Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SNP-based genomic variation in candidate disease resistance genes of Theobroma cacao

Gabriela Nicolle Ramos-Lizardo
State University of Santa Cruz
Konstantinos G. Alexiou
Center for Research in Agricultural Genomics
Jonathan Javier Mucherino-Muñoz
Federal University of Rio de Janeiro
María José Aranzana
Center for Research in Agricultural Genomics
Ronan Xavier Corrêa
ronanxc@uesc.br

State University of Santa Cruz

Article

Keywords: Disease resistance, Cocoa, Quantitative trait loci (QTLs), Plant–pathogen interaction, Marker-assisted selection, Resistance gene analogs

Posted Date: June 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9441324/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Theobroma cacao is a species with high economic value whose production is impaired by multiple factors, including yield losses caused by pathogens such as *Moniliophthora perniciosa*, *Phytophthora* spp., *Ceratocystis cacaofunesta*, and *Moniliophthora roreri*. In this context, knowledge of the genetic variability underpinning esistance is crucial to elucidate plant–pathogen interactions and develop resistant cultivars. In this study, we employed bioinformatics tools to predict candidate resistance genes and analyze their genomic variability based on SNPs in the *T. cacao* genome. A total of 4,523 candidate genes encoding 6,193 proteins were identified, among which 196 genes remain uncharacterized but contain domains associated with defense-related functions, including NBS-LRR domains, kinases and pathogenesis-related proteins. Analysis of coding regions revealed 17,576 SNPs, of which 9,014 were classified as missense variants, 8,418 were synonymous variants, and 156 were high-impact variants, including start-loss, stop-loss, and stop-gain mutations. Enrichment analysis of these coding-region genes identified biological pathways and processes involved in plant defense, including oxidative phosphorylation, plant–pathogen interaction, MAPK signaling pathways in plants and phenylpropanoid biosynthesis. Furthermore, five genes were located within quantitative trait loci previously associated with resistance to *Phytophthora* spp., and that encode proteins such as class I chitinase, protease, ACD11, NADH oxidoreductase (NADH-OxR) and aquaporin NIP6-1, which are directly involved in defense responses. These results broaden understanding of defense mechanisms in *T. cacao* genotypes, identify potential candidates for future functional validation, and can be applied in marker-assisted breeding programs focused on disease resistance.

Introduction

Cocoa (*T. cacao L.*) is a perennial fruit tree typical of tropical regions (Lahive et al., 2019), with high economic importance due to the use of its seeds as raw material for the production of various products used in the food, cosmetic and pharmaceutical industries (Kongor et al., 2024). According to the 2023/24 bulletin of the International Cocoa Organization (ICCO), Brazil ranks sixth in global cocoa production. Data from the Brazilian Institute of Geography and Statistics (IBGE) for 2024 indicated that Brazil reached a total production of 297,509 tons, with the states of Pará (137,460 t) and Bahia (137,030 t) being the main national producers. In recent decades, cocoa production in several countries has been hampered by various biotic and abiotic factors that compromise crop development. Among the biotic factors are diseases caused by *M. perniciosa* (witches' broom), *Phytophthora* spp. (black pod rot), *C. cacaofunesta* (ceratocystis wilt), and *M. roreri* (frosty pod rot) (Marelli J.P. *et al*., 2019; Carrero-Gutiérrez M.L. *et al*., 2025). These diseases negatively impact crop productivity, posing a major challenge to the sustainability of cocoa cultivation.

Currently, several strategies are employed to mitigate the impact of pathogens that attack cocoa plants, including cultural management practices, the use of antagonistic microorganisms as biological control agents, chemical control, and the use of resistant genotypes obtained through plant breeding (Sudarma, I. M. et al. 2023, Wijaya, G. et al. 2023, Jibat, M. *et al.*, 2023, Adeniran. A.O. *et al.*, 2024, Pires, J. L. et al.,

2025). In addition to these approaches, plants possess an innate immune system that enables them to recognize and respond to pathogen attack. During the initial interaction, plants detect conserved molecular patterns associated with microorganisms (PAMPs) through pattern recognition receptors (PRRs) located in the plasma membrane, activating the first line of defense, called pattern-triggered immunity (PTI). At this stage, several defense pathways are rapidly activated, including the production of reactive oxygen species (ROS), the activation of MAPK cascades, and the deposition of callose in cell walls (Yuan, M. et al., 2021).

In the second line of defense, known as effector-triggered immunity (ETI), pathogens secrete proteins or small molecules called effectors, which are recognized by plants through intracellular nucleotide-binding leucine-rich repeat (NLR) receptors. Activation of these receptors generally triggers a more intense defense response, often associated with a localized hypersensitive response (HR) (Yuan, M., et al 2021[1], Yuan, M. et al., 2021[2]). Recent studies demonstrate that these two lines of defense in plant immunity do not act independently, and there is considerable interconnection between them. Thus, components activated during PTI, such as the MAPKs MPK3/MPK6, are essential for modulating and amplifying ETI responses. Similarly, signals activated by NLRs during ETI can potentiate PTI (Wang, D. et al., 2023). These mechanisms are highly relevant because different pathogens secrete effector proteins that suppress PTI, thereby facilitating host colonization. Consequently, the identification of genes involved in pathogen recognition and signal transduction is essential for the development of resistant cultivars (Li, Q. et al., 2024).

The availability of genomic and RNA-seq data for *T. cacao* has enabled the identification of genes associated with agronomic traits and genetic improvement (Kulesza, E. et al., 2024). Genotyping studies of *T. cacao* using SNP markers in germplasm collections and diversity panels have demonstrated that allelic variation detected by these markers constitutes a robust tool for inferring genetic diversity, detecting signatures of selection, and identifying loci associated with disease resistance and productivity (Nieves-Orduña, H.E. *et al.*, 2024). In this context, the integration of polymorphism (SNP) analyses, functional annotation of protein domains, and QTL and GWAS based mapping approaches has allowed the convergence of multiple lines of evidence, revealing the co-localization of SNPs, candidate genes, and QTLs. This integrative strategy strengthens the prioritization of candidate genes for functional validation and their application in marker-assisted selection programs (Fernandes et al., 2020).

This study aimed to predict genes potentially associated with resistance, functionally characterize them through the identification of conserved defense-related domains, analyze their genomic variability and the potential functional impacts caused by mutations present in these genes, and confirm their co-localization with previously identified resistance QTLs.

Methodology

In this study, we used the genome and proteome of *T. cacao* Criollo B97-61/B2 v.2 as references. These data are available in the GenBank of the National Center for Biotechnology Information (NCBI) under accession GCF_000208745.1. A total of 112 Whole-genome resequencing (WGS) libraries deposited in the NCBI Sequence Read Archive (SRA) (https://www.ncbi.nlm.nih.gov/sra) were analyzed, originating from BioProjects PRJNA558793, PRJNA734904, and PRJNA880462, as detailed in **Table S1**. These libraries represent different *T. cacao* genotypes from various **growing** regions. The different steps and tools used in each phase of data processing and annotation is summarized in the workflow shown in **Fig. 1.**

Selection and processing of WGS libraries

The selection of the 112 WGS genomic DNA libraries from the SRA was carried out to achieve broad genetic diversity and representativeness of the main populations and genotypes of *T. cacao*. The selection criteria included: the availability of data in FASTQ format, paired-end sequencing performed on the Illumina platform, and complete metadata analysis allowing the identification of the biological origin, cocoa variety, experimental conditions, and genotype resistance or susceptibility. Manual curation of the metadata was performed based on the SRA reports to classify each library according to its specific genotype and population of origin, including Amelonado, Criollo, Contamana, Nanay, Marañon, Guiana, Iquitos, Curaray, Nacional, Matina, and admixed genotypes. Additionally, reference genotypes were selected, both resistant (SCA6, SCA11, IMC65, NA672, NA127, MO76, RB41, RB43, RB46) and susceptible (LIB3, CATONGO, LCTEEN, EBC, TIWI, PAL8-A1) (Osorio-Guarín, J. *et al.,* 2020, Hämälä, T. *et al*, 2021, Argout, X. *et al.,* 2023) (**Table S1**). Subsequently, the SRA files were downloaded and converted to FASTQ format using the SRA Toolkit v.3.1.1 (Team 2021) with the “fastq-dump --split-files” command, ensuring high quality and uniformity of paired-end reads for each library.

Quality control and alignment

The quality of the reads was assessed using FastQC v0.11.9 (**Andrews S., 2010)**, considering standard parameters for read integrity, GC content, and per-base quality distribution. Reads containing residual adapters or low-quality bases were filtered with TrimGalore v0.4.5, retaining only fragments with a minimum length of 50 bp and a Phred score ≥ 30. The *T. cacao* reference genome (GenBank accession: GCA_000208745.2) was indexed using BWA index v0.7.17-r1188 (Li H & Durbin R., 2009), and the CreateSequenceDictionary module from Picard tools v1.110 (https://broadinstitute.github.io/picard/) was used to generate the sequence dictionary.

Paired-end reads were aligned using BWA-MEM (Li H & Durbin R., 2009), implemented in the Clara Parabricks v4.0.1 environment (NVIDIA) (Pham, M. *et al.,* 2023) with GPU acceleration. The resulting files were sorted by genomic coordinates and indexed with SAMtools v1.15 (Li H. *et al.,* 2009). PCR duplicates were identified using the MarkDuplicates module of Picard. Mapping quality was assessed through alignment metrics (CollectAlignmentSummaryMetrics) and mapping statistics (samtools flagstat) to confirm the integrity and coverage of the samples across the genome.

Variant calling and filtering

The identification of genetic variants, including SNPs and INDELs, was performed using the Genome Analysis Toolkit (GATK v4.1.6.0) (McKenna A. *et al*., 2010). BAM files were processed with the HaplotypeCaller module in gVCF format, allowing up to three alternative alleles per site (--max-alternate-alleles 3) and applying a minimum confidence limit of 20 (--stand-call-conf 20). Individual gVCFs were combined using CombineGVCFs, followed by the use of genotype gvcf to generate the raw VCF file. Variants were subjected to hard filtering separately for SNPs and INDELs, using predefined GATK VariantFiltration parameters. For SNPs, the following criteria were applied: Quality Depth (QD) < 2.0, Strand Odds Ratio (SOR) > 3.0, Fisher Strand Bias (FS) > 60.0, Mapping Quality (MQ) < 40.0, MQRankSum < -12.5, ReadPosRankSum < -8.0, and ExcessHet > 54.69. For INDELs, the applied filters were: QD < 2.0, QUAL < 30.0, FS > 200.0, ReadPosRankSum < -20.0, and ExcessHet > 54.69.

After the initial filtering, only biallelic variants were retained (--restrict-alleles-to BIALLELIC). Genotypes with low quality (GQ < 20) or read depth outside the acceptable range (DP ≥ 4 or DP > 300) were set to missing (--set-filtered-gt-to-nocall). Non-polymorphic variants were removed, and phased genotypes were collapsed to standardize annotation. Finally, variants with missing data were excluded using VCFtools v0.1.16 (--max-missing 1) (Danecek, P. *et al*., 2011). This process ensured the retention of highly reliable loci, with genotypes of sufficient quality and coverage for subsequent analyses.

Prediction of resistance genes in *T. cacao*

The prediction of resistance genes associated with the *T. cacao* B97-61/B2 v.2 genome was performed using data obtained from the Phytozome v.14 database (https://phytozome-next.jgi.doe.gov). For this purpose, the *Arabidopsis thaliana* TAIR10 genome was used as a reference, employing the keyword “resistance to disease” to retrieve genes related to plant defense, including representative families such as NBS-LRRs, pathogenesis-related (PR) proteins, defensins, and chitinases. The protein sequences corresponding to the obtained identifiers were downloaded using the TAIR Bulk Download tool (https://v2.arabidopsis.org/tools/bulk/sequences/index.jsp) and were used as input for BLASTp v2.12.0 analyses against the *T. cacao* B97-61/B2 v.2 proteome (GCA_000208745.1 v2). Strict filtering criteria were applied, considering coverage ≥ 75%, identity ≥ 60%, and E-value ≤ 1e-5, to identify putative homologs potentially associated with resistance. The proteins resulting from the BLASTp analyses were used to extract genomic coordinates, corresponding genes, and functional annotations, based on the *T. cacao* genome annotation (.gff) file.

Identification of functional domains

Proteins from *T. cacao* that were uncharacterized after the BLASTp analysis were subjected to conserved domain and motif annotation using InterProScan v.5.75-106.0, HMMER v.3.4, and DRAGO 3, a plant resistance gene prediction and annotation tool available on the PRGdb platform (Blum M, *et al.,* 2025; García J.C *et al.,* 2022), to identify functional regions and infer potential biological roles in plant defense.

Additionally, *T. cacao* proteins with previous functional annotation were also analyzed using DRAGO 3 to detect structural domains associated with resistance and pathogen recognition mechanisms.

Filtering of SNPs in resistance-associated genes

A BED file containing the genomic coordinates of the predicted resistance genes was generated, including both functionally annotated and non-annotated genes in the *T. cacao* genome. These coordinates were used to filter the hard-filtered SNP list using VCFtools v0.1.16 (Danecek, P. *et al.,* 2011), resulting in a set of variants located in genomic regions associated with resistance. Genomic variability was visualized using the Integrative Genomics Viewer (IGV) (Robinson J.T. *et al*., 2017) and GenoVarView (https://bioinformatics.cragenomica.es/projects/genovarview/), employing the *T. cacao* reference genome and the corresponding filtered SNP VCF file. IGV allowed the inspection and comparison of the presence or absence of variants across the 112 analyzed libraries.

Functional prediction and classification of variants

The functional effects of variants associated with putative resistance genes were predicted using the Ensembl Variant Effect Predictor for Plants (VEP, v113.0) (McLaren, W. *et al*., 2016). This prediction considered the impact of variants on coding regions and their potential functional consequences. Variants were classified according to their functional impact as**high**(**stop-gain, start-loss, stop-loss**),**moderate**(**missense**),or **low**(**synonymous**).Missense variants were subsequently analyzed with the SIFT tool to predict possible deleterious effects on the encoded proteins. SNPs with high and moderate impact consequences were filtered based on the observed allele frequencies across the 112 libraries and categorized as rare (<20%), intermediate (20–80%), or common (>80%).

Genetic diversity and enrichment analysis

Genetic diversity was estimated using SNPs located within the coding regions of the 2,460 previously predicted resistance-associated genes. Nucleotide diversity (π) and Tajima's D neutrality index (Nei, M., & Li, W., 1979; Tajima, F., 1983) were calculated using VCFtools v0.1.16 (Danecek, P. *et al.,* 2011) with a 1,000 bp sliding window. Regions potentially suitable for selection were considered those with Tajima's D values above the 95th percentile, indicating balancing selection, or below the 5th percentile, indicating positive selection across **the genome-wide distribution**. Windows exceeding these thresholds were classified as atypical regions and grouped according to the overlapping functional genes. To assess the significance of the clustering of these regions (hotspots), an empirical permutation test with 10,000 iterations was performed, comparing the observed odds ratio (OR) for the proportion of atypical regions per chromosome with the expected proportion under the null hypothesis of random distribution. A significance threshold of $P < 0.05$ was adopted.

Functional analysis of genes with coding SNPs

Genes containing SNPs in coding regions of high or moderate impact were subjected to functional enrichment analysis using the ShinyGO v0.85 platform (Ge, S. X. *et al*., 2020) and employing the KEGG

database (Kyoto Encyclopedia of Genes and Genomes, https://www.genome.jp/kegg/) to identify enriched metabolic pathways and associated biological processes. Enrichment results were filtered using a false discovery rate (FDR) threshold <0.05 and ranked according to the fold enrichment for each functional term (Kanehisa, M. *et al*., 2021). Graphical representations of these results were subsequently generated in R (v.4.4.3) using the ggplot2 package (Wickham, 2016), based on the enrichment outputs.

Genes associated with resistance QTLs

SNPs annotated by VEP as high-impact consequences were used as input. For each SNP, the coordinates were **retrieved** and the corresponding genes within the *T. cacao* Criollo v2 genome were identified. To locate candidate resistance genes, these SNPs were intersected with physical QTL intervals previously reported for resistance to *Phytophthora* spp. and *M. perniciosa*. The QTLs considered were obtained from the studies of Wong *et al*. (2022) and Mucherino-Muñoz *et al*. (2021), which provide genomic coordinates flanking the QTLs on chromosomes in the reference genome.

The association between SNPs and QTLs was performed using an in-house AWK script, which loaded the QTL intervals and compared each SNP, **identifying as associated those whose positions showed partial or complete overlap with a QTL interval.** Overlapping SNPs were considered as belonging to the reported QTLs, and genes containing these SNPs were considered as candidates for resistance.

Results

Prediction of genes associated with resistance

The search for genes associated with disease resistance in *A. thaliana* using the Phytozome database resulted in the identification of 6,482 candidate genes, including 114 genes encoding NBS-LRR–type proteins, 357 kinases, 43 defensin-like (DEFL) and 24 genes encoding thaumatin-like PR proteins (**Table S2**). After alignment of these sequences with the *T. cacao* proteome using BLASTp and subsequent filtering of the results, 4,523 genes were identified, encoding 6,193 proteins potentially associated with resistance. Among these, 4,294 genes had functional annotations related to defense mechanisms, comprising 301 genes encoding kinase-type proteins, 13 thaumatin-like proteins, one DEFL protein, one osmotin-like protein, and 226 genes without functional annotation (**Table S3**).

Functional domain annotation

The proteins encoded by the 4,294 annotated genes were analyzed using the DRAGO tool to identify resistance-associated domains. Of these, 272 genes contained a total of 2,025 domains, distributed across 389 encoded proteins. Among the identified structural and functional domain classes, CNL, RLP, RLK, TRAN, CN, CNLK, KIN, LEC, LECRK, LYK, N, CL, CK, and C were predominant, comprising the seven main types of conserved domains: CC, TM, kinase, LRR, NBS, LysM, and LECM. The KIN and RLK classes had the highest number of domains, with 818 and 588, respectively. Within the KIN class, kinase (551) and TM (264) domains were present in 77 proteins, while in the RLK class, TM (278), kinase (218), and

LRR (159) domains were distributed among 58 proteins. These results indicate that the KIN and RLK classes play a central role in the molecular architecture of the predicted resistance genes in *T. cacao* (**Table S4**).

Functional analysis of the 227 previously uncharacterized resistance genes revealed a diversity of relevant structural domains. The DRAGO tool identified TM, Kinase, NBS, and LECM domains in seven proteins encoded by seven genes (Table 1, S5**)**. InterProScan detected a total of 719 domains across 252 proteins encoded by 196 genes. Of these, 631 domains were distributed among NB-ARC (PF00931), EF-hand (IPR002048), kinase (IPR011009, SSF56112, PF07714, SM00220), coiled-coil (SM00320), and NBS-LRR (SSF52540, IPR003593, SM00382, PF00004). In addition, 88 putative domains and 15 DUFs (domains of unknown function) were identified (Table 1, S6).

HMMER identified 2,917 domains distributed across 265 proteins encoded by 209 genes, including NB-ARC (PF00931), TIR (PF01582, PF13676), kinase (PF00069, PF17667), two PR protein families (PF01535, PF12854, PF13041, PF13812), EF-hand (PF00036, PF13202, PF13405, PF13499, PF13833), and 197 DUFs. These results highlight a diverse set of functional domains potentially involved in pathogen recognition and plant defense signaling (Table 1, S7**).**

Table 1
Distribution of functional domains in genes associated with *T. cacao* resistance

Tool	Total Domains	Proteins/Genes	Domain Family	Number of Domains
			TM	**13**
Drago	27	10/5	Kinase	9
			NBS	3
			LECM	2
			NB-ARC	1
			EF-hand	10
			Kinase	13
InterProScan	719	252/196	Coiled-Coil	13
			NBS-LRR	18
			Putative	88
			DUFs	15
			NB-ARC	5
			TIR	3
Hmmer	2917	265/209	Kinase	20
			PRs	56
			EF-hand	60
			DUFs	197

Functional annotation and genomic distribution of SNPs

After applying the relevant quality and coverage criteria, we identified a total of 11,439,669 SNPs and 1,069,309 INDELs across the 112 WGS libraries of *T. cacao*. The SNPs were filtered based on the genomic coordinates of 4,523 predicted genes with potential resistance-related functions. Functional annotation using VEP revealed 42,394 annotated SNPs distributed across 4,882 genes (Fig. 2a, **Table S8**). The genomic distribution showed a higher concentration of SNPs on chromosomes 2 (8,228), 5 (8,095), 9 (7,381), and 3 (6,885), while chromosome 1 showed the lowest number (440 SNPs) (Fig. 2a, b). Normalization by chromosome length confirmed this pattern, with SNP densities of 20.56, 19.95, 19.11, and 18.93 SNPs/Mb on chromosomes 5, 2, 9, and 3, respectively, within genes predicted to have resistance-related functions, and only 1.18 SNPs/Mb on chromosome 1. In agreement with this

distribution, the highest number of genes was observed on chromosomes 9 (955 genes), 2 (888), 3 (866), and 5 (848), distributed across both coding and non-coding regions (Fig. 2b, **Table S8**).

Of the 42,394 annotated SNPs, 41.5% (17,576 SNPs) were located within coding regions, whereas 58.44% (24,735 SNPs) were found in non-coding regions (Fig. 2c, **Table S9**). Within coding regions, different types of consequences were identified, comprising three possible impact levels and affecting a total of 2,460 genes. Among these consequences were missense variants, with moderate impact, representing 51.28%, corresponding to 9,014 SNPs. Low-impact variants, including synonymous and stop-retained variants, represented 47.89% (8,418 SNPs). High-impact variants, comprising stop-gained, start-lost, stop-lost, and stop-gained variants in splice regions, totaled 156 SNPs, corresponding to 0.37% of all annotated SNPs and 0.86% of the variants located in coding regions (Fig. 2c, d, **Table S9**).

The distribution of these SNPs in the coding regions by chromosome showed that chromosomes 2, 5, 9, and 3 harbored the highest numbers of SNPs, representing 19.22%, 18.76%, 18.12%, and 16.17%, respectively, of the total variants. These chromosomes also exhibited the largest number of missense variants. Chromosomes 2, 5, and 3 showed the highest numbers of high-impact variants, with chromosome 2 containing 23.18% (35 SNPs), chromosome 5 at 21.19% (32 SNPs), and chromosome 3 at 13.91% (21 SNPs) of the total high-impact variants (Fig. 2d).

The analysis of allele frequency for high and moderate impact variants located in coding regions revealed that 80% (125 SNPs) of the high-impact variants, and 74.26% (6,694 SNPs) of the moderate-impact variants exhibited allele frequencies below 20%, and were therefore classified as rare variants, being poorly represented among the analyzed genotypes. Variants with allele frequencies between 20% and 80% were considered intermediate-frequency variants, showing 13.93% (22 SNPs) in high impact variants and 19.26% (1,736 SNPs) of moderate-impact variants. In contrast, variants considered common (allele frequency > 80%) represented 5.72% (9 SNPs) of high-impact SNPs and 6.49% (585 SNPs) of moderate-impact SNPs (**Table S10**).

Patterns of selection in coding regions

The analysis of Tajima's D index in coding regions across the different chromosomes of *T. cacao* identified a total of 3,274 windows of 1,000 bp each. Among these, 328 windows showed evidence of potential selective pressures, with 50% being associated with positive selection and 50% with balancing selection, distributed across all ten chromosomes. The number of SNPs per window ranged from 1 to 69, while Tajima's D values varied from − 2.05 to 3.64. Positive Tajima's D values, especially those above the 95th percentile (D > 1.72), were associated with regions under balancing selection, whereas negative values below the 5th percentile (D < − 1.16) indicated regions under positive selection (Fig. 3a, **Table S11**).

The nucleotide diversity (π) in the windows identified as being under potential selection revealed that 91% had values below 0.001, indicating low genetic variability consistent with possible recent positive selection events, whereas 9% of the windows showed π values between 0.001 and 0.004, suggesting

maintenance of allelic diversity compatible with balancing selection (Fig. 3b, **Table S11**). Enrichment analysis identified significant hotspots only on chromosomes 6 and 7, with odds ratios (OR) of 1.63 ($P \leq 0.03$) and 1.60 ($P \leq 0.025$), respectively, indicating a significant concentration of atypical regions on these chromosomes. Chromosomes 2, 5, and 9 showed the highest absolute numbers of regions under selection. However, these values were not statistically significant ($P > 0.05$), suggesting their distribution follows a random pattern proportional to genome size (**Table S12**).

On chromosome 6, eight genes (Tc06v2_g002950, Tc06v2_g003200, Tc06v2_g006270, Tc06v2_g000370, Tc06v2_g002090, Tc06v2_g004180, Tc06v2_g004120, and Tc06v2_g006240) (**Table 13**), were identified as candidates for balancing selection. All of these genes exhibited intermediate allele frequencies in the studied population and harbored missense variants. Among them, the gene Tc06v2_g002950 (XP_017979048.1) is noteworthy, as it harbors a missense variant predicted to be deleterious and is associated with biological processes such as cellular metabolic process (GO:0044237), cellular lipid metabolic process (GO:0044255), and triterpenoid biosynthetic process (GO:0016104) **(Table S13**).

Furthermore, two of these genes are located within QTLs previously associated with resistance of *T. cacao* to *Phytophthora* spp. Within the QTL BARq1BPPctCHR6, the gene Tc06v2_g000370, which encodes a chitinase (XP_007022927.2) involved in biological processes such as response to stimulus (GO:0050896), response to chemicals (GO:0042221), small molecule metabolic process (GO:0044281), plant organ development (GO:0099402), plant organ morphogenesis, beta-glucan metabolic process (GO:1905392), and response to heat (GO:0051273). The second gene, Tc06v2_g003200, located within the QTL AKAprrICS100CHR62, presents a missense variant in an aquaporin NIP6-1 protein (XP_007023401.2), which is associated with biological processes such as localization (GO:0051179), establishment of localization (GO:0051234), transport (GO:0006810), and transmembrane transport (GO:0055085).

On chromosome 7, two genes (Tc07v2_g007600 and Tc07v2_g007420) were identified within regions under positive selection. On this chromosome, nine genes (Tc07v2_g002310, Tc07v2_g003890, Tc07v2_g003960, Tc07v2_g001630, Tc07v2_g002360, Tc07v2_g008920, Tc07v2_g006340, Tc07v2_g007680, and Tc07v2_g005890) were located in regions under balancing selection, all exhibiting intermediate allele frequencies and missense variants (**Table S13**). Among the candidates under balancing selection, two genes exhibited missense variants predicted to be deleterious. The first was Tc07v2_g001630, which encodes an ABC transporter protein (XP_017980613.1) and is associated with localization (GO:0051179), establishment of localization (GO:0051234), transport (GO:0006810), and transmembrane transport (GO:0055085). The second gene, Tc07v2_g008920, encodes the enzyme uroporphyrinogen decarboxylase (UROD) (XP_017980720.1), involved in the biosynthetic process (GO:0009058), tetrapyrrole metabolic process (GO:0033013), porphyrin-containing compound metabolic process (GO:0006778), and tetrapyrrole biosynthetic process (GO:0033014) (**Table S13, 14**).

Functional enrichment

Functional enrichment analysis of the 2,460 genes located in coding regions, annotated by VEP, revealed a total of 39 significantly enriched KEGG pathways, of which 20 were associated with central metabolic and defense processes. Among these, metabolic pathways (127 genes) and secondary metabolite biosynthesis (66 genes) were prominent. In addition, pathways directly related to pathogen response were identified, including oxidative phosphorylation (46 genes), plant–pathogen interaction (31 genes), MAPK signaling pathway–plant (20 genes), phenylpropanoid biosynthesis (20 genes), and glutathione metabolism (14 genes) (Fig. 3c, **Table S15**).

Analysis of the variants in the genes associated with these five pathways indicated that most variants had moderate impact, with only a single high-impact variant being located in the glutathione metabolism pathway. This variant exhibited a low allele frequency and was classified as rare. Furthermore, the enrichment analysis of biological processes identified 764 significantly enriched terms, with notable representation of response to stimulus (434 genes), localization (344 genes), transport (326 genes), small molecule metabolic process (211 genes), and response to oxidative stress (75 genes), showing FDR values ranging from 1.10E-42 to 3.60E-28 (Fig. 3d, **Table S14**).

Structural distribution of SNPs with high and moderate functional impact

In the structural distribution analysis, 21 genes harboring genetic variants with high and moderate functional impacts as well as intermediate allele frequencies were prioritized. Analysis of the proteins encoded by these genes revealed 11 high-impact start-lost variants, 10 stop-gained variants, and 57 missense variants of moderate impact, distributed across the 21 proteins. These proteins contain domains associated with diverse biological processes, including transport, metabolism, signaling, and macromolecule modification (**Table S16**).

Among these genes, eight stood out because they not only harbored high-impact variants but were also located in atypical genomic regions. On chromosome 5, two genes were found in atypical regions under putative balancing selection, and are involved in enriched biological processes related to localization (GO:0051179), transport (GO:0006810), and transmembrane transport (GO:0055085). One of these genes, Tc05v2_g005990, encodes a protein (XP_017977894.1) containing a Nodulin domain and carries a stop-gained mutation at position 327. Five missense variants predicted to be deleterious are located within the functional domain at residues 101, 108, 131, 142, and 187 (Fig. 4d, **Table S14, S16**). The gene Tc05v2_g004320, which encodes a protein with an ABC transporter domain (XP_017976290.1), with a stop-gained at position 327, and two of the missense variants are predicted to be deleterious at residues 360 and 481 within the functional domain (Fig. 4c).

On chromosome 2, two genes were identified in regions under putative balancing selection. One of them was the Tc02v2_g012540 gene, which encodes a protein containing PKinase and PASS domains (XP_017971484.1), participating in different biological processes, including phosphate-containing compound metabolic process (GO:0006796), protein modification process (GO:0036211), and phosphorylation (GO:0016310) (Fig. 4b). The other gene in this region is Tc02v2_g026000, encoding the

Na^+/H^+ ion exchange protein (XP_017970494.1). Both proteins harbor start-lost variants as well as tolerated missense variants in their sequences (Fig. 4a, **Table S14, S16**).

On the other hand, four genes were found in regions under putative positive selection.

One of these is located on chromosome 2, Tc02v2_g009870, which encodes a protein containing a coiled-coil domain (XP_007042686.2) and has a stop-gained variant at residue 113, as well as a tolerated missense variant (Fig. 4g, **Table S14, S16**). A second gene is located on chromosome 3, Tc03v2_g009380, which is involved in several biological processes, including the macromolecule metabolic process (GO:0043170) and cellular biosynthetic process (GO:0044249). This gene encodes a protein (XP_007037753.2) containing a SWAP domain and presents a stop-gained variant at residue 42, and a tolerated missense variant (Fig. 4f). On chromosome 9, the third gene Tc09v2_g011380 encoded a protein (XP_017983591.1), which is associated with transport and localization processes. This protein contains Longin/Synaptobrevin domains and harbors a start-lost variant as well as a tolerated missense variant (Fig. 4e). Chromosome 10 has the fourth gene located in a region under putative positive selection, Tc10v2_g009700, which encodes a glycosyltransferase (XP_017984892.1), with a stop- gained at residue 483 and a tolerated missense variant (Fig. 4h, **Table S14, S16**).

The remaining 13 proteins are located in genes present in neutral regions, according to Tajima's D analysis (Fig. 4i-l, **Fig. S1**). Among these, four genes stand out, with the Tc02v2_g001040 gene located on chromosome 2, and is associated with enriched biological processes related to localization and transport. This gene encodes a protein (XP_007041139.2) containing ACD11/GLTP domains, which in addition to participating in lipid transport, has also been associated with the regulation of programmed cell death in plants. This protein harbors a start-lost variant and a deleterious missense variant at position 74 within its functional domains (Fig. 4k).

The other three genes are located on chromosome 9. One of them, Tc09v2_g010890, is associated with different enriched biological processes, including response to stimulus (GO:0050896), response to oxidative stress (GO:0006979), catabolic process (GO:0009056), and reactive oxygen species metabolic process (GO:0072593). This gene encodes a protein (XP_017983855.1) containing a peroxidase domain, presenting a start-lost variant in the sequence (Fig. 4i). The second gene, Tc09v2_g030370, is associated with enriched biological processes related to response to stimulus (GO:0050896), signal transduction (GO:0007165), cell communication (GO:0007154), and signaling (GO:0023052). It encodes a protein (XP_007017047.2) containing Pkinase, ATP-binding, and NAF domains, and presents a stop- gained variant at position 302, as well as a deleterious missense variant at position 120 within the functional Pkinase domain. These variants may negatively affect the cellular signaling cascades regulated by this kinase (Fig. 4j). The third gene, Tc09v2_g027150, encodes the protein (XP_007016442.2), with HSP20/α-crystallin domains, involved in stress response and protein homeostasis. This protein has a stop-gained variant at position 112 and one of the missense and deleterious variants at position 90, located within its functional domains (Fig. 4l, **Table S14, S16**).

Analysis of QTLs associated with disease resistance in T. cacao

Of the total missense variants identified in coding regions, 794 SNPs were detected in 170 genes located within the coordinates of nine QTLs associated with resistance to *Phytophthora* spp. and two QTLs related to resistance to *M. perniciosa* in *T. cacao*. Among these genes, 84 presented 161 missense variants with intermediate allele frequencies, distributed across nine QTLs, eight associated with resistance to *Phytophthora* spp. and one associated with resistance to *M. perniciosa* (**Table S17**).

Of particular note is the Tc02v2_g008430 gene, located on chromosome 2, which is colocalized with qSINT2 QTL, previously associated with resistance to *M. perniciosa*. This gene encodes the XP_007042477.2 protein, which has a functional domain characteristic of E3 ubiquitin-protein ligase (PF23113), suggesting a potential role in the ubiquitination-mediated immune response (**Table S17**). Among the QTLs associated with *T. cacao* resistance to *Phytophthora* spp., five candidate genes stand out, containing variants with high or moderate impact distributed across four QTLs. On chromosome 2, the gene Tc02v2_g001040 carries a high-impact variant (start-lost) located within the AKAprrICS95CHR2 QTL. This gene encodes the protein XP_007041139.2, which is associated with the regulation of plant cell death and shows enrichment for biological processes related to localization and transport (Fig. 4k, **Table S14, S17**).

On chromosome 6, three candidate genes were identified. Tc06v2_g000370, harboring moderate-impact variants within the BARq1BPPctCHR6 QTL, encodes a class I chitinase (PF00182) and is involved in enriched biological processes such as response to stimulus and a small molecule metabolic process (**Table S14, S17**). Also, on chromosome 6, two additional genes have SNPs located within the AKAprrICS100CHR62 QTL. The gene Tc06v2_g003440, which encodes the protein XP_007023452.2, contains a catalytic NADH-OxR domain and presents a stop-gained variant within the functional domain, suggesting potential loss of function associated with redox metabolism and reactive oxygen species (ROS) production during the immune response. In contrast, the gene Tc06v2_g003200 carries missense variants in the aquaporin NIP6-1 protein (XP_007023401.2), which is involved in cellular transport and localization processes (**Table S14, S16, S17**).

On chromosome 8, the gene Tc08v2_g004590 harbors a start-lost variant located within the BROphy2CHR8 QTL. This gene encodes a protease **(**XP_007017945.2) containing Pept_S9A-N and AB-hydrolase domains, and is associated with enriched biological processes such as protein metabolic process (GO:0019538), macromolecule metabolic process (GO:0043170), and proteolysis (GO:0006508) (**Fig. S1, Table S14, S16, S17**).

Discussion

This study combined the prediction of genes associated with resistance through genetic variability analysis across 112 WGS libraries, aiming to support strategies for the genetic improvement of *T. cacao* resistance to diseases. The identification of 4,523 candidate resistance-associated genes revealed a diverse functional repertoire, with a predominance of families involved in defense recognition and signaling, including kinases, RLK/RLP receptors, and NBS-LRR proteins (**Table S3**). These classes

represent central components of plant immunity and have been widely associated with pathogen resistance and tolerance to abiotic stresses (Kim et al., 2024).

The presence of proteins related to the classical defense response, such as thaumatin-like, defensin-like, and osmotin-like proteins, supports the functional potential of the candidate genes, since these proteins play important roles in antifungal responses, the regulation of reactive oxygen species, and adaptation to adverse conditions (De Oliveira, I.B. *et al*., 2024; Singh et al., 2024) (**Table S3**). We identified genes without previous functional annotation in the *T. cacao* genome, some of which contain typical defense-related domains. Among them is Tc05v2_g022100, which encodes two proteins with PR1/PR3 domains located on chromosome 5. This result is consistent with previous analyses (Lanaud et al., 2009), which reported PR2 proteins colocalized within a QTL associated with resistance to *Phytophthora* spp. on the same chromosome, suggesting that this genomic region harbors multiple components of the defense response.

Furthermore, the presence of genes containing NBS-LRR domains on chromosomes 1, 2, 4, 7, 9, and 10, along with kinase domains on chromosomes 2, 4, 9, and 10 (Table 1, S5, S6, S7), is consistent with previous functional studies of *T. cacao* resistance to *P. palmivora* (Mucherino-Muñoz *et al*.,2021), where these chromosomes were also identified as carrying genes containing NBS, LRR, and RLK-type kinase domains, corroborating their role in pathogen recognition and activation of defense responses. On the other hand, the detection of DUF domains distributed across multiple chromosomes suggests that the *T. cacao* genome has genetic components that are still poorly characterized. These domains, often associated with unknown functions, may represent previously non-described defense elements, indicating the need for further functional studies.

SNP annotation using VEP revealed that 41.5% of the variants are located in coding regions. Among these, 51.28% correspond to missense mutations, which are classified as having a moderate impact because they potentially alter the amino acid composition of proteins (Fig. 2d). This type of substitution may affect protein stability or activity, resulting in partial or complete loss of function when deleterious, although some of these variants may be tolerated without significantly compromising protein functionality. We also identified 0.86% high-impact variants (156 SNPs), distributed across 129 genes (Fig. 2d). These mutations included loss of start codons (start-lost), premature introduction of stop codons (stop-gained), and loss of stop codons (stop-lost), alterations that generally result in truncated, extended, or structurally unstable proteins that are frequently nonfunctional and may irreversibly compromise the function of the affected genes. The low frequency of high-impact variants observed in our dataset is consistent with studies of other plant species, such as *Prunus*, in which high-impact mutations are rare, while missense variants with moderate impact constitute the majority of annotated SNPs (Guajardo et al., 2020).

Functional enrichment confirmed that a large proportion of the genes under genetic variation encode proteins involved in essential defense processes, including plant–pathogen interaction, MAPK signaling, glutathione metabolism, and phenylpropanoid biosynthesis (Fig. 3c). These results are consistent with

transcriptomic analyses of *T. cacao* leaves infected by *Phytophthora palmivora*, the causal agent of black pod disease, in which pathways such as MAPK signaling, hormone signal transduction, and defense-related metabolite biosynthesis (phenylpropanoids, α-linolenic acid, ethylene, and terpenoids) have been found to be strongly induced following infection (Baruah I.K. et al., 2022). Consistently, functional enrichment of the previously predicted candidate genes identified biological processes related to metabolism, signaling, and transport (Fig. 3d).

Within the localization and transport categories, six genes showed high-impact mutations: Tc02v2_g026000, Tc05v2_g004320, Tc09v2_g011380, Tc05v2_g005990, Tc06v2_g003200, and Tc02v2_g001040. The last, Tc02v2_g001040, encodes an ACD11-like protein, which is involved in lipid transport and the regulation of programmed cell death. This protein exhibits a start-loss mutation at residue 1 and a deleterious missense variant within its functional domain (Fig. 4k), suggesting potential impairment of its activity, and consequently a possible impact on cell death–mediated defense responses.

In *A. thaliana*, proteins containing the ACD11 domain participate in important functions in plant defense against pathogens. Deleterious mutations affecting the GLTP domain within this protein can inhibit its activity, destabilizing its structure and compromising its catalytic function (Rifat M.H. *et al*., 2022). Among the genes related to transport, Tc07v2_g001630 encodes a protein containing an ABC-Transp domain (Fig. 4c). Proteins with these functional domains have been studied for their participation in detoxification processes, responses to abiotic stresses, and pathogen resistance (Dahuja A. et al., 2021). The gene Tc05v2_g005990 encodes a protein containing a nodulin transporter domain (Fig. 4d) and represents another candidate gene associated with transport processes. This type of domain has previously been linked to plant defense responses, including resistance of banana against *Xanthomonas campestris pv. musacearum* (Ntui, V.O. *et al*., 2024).

In addition, genes associated with response to stimulus, small molecule metabolic processes, and response to oxidative stress also harbored high-impact mutations. Among them, the genes Tc02v2_g012540 and Tc09v2_g030370 encode proteins with kinase domains, affected by stop-gained and start-lost mutations (Fig. 4b, j). In *A. thaliana*, proteins containing kinase domains are essential for the activation of plant defense pathways, functioning in the transduction of extracellular signals and in the induction of PTI responses (Yan H. et al., 2024). We also identified a glycosyltransferase encoded by Tc10v2_g009700, which has not yet been functionally characterized in *T. cacao*, with a UGT/GlycPhos domain and harboring high and moderate impact mutations (Fig. 4h). Glycosyltransferase proteins contribute to the detoxification of mycotoxins secreted by pathogens. In grasses, these genes have been associated with the detoxification of toxins produced by *Fusarium* spp. (Bethke G. et al., 2023; Della Gala V. et al., 2025).

On the other hand, in the response to oxidative stress, the gene Tc09v2_g010890 encodes a peroxidase protein harboring a start-lost mutation (Fig. 4i**).** Peroxidases have been extensively studied in various plant species due to their important role in maintaining reactive oxygen species (ROS) homeostasis and

in mediating plant immune responses (Lu Y. et al., 2025). In poplar, for example, the gene PdePRX12 modulates resistance by regulating hydrogen peroxide (H_2O_2) levels after infection by *Botryosphaeria dothidea* and *Alternaria alternata* (Cai G. et al., 2023). The SNPs identified in these genes may alter the efficiency of defense signaling and transport of diverse molecules, including defense proteins, phytohormones and metabolites, thereby modifying or completely eliminating their functions and potentially rendering *T. cacao* plants more susceptible to pathogen attack.

Some of these genes were also located within QTLs previously associated with resistance of *T. cacao* to *Phytophthora* spp. On chromosome 2, within the QTL AKAprrICS95CHR2, Tc02v2_g001040 was identified. This is involved in transport processes and in the regulation of plant programmed cell death. Three other genes were identified within QTLs located on chromosome 6. Within the QTL AKAprrICS100CHR62, the genes Tc06v2_g003200 and Tc06v2_g003440 were associated with transport processes and response to oxidative stress, encoding proteins containing aquaporin NIP6-1 and NADH oxidoreductase domains, respectively. These two proteins participate in the interaction of plants against different pathogens, with NADH enzymes contributing to ROS activation, while aquaporins act as regulators and participate in the activation of kinase-mediated signaling pathways and PR-domain–containing defense proteins, ultimately triggering PTI in plants (Li G. et al., 2020, Yuan M. et al., 2021, Dimitris P.J. *et al*., 2025).

In the QTL BARq1BPPctCHR6, we identified the gene Tc06v2_g000370, which encodes a class I chitinase (PF00182). The presence of chitinases in plants is of particular importance, as these enzymes play a key role in defense against fungal pathogens. In the case of barley, a purification study of a chitinase protein was shown to significantly inhibit the growth of several phytopathogenic fungi, including *Alternaria solani*, *Fusarium* spp., *Rhizoctonia solani*, and *Verticillium dahliae* (Toufiq N. et al., 2018; Vaghela B. et al., 2022). The gene Tc08v2_g004590, located on chromosome 8, was identified within the QTL BROphy2CHR8. This gene encodes a Peptidase S9A family protease, which is involved in multiple signaling processes, particularly in plant–pathogen interactions, contributing to the modulation of plant defense responses through the induction of HR (Fernández-Fernández Á.D. et al., 2023). The high variability of high and moderate impact variants detected in these regions is consistent with previous studies in which these chromosomes were identified as harboring QTLs associated with disease resistance, suggesting that these genomic regions represent promising targets for the characterization of resistance-associated alleles in *T. cacao* (Mucherino-Muñoz J.J. *et al*., 2021; Lanaud C. et al., 2009).

Our nucleotide variation and selection analysis performed on genes previously predicted to be located in coding regions revealed a hypothetical distribution of atypical regions across the ten chromosomes. Chromosome 6 presented eight genes under putative balancing selection, associated with diverse biological processes, including metabolic processes, transport, and response to stimuli, exhibiting missense variants. Among these genes, Tc06v2_g000370 and Tc06v2_g003200 stand out, since they not only were found to participate in biological processes associated with plant resistance, but were also identified within QTLs BARq1BPPctCHR6 and AKAprrICS100CHR62, associated with *T. cacao* resistance to *Phytophthora* (Mucherino Muñoz, J.J. *et al*., 2021). On chromosome 7, two genes were found in

regions under putative positive selection, and nine genes were located in regions under balancing selection. Among these, Tc07v2_g001630 and Tc07v2_g008920, both under balancing selection, are involved in enriched transport and metabolism processes.

Chromosomes 2, 3, 5, and 9 not only harbored the highest numbers of high-impact SNPs, but also contained variants located in hypothetically atypical regions under putative positive and balancing selection. On chromosomes 2 and 5 (Fig. 4a-d), four genes under putative balancing selection were identified, participating in diverse biological processes related to transport, localization, and phosphorylation. Within these genes, a high number of missense variants were observed, consistent with the expected pattern of balancing selection, which maintains old alleles over extended evolutionary periods and contributes to the preservation of genetic diversity, since distinct haplotypes may confer adaptive characteristics under multiple pathogen pressures or environmental conditions (Sutherland C.A. et al., 2024). The presence of putative balancing selection on chromosomes 2 and 5 is particularly relevant, as both chromosomes have previously been associated with quantitative resistance to witches' broom and frosty pod diseases (Brown *et al*., 2005; Lanaud et al., 2009).

On the other hand, genes under putative positive selection, such as the four candidate genes identified on chromosomes 2, 3, 9, and 10, exhibited a reduced number of missense variants (Fig. 4e-h). This pattern suggests the elimination of alternative haplotypes under positive selective pressure, leading to reduced genetic diversity, likely reflecting recent adaptive responses to selective pressures imposed by emerging pathogen populations in recent decades. It is important to emphasize that the signals of putative selection identified in this study are exploratory and restricted to previously predicted resistance-associated genes, so further validation studies are necessary to confirm these results.

Conclusions

Our results revealed a repertoire of genes potentially associated with resistance in *T. cacao*, characterized by classical defense-related functional domains. The integration between functional prediction, SNP impact, and co-localization with previously described QTLs reinforces the biological relevance of the identified candidate genes, including genes linked to MAPK signaling, transport, oxidative metabolism, and pathogen recognition. The presence of high- and moderate-impact variants in key genes and the identification of regions under putative positive and balancing selection suggest that heterogeneous evolutionary pressures shape the immune response of this species. These findings provide a robust foundation for prioritizing alleles of interest in breeding programs and for functional studies aimed at developing disease-resistant cacao cultivars.

Perspectives

These results open the possibility of future investigations focused on the functional validation of the identified genes, through expression analyses, genetic assays, and phenotyping associated with resistance. Furthermore, approaches such as marker-assisted selection and gene editing are promising

avenues to accelerate the development of resistant cultivars against the main pathogens affecting *T. cacao*.

Declarations

Author Contributions

The original draft was written by Gabriela Nicolle Ramos-Lizardo. The methodology was designed and conducted by Gabriela Nicolle Ramos-Lizardo, Konstantinos Alexiou, and María José Aranzana. Formal data analysis was performed by Gabriela Nicolle Ramos-Lizardo. Conceptualization, review, and editing were carried out by Ronan Xavier Corrêa and Jonathan Javier Mucherino-Muñoz. The research was conducted partly in María José Aranzana's laboratory in Barcelona and partly in Ronan Xavier Corrêa's laboratory in Brazil. All authors read and approved the final version of the manuscript.

Acknowledgments

We thank the Center for Biotechnology and Genetics (CBG, UESC) and the Center for Research in Agricultural Genomics (CRAG, Spain) for providing laboratory infrastructure, and the Postgraduate Program in Genetics and Molecular Biology (PPGGBM, UESC) for the opportunity to carry out this study. The work was funded by the Brazilian National Council for Scientific and Technological Development (CNPq) under grant no. 315913/2023-1. This study was also partially funded by the Office to Coordinate Improvement of Higher Education Personnel (CAPES), Brazil, including sandwich doctorate fellowships and a scholarship awarded to G.N.R.-L. (Financial Code 001).

Funding

This study was supported by the National Council for Scientific and Technological Development (CNPq), grant number 315913/2023-1. It was also partially financed by the Office to Coordinate Improvement of Higher Education Personnel CAPES), Finance Code 001.

Conflicts of Interest

The authors declare no conflicts of interest.

Availability of data

All data supporting the findings of this study are available within the paper and its Supplementary Information

References

1. **Adeniran, A. O., Onanaye, A. S., & Adeleke, O. J. (2024).** Optimal control of cocoa black pod disease: A multi-pronged approach. *Franklin Open*, **7**, 100100.

https://doi.org/10.1016/j.fraope.2024.100100

2. **Andrews, S. (2010).** FastQC: A quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
3. **Argout, X., *et al*. (2023).** Pangenomic exploration of *Theobroma cacao*: New insights into gene content diversity and selection during domestication. *bioRxiv*. https://doi.org/10.1101/2023.11.03.565324
4. **Baruah, I. K., Ali, S. S., Shao, J., Lary, D., & Bailey, B. A. (2022).** Changes in gene expression in leaves of cacao genotypes resistant and susceptible to *Phytophthora palmivora* infection. *Frontiers in Plant Science*, **12**, 780805. https://doi.org/10.3389/fpls.2021.780805
5. **Bethke, G., Huang, Y., Hensel, G., *et al*. (2023).** UDP-glucosyltransferase HvUGT13248 confers type II resistance to *Fusarium graminearum* in barley. *Plant Physiology*, **193**(4), 2691–2710. https://doi.org/10.1093/plphys/kiad467
6. **Blum, M., *et al*. (2025).**InterPro: The protein sequence classification resource in 2025. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkae1082
7. **Brown, J. S., Phillips-Mora, W., Power, E. J., Krol, C., Cervantes-Martínez, C., Motamayor, J. C., & Schnell, R. J. (2007).** Mapping QTLs for resistance to frosty pod and black pod diseases and horticultural traits in *Theobroma cacao* L. *Crop Science*, **47**, 1851–1858. https://doi.org/10.2135/cropsci2006.11.0753
8. **Cai, G., Zhang, Y., Huang, L., & Wang, N. (2023).**Uncovering the role of PdePrx12 peroxidase in enhancing disease resistance in poplar trees. *Journal of Fungi*, **9**(4), 410. https://doi.org/10.3390/jof9040410
9. **Carrero-Gutiérrez, M. L., González-Sayer, S., Jaimes-Suárez, Y., *et al*. (2025).** *Ceratocystis cacaofunesta* is responsible for cocoa crop wilt in Colombia: Morphological and molecular characterization of isolates. *Journal of Plant Pathology*, **107**, 795–810. https://doi.org/10.1007/s42161-024-01798-7
10. **Dahuja, A., Kumar, R. R., Sakhare, A., Watts, A., Singh, B., Goswami, S., Sachdev, A., & Praveen, S. (2021).**Role of ATP-binding cassette transporters in maintaining plant homeostasis under abiotic and biotic stresses. *Physiologia Plantarum*, **171**(4), 785–801. https://doi.org/10.1111/ppl.13302
11. **Danecek, P., *et al*. (2011).**The variant call format and VCFtools. *Bioinformatics*, **27**(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330
12. **De Oliveira, I. B., *et al*. (2024).**Apoplastomes of contrasting cacao genotypes to witches' broom disease reveal differential accumulation of PR proteins. *Frontiers in Plant Science*, **15**, 1387153. https://doi.org/10.3389/fpls.2024.1387153
13. **Della Gala, V., Dato, L., Wiesenberger, G., Jæger, D., Adam, G., Hansen, J., & Welner, D. H. (2025).** Plant-derived UDP-glycosyltransferases for glycosylation-mediated detoxification of deoxynivalenol: Enzyme discovery, characterization, and in vivo resistance assessment. *Toxins*, **17**(4), 153. https://doi.org/10.3390/toxins17040153

14. Dimitris P. J. *et al*. (2025). Comprehensive mechanism of biological and health effects of anthropogenic extremely low frequency and wireless communication electromagnetic fields. *Frontiers in Public Health*.13. https://doi.org/10.3389/fpubh.2025.1585441 ISSN=2296-2565
15. **Fernandes D. S., L., Correa, F. M., Ingram, K. T., de Almeida, A.-A. F., & Royaert, S. (2020).** QTL mapping and identification of SNP haplotypes affecting yield components of *Theobroma cacao* L. *Horticulture Research*, **7**, 26. https://doi.org/10.1038/s41438-020-0250-3
16. **Fernández-Fernández, Á. D., Stael, S., & Van Breusegem, F. (2023).** Mechanisms controlling plant proteases and their substrates. *Cell Death & Differentiation*, **30**(4), 1047–1058. https://doi.org/10.1038/s41418-023-01120-5
17. **García, J. C., *et al*. (2022).**PRGdb 4.0: An updated database dedicated to genes involved in plant disease resistance processes. *Nucleic Acids Research*, **50**(D1), D1483–D1490. https://doi.org/10.1093/nar/gkab1087
18. **Ge, S. X., Jung, D., & Yao, R. (2020).** ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, **36**(8), 2628–2629. https://doi.org/10.1093/bioinformatics/btz931
19. **Goodstein, D. M., *et al*. (2012).** Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Research*, **40**(D1), D1178–D1186.https://doi.org/10.1093/nar/gkr944
20. **Guajardo, V., Solís, S., Almada, R., *et al*. (2020).** Genome-wide SNP identification in *Prunus* rootstock germplasm collections using genotyping-by-sequencing: Phylogenetic analysis, distribution of SNPs, and prediction of their effect on gene function. *Scientific Reports*, **10**, 1467. https://doi.org/10.1038/s41598-020-58271-5
21. **Hämälä, T., Wafula, E. K., Guiltinan, M. J., Ralph, P. E., dePamphilis, C. W., & Tiffin, P. (2021).** Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proceedings of the National Academy of Sciences of the United States of America*, **118**(35), e2102914118. https://doi.org/10.1073/pnas.2102914118
22. **Instituto Brasileiro de Geografia e Estatística (IBGE). (2024).** Cacau: Produção agropecuária. https://www.ibge.gov.br/explica/producao-agropecuaria/cacau/br
23. **International Cocoa Organization (ICCO). (2024).** Cocoa market and economy. https://www.icco.org/economy/#market
24. **Jibat, M., & Alo, S. (2023).** Integrated management of black pod (*Phytophthora palmivora*) disease of cocoa through fungicides and cultural practices in southwestern Ethiopia. *International Journal on Food, Agriculture and Natural Resources*, **4**(3), 43–45. https://doi.org/10.46676/ij-fanres.v4i3.150
25. **Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021).** KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Research*, **49**(D1), D545–D551. https://doi.org/10.1093/nar/gkaa970
26. **Kim, H.-J., Jang, J. W., Pham, T., *et al*. (2024).** OsLRR-RLP2 gene regulates immunity to *Magnaporthe oryzae* in japonica rice. *International Journal of Molecular Sciences*, **25**, 2216. https://doi.org/10.3390/ijms25042216

27. **Kongor, J. E., Owusu, M., & Oduro-Yeboah, C. (2024).** Cocoa production in the 2020s: Challenges and solutions. *CABI Agriculture and Bioscience*, **5**, 102. https://doi.org/10.1186/s43170-024-00310-6
28. **Krueger, F. (2019).** TrimGalore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
29. **Kulesza, E., Thomas, P., Prewitt, S. F., *et al*. (2024).** The cacao gene atlas: A transcriptome developmental atlas reveals highly tissue-specific and dynamically regulated gene networks in *Theobroma cacao* L. *BMC Plant Biology*, **24**, 601. https://doi.org/10.1186/s12870-024-05171-9
30. **Lahive, F., Hadley, P., & Daymond, A. J. (2019).** The physiological responses of cacao to the environment and the implications for climate change resilience: A review. *Agronomy for Sustainable Development*, **39**, 5. https://doi.org/10.1007/s13593-018-0552-0
31. **Li, Q., Zhu, H., Ai, G., *et al*. (2024).** Plant genes related to *Phytophthora* pathogen resistance. *Phytopathology Research*, **6**, 15. https://doi.org/10.1186/s42483-024-00229-w
32. **Li, G., Chen, T., Zhang, Z., Li, B., & Tian, S. (2020).** Roles of aquaporins in plant–pathogen interaction. *Plants*, **9**(9), 1134. https://doi.org/10.3390/plants9091134
33. **Li, H., & Durbin, R. (2009).** Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324
34. **Li, H., Handsaker, B., Wysoker, A., *et al*. (2009).** The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352
35. **Lanaud, C., Fouet, O., Clément, D., *et al*. (2009).** A meta-QTL analysis of disease resistance traits of *Theobroma cacao* L. *Molecular Breeding*, **24**, 361–374. https://doi.org/10.1007/s11032-009-9297-4
36. **Lu, Y., Ma, R., Wu, K., *et al*. (2025).** Genome-wide analysis of the class III peroxidase gene family in apple (*Malus domestica*). *PeerJ*, **13**, e19741.https://doi.org/10.7717/peerj.19741
37. Marelli J.P., *et al*. (2019). Chocolate under threat from old and new cacao diseases. *Phytopathology* **109**:1331–1343. https://doi.org/10.1094/PHYTO-12-18-0477-RVW
38. McKenna A *et al*. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**:1297–1303. https://doi.org/10.1101/gr.107524.110
39. McLaren W *et al*. (2016). The Ensembl Variant Effect Predictor. *Genome Biology* **17**:122. https://doi.org/10.1186/s13059-016-0974-4
40. Mucherino-Muñoz J.J., *et al*. (2021). Structural and functional genomics of the resistance of cacao to *Phytophthora palmivora*. *Pathogens* **10**:961. https://doi.org/10.3390/pathogens10080961
41. Nei M. & Li W. H., (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America* **76**:5269–5273. http://www.jstor.org/stable/70430.
42. Nieves-Orduña H. E., *et al*. (2024). Genotyping of cacao (*Theobroma cacao* L.) germplasm resources with SNP markers linked to agronomic traits reveals signs of selection. *Tree Genetics & Genomes* **20**:13. https://doi.org/10.1007/s11295-024-01646-w

43. Ntui V. O., *et al*. (2024). Targeted knockout of early nodulin-like 3 (*MusaENODL3*) reveals its function in resistance to Xanthomonas wilt disease. *Plant Biotechnology Journal* **22**:1101–1112. https://doi.org/10.1111/pbi.14248

44. Osorio-Guarín J., *et al*. (2020). Genome-wide association study reveals novel candidate genes associated with productivity and disease resistance to *Moniliophthora* spp. in cacao (*Theobroma cacao* L.). *G3: Genes, Genomes, Genetics* **10**:1713–1725. https://doi.org/10.1534/g3.120.401153

45. Pham M, Tu Y, Lv X (2023). Accelerating BWA-MEM read mapping on GPUs. In: *Proceedings of the 37th ACM International Conference on Supercomputing (ICS '23)*. Association for Computing Machinery, New York, pp 155–166. https://doi.org/10.1145/3577193.3593703

46. Pires JL, Pires Melo GR, Yamada MM, Martins ED, Luz N (2025). Resistance of cocoa to *Moniliophthora perniciosa*: progenitors and progenies selection. *Revista Ceres* **69**:544–552. https://ojs.ceres.ufv.br/ceres/article/view/8039.

47. **R Core Team.** (2025). *R: A language and environment for statistical computing* (Version 4.4.3). R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

48. Rifat MH., *et al*. (2022) Prediction and expression analysis of deleterious nonsynonymous SNPs of *Arabidopsis ACD11* gene by combining computational algorithms and molecular docking. *PLOS Computational Biology* **18**:e1009539. https://doi.org/10.1371/journal.pcbi.1009539

49. Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP (2017). Variant review with the Integrative Genomics Viewer. *Cancer Research* **77**:e31–e34. https://doi.org/10.1158/0008-5472.CAN-17-0337

50. Singh V, Hallan V, Pati PK (2024). *Withania somnifera* osmotin (WsOsm) confers stress tolerance in tobacco and establishes novel interactions with the defensin protein (WsDF). *Physiologia Plantarum* **176**:e14513. https://doi.org/10.1111/ppl.14513

51. Sudarma, I. M., Adnyana, I. W., & Darmiati, N. (2023). *Trichoderma* sp. as endophytic fungi as a candidate for cocoa pod rot disease control. *GPH-International Journal of Agriculture and Research* **6**:88–96. https://doi.org/10.5281/zenodo.10441163

52. Sutherland CA, Prigozhin DM, Monroe JG, Krasileva KV. (2024). High allelic diversity in *Arabidopsis* NLRs is associated with distinct genomic features. *EMBO Reports* **25**:2306–2322. https://doi.org/10.1038/s44319-024-00122-9

53. Tajima F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**:437–460. https://doi.org/10.1093/genetics/105.2.437

54. Toufiq N, Tabassum B, Bhatti MU, Khan A, Tariq M, Shahid N, Nasir IA, Husnain T. (2018). Improved antifungal activity of barley-derived chitinase I overexpressing a 32 kDa recombinant chitinase in *Escherichia coli*. *Brazilian Journal of Microbiology* **49**:414–421. https://doi.org/10.1016/j.bjm.2017.05.007

55. Vaghela B, Vashi R, Rajput K, Joshi R. (2022). Plant chitinases and their role in plant defense: a comprehensive review. *Enzyme and Microbial Technology* **159**:110055. https://doi.org/10.1016/j.enzmictec.2022.110055

56. Wang D, Wei L, Liu T, Ma J, Huang K, Guo H, Huang Y, Zhang L, Zhao J, Tsuda K, Wang Y. (2023). Suppression of ETI by PTI priming to balance plant growth and defense through an MPK3/MPK6–WRKYs–PP2Cs module. *Molecular Plant* **16**:903–918. https://doi.org/10.1016/j.molp.2023.04.004
57. Wijaya G, Wardani AK, Eris D. (2023). Biocontrol activity of endophytic bacteria from cocoa against *Phytophthora* sp. and *Colletotrichum* sp. *Menara Perkebunan* **91**:72–86. https://doi.org/10.22302/iribb.jur.mp.v91i1.522
58. **Wickham, H.** (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org
59. Wong, C. J.A., Clement, D.P.L., Mournet, P. *et al.*(2022). A high-density genetic map from a cacao F2 progeny and QTL detection for resistance to witches' broom disease. *Tree Genetics & Genomes* 18, 31. https://doi.org/10.1007/s11295-022-01562-x.
60. Yan H, Yuan Y, Yang R, Gou X, Dai S, Zhou J, Guo J, Shen J, Lu Y, Liu Y, Cai Y. (2024). A large-scale screening identifies receptor-like kinases with common features in kinase domains potentially related to disease resistance in planta. *Frontiers in Plant Science* **15**. https://doi.org/10.3389/fpls.2024.1503773
61. Yuan M, Ngou BPM, Ding P, Xin X-F. (2021). PTI–ETI crosstalk: an integrative view of plant immunity. *Current Opinion in Plant Biology* **62**:102030. https://doi.org/10.1016/j.pbi.2021.102030
62. Yuan M, Jiang Z, Bi G *et al*. (2021). Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature* **592**:105–109. https://doi.org/10.1038/s41586-021-03316-6

Figures

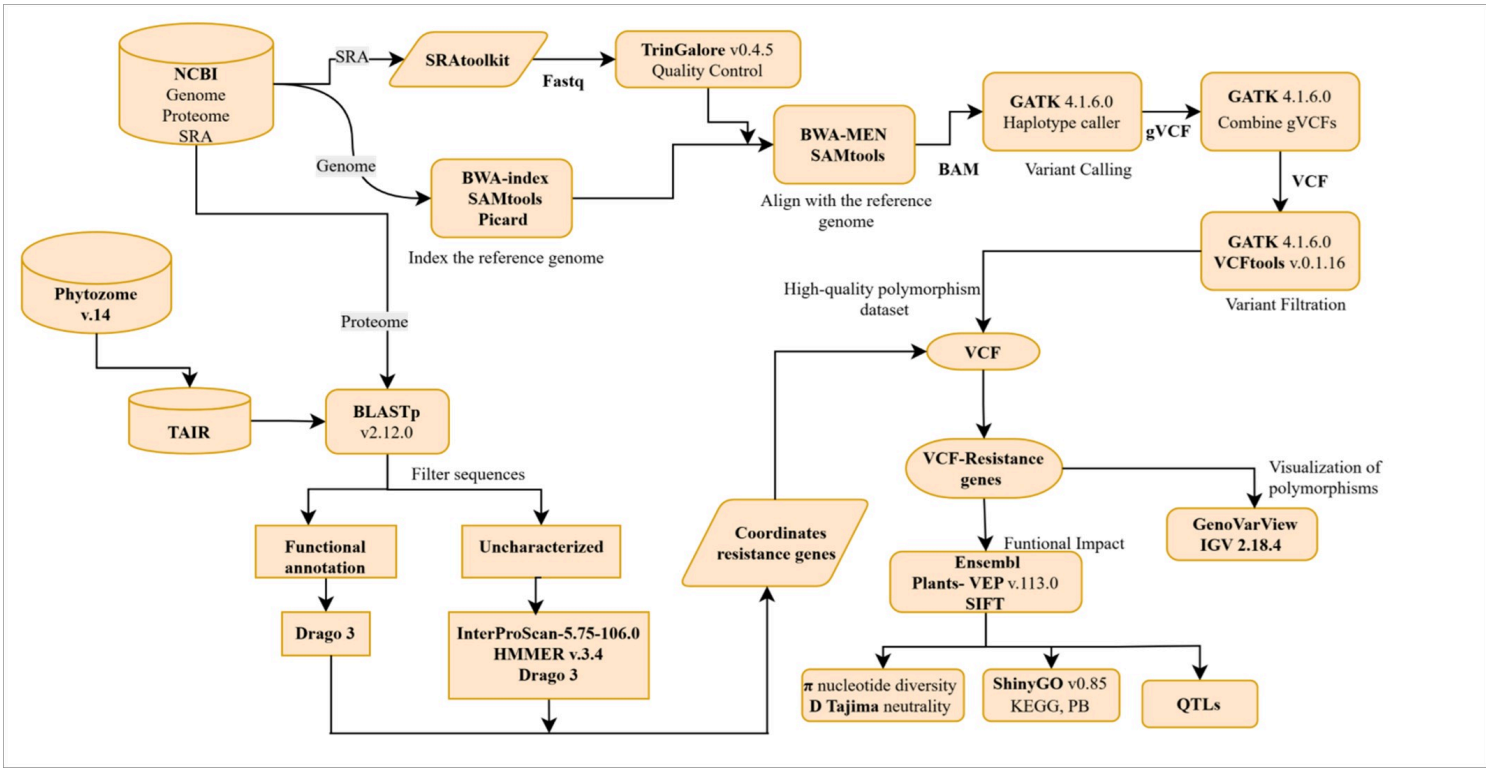

Figure 1

The workflow shows the different tools used for variant analysis, prediction, and functional annotation of genes associated with resistance in T. cacao.

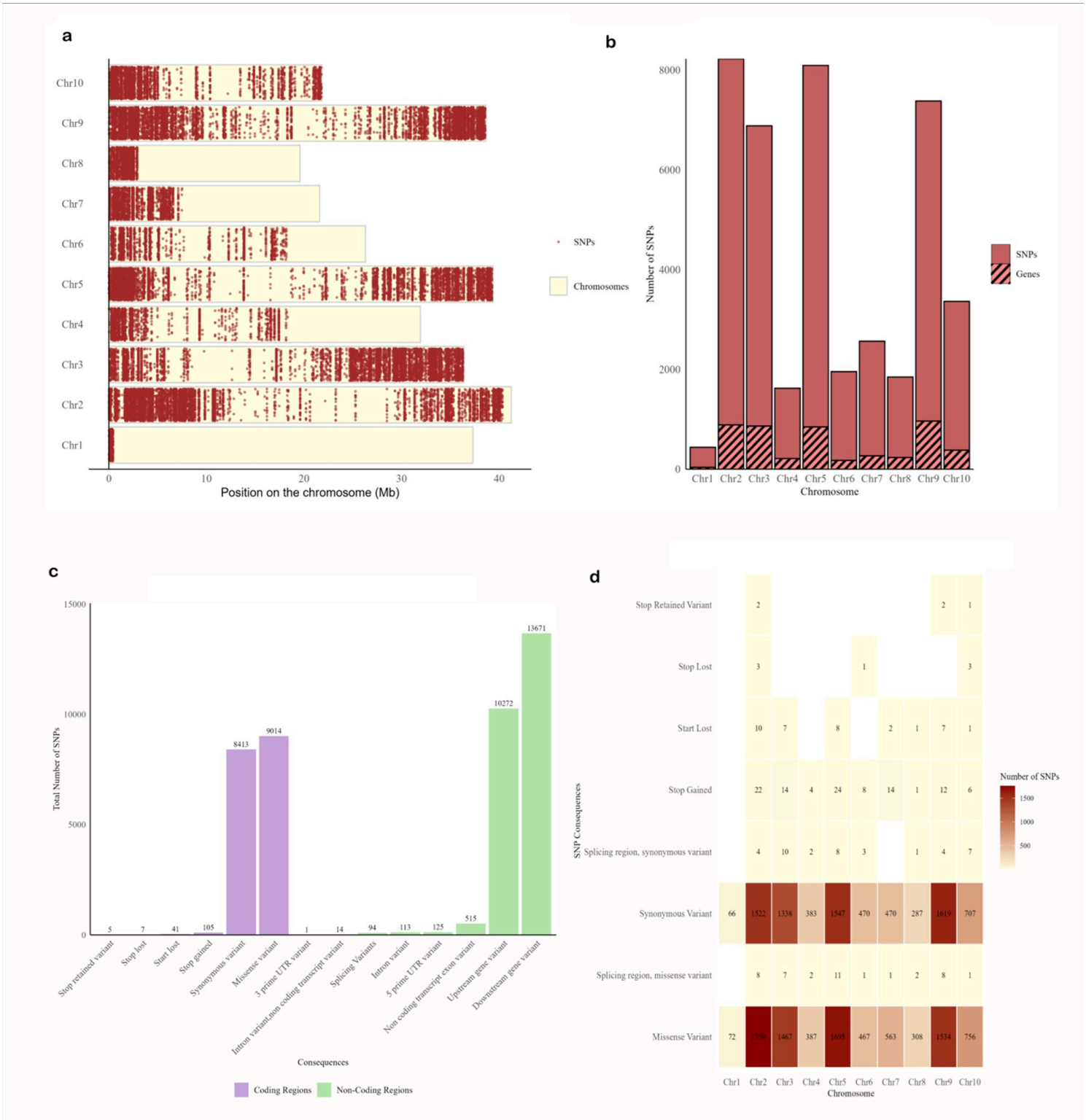

Figure 2

Genomic distribution of resistance-associated SNPs in *T. cacao* and classification of variants by predicted consequence. (**a**) Distribution of SNPs across the ten chromosomes; red dots indicate the physical positions of SNPs, and yellow bars represent chromosome length. (**b**) Number of SNPs and annotated genes per chromosome. Red bars represent the total number of SNPs, and superimposedlines indicate the number of annotated genes. (**c**) Distribution of genomic variants annotated by Variant Effect Predictor (VEP) according to functional consequence. Purple bars represent variants located in coding regions, including missense, synonymous, stop gained, start lost, and stop lost variants, while green bars correspond to variants located in non-coding regions, such as UTRs, introns, and intergenic regions. (**d**) Heatmap showing the distribution of variant consequences within coding regions across chromosomes. Color intensity represents the number of SNPs per chromosome, with darker brown indicating higher values.

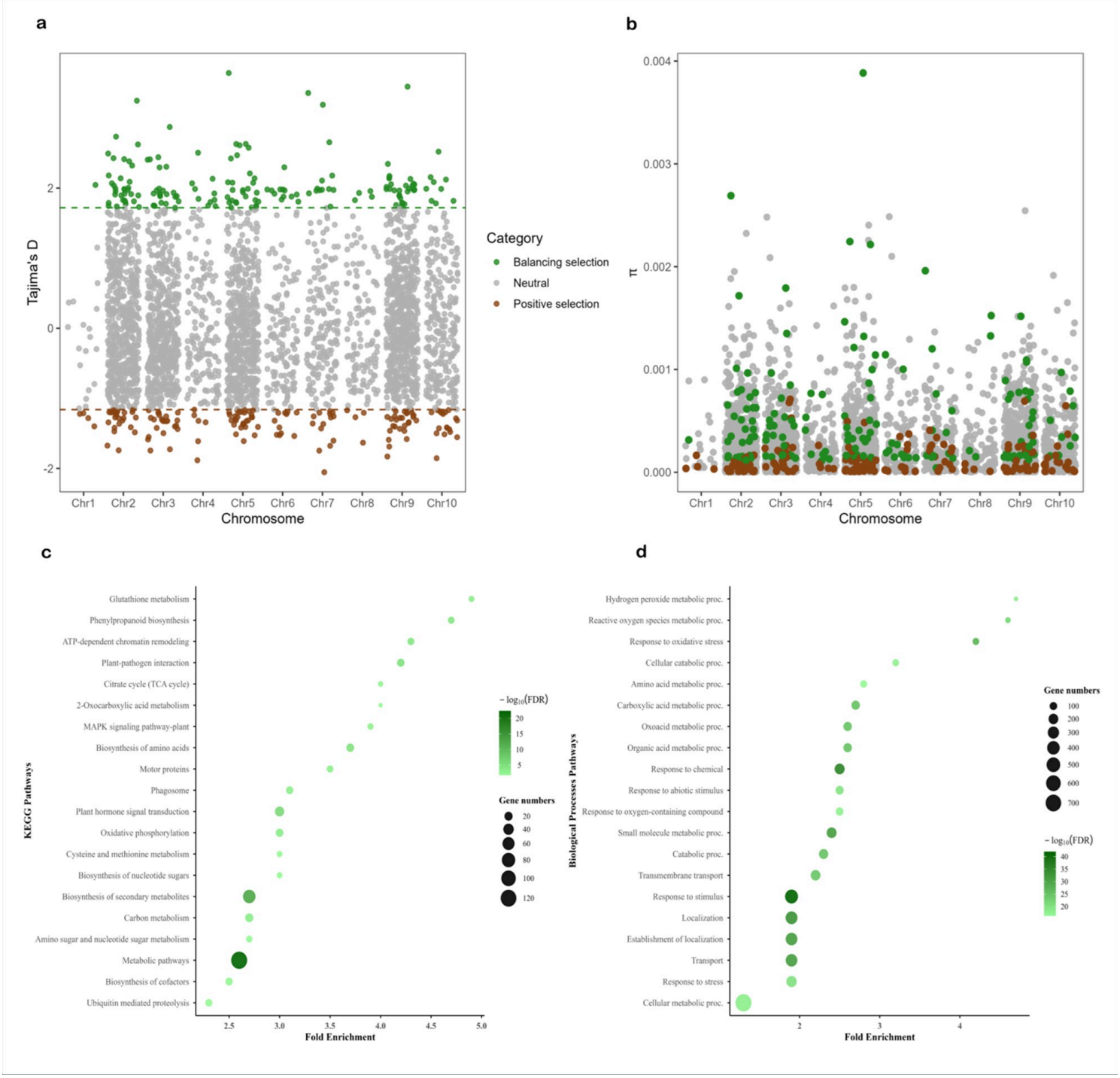

Figure 3

Distribution of Tajima's D and nucleotide diversity (π) in 1,000 bp windows of genes previously predicted as resistance-associated within coding regions of the *T. cacao* genome, and functional enrichment analysis of candidate exonic genes. (**a**) Chromosomal distribution of Tajima's D values; each point represents a 1,000 bp window. Windows were classified as putative balancing selection (green), neutral (gray), or putative positive selection (brown). The dashed green line represents the upper threshold (95th percentile; D = 1.72), indicating windows under potential balancing selection, while the dashed brown line represents the lower threshold (5th percentile; D = −1.16), indicating windows under

potential positive selection. **(b**) Nucleotide diversity (π) per chromosome calculated using the same windows and classification criteria as in panel (a). Windows under putative balancing selection show higher π values (green), windows under putative positive selection show lower π values (brown), and neutral windows display intermediate values (gray). (**c**) KEGG pathway enrichment analysis of candidate genes. Each point corresponds to a significantly enriched pathway; the x-axis indicates fold enrichment. Point size reflects the number of genes per pathway, and the green color intensity indicates statistical significance ($-\log_{10}$ FDR). (**d**) Gene Ontology (GO) biological process enrichment analysis of candidate genes, using the same visual representation as in panel (c).

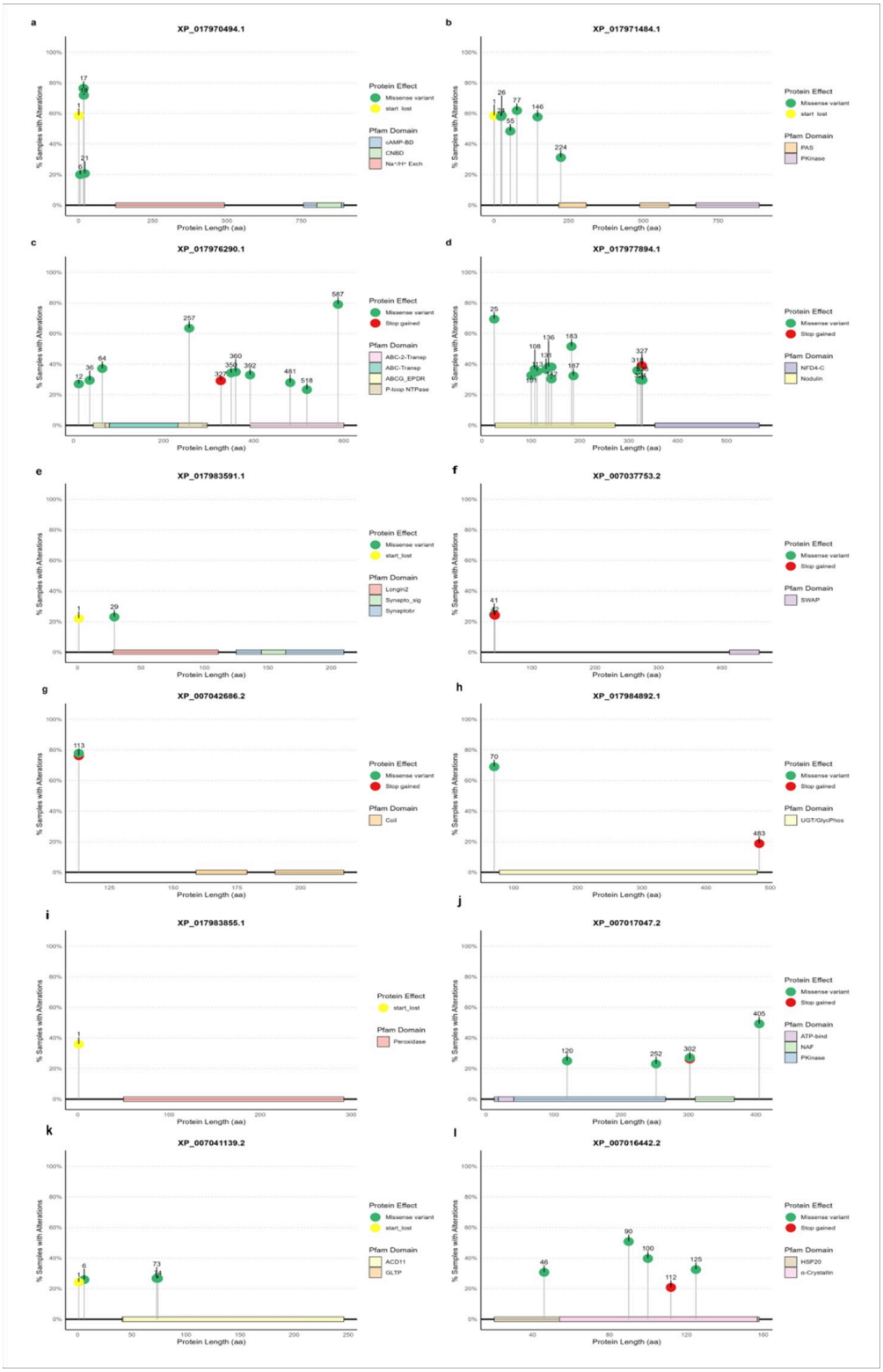

Figure 4

Lollipop plots showing high and moderate impact variants identified in genes located within windows under putative balancing selection (**a–d**), **putative positive selection (e–h), and neutral regions** (**i–l**). Each panel represents an individual protein, with total protein length displayed on the x-axis. Variant frequency across genotypes is shown on the y-axis. Circles along the protein sequence indicate the positions of affected amino acid residues and the type of mutation: missense variants (green), stop-

gained (red), and start-lost (yellow). Colored boxes along the protein sequences represent predicted functional domains annotated using Pfam and InterProScan. Numbers above each lollipop indicate the amino acid position of the corresponding variant.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables1117.xlsx
- Fig.S1.pdf
- SupplementaryTables110b.xlsx