. A tandem cluster of PR-14 and PR-16 genes was also detected, containing seven genes in C174 and two in C1074 (Figure 5D; Table S7).

Ethylene-related genes were also identified within the QTL region. These include segmental duplicates from the ethylene-responsive factor (ERF) and dehydration-responsive element-binding (DREB) subfamily (*TgrandC174G00000023718* and *TgrandC1074G00000024136*), as well as ethylene-regulated nuclear protein singletons (*TgrandC174G00000024036* and

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

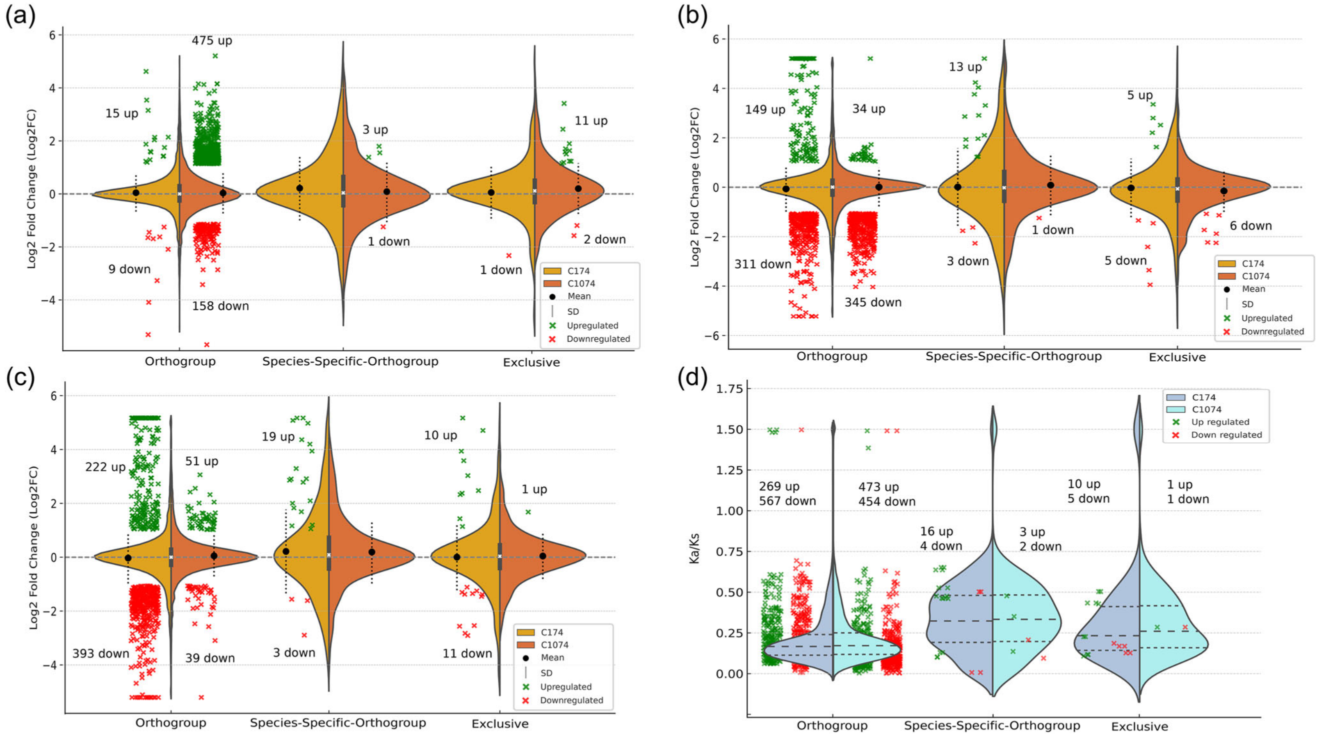

FIGURE 5 Violin plots showing the distribution of Log2 fold change (Log_2FC) values, differentially expressed genes (DEGs), and relative rates of synonymous and nonsynonymous substitutions (Ka/Ks) at exclusive genes, orthogroup-specific genes, or shared genes in C174 and C1074 genomes along time points after *Moniliophthora perniciosa* inoculation. (a) Log_2FC values and DEGs of 24 h after inoculation (hai). (b) Log_2FC values and DEGs during the 24–48 hai transition. (c) Log_2FC values and DEGs at 48 hai. Distinct gene expression patterns emerge over time between the two genotypes, with marked reversals in differential gene regulation observed across the analyzed time points. (d) Ka/Ks landscape; overlaid points mark DEGs (green = upregulated, red = downregulated).

TgrandC1074G00000024433). One bHLH and one WRKY TF were also detected in both genotypes, though with low expression levels ($\log_2$FC < 1) under the conditions tested (Table S7A,B).

Within the same QTL region, two C174-specific *DUF4220/DUF594* duplicates (*TgrandC174G00000023870* and *TgrandC174G00000023871*) showed strong induction during infection and exhibited Ka/Ks > 1. BLAST analyses indicated that these genes are orthologous to a functionally characterized *Z. mays* gene (LOC103634932) previously reported as a fungal-responsive gene (Miranda et al., 2017; Zhu et al., 2017), suggesting that these C174-specific duplicates are plausible candidates for involvement in plant–pathogen interactions.

Phylogenetic and branch-site analyses (Figure S5) revealed episodic positive selection restricted to the C174 (aBSREL $p < 0.005$). MEME and RELAX tests supported intensified selection on *TgrandC174G00000023871* ($K = 2.24$, $p = 0.0027$), whereas *TgrandC174G00000023870* remained under purifying constraints. Collectively, these findings identify *TgrandC174G00000023871* as a promising candidate for functional studies on the molecular basis of WBD resistance in cupuassu.

3.10 | Coordinated peroxidase–chitinase expression patterns in C174 and C1074

A total of 394 PR genes were identified in C174 and 390 in C1074, with both genotypes sharing a comparable number of genes across PR families (Figures S6 and S7; Table 2; Table S7). Most PR genes, including the DEGs, appear to have originated predominantly from tandem and segmental duplications (Figure S7). Overall, C174 showed more transcriptional plasticity and greater changes in PR gene expression, with 29 DEGs, whereas C1074 displayed only 12 DEGs (Figure 7A). While PRs that are DEGs in both genotypes are under purifying selection, the ones from C174 displayed slightly higher Ka/Ks ratios (mean ~0.18 vs. ~0.14 in C1074, $p = 0.05022$) (Figure S7).

Both genotypes upregulate one chitinase GH19 gene at 24 hai (*TgrandC174G00000008836* and *TgrandC1074G00000014930*), but expression of other chitinase-related genes diverges sharply. Notably, C174 activates several *PR-3* (GH18), *PR-5* (thaumatin-like/osmotin), other chitinases *PR-4/8/11* (GH19), and *PR-9* (peroxidase) genes during the 24–48 hai transition and at 48 hai (Figure 7A; Table S7).

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

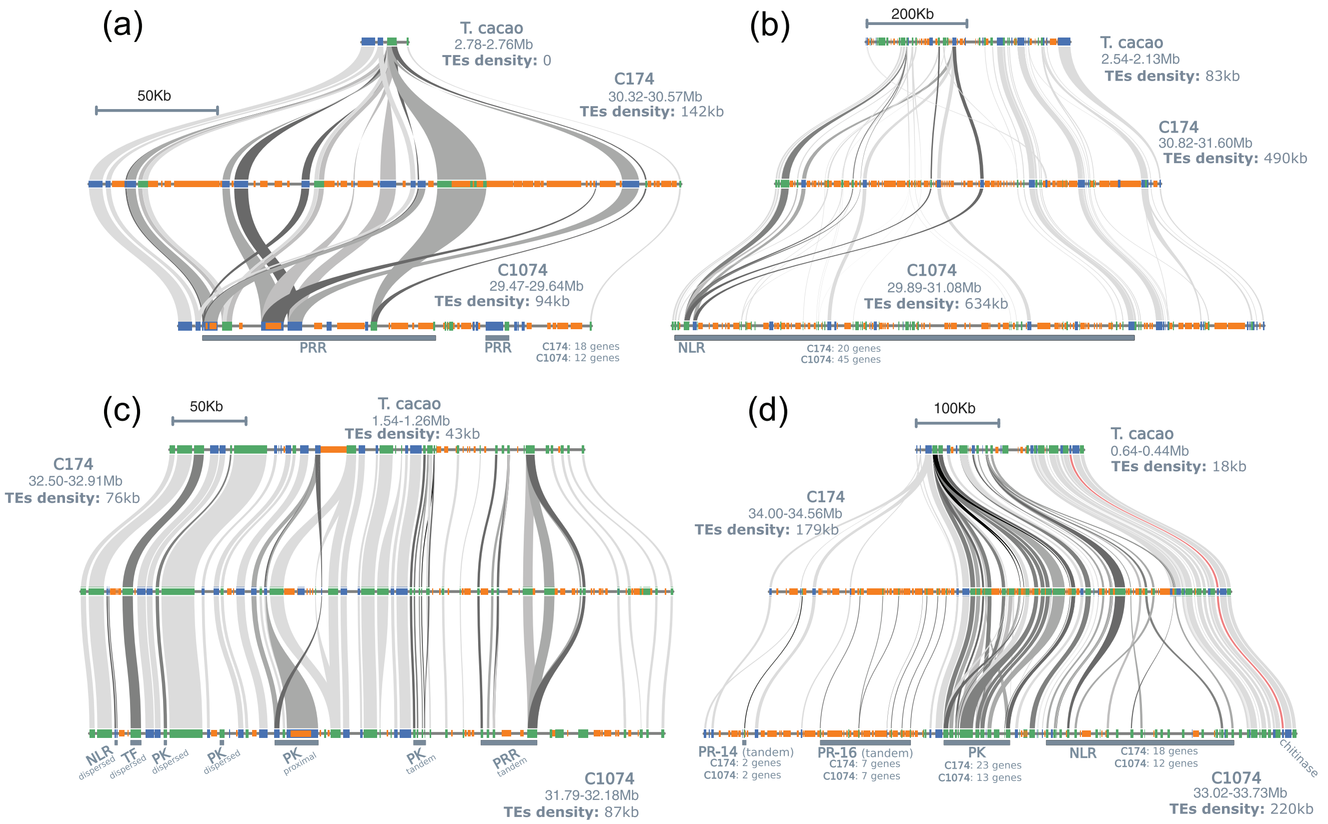

FIGURE 6 Microsynteny analysis of the witches' broom disease resistance quantitative trait locus (QTL) genomic region in the C174 and C1074 genotypes (Mournet et al., 2020) compared to *Theobroma cacao*. (a) Cluster 1, composed mostly of pattern-recognition receptors (PRR) genes. (b) Cluster 2, which is predominantly nucleotide-binding domain leucine-rich repeat receptor (NLR) genes that are expanded in the C1074 genome. (c) Cluster 3, primarily featuring protein kinases (PK) genes, indicates a high conservation across cupuassu and *T. cacao* genomes. (d) Cluster 4, containing NLR, PRR, PK, and pathogenesis-related (PR) genes, including the *PR-8/PR-11* (chitinase, GH19) family. Cluster 4 exhibits an expansion of NLR genes in the C1074 genotype compared to the C174 genotype. All clusters exhibit both shared and genotype-specific transposable element (TE) expansions. Forward-strand genes are shown in blue, reverse-strand genes in green, and TE-related sequences in orange.

C174 tandem-duplicated *PR-3* genes *TgrandC174G00000036098* and *TgrandC174G00000036100* are upregulated at 48 hai ($\log_2$FC > 4.2). Another tandem *PR-3* gene, *TgrandC174G00000003748*, is also upregulated at 48 hai ($\log_2$FC > 1.8), while its paralog—previously annotated as *TgPR8* and potentially associated with WBD resistance (Santana Silva et al., 2020)—*TgrandC174G00000003747* remains non-differentially expressed.

The top DEG in C174 is *TgrandC174G00000011264*, is a *PR-14* gene (lipid transfer protein), highly upregulated at 48 hai ($\log_2$FC = 6.5). The *PR-5* gene *TgrandC174G00000012658* is downregulated at 48 hai ($\log_2$FC = −1.45), suggesting early constitutive expression. *TgPR5* (*TgrandC174G00000012660*), previously associated with WBD resistance (Santana Silva et al., 2020), is part of a six-gene osmotin cluster, and it shows high, but not statistically significant expression at the 24–48 h transition ($\log_2$FC = 4.2) and at 48 hai ($\log_2$FC = 3.6) (Table S7).

In contrast, C1074 presents a conservative response. Only one peroxidase gene and two *PR-5* genes (*TgrandC1074G00000004764* and *TgrandC1074G00000012662*) are upregulated at 24 hai. The most highly expressed DEG is *TgrandC1074G00000036493*, a PR-6 serine protease inhibitor ($\log_2$FC > 5 at 24–48 hai), part of a six-gene tandem cluster. This DEG may reflect early protease-inhibitory activity rather than implying a defined regulatory mechanism. In C174, the corresponding cluster contains eight genes, none of which are differentially expressed, suggesting distinct transcriptional regulation of this locus (Table S7).

TgPR5 (*TgrandC1074G00000010022*) exhibits moderate expression during the 24–48 hai transition ($\log_2$FC = 2.4) and at 48 hai ($\log_2$FC = 2.9) in C1074, lower than in C174 and not statistically significant. *TgPR5* belongs to an eight-gene tandem osmotin cluster. *TgPR8* (*TgrandC1074G00000001104*) and its tandem duplicate show low, non-differential expression across all time points (Table S7).

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Comparative census of defense-related gene families in the witches' broom disease (WBD) resistant cupuassu genotype C174 and the susceptible genotype C1074. List of the total number of genes annotated as pathogenesis-related (PR), nucleotide-binding leucine-rich repeat receptor (NLR) immune receptors, pattern-recognition receptors (PRR), protein kinases (PKs), and transcription factors (TF) for each genotype, followed by the subset that is significantly differentially expressed gene (DEG; $|\log_2 FC| \geq 1$, FDR ≤ 0.05) during WBD infection. The same categories are then given for the chromosome-6 WBD-resistance quantitative trait locus (QTL) (Mournet et al., 2020), together with the number of QTL-resident genes that are DEGs.

Category	Possible role against WBD	C174	C1074	DEGs C174	DEGs C1074	C174 (QTL)	C1074 (QTL)	DEGs C174 (QTL)	DEGs C1074 (QTL)
PR genes									
PR-1 (cysteine-rich secretory protein)	General antifungal	12	12	1	1	–	–	–	–
PR-2 (glucanases, GH17)	Fungal cell wall degradation	43	43	2	2	1	1	–	–
PR-3 (chitinases, GH18)	Chitin hydrolysis	35	34	3	–	–	–	–	–
PR-4 (chitinases, GH19–barwin domain)	Chitin-targeted defense	13	12	3	2	1	1	–	–
PR-5 (thaumatin-like/osmotin family)	Osmotic antifungal	28	29	2	2	–	–	–	–
PR-6 (cystatin-like protease inhibitors)	Pathogen protease inhibition	13	11	1	1	–	–	–	–
PR-7 (aspartyl protease family)	Cell lysis and defense	10	10	–	–	–	–	–	–
PR-8/PR-11 (chitinases, GH19)	Chitin degradation	5	5	–	–	1	1	–	–
PR-9 (peroxidases)	ROS generation	85	84	8	1	–	–	–	–
PR-10 (Bet v1-like superfamily/ribonucleases)	RNase activity / signaling	5	5	–	–	–	–	–	–
PR-12 (defensin-like)	Antifungal pore formation	6	5	–	–	–	–	–	–
PR-14 (lipid transfer proteins, LTPs)	Cuticle defense / antifungal	90	91	6	2	2	2	–	–
PR-16 (germin family)	H_2O_2 production	49	49	3	1	9	9	–	–
PRR genes									
RLK (receptor-like kinases)	Signal transduction	581	594	21	31	2	2	–	–
RLP (receptor-like proteins)	Pathogen recognition	491	518	10	11	19	17	2 (up)	–
NLR genes									
CN (coiled-coil + NB-ARC)	Partial immune sensor	71	82	6	–	13	32	–	–
CNL (coiled-coil + NB-ARC + LRR)	Pathogen effector receptor	276	360	28	3	14	33	1 (up)	–
N (NB-ARC)	Truncated/partial immune module	45	46	4	–	4	8	–	–
NL (NB-ARC + LRR)	Effector detection	15	18	3	–	–	1	–	–
RNL (RPW8-like + NB-ARC + LRR)	Helper NLR for immune signaling	4	4	–	1	–	–	–	–
TN (toll-1 receptor + NB-ARC)	Possible regulatory function	3	3	–	–	–	–	–	–
TNL (toll-1 receptor) + NB-ARC + LRR)	Effector recognition	20	28	–	2	8	10	–	–

(Continues)

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 (Continued)

Category	Possible role against WBD	C174	C1074	DEGs C174	DEGs C1074	C174 (QTL)	C1074 (QTL)	DEGs C174 (QTL)	DEGs C1074 (QTL)
PK genes	Defense signaling	610	620	36	77	22	19	2 (up)	–
TF and TRs									
AP2/ERF (APETALA2/ethylene responsive factor)	Hormone signaling	118	118	15	29	1	1	–	–
AP2 subfamily	Developmental regulation	15	14	1	–	–	–	–	–
ERF + DREB subfamily	Stress response	101	102	13	27	1	1	–	–
RAV subfamily	Growth–defense balance	2	2	1	2	–	–	–	–
MYB (myeloblastosis related)	Secondary metabolism	180	182	10	15	–	–	–	–
bHLH (basic helix-loop-helix)	Hormonal crosstalk	109	109	9	16	1	1	–	–
NAC (no apical meristem activation factor)	Cell wall regulation	107	112	9	9	–	–	–	–
WRKY (basic leucine zipper)	Defense transcription	61	59	4	17	1	1	–	–
bZIP (basic leucine zipper)	Abiotic/biotic response	51	51	2	2	–	–	–	–

Abbreviation: DREB, dehydration-responsive element-binding; ROS, reactive oxygen species.

3.11 | Differential timing of PRR activation in C174 and C1074

We identified 1079 PRRs in C174 and 1119 in C1074, with most genes and DEGs categorized as dispersed or tandem duplicates (Figures S6 and S8; Table ; Table S7). C174 exhibited 31 PRRs that are DEGs, while C1074 showed 42 (Figure 7B). Notably, C1074 displayed an early and strong transcriptional response at 24 hai. However, the pattern shifted during the 24–48 hai transition and at 48 hai, when C174 displayed coordinated activation.

In C174, the RLK *TgrandC174G00000011298* exhibited the highest expression level among PRRs during the 24–48 hai transition ($\log_2$FC = 8.07) and at 48 hai ($\log_2$FC = 5.69) (Figure S8). In contrast, nine PRRs were downregulated in C174 at 48 hai (Figure S8); these may represent early PTI regulators whose expression dynamics differ between genotypes and constitute strong candidates for functional validation in future studies.

3.12 | NLR expression patterns and potential ETI-associated signatures in C174 and C1074

A total of 434 NLR genes were identified in C174 and 541 in C1074; the higher number in C1074 suggests recent gene expansion events, primarily mediated by tandem and proximal duplications (Figures S6 and S9; Table 2; Table S7). Considering that NLRs from the CNL subfamily are core components of resistosome formation (M. Smith et al., 2025), the differential expression analysis and patterns (Figure 7C) are consistent with ETI-associated transcriptional activation in C174. C174 displayed 14 downregulated NLRs at 48 hai, which may correspond to early expressed genes whose induction occurred before this time point (Figure 7C). Among the most highly induced NLRs in C174 at 48 hai were *TgrandC174G00000036265*, *TgrandC174G00000026898*, and *TgrandC174G00000035888* (Table S7; Figure S9).

Ka/Ks analysis revealed slightly higher values for C1074 NLRs (mean Ka/Ks ~0.44) than for C174 (mean Ka/Ks ~0.39), a statistically significant difference (p = 0.0034), suggesting distinct selective constraints in the susceptible genotype (Figure S9). Notably, only one NLR in C1074 (*TgrandC1074G00000015673*) was both differentially expressed and under positive selection—it was upregulated at 24 hai but subsequently downregulated during the 24–48 hai transition (Table S7). This DEG therefore shows sequence divergence consistent with functional diversification, although its transient activity limits inference about its contribution to defense efficiency. The relative expansion of CNLs in C1074 may reflect a reservoir of immune potential, although their low inducibility suggests incomplete regulatory integration during infection.

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.13 | RLK and MAPK transcriptional dynamics reveal genotype-specific timing

C174 and C1074 harbored 610 and 620 PKs, respectively. Most PKs and their DEGs are associated with dispersed and segmental duplications (Figures S6 and S10; Table 2; Table S7). At 24 hai, C1074 presented 22 upregulated PKs, while C174 showed no DEGs. However, at 48 hai, C174 exhibited downregulation of 10 PKs, a pattern consistent with constitutive expression and subsequent feedback regulation of immune signaling (Table S7). These include *TgrandC174G00000000085* and *TgrandC174G00000030428* (receptor-like cytoplasmic kinases), *TgrandC174G00000017255* (a cell wall-associated kinase), *TgrandC174G00000021672* (an abiotic stress-responsive RLK), and *TgrandC174G00000033544* (a MAPK involved in immune signaling) (Figure 7D).

In C1074, the most upregulated DEGs at 24 hai included *TgrandC1074G00000022354* (an RLK linked to stress and hormonal signaling) and *TgrandC1074G00000027358* (a MAPKKK initiating MAPK signaling), both showing $\log_2$FC > 4. In contrast, at 48 hai, two highly upregulated RLKs were detected in C174—*TgrandC174G00000023969* and *TgrandC174G00000023974*—both LRR-containing kinases associated with PTI and MAPK activation.

MAPK DEGs (*TgrandC174G00000030702* and *TgrandC1074G00000033544*) were downregulated in both genotypes during the 24–48 hai transition and at 48 hai, suggesting negative modulation of MAPK signaling. In C1074, MAPK downregulation was slightly stronger ($\log_2$FC ≈ 1.55) than its paralog from the C174 genotype ($\log_2$FC ≈ 1.35). These trends are consistent with genotype-specific regulatory dynamics of the MAPK pathway rather than direct mechanistic differences.

3.14 | Temporal profiles of defense-related TFs

We investigated the six TF families known to play central roles in plant immune responses (Table 2). Segmental duplications were the most prevalent among these TFs, followed by dispersed and tandem duplications (Figure S11). C1074 exhibited a sharp induction of TFs at 24 hai, followed by a decline during the 24–48 hai transition, and at 48 hai (Figure 7E). Conversely, C174 showed a delayed but gradually intensifying response, with increasing TF expression during the 24–48 hai transition and peaking at 48 hai. This temporal pattern indicates a phased activation sequence in C174 rather than an immediate burst.

Notably, three TFs in C174 were highly upregulated ($\log_2$FC > 6) at 48 hai: an ERF (*TgrandC174G00000007236*) and two NAC factors (*TgrandC174G00000015577* and *TgrandC174G00000021289*). In addition, one MYB TF (*TgrandC174G00000009424*) was strongly downregulated ($\log_2$FC < −5.5).

3.15 | GO enrichment analysis of DEGs

Gene ontology enrichment (Figure 8) highlighted contrasting temporal responses between genotypes (FDR ≤ 0.05, Benjamini–Hochberg correction). In C1074, upregulated genes at 24 hai were enriched for salicylic acid and ethylene signaling, transcriptional regulation, and trehalose biosynthesis. Downregulated genes during the 24–48 hai transition involved jasmonic acid and flavonoid biosynthesis, and at 48 hai, suppression of cell wall and nitrate metabolism was evident. In C174, 24 hai samples showed repression of reactive nitrogen species metabolism, followed by strong enrichment for defense-related terms during the 24–48 hai transition and at 48 hai, including sesquiterpene and cadinene synthase activity and response to hydrogen peroxide. These patterns indicate distinct temporal signatures of immune activation between C174 and C1074 and extend previous transcriptomic findings (Falcão et al., 2022).

4 | DISCUSSION

To investigate the genetic basis of WBD resistance in *T. grandiflorum*, we generated the first chromosome-scale assembly of the resistant genotype C174 and compared it with the susceptible genotype C1074. By integrating SV analyses, TE profiling, gene-duplication patterns, and time-resolved transcriptomic data during infection, we identified structural and transcriptional contrasts that help explain their distinct responses to *M. perniciosa*. Together, these datasets provide a mechanistic framework for understanding how genome structure, regulation, and evolution jointly shape disease resistance in *T. grandiflorum*.

Both genotypes were originally selected from natural planting stands and have undergone few breeding cycles (Alves & Chaves, 2023), implying that their genomic contrasts reflect naturally occurring variation preserved within Amazonian populations. The observed signatures—positive selection on defense-related genes, recent tandem duplications, and nonrandom TE accumulation near immunity loci—thus represent preexisting differences rather than breeding artifacts. Documenting these at chromosome-scale resolution offers the first detailed view of the genomic architecture underlying the contrasting resistance phenotypes of C174 and C1074.

Although globally similar (Figure 1A,B), the two genomes differ significantly in local architecture. *K*-mer and self-mapping SNP analyses indicate broader allelic diversity in C174, consistent with higher baseline heterozygosity. Uneven

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

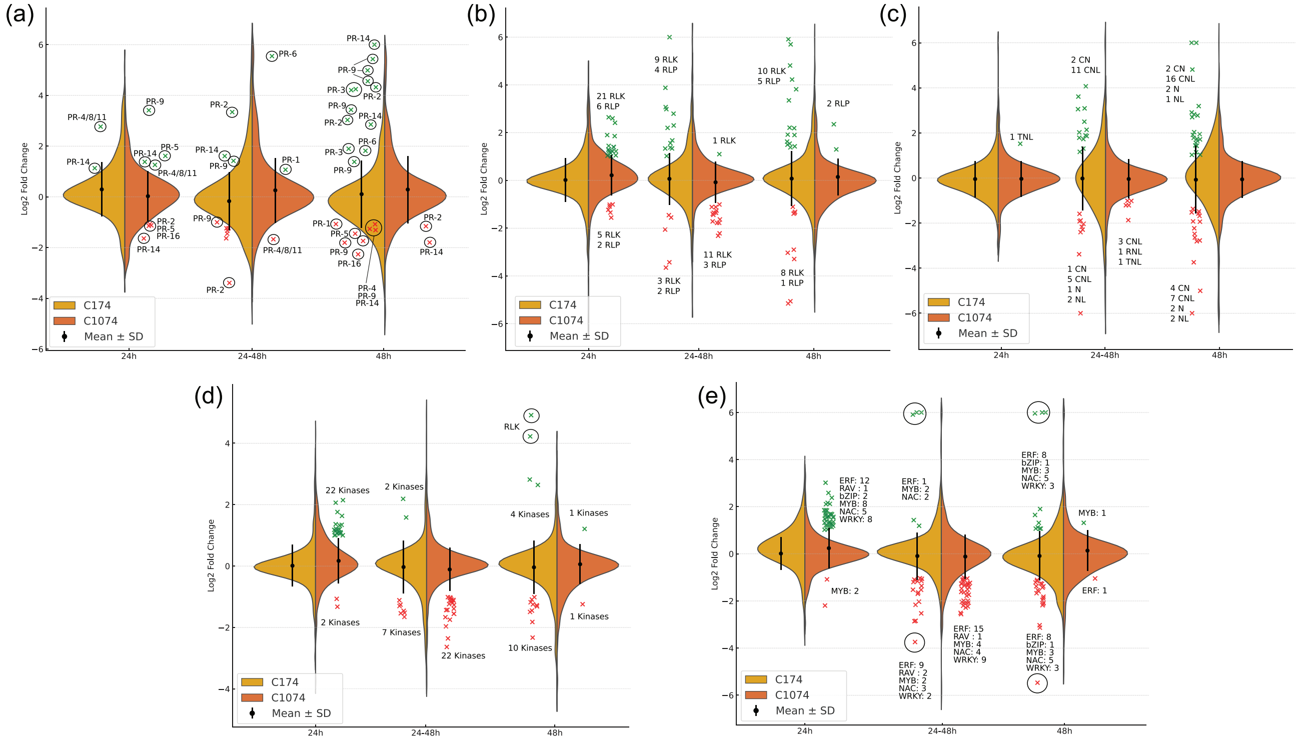

FIGURE 7 Violin plots illustrating the distribution of Log2 fold change (Log2FC) values for differentially expressed genes (DEGs) of specific categories at various time points after infection with *Moniliophthora perniciosa*. (a) Pathogenesis-related (PRs). (b) Pattern-recognition receptors (PRRs). (c) Nucleotide-binding domain leucine-rich repeat receptors (NLRs). (d) Protein kinase (PK). (e) Defense-related transcription factors (TFs). The plots encompass DEGs at 24 h after inoculation (hai) with *M. perniciosa*, 24 hai versus 48 hai, and 48 at hai.

SNP landscapes, as seen here, are common in rice, barley, and other crops, where regions of elevated polymorphism often arise from introgression, recombination-rate variation, or local haplotype differentiation (Borevitz et al., 2007; Chai et al., 2018; Mago et al., 2014; Yeo et al., 2016). For comparison, *T. cacao* shows an average intraspecies diversity of ∼5 SNP kb^{-1} (Cornejo et al., 2018), similar to the divergence observed between C174 and C1074, although here it reflects a single pairwise contrast.

At the structural level, both genomes harbor abundant TE-associated insertions and deletions, but their distributions are genotype-specific (Figure 1A,C). These differences coincide with local variation in TE expression (Figure 2A,C), suggesting spatial coupling between structural turnover and transcriptional variation. Across the genome, the distribution of major defense-gene classes (NLRs, PRs, PRRs, PKs, and TFs) is largely conserved—synteny reaches 97%—yet local deviations, especially within the Chromosome 6 resistance QTL, reveal genotype-specific TE enrichment within ±10 kb of defense genes in C174, whereas PRR neighborhoods remain comparatively stable (Figures 3 and 4). Because many plant LTRs carry WRKY, ERF, and bZIP binding motifs (Deneweth et al., 2022), such local TE accumulation could influence transcriptional responsiveness, as illustrated by an infection-induced *Copia*-Ivana element in C174 (Figure 2C).

Evidence indicates that heterozygosity, TE activity, and gene duplication act synergistically to diversify plant immune systems. High heterozygosity broadens allelic repertoires (Dievart et al., 2020; Gong & Han, 2021; Liu et al., 2017; Ngou et al., 2022); TEs introduce *cis*-regulatory motifs modulating expression in time and space (Bourque et al., 2018; Domínguez et al., 2020; Hassan et al., 2024; McDowell & Meyers, 2013); and tandem duplications generate new copies that may undergo neo- or subfunctionalization (Flagel & Wendel, 2009). In C174, these processes coexist, creating a genomic landscape enriched in structural and regulatory diversity that may underpin more dynamic immune responses.

Although total TE content is similar between genotypes (Figure 2A), C174 carries a distinct subset of transcriptionally active *Copia*-Ivana insertions with no detectable LTR divergence (Figure 2C). Combined with infection-responsive expression of older LTR-RTs (>0.1 Ma), these results indicate that TEs have contributed to both long-term genome remodeling and to condition-specific expression during infection. Their influence thus operates on dual timescales: structural modification through historical insertions and short-term transcriptional responsiveness to biotic stress. *Copia* elements in particular have been associated with *cis*-regulatory rewiring and even horizontal gene transfer in plants (S. Li et al., 2022; Ma et al., 2019; Orozco-Arias et al., 2023).

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

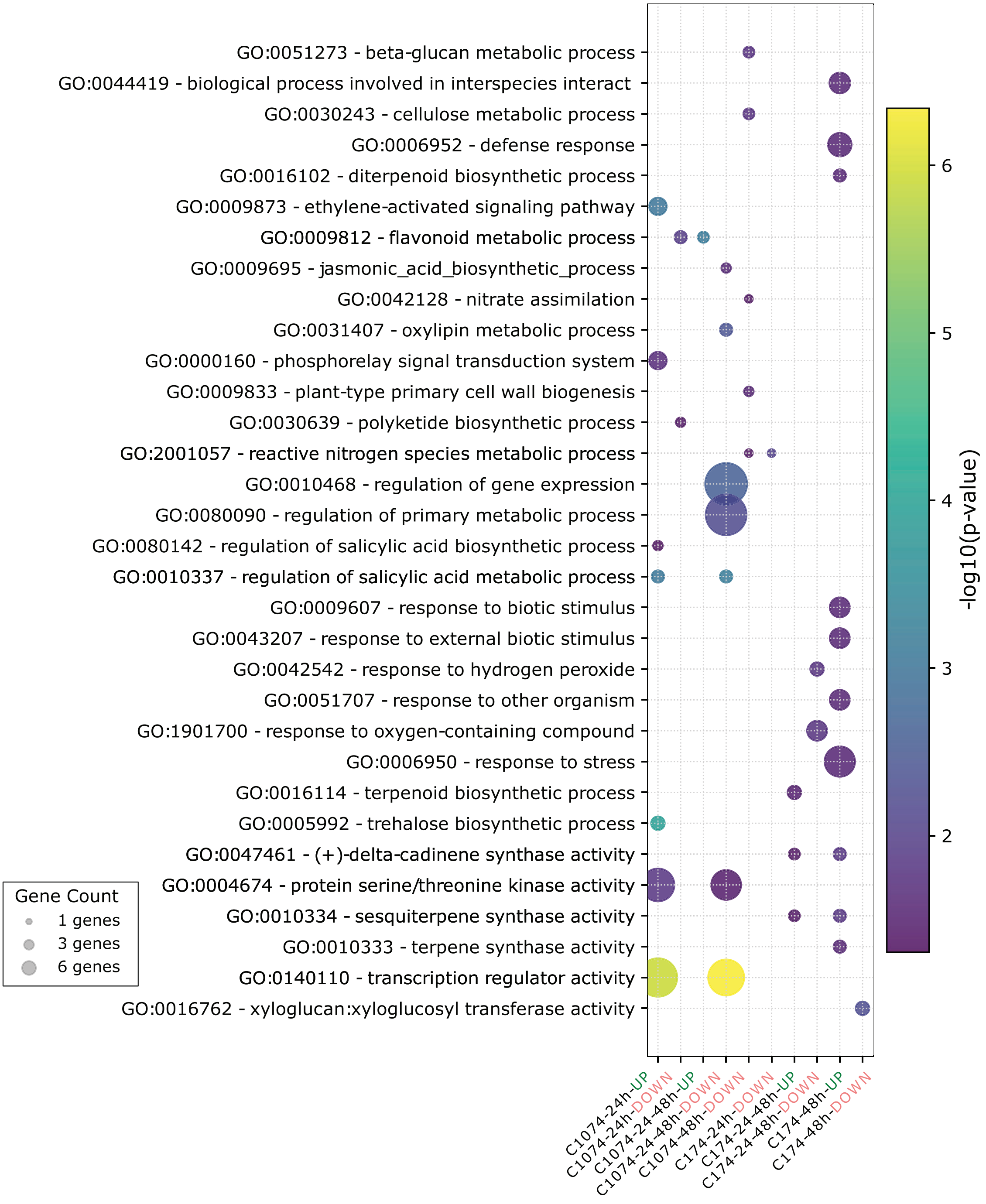

FIGURE 8 Gene ontology (GO) enrichment analysis for differentially expressed genes (DEGs) at various time points after infection with *Moniliophthora perniciosa*. DEGs upregulated or downregulated for each genotype (C174 and C1074) 24 h after *M. perniciosa* inoculation (hai), 24 hai versus 48 hai, and 48 at hai. Enriched GO terms highlight distinct biological processes and molecular functions, with C1074 DEGs predominantly associated with hormonal signaling and transcriptional regulation, and C174 DEGs enriched for defense response and secondary metabolism pathways.

In cupuassu, their enrichment near defense loci (Figures 3 and 4) and co-expression with terpene-biosynthetic and ROS-metabolism genes (Figure 8) suggest regulatory coupling between TEs and defense pathways.

Within the resistance QTL, NLRs, PRs, PRRs, and TFs are frequently flanked by TEs (Figure 6), echoing TE-centered feedback loops reported in species such as rice, tomato, soybean, cotton, and wheat (Castanera et al., 2023; Domínguez et al., 2020; Panchy et al., 2016; Xiao et al., 2008). In *Capsella rubella*, post-bottleneck TE insertions near promoters altered adaptive-trait expression (Niu et al., 2019), illustrating how genomic context shapes regulatory potential.

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In cupuassu, despite C1074 harboring more NLRs (541 vs. 434), none are upregulated, whereas 14 are induced in C174 upon infection (Figure 7). Thus, effective resistance depends more on regulatory integration than on gene number, consistent with *pan-NLRome* analyses (Barragan & Weigel, 2021), and recent findings in sorghum showing stronger NLR activation in resistant lines despite comparable member counts (Zhang et al., 2025). This supports a model in which elevated heterozygosity, localized duplication, and TE-mediated regulatory diversity converge in C174 to enable flexible immune activation.

The Chromosome 6 QTL also exhibits enriched SNP density, indicating a highly differentiated haplotype in C174. Elevated nucleotide diversity at resistance loci is a recurring hallmark of balancing selection, as seen in the *Sr2* locus in wheat (Mago et al., 2014), the *Rphq2* QTL in barley (Yeo et al., 2016), and clustered *R*-genes in *Arabidopsis* (Bakker et al., 2006; Borevitz et al., 2007). The pattern in cupuassu is consistent with long-term maintenance of divergent haplotypes underlying quantitative resistance. Notably, two C174-specific DUF4220/DUF594 gene duplicates (*TgrandC174G00000023870* and *TgrandC174G00000023871*) are strongly induced at 24–48 hai. Phylogenetically, they cluster with the maize homolog LOC103634932 (Figure S6), a membrane-associated, pathogen-inducible protein linked to carbohydrate partitioning and fungal defense (Miranda et al., 2017; Zhu et al., 2017). Expression and selection analyses suggest recent functional diversification, making these genes promising candidates for promoter-reporter or gain/loss-of-function validation.

Reanalysis of the time-course transcriptomes from Falcão et al. (2022) integrated with our new assemblies reinforces their model of early salicylic- and ethylene-signaling in susceptible tissues versus delayed ROS-centered activation in resistant ones. The genome assemblies now connect these transcriptional trajectories to their genomic contexts. C1074 exhibits a transient, SA/ET-enriched response at 24 hai that declines by 48 hai (Figure 8), consistent with feedback-suppressed responses in vulnerable hosts (López-Sánchez et al., 2016). Conversely, C174 shows a delayed but coordinated activation encompassing PRRs, MAPKs, ROS-responsive TFs, and inducible NLRs, culminating in strengthened defense at 48 hai (Figures 5, 7, and 8). Whether these temporal shifts result from gene duplication or nearby TE insertions warrants further functional analysis.

Despite C1074's larger nominal NLR inventory, its defense response is less coherent, with weaker induction of signaling and downstream pathways. TE insertions in C1074 occur more heterogeneously around immune loci, while in C174 they cluster within defined neighborhoods (Figures 3 and 4), potentially affecting chromatin configuration and expression potential. Comparable expression polymorphisms in *Arabidopsis R*-genes (Bakker et al., 2006) suggest that such regulatory variation is an evolutionarily conserved mechanism for modulating immune activation.

In summary, C174 displays a genomic architecture enriched for heterozygosity, tandem duplication, and TE-associated regulatory elements, yielding a flexible and coordinated immune response, whereas C1074 possesses a numerically larger but transcriptionally inert defense repertoire. These results align with patterns observed in rice, tomato, soybean, and *Arabidopsis*, where TE-mediated regulatory remodeling—rather than gene count—influences durable resistance (Barragan & Weigel, 2021; Bhattacharyya et al., 1997; Castanera et al., 2023; Domínguez et al., 2020; Gao & Bhattacharyya, 2008).

By providing the first chromosome-scale genome of a WBD-resistant genotype, alongside genotype-resolved transcriptome data (Table S7A,B; Figure S12), a comprehensive TE–gene neighborhood map, and a curated catalog of ∼320 candidate genes (Tables S8 and S9), this study establishes a multilayered resource for the *Theobroma* research community. Beyond these two contrasting genotypes, the integrative framework developed here offers a foundation for future pan-genome efforts to capture the full structural and regulatory spectrum of *T. grandiflorum*. Combining such analyses with high-resolution methylomes, histone-modification maps, and single-cell expression data will further reveal the evolutionary and epigenomic bases of disease resistance—supporting both fundamental discovery and the sustainable genetic improvement of tropical perennial crops central to Amazonian biodiversity and bioeconomy.

AUTHOR CONTRIBUTIONS

Vinicius A. C. de Abreu: Conceptualization; data curation; formal analysis; funding acquisition; investigation; methodology; project administration; resources; software; supervision; validation; visualization; writing—original draft; writing—review and editing. **Rafael Moysés Alves**: Conceptualization; funding acquisition; investigation; methodology; project administration; resources; supervision; writing—original draft; writing—review and editing. **Mauro de Medeiros Oliveira**: Data curation; formal analysis; validation; writing—original draft; writing—review and editing. **Vitor Trinca**: Formal analysis; investigation; software; validation; visualization; writing—original draft; writing—review and editing. **Loeni Ludke Falcão**: Investigation; resources; writing—original draft; writing—review and editing. **Lucilia Helena Marcellino**: Investigation; resources; writing—original draft. **Antonio Figueira**: Formal analysis; investigation; validation; writing—original draft; writing—review and editing. **Douglas S. Domingues**: Investigation; methodology; writing—original draft; writing—review and editing. **Alessandro M. Varani**: Conceptualization; data curation; formal analysis; funding acquisition; investigation;

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

methodology; project administration; resources; software; supervision; validation; visualization; writing—original draft; writing—review and editing.

ACKNOWLEDGMENTS

The authors would like to thank the Arizona Genomics Institute (Tucson, AZ) for providing support with HMW DNA extraction and PacBio Sequel IIe sequencing. We acknowledge the Universidade Federal do Pará (UFPA) for its general support. This study was funded by the "Fundação de Amparo à Pesquisa do Estado de São Paulo"—FAPESP, grant number 2019/25176-0 to Alessandro M. Varani and "Fundação Amazônia de Amparo a Estudos e Pesquisas"—FAPESPA, grant number 075/2020 to Vinicius A. C. de Abreu and Rafael Moysés Alves. Alessandro M. Varani, Antonio Figueira, and Douglas S. Domingues are currently supported by the Brazilian National Council for Scientific and Technological Development (CNPq) productivity grants (304367/2022-2, 313174/2022, and 310645/2021). Mauro de Medeiros Oliveira and Vitor Trinca are supported by FAPESP grants 2023/04372-1 and 2023/10314-4, respectively. These funding agencies had no role in the study design, data collection, analysis, interpretation, or manuscript writing.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The *T. grandiflorum* samples (GenBank BioSample SAMN17274453) were included at the National Genetic Heritage and Associated Traditional Knowledge Management System (SisGen) under accession #A2A72C6. The complete genome was deposited in GenBank under BioProject PRJNA691024 (CP142363–CP142372); the raw reads are available at GenBank Sequence Read Archive (SRA) under the following accession numbers: SRR28297999 and SRR26316972. All genome assemblies, gene models, and functional annotation files (GFF3 and FASTA formats) are accessible via our genome browser at https://plantgenomics.ncc.unesp.br/Genomes/TheobromaDB/. The Cupuassu Integrative Genomics Dashboard (https://plantgenomics.ncc.unesp.br/Genomes/TheobromaDB/Genes_Tables/) provides an interactive interface compiling the full content of Tables S7A,B and S8, including all annotated genes, expression data, and candidate loci within the resistance QTL. Together, these resources constitute the complete *Theobroma grandiflorum* multi-omics dataset generated in this study, offering open, dynamic access to the data underlying all analyses.

ORCID

Vinicius A. C. de Abreu https://orcid.org/0000-0002-4243-2421
Rafael Moysés Alves https://orcid.org/0000-0002-9826-4690
Mauro de Medeiros Oliveira https://orcid.org/0000-0002-2048-6664
Vitor Trinca https://orcid.org/0000-0002-9730-7492
Loeni Ludke Falcão https://orcid.org/0000-0002-3395-1254
Lucilia Helena Marcellino https://orcid.org/0000-0002-0430-2490
Antonio Figueira https://orcid.org/0000-0001-8641-2556
Douglas S. Domingues https://orcid.org/0000-0002-1290-0853
Alessandro M. Varani https://orcid.org/0000-0002-8876-3269

REFERENCES

Alves, R. M., & Chaves, S. F. S. (2023). Selection of *Theobroma grandiflorum* clones adapted to agroforestry systems using an additive index. *Acta Scientiarum Agronomy*, *45*, e57519. https://doi.org/10.4025/actasciagron.v45i1.57519

Alves, R. M., de Abreu, V. A. C., Oliveira, R. P., Almeida, J. V. D. A., de Oliveira, M. D. M., Silva, S. R., Paschoal, A. R., de Almeida, S. S., de Souza, P. A. F., Ferro, J. A., Miranda, V. F. O., Figueira, A., Domingues, D. S., & Varani, A. M. (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

Alves, R. M., & Resende, M. D. V. (2008). Avaliação genética de indivíduos e progênies de cupuaçuzeiro no estado do Pará e estimativas de parâmetros genéticos. *Revista Brasileira de Fruticultura*, *30*, 696–701. https://doi.org/10.1590/S0100-29452008000300023

Argout, X., Martin, G., Droc, G., Fouet, O., Labadie, K., Rivals, E., Aury, J. M., & Lanaud, C. (2017). The cacao Criollo genome v2.0: An improved version of the genome for genetic and functional genomic studies. *BMC Genomics*, *18*(1), 730. https://doi.org/10.1186/s12864-017-4120-9

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

Bakker, E. G., Toomajian, C., Kreitman, M., & Bergelson, J. (2006). A genome-wide survey of *R* gene polymorphisms in *Arabidopsis*. *The Plant Cell*, *18*(8), 1803–1818. https://doi.org/10.1105/tpc.106.042614

Barbosa, C. S., Fonseca, R. R. D., Batista, T. M., Barreto, M. A., Argolo, C. S., Carvalho, M. R. D., Amaral, D. O. J. D., Silva, E. M. D. A., Arévalo-Gardini, E., Hidalgo, K. S., Franco, G. R., Pirovani, C. P., Micheli, F., & Gramacho, K. P. (2018). Genome sequence and effectorome of *Moniliophthora perniciosa* and *Moniliophthora roreri* subpopulations. *BMC Genomics*, *19*(1), 509. https://doi.org/10.1186/s12864-018-4875-7

Barragan, A. C., & Weigel, D. (2021). Plant NLR diversity: The known unknowns of pan-NLRomes. *The Plant Cell*, *33*(4), 814–831. https://doi.org/10.1093/plcell/koaa002

Bhattacharyya, M. K., Gonzales, R. A., Kraft, M., & Buzzell, R. I. (1997). A copia-like retrotransposon Tgmr closely linked to the Rps1-k allele that confers race-specific resistance of soybean to

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Phytophthora sojae. *Plant Molecular Biology*, *34*(2), 255–264. https://doi.org/10.1023/A:1005851623493

Borevitz, J. O., Hazen, S. P., Michael, T. P., Morris, G. P., Baxter, I. R., Hu, T. T., Chen, H., Werner, J. D., Nordborg, M., Salt, D. E., Kay, S. A., Chory, J., Weigel, D., Jones, J. D. G., & Ecker, J. R. (2007). Genome-wide patterns of single-feature polymorphism in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(29), 12057–12062. https://doi.org/10.1073/pnas.0705323104

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1), 199. https://doi.org/10.1186/s13059-018-1577-z

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Buschiazzo, E., Ritland, C., Bohlmann, J., & Ritland, K. (2012). Slow but not low: Genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evolutionary Biology*, *12*(1), 8. https://doi.org/10.1186/1471-2148-12-8

Cabanettes, F., & Klopp, C. (2018). D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*, e4958. https://doi.org/10.7717/peerj.4958

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Castanera, R., Morales-Diaz, N., Gupta, S., Purugganan, M., & Casacuberta, J. M. (2023). Transposons are important contributors to gene expression variability under selection in rice populations. *Elife*, *12*, RP86324. https://doi.org/10.7554/eLife.86324.3

Chai, C., Shankar, R., Jain, M., & Subudhi, P. K. (2018). Genome-wide discovery of DNA polymorphisms by whole genome sequencing differentiates weedy and cultivated rice. *Scientific Reports*, *8*(1), 14218. https://doi.org/10.1038/s41598-018-32513-z

Chen, Y., Zhang, Y., Wang, A. Y., Gao, M., & Chong, Z. (2021). Accurate long-read de novo assembly evaluation with inspector. *Genome Biology*, *22*(1), 312. https://doi.org/10.1186/s13059-021-02527-4

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., & Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Cornejo, O. E., Yee, M.-C., Dominguez, V., Andrews, M., Sockell, A., Strandberg, E., Livingstone, D., Stack, C., Romero, A., Umaharan, P., Royaert, S., Tawari, N. R., Ng, P., Gutierrez, O., Phillips, W., Mockaitis, K., Bustamante, C. D., & Motamayor, J. C. (2018). Population genomic analyses of the chocolate tree, *Theobroma cacao* L., provide insights into its domestication process. *Communications Biology*, *1*(1), 167. https://doi.org/10.1038/s42003-018-0168-6

Costa, R. S. D., Santos, O. V. D., Lannes, S. C. D. S., Casazza, A. A., Aliakbarian, B., Perego, P., Ribeiro-Costa, R. M., Converti, A., & Silva Júnior, J. O. C. (2020). Bioactive compounds and value-added applications of cupuassu (*Theobroma grandiflorum* Schum.) agroindustrial by-product. *Food Science and Technology*, *40*, 401–407. https://doi.org/10.1590/fst.01119

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

de Matos, J. P., Ribeiro, D. F., da Silva, A. K., de Paula, C. H., Cordeiro, I. F., Lemes, C. G. D. C., Sanchez, A. B., Rocha, L. C. M., Garcia, C. C. M., Almeida, N. F., Alves, R. M., de Abreu, V. A. C., Varani, A. M., & Moreira, L. M. (2024). Diversity and potential functional role of phyllosphere-associated actinomycetota isolated from cupuassu (*Theobroma grandiflorum*) leaves: Implications for ecosystem dynamics and plant defense strategies. *Molecular Genetics and Genomics*, *299*(1), 73. https://doi.org/10.1007/s00438-024-02162-1

Deneweth, J., Van de Peer, Y., & Vermeirssen, V. (2022). Nearby transposable elements impact plant stress gene regulatory networks: A meta-analysis in *A. thaliana* and *S. lycopersicum*. *BMC Genomics*, *23*(1), 18. https://doi.org/10.1186/s12864-021-08215-8

Dievart, A., Gottin, C., Périn, C., Ranwez, V., & Chantret, N. (2020). Origin and diversity of plant receptor-like kinases. *Annual Review of Plant Biology*, *71*, 131–156. https://doi.org/10.1146/annurev-arplant-073019-025927

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2012). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J. M., Colot, V., & Quadrana, L. (2020). The impact of transposable elements on tomato diversity. *Nature Communications*, *11*(1), 4058. https://doi.org/10.1038/s41467-020-17874-2

dos Santos, C., & Franco, O. L. (2023). Pathogenesis-related proteins (PRs) with enzyme activity activating plant defense responses. *Plants*, *12*(11), 2226. https://doi.org/10.3390/plants12112226

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., & Aiden, E. L. (2017). De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. https://doi.org/10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems*, *3*(1), 99–101. https://doi.org/10.1016/j.cels.2015.07.012

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems*, *3*(1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002

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. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Falcão, L. L., Silva-Werneck, J. O., Albuquerque, P. S. B., Alves, R. M., Grynberg, P., Togawa, R. C., Costa, M. M. D. C., Brigido, M. M., & Marcellino, L. H. (2022). Comparative transcriptomics of cupuassu (*Theobroma grandiflorum*) offers insights into the early defense mechanism to *Moniliophthora perniciosa*, the causal agent of witches' broom disease. *Journal of Plant Interactions*, *17*(1), 991–1005. https://doi.org/10.1080/17429145.2022.2144650

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Flagel, L. E., & Wendel, J. F. (2009). Gene duplication and evolutionary novelty in plants. *New Phytologist*, *183*(3), 557–564. https://doi.org/10.1111/j.1469-8137.2009.02923.x

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Gao, H., & Bhattacharyya, M. K. (2008). The soybean-phytophthora resistance locus Rps1-k encompasses coiled coil-nucleotide binding-leucine rich repeat-like genes and repetitive sequences. *BMC Plant Biology*, *8*(1), 29. https://doi.org/10.1186/1471-2229-8-29

Gong, Z., & Han, G.-Z. (2021). Flourishing in water: The early evolution and diversification of plant receptor-like kinases. *The Plant Journal*, *106*(1), 174–184. https://doi.org/10.1111/tpj.15157

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., & White, O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., & … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hassan, A. H., Mokhtar, M. M., & El Allali, A. (2024). Transposable elements: Multifunctional players in the plant genome. *Frontiers in Plant Science*, *14*, 1330127. https://doi.org/10.3389/fpls.2023.1330127

Jain, C., Rhie, A., Hansen, N. F., Koren, S., & Phillippy, A. M. (2022). Long-read mapping to repetitive reference sequences using Winnowmap2. *Nature Methods*, *19*(6), 705–710. https://doi.org/10.1038/s41592-022-01457-8

Jin, Y., Tam, O. H., Paniagua, E., & Hammell, M. (2015). TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, *31*(22), 3593–3599. https://doi.org/10.1093/bioinformatics/btv422

Jones, J. D. G., & Dangl, J. L. (2006). The plant immune system. *Nature*, *444*(7117), 323–329. https://doi.org/10.1038/nature05286

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., & Tang, H. (2018). GOATOOLS: A python library for gene ontology analyses. *Scientific Reports*, *8*(1), 10872. https://doi.org/10.1038/s41598-018-28948-z

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., & Zdobnov, E. M. (2018). OrthoDB v10: Sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, *47*(D1), D807–D811. https://doi.org/10.1093/nar/gky1053

Lars, G., Tomáš, B., Katharina, J. H., Matthis, E., Alexandre, L., Mark, B., & Mario, S. (2023). BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv. https://doi.org/10.1101/2023.06.10.544449

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics*, *37*(23), 4572–4574. https://doi.org/10.1093/bioinformatics/btab705

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, S.-F., She, H.-B., Yang, L.-L., Lan, L.-N., Zhang, X.-Y., Wang, L.-Y., Zhang, Y.-L., Li, N., Deng, C.-L., Qian, W., & Gao, W.-J. (2022). Impact of LTR-retrotransposons on genome structure, evolution, and function in Curcurbitaceae species. *International Journal of Molecular Sciences*, *23*(17), 10158. https://doi.org/10.3390/ijms231710158

Li, X., Ma, L., Wang, Y., Ye, C., Guo, C., Li, Y., Mei, X., Du, F., & Huang, H. (2023). PlantNLRatlas: A comprehensive dataset of full- and partial-length NLR resistance genes across 100 chromosome-level plant genomes. *Frontiers in Plant Science*, *14*, 1178069. https://doi.org/10.3389/fpls.2023.1178069

Lin, Y., Ye, C., Li, X., Chen, Q., Wu, Y., Zhang, F., Pan, R., Zhang, S., Chen, S., Wang, X., Cao, S., Wang, Y., Yue, Y., Liu, Y., & Yue, J. (2023). quarTeT: A telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research*, *10*(8), uhad127. https://doi.org/10.1093/hr/uhad127

Liu, P.-L., Du, L., Huang, Y., Gao, S.-M., & Yu, M. (2017). Origin and diversification of leucine-rich repeat receptor-like protein kinase (*LRR-RLK*) genes in plants. *BMC Evolutionary Biology*, *17*(1), 47. https://doi.org/10.1186/s12862-017-0891-5

López-Sánchez, A., Stassen, J. H. M., Furci, L., Smith, L. M., & Ton, J. (2016). The role of DNA (de)methylation in immune responsiveness of *Arabidopsis*. *The Plant Journal*, *88*(3), 361–374. https://doi.org/10.1111/tpj.13252

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, J., Rincon, N., Wood, D. E., Breitwieser, F. P., Pockrandt, C., Langmead, B., Salzberg, S. L., & Steinegger, M. (2022). Metagenome analysis using the Kraken software suite. *Nature Protocols*, *17*(12), 2815–2839. https://doi.org/10.1038/s41596-022-00738-y

Ma, B., Kuang, L., Xin, Y., & He, N. (2019). New insights into long terminal repeat retrotransposons in mulberry species. *Genes*, *10*(4), 285. https://doi.org/10.3390/genes10040285

Mago, R., Tabe, L., Vautrin, S., Šimková, H., Kubaláková, M., Upadhyaya, N., Berges, H., Kong, X., Breen, J., Doležel, J., Appels, R., Ellis, J. G., & Spielmeyer, W. (2014). Major haplotype divergence including multiple germin-like protein genes, at the wheat Sr2 adult plant stem rust resistance locus. *BMC Plant Biology*, *14*(1), 379. https://doi.org/10.1186/s12870-014-0379-z

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, *1*(12), e323. https://doi.org/10.1002/cpz1.323

McDowell, J. M., & Meyers, B. C. (2013). A transposable element is domesticated for service in the plant immune system. *PNAS*, *110*(37), 14821–14822. https://doi.org/10.1073/pnas.1314089110

Meng, X., & Zhang, S. (2013). MAPK cascades in plant disease resistance signaling. *Annual Review of Phytopathology*, *51*, 245–266. https://doi.org/10.1146/annurev-phyto-082712-102314

Miranda, V. D. J., Porto, W. F., Fernandes, G. D. R., Pogue, R., Nolasco, D. O., Araujo, A. C. G., Cota, L. V., Freitas, C. G. D., Dias, S. C., & Franco, O. L. (2017). Comparative transcriptomic analysis indicates genes associated with local and systemic resistance to *Colletotrichum graminicola* in maize. *Scientific Reports*, *7*(1), 2483. https://doi.org/10.1038/s41598-017-02298-8

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mournet, P., de Albuquerque, P. S. B., Alves, R. M., Silva-Werneck, J. O., Rivallan, R., Marcellino, L. H., & 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 Genetics & Genomes*, *16*(6), 89. https://doi.org/10.1007/s11295-020-01479-3

Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K., & Kosakovsky Pond, S. L. (2012). Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*, *8*(7), e1002764. https://doi.org/10.1371/journal.pgen.1002764

Ngou, B. P. M., Ding, P., & Jones, J. D. G. (2022). Thirty years of resistance: Zig-zag through the plant immune system. *The Plant Cell*, *34*(5), 1447–1478. https://doi.org/10.1093/plcell/koac041

Niu, X. M., Xu, Y. C., Li, Z. W., Bian, Y. T., Hou, X. H., Chen, J. F., Zou, Y. P., Jiang, J., Wu, Q., Ge, S., Balasubramanian, S., & Guo, Y. L. (2019). Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *PNAS*, *116*(14), 6908–6913. https://doi.org/10.1073/pnas.1811498116

O'Donnell, S., & Fischer, G. (2020). MUM&Co: Accurate detection of all SV types through whole-genome alignment. *Bioinformatics*, *36*(10), 3242–3243. https://doi.org/10.1093/bioinformatics/btaa115

Orozco-Arias, S., Dupeyron, M., Gutiérrez-Duque, D., Tabares-Soto, R., & Guyot, R. (2023). High nucleotide similarity of three *Copia* lineage LTR retrotransposons among plant genomes. *Genome*, *66*(3), 51–61. https://doi.org/10.1139/gen-2022-0026

Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR assembly index (LAI). *Nucleic Acids Research*, *46*(21), e126. https://doi.org/10.1093/nar/gky730

Ou, S., & Jiang, N. (2017). LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiology*, *176*(2), 1410–1422. https://doi.org/10.1104/pp.17.01310

Ou, S., Scheben, A., Collins, T., Qiu, Y., Seetharam, A. S., Menard, C. C., Manchanda, N., Gent, J. I., Schatz, M. C., Anderson, S. N., Hufford, M. B., & Hirsch, C. N. (2024). Differences in activity and stability drive transposable element variation in tropical and temperate maize. *Genome Research*, *34*(8), 1140–1153. https://doi.org/10.1101/gr.278131.123

Panchy, N., Lehti-Shiu, M., & Shiu, S.-H. (2016). Evolution of gene duplication in plants. *Plant Physiology*, *171*(4), 2294–2316. https://doi.org/10.1104/pp.16.00523

Pertea, G., & Pertea, M. (2020). GFF utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. *F1000 Research*, *9*(304). https://doi.org/10.12688/f1000research.23297.2

Pugliese, A. G., Tomas-Barberan, F. A., Truchado, P., & Genovese, M. I. (2013). Flavonoids, proanthocyanidins, vitamin C, and antioxidant activity of *Theobroma grandiflorum* (cupuassu) pulp and seeds. *Journal of Agricultural and Food Chemistry*, *61*(11), 2720–2728. https://doi.org/10.1021/jf304349u

Qiao, X., Li, Q., Yin, H., Qi, K., Li, L., Wang, R., Zhang, S., & Paterson, A. H. (2019). Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biology*, *20*(1), 38. https://doi.org/10.1186/s13059-019-1650-2

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications*, *11*(1), 1432. https://doi.org/10.1038/s41467-020-14998-3

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology*, *21*(1), 245. https://doi.org/10.1186/s13059-020-02134-9

Robert-Seilaniantz, A., Grant, M., & Jones, J. D. G. (2011). Hormone crosstalk in plant disease and defense: More than just jasmonate-salicylate antagonism. *Annual Review of Phytopathology*, *49*, 317–343. https://doi.org/10.1146/annurev-phyto-073009-114447

Rosa, J. S. D., Oliveira Moreira, P. I., Carvalho, A. V., & Freitas-Silva, O. (2024). Cupuassu fruit, a non-timber forest product in sustainable bioeconomy of the Amazon—A mini review. *Processes*, *12*(7), 1353. https://doi.org/10.3390/pr12071353

Santana Silva, R. J., Alves, R. M., Peres Gramacho, K., Marcellino, L. H., & Micheli, F. (2020). Involvement of structurally distinct cupuassu chitinases and osmotin in plant resistance to the fungus *Moniliophthora perniciosa*. *Plant Physiology and Biochemistry*, *148*, 142–151. https://doi.org/10.1016/j.plaphy.2020.01.009

Seo, E., Choi, D., & Choi (2015). Functional studies of transcription factors involved in plant defenses in the genomics era. *Briefings in Functional Genomics*, *14*(4), 260–267. https://doi.org/10.1093/bfgp/elv011

Smith, M., Jones, J. T., & Hein, I. (2025). Resistify: A novel NLR classifier that reveals Helitron-associated NLR expansion in Solanaceae. *Bioinformatics and Biology Insights*, *19*, 1–9. https://doi.org/10.1177/11779322241308944

Smith, M. D., Wertheim, J. O., Weaver, S., Murrell, B., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution*, *32*(5), 1342–1353. https://doi.org/10.1093/molbev/msv022

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*, W609–W612. https://doi.org/10.1093/nar/gkl315

Tamura, K., & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and chimpanzees. *Molecular Biology and Evolution*, *10*(3), 512–526. https://doi.org/10.1093/oxfordjournals.molbev.a040023

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., & Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science*, *320*(5875), 486–488. https://doi.org/10.1126/science.1153917

Teresi, S. J., Teresi, M. B., & Edger, P. P. (2022). TE density: A tool to investigate the biology of transposable elements. *Mobile DNA*, *13*(1), 11. https://doi.org/10.1186/s13100-022-00264-4

Vuruputoor, V. S., Monyak, D., Fetter, K. C., Webster, C., Bhattarai, A., Shrestha, B., Zaman, S., Bennett, J., McEvoy, S. L., Caballero, M., & Wegrzyn, J. L. (2023). Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. *Applications in Plant Sciences*, *11*(4), e11533. https://doi.org/10.1002/aps3.11533

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., & Yu, J. (2010). KaKs_Calculator 2.0: A toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, Proteomics & Bioinformatics*, *8*(1), 77–80. https://doi.org/10.1016/s1672-0229(10)60008-3

Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L., & Scheffler, K. (2014). RELAX: Detecting relaxed selection in a phylogenetic framework. *Molecular Biology and Evolution*, *32*(3), 820–832. https://doi.org/10.1093/molbev/msu400

Wong, T. K. F., Ly-Trong, N., Ren, H., Baños, H., Roger, A. J., Susko, E., Bielow, C., Maio, N. D., Goldman, N., Hahn, M. W., Huttley, G., Lanfear, R., & Minh, B. Q. (2025). IQ-TREE 3: Phylogenomic inference software using complex evolutionary models. *EcoEvoRxiv*. https://doi.org/10.32942/X2P62N

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., & van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science & Culture*, *319*(5869), 1527–1530. https://doi.org/10.1126/science.1153040

Yeo, F. K. S., Wang, Y., Vozabova, T., Huneau, C., Leroy, P., Chalhoub, B., Qi, X. Q., Niks, R. E., & Marcel, T. C. (2016). Haplotype divergence and multiple candidate genes at *Rphq2*, a partial resistance QTL of barley to *Puccinia hordei*. *Theoretical and Applied Genetics*, *129*(2), 289–304. https://doi.org/10.1007/s00122-015-2627-5

Zhang, J. W., Li, J. Y., Yu, Z. F., Chang, X. Y., Han, J. R., Xia, J. Y., Kami, Y. B., Sun, Y. T., Li, L., Wang, S. T., Ni, X. L., Wang, H., Li, Y., & Wang, W. M. (2025). Comparative genomic analysis reveals the difference of NLR immune receptors between anthracnose-resistant and susceptible sorghum cultivars. *Phytopathology Research*, *7*(1), 29. https://doi.org/10.1186/s42483-025-00318-4

Zheng, Y., Jiao, C., Sun, H., Rosli, H. G., Pombo, M. A., Zhang, P., Banf, M., Dai, X., Martin, G. B., Giovannoni, J. J., Zhao, P. X., Rhee, S. Y., & Fei, Z. (2016). iTAK: A program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Molecular Plant*, *9*(12), 1667–1670. https://doi.org/10.1016/j.molp.2016.09.014

Zhu, X., Shen, W., Huang, J., Zhang, T., Zhang, X., Cui, Y., Sang, X., Ling, Y., Li, Y., Wang, N., Zhao, F., Zhang, C., Yang, Z., & He, G. (2017). Mutation of the OsSAC1 gene, which encodes an endoplasmic reticulum protein with an unknown function, causes sugar accumulation in rice leaves. *Plant and Cell Physiology*, *59*(3), 487–499. https://doi.org/10.1093/pcp/pcx203

Zimin, A. V., Puiu, D., Luo, M. C., Zhu, T., Koren, S., Marcais, G., Yorke, J. A., Dvorak, J., & Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of *Aegilops tauschii*, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research*, *27*(5), 787–792. https://doi.org/10.1101/gr.213405.116

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: de Abreu, V. A. C., Alves, R. M., Oliveira, M. M., Trinca, V., Falcão, L. L., Marcellino, L. H., Figueira, A., Domingues, D. S., & Varani, A. M. (2026). Integrative chromosome-scale genome analysis of cupuassu provides insights into witches' broom disease resistance and expands genomic resources for *Theobroma*. *The Plant Genome*, *19*, e70196. https://doi.org/10.1002/tpg2.70196

19403372, 2026, 1, Downloaded from https://acsess.onlinelibrary.wiley.com/doi/10.1002/tpg2.70196 by Capes, Wiley Online Library on [06/03/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License