

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

Gabriela Nicolle Ramos-Lizardo^{ID} · Konstantinos G. Alexiou · Jonathan Javier Mucherino-Muñoz · María José Aranzana · Ronan Xavier Corrêa^{ID}

Received: 16 April 2026 / Accepted: 30 July 2026

© The Author(s) 2026

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 resistance 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 typically associated with plant defense functions, including NBS-LRR domains, kinases and pathogenesis-related proteins. Functional annotation of coding regions variants identified 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, intervals previously reported in studies of resistance to *Phytophthora* spp., encoding proteins including class I chitinase, protease, ACD11, NADH oxidoreductase (NADH-OxR) and aquaporin NIP6-1, all of which have been associated with biological functions relevant to plant defense responses. These findings contribute to the understanding of potential defense-related mechanisms in *T. cacao* and identify promising candidate genes and variants for future functional validation. Following experimental validation, these candidates may contribute to the development of molecular markers for disease-resistance breeding programs.

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

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.



Introduction

Cocoa (*T. cacao* L.) is a perennial fruit tree typical of tropical regions¹, 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 industries². 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. pernicioso* (witches' broom), *Phytophthora* spp. (black pod rot), *C. cacaofunesta* (ceratocystis wilt), and *M. roreri* (frosty pod rot)^{3,4}. 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^{5–9}. 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^{10,11}.

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)^{10,11}. 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 PTI¹². 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 cultivars¹³.

The availability of genomic and RNA-seq data for *T. cacao* has enabled the identification of genes associated with agronomic traits and genetic improvement¹⁴. 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 productivity¹⁵. 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 programs¹⁶.

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, 2. 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

A total of 112 whole-genome sequencing (WGS) libraries were selected from three publicly available BioProjects (PRJNA558793, PRJNA734904, and PRJNA880462) Table S1, 2 which were chosen because they are associated with published studies addressing cacao genetic diversity, disease resistance, and genome-wide variation^{17–19}. Library selection was performed through manual curation of the associated metadata. Inclusion criteria required: genomic DNA sequencing data, paired-end Illumina sequencing, WGS libraries, FASTQ data availability, and sufficient metadata to identify genotype identity, population assignment, and available phenotypic information. Libraries lacking adequate metadata or not meeting these criteria were excluded.

Manual curation of the metadata was performed based on SRA records and the associated publications to classify each library according to its genotype identity and population of origin, including Amelonado, Criollo, Contamana, Nanay, Mara  n, Guiana, Iquitos, Curaray, Nacional, Matina, and admixed genetic groups. Historical disease resistance and susceptibility classifications were assigned only when supported by published phenotypic information or well-characterized reference genotypes.

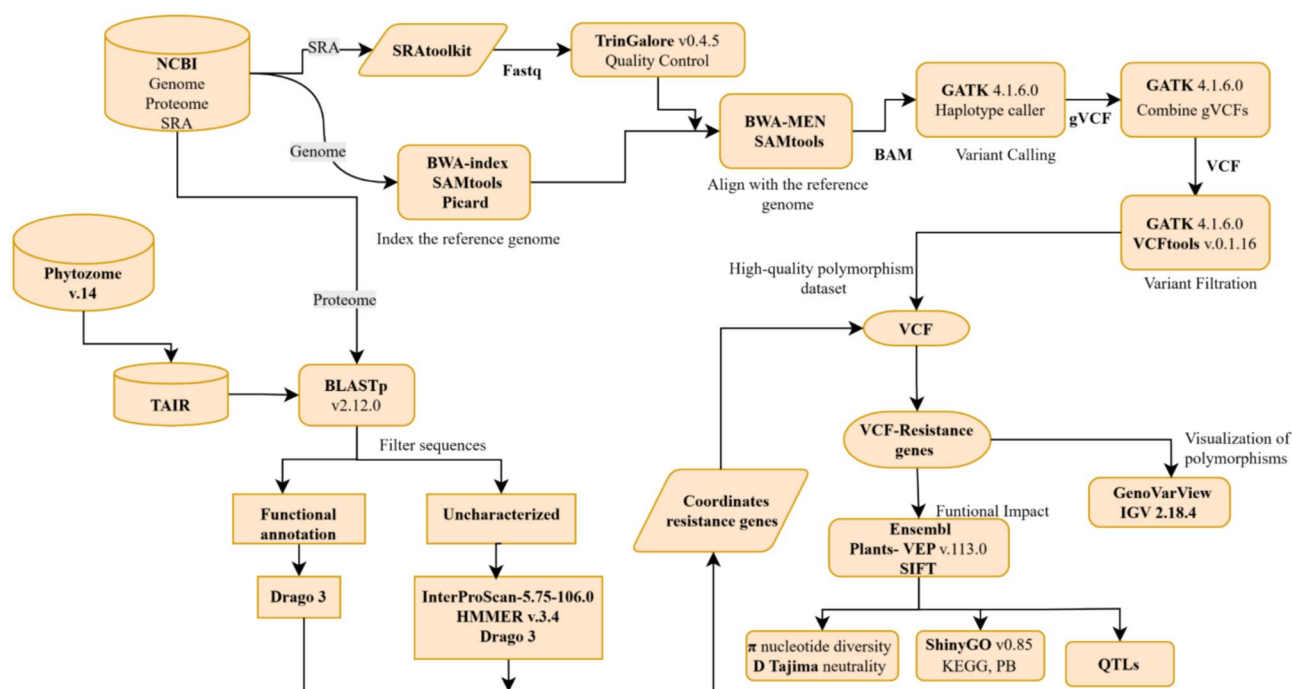


Fig. 1 The workflow shows the different tools used for variant analysis, prediction, and functional annotation of genes associated with resistance in *T. cacao*.

These included resistant genotypes (SCA6, SCA11, IMC65, NA672, NA127, MO76, RB41, RB43, and RB46) and susceptible genotypes (LIB3, CATONGO, LCTEEN, EBC, TIWI, and PAL8-A1)^{17–19} (Table S1, 2). Accessions lacking phenotypic information were retained to maximize representation of the genetic diversity of *T. cacao* and were not assigned to resistance categories. Importantly, these qualitative classifications were used strictly for contextual purposes, no formal comparative analysis or genome-wide association mapping was executed between resistant and susceptible phenotypic groups.

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.

Quality control and alignment

The quality of the reads was assessed using FastQC v0.11.9²⁰, 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 standard quality parameters, including a minimum length of 20 bp and a Phred score threshold of ≥ 20 , as pre-defined in TrimGalore's default settings for paired-end reads.

The *T. cacao* reference genome (GenBank accession: GCA_000208745.2) was indexed using BWA index v0.7.17-r1188²³, and SAMtools v1.15²¹, and its sequence dictionary was generated using the Create Sequence Dictionary module from Picard tools v1.110 (<https://broadinstitute.github.io/picard/>) was used to generate the sequence dictionary. Raw paired-end reads from all 112 libraries were aligned to the reference genome, sorted, indexed and processed for duplicates were systematically identified using fq2bam module implemented in the NVIDIA Clara Parabricks v4.0.1 environment with GPU acceleration²². This integrated end-to-end pipeline utilizes a GPU-accelerated version of BWA-MEM v0.7.17-r1188²³ and GATK-compatible duplicate marking algorithms (available at, <https://github.com/gnramliz/cacao-genomic-variability>).

Read groups (@RG) were assigned to each library during alignment to preserve sample identity and enable accurate joint processing of datasets originating from different BioProjects. Final alignment performance and mapping metrics were independently verified using Picard tools (Collect Alignment Summary Metrics) and SAMtools v1.15 (samtools flagstat). Sequencing effort varied substantially among libraries, ranging from 518,854 to 105,057,364 reads per dataset and corresponding to a theoretical genome coverage range of approximately 0.29 $\times$ to 93.22 $\times$ (Table S2). To reduce extreme differences in sequencing effort, a subset of high-depth libraries ($n = 23$) was randomly downsampled using seqtk prior to alignment, yielding datasets of approximately 40 million reads. This procedure standardized these high-depth libraries to approximately 37 $\times$ theoretical coverage relative to the 324.7 Mb *T. cacao* reference genome. No libraries were excluded following quality-control assessment, and all 112 datasets were retained for downstream analyses.

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)²⁴. 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 GenotypeGVCFs 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)²⁵. This process ensured the retention of high-confidence loci that satisfied the predefined quality, depth, and missing-data criteria. To further reduce potential biases associated with heterogeneous sequencing depth among libraries, downstream population-genetic analyses were conducted using variants that passed identical genotype-quality, depth, biallelic, and missing-data filters across all samples (available at, <https://github.com/gnramliz/cacao-genomic-variability>).

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 an initial reference framework, because it represents the most extensively characterized model species for plant immunity and disease resistance. Genes associated with plant defense were retrieved using the keyword “resistance to disease” 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).

To balance sensitivity and specificity in the detection of homologous proteins, stringent filtering criteria were applied, considering a query coverage $\geq 75\%$, identity $\geq 60\%$, and E-value $\leq 1e-5$. The resulting proteins 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

To complement the homology-based prediction and provide independent structural evidence supporting resistance-associated functions, both uncharacterized and previously annotated *T. cacao* proteins were subjected to conserved domain and motif analyses. Proteins that remained uncharacterized after the BLASTp analysis were analyzed using InterProScan v.5.75-106.0, HMMER v.3.4, and DRAGO 3, a specialized plant resistance gene prediction and annotation tool available through the PRGdb platform, run with default settings^{26,27}, to identify conserved domains and infer potential biological roles in plant defense.

InterProScan analyses were performed using the Pfam, SMART, CDD, ProSiteProfiles, ProSitePatterns, TIGRFAM, Coils, and SUPERFAMILY databases, with Gene Ontology, InterPro, and pathway annotations enabled. HMMER analyses were conducted against the Pfam-A database using hmmscan with six CPU threads. Additionally, *T. cacao* proteins with previous functional annotations were also analyzed using DRAGO 3 to identify and validate structural architectures associated with resistance and pathogen recognition, including NBS-LRR, RLK, and RLP classes.

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 v0.1.16²⁵, resulting in a set of variants located in genomic regions associated with resistance. Genomic variability was visualized using the Integrative Genomics Viewer (IGV)²⁸ 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)²⁹, using the *T. cacao* Criollo_cocoa_genome_V2 reference annotation with gene symbols, transcript versions, and transcript biotypes enabled, a 5 kb upstream/downstream window, a buffer size of 5000, and no right-alignment of variants. Variants were classified according to their predicted functional impact as high (stop-gain, start-loss, stop-loss), moderate (missense), or low (synonymous).

Missense substitutions identified in genes harboring high-impact coding variants were additionally evaluated using the SIFT web server to assess their potential effects on protein function. SIFT analyses were performed using a median conservation threshold of 3.0 and exclusion of sequences sharing more than 90% identity with the query protein. 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 candidate resistance-associated genes. Nucleotide diversity (π) and Tajima's D neutrality index^{30,31} were calculated using VCFtools v0.1.16²⁵ 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

A total of 2,460 genes containing high or moderate impact SNPs within their exonic regions were subjected to functional enrichment analysis using the ShinyGO v0.85 platform³². Enrichment was evaluated for both Gene Ontology (GO) Biological Processes and KEGG pathways (Kyoto Encyclopedia of Genes and Genomes, <https://www.genome.jp/kegg/>) to identify enriched functions and metabolic pathways. The statistical background for the hypergeometric test was defined as the complete set of annotated genes from the *T. cacao* Criollo B97-61/B2 v.2 genome reference genome available in the ShinyGO database. Enrichment results were filtered using a false discovery rate (FDR) threshold <0.05. Pathway size constraints restricted to a minimum of 5 and a maximum of 5000 genes to eliminate statistical noise, and redundant GO terms were automatically removed to optimize specificity. Enriched categories were ranked according to their fold enrichment³³. Graphical representations of these results were subsequently generated in R (v.4.4.3) using the ggplot2 package³⁴.

Genes associated with resistance QTLs

SNPs annotated by VEP as having moderate or high predicted impact consequences were used as input. For each SNP, genomic coordinates were retrieved and the corresponding genes in the *T. cacao* Criollo v2 genome were identified. To locate candidate resistance genes, SNP positions were intersected with physical QTL intervals previously reported for resistance to *Phytophthora* spp. and *M. perniciosa*. The QTL coordinates were obtained from the studies of Wong et al.³⁵ and Mucherino-Muñoz et al.³⁶, 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 (available at, <https://github.com/gnramliz/cacao-genomic-variability>), which loaded the QTL coordinates and compared them with the genomic position of each SNP, identifying as associated those whose positions fell within the corresponding QTL interval. SNPs located within QTL intervals 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 S3). 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 S4).

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 provide structural evidence supporting the classification of the predicted genes as resistance-associated candidates in *T. cacao*, particularly through the predominance of kinase and receptor like kinase-related architectures (Table S5).

Functional analysis of the 227 previously uncharacterized candidate resistance genes revealed a diversity of conserved domains commonly associated with plant immune receptors and defense-related proteins. The DRAGO tool identified TM, Kinase, NBS, and LECM domains in seven proteins encoded by seven genes (Table 1· Table S6). 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· Table S7).

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. Collectively, the DRAGO, InterProScan, and HMMER analyses identified multiple domain families frequently associated with plant immunity, including NB-ARC, NBS-LRR, TIR, kinase, and PR-related domains, providing independent structural support for the resistance-associated nature of these candidate genes (Table 1, Table S8).

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

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

Functional annotation and genomic distribution of SNPs

After applying the relevant quality and coverage criteria, a total of 11,439,669 SNPs and 1,069,309 INDELs was retained across the 112 WGS libraries of *T. cacao*. SNPs were subsequently filtered using a BED file containing the genomic coordinates of 4,523 predicted genes with potential resistance-related functions. Functional annotation with VEP identified 42,394 located within these genes (Fig. 2a, Table S9). 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 S9).

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 S10). Within coding regions, different types of variant consequences were identified, affecting a total of 2,460 genes and classified into three impact levels. 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 S10).

The distribution of these SNPs in the coding regions by chromosome showed that chromosomes 2, 5, 9, and 3 contained 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 moderate-impact variants exhibited allele frequencies below 20% and were classified as rare variants, being poorly represented among the analyzed genotypes. Variants with allele frequencies between 20% and 80% were considered intermediate-frequency

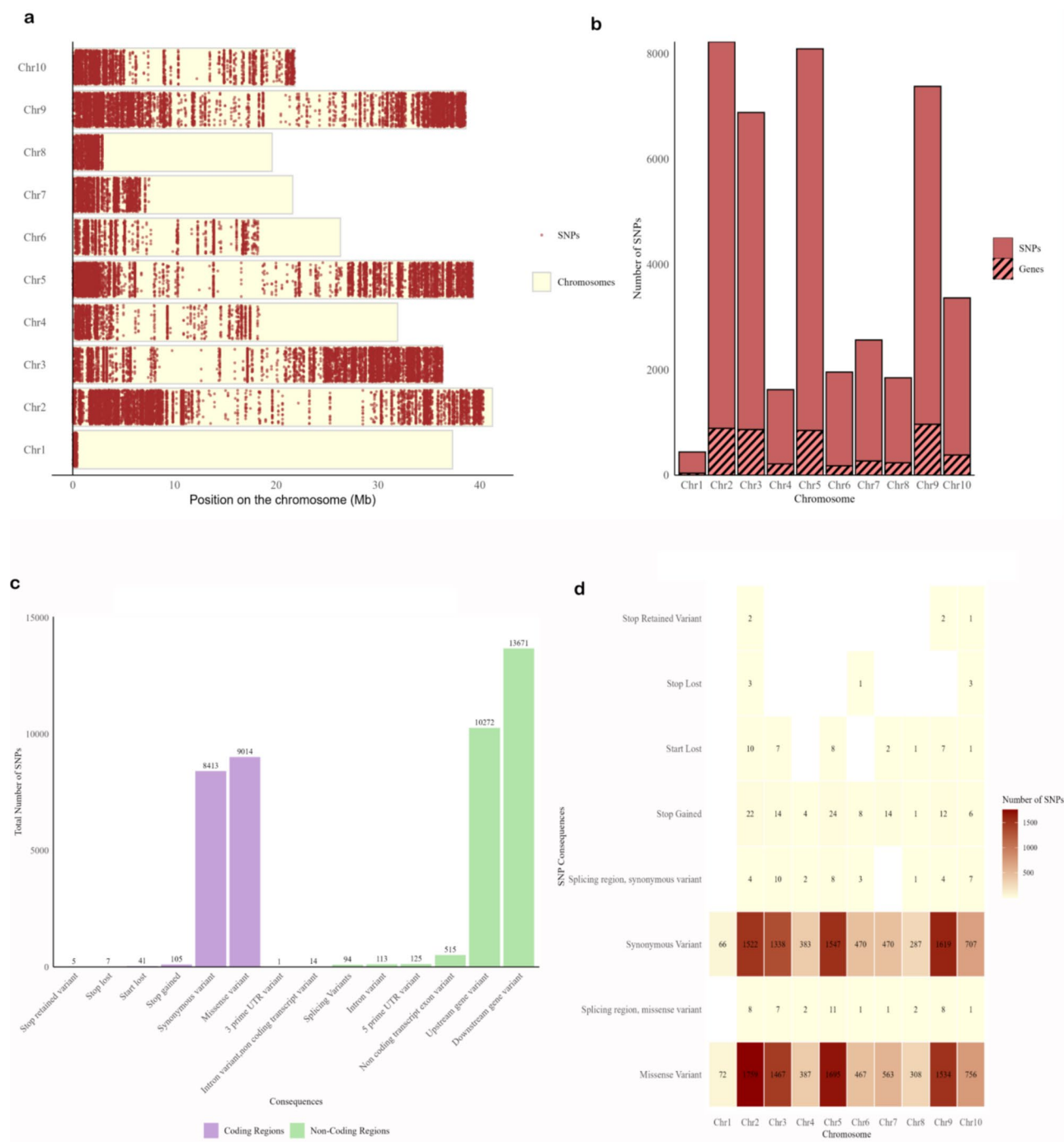


Fig. 2 Genomic distribution of resistance-associated SNPs in *T. cacao* and classification of variants by predicted consequence. **(a)** Distribution of SNPs across the ten chromosomes; red dots indicate the physical positions of SNPs, and yellow bars represent chromosome length. **(b)** Number of SNPs and annotated genes per chromosome. Red bars represent the total number of SNPs, and superimposed lines 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.

variants, representing 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 S11).

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 exhibited atypical Tajima's D values that may represent putative signatures of selection, with 50% corresponding to putative signals of positive selection and 50% to putative signals of 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 ($D > 1.72$), correspond to genomic windows that may represent putative evidence of balancing selection, whereas negative values below the 5th percentile ($D < -1.16$) correspond to genomic windows that may represent putative signatures of positive selection (Fig. 3a, Table S12). However, these atypical Tajima's D values should be interpreted cautiously, as they may reflect selection but can also result from population structure, admixture, demographic history, uneven sampling, or technical filtering. Tajima's D alone does not provide conclusive evidence of selection, and these signals require validation using complementary population-genomic approaches.

The nucleotide diversity (π) in windows showing putative evidence of selection revealed that 91% had values below 0.001, indicating low genetic variability, whereas 9% of the windows showed π values between 0.001 and 0.004, reflecting comparatively higher levels of allelic diversity (Fig. 3b, Table S12). 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 statistically significant enrichment of atypical Tajima's D windows on these chromosomes. Chromosomes 2, 5, and 9 showed the highest absolute numbers of genomic windows with atypical Tajima's D values. However, these values were not statistically significant ($P > 0.05$), suggesting their distribution follows a random pattern proportional to genome size (Table S13).

On chromosome 6, eight genes (Tc06v2_g002950, Tc06v2_g003200, Tc06v2_g006270, Tc06v2_g000370, Tc06v2_g002090, Tc06v2_g004180, Tc06v2_g004120, and Tc06v2_g006240) (Table S14), were located within genomic windows exhibiting putative signals of balancing selection. All of these genes exhibited intermediate allele frequencies across the analyzed genotypes and harbored missense variants. Among them, the gene Tc06v2_g002950 (XP_017979048.1) was selected for further examination because 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 S14).

Furthermore, two of these genes are located within previously reported QTLs intervals associated with resistance to *Phytophthora* spp in *T. cacao*. 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 AKAprIC-S100CHR62, 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 genomic windows exhibiting putative evidence of 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 genomic windows showing putative signals of balancing selection, all exhibiting intermediate allele frequencies and missense variants (Table S14). Among the candidate's genes located within windows showing putative signatures of balancing selection, two genes harbored

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 S14, 15).

Functional enrichment

Functional enrichment analysis of the 2,460 genes containing exonic variants revealed a total of 39 significantly enriched KEGG pathways, comprising both broad housekeeping metabolic networks and specific defense-related pathways. Among the general categories, metabolic pathways (127 genes) and secondary metabolite biosynthesis (66 genes) were highly prominent. Importantly, key pathways directly implicated in plant immunity and pathogen response were significantly enriched, 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 S16).

Analysis of the variants within these five defense-related pathways indicated that most mutations had a moderate impact; only a single high-impact variant was detected (located in the glutathione metabolism pathway), which exhibited a low allele frequency and was classified as rare. Furthermore, the enrichment analysis of GO Biological Processes identified 764 significantly enriched terms. Reflecting general cellular maintenance and systemic baseline functions, broad categories such as response to stimulus (434 genes), localization (344 genes), transport (326 genes), and small molecule metabolic processes (211 genes) showed the highest gene representation. However, defense-specific functional evidence was also prominent, as highlighted by the significant enrichment of response to oxidative stress (75 genes), with overall FDR values ranging from $1.10\text{E-}42$ to $3.60\text{E-}28$ (Fig. 3d, Table S15).

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 S17).

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 S15, S17). 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).

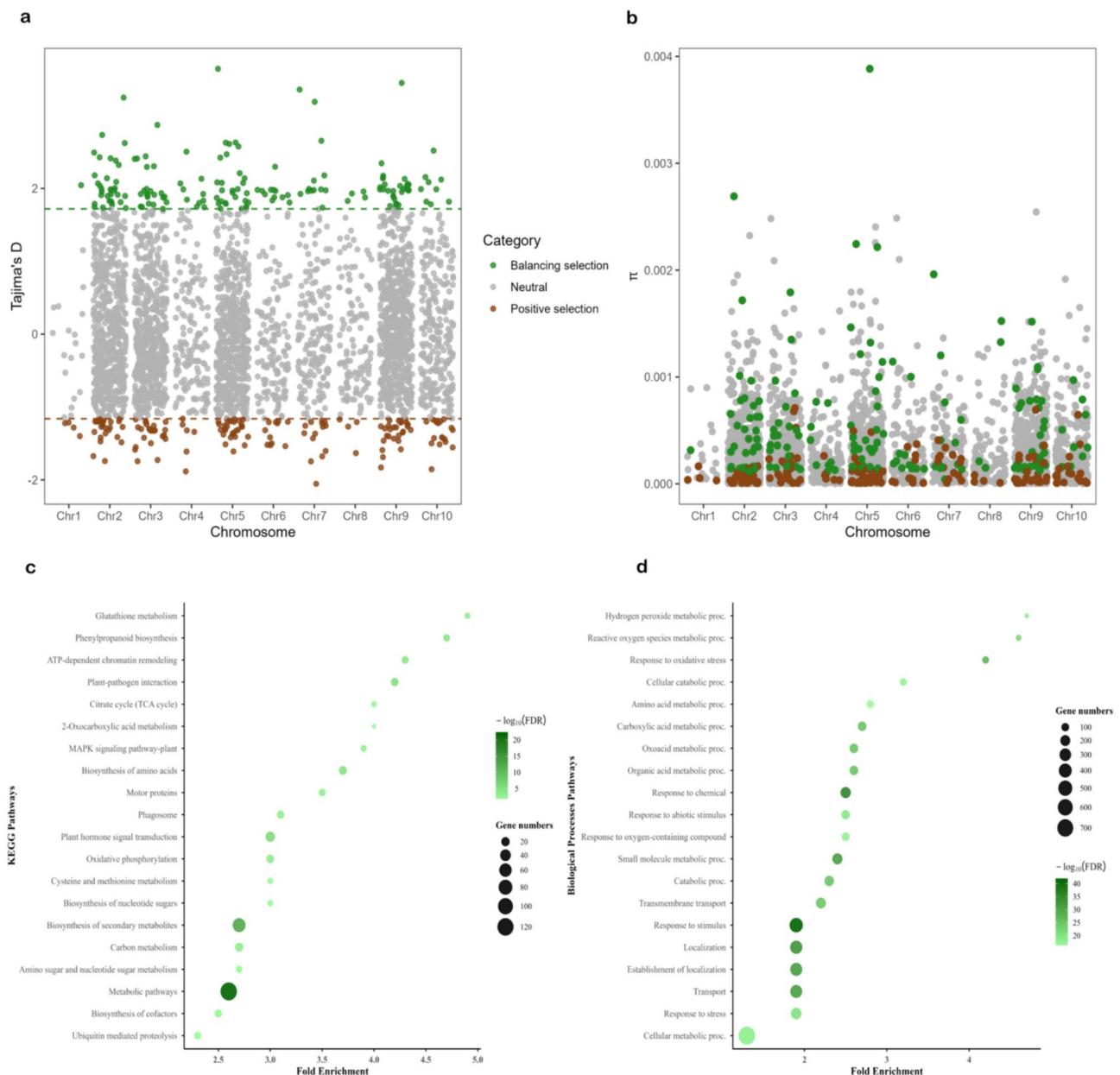
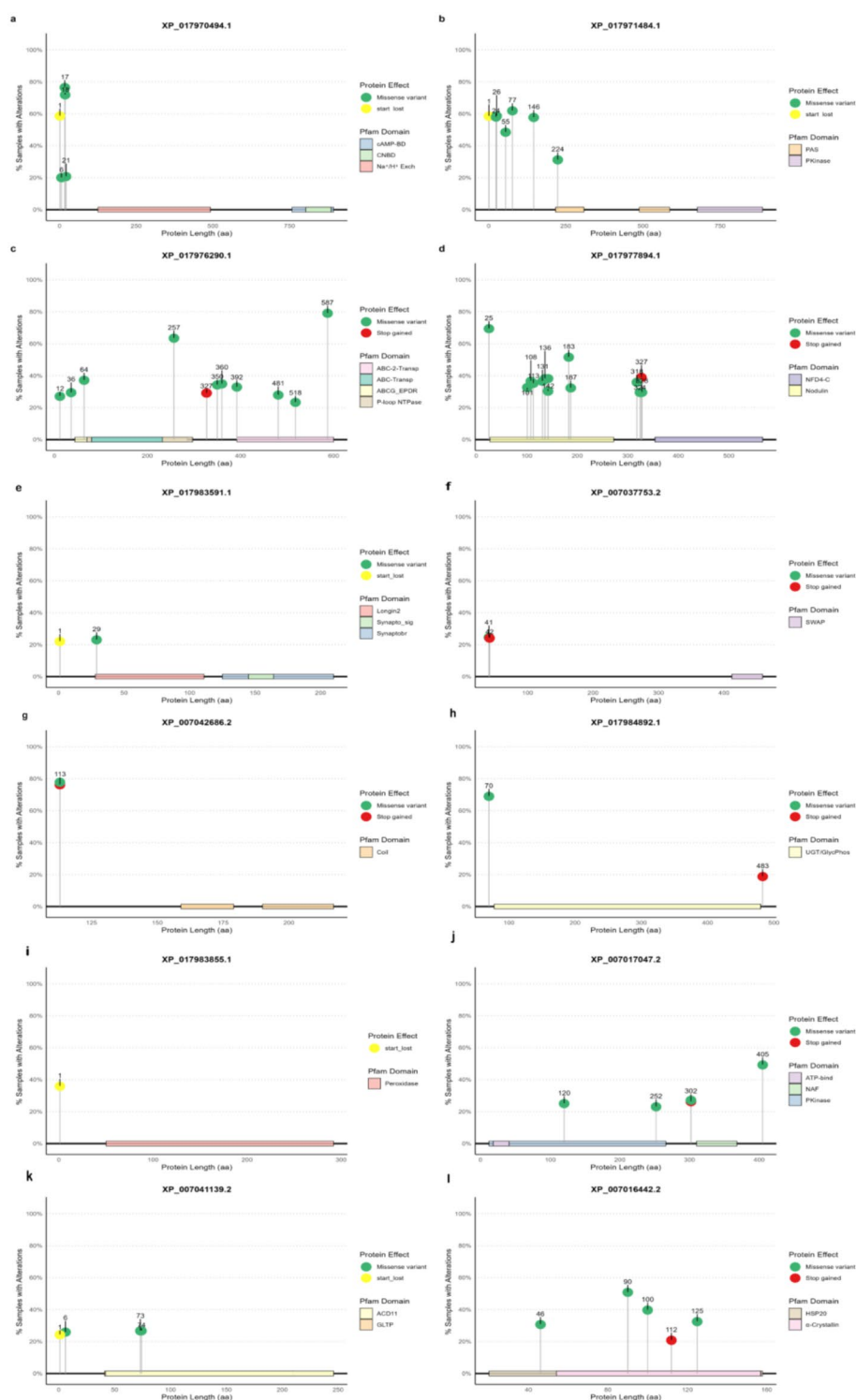


Fig. 3 Distribution of Tajima's D and nucleotide diversity (π) across 1,000 bp windows of genes previously identified as putative resistance-related candidates within coding regions of the *T. cacao* genome, and functional enrichment analysis of exonic candidate genes. **(a)** Chromosomal distribution 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) based on empirical thresholds. The dashed green line indicates the upper threshold (95th percentile; $D = 1.72$), corresponding to putative balancing selection, while the dashed brown line indicates the lower threshold (5th percentile; $D = -1.16$), corresponding to putative positive selection. **(b)** Nucleotide diversity (π) per chromosome calculated for the same windows and classification criteria as in panel (a). Windows classified as putative balancing selection show higher π values (green), whereas those classified as 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 ($-\log_{10} \text{FDR}$). **(d)** Gene Ontology (GO) biological process enrichment analysis of candidate genes, using the same visual representation as in panel (c).

Fig. 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.



Both proteins harbor start-lost variants as well as tolerated missense variants in their sequences (Fig. 4a, Table S15, S17).

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 S15, S17). 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 S15, S17).

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 potentially 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 S15, S17).

Analysis of QTLs associated with disease resistance in *T. cacao*

Of the total missense variants identified in coding regions, 797 SNPs were detected in 170 genes located within the coordinates of 11 previously reported QTL intervals in *T. cacao*, including nine intervals identified in studies of resistance to *Phytophthora* spp., and two intervals identified in studies of resistance to *M. perniciosa*. Among these genes, 84 presented 161 missense variants with intermediate allele frequencies, distributed across nine QTL intervals, including eight intervals previously reported in studies of resistance to *Phytophthora* spp. and one interval previously reported in studies of resistance to *M. perniciosa* (Table S19, S20, S21).

Of particular note is the gene Tc02v2_g008430, located on chromosome 2, which is co-localized with the previously reported QTL interval qSINT2, identified in studies of resistance to *M. perniciosa* in *T. cacao*. 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 candidate genes located within previously reported QTL intervals identified in studies of *T. cacao* resistance to *Phytophthora* spp., five genes stood out because they harbored variants with high or moderate predicted impact distributed across four QTL intervals. On chromosome 2, the gene Tc02v2_g001040 carries a high-impact variant (start-lost) located within the previously reported QTL interval AKAprrICS95CHR2. This gene encodes the protein XP_007041139.2, which is involved with the regulation of plant cell death and shows enrichment for biological processes related to localization and transport (Fig. 4k, Table S15, S18).

On chromosome 6, three candidate genes were identified. Tc06v2_g000370, harboring moderate-impact variants within the previously reported QTL interval BARq1BPPctCHR6, 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 S15, S18, S19). Also, on chromosome 6, two additional genes harbored SNPs located within the previously reported QTL interval AKAprrICS100CHR62. 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 a potential impact on protein function, may affect redox metabolism and reactive oxygen species (ROS) production during 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 S15, S17, S18, S19).

On chromosome 8, the gene Tc08v2_g004590 harbors a start-lost variant located within the previously reported QTL interval BROphy2CHR8. This gene encodes a protease (XP_007017945.2) containing Pept_S9A-N and AB-hydrolase domains and is enriched for biological processes such as protein metabolic process (GO:0019538), macromolecule metabolic process (GO:0043170), and proteolysis (GO:0006508) (Fig. S1, Table S15, S17, S18, S19).

Discussion

This study combined the prediction of genes associated with resistance through genetic variability analysis across 112 WGS libraries, aiming to provide candidate targets for future validation in strategies for the genetic improvement of *T. cacao* resistance to diseases. The identification of 4,523 candidate resistance-associated genes revealing a diverse functional repertoire, with a predominance of families putatively annotated as candidates involved in defense recognition and signaling, including kinases, RLK/RLP receptors, and NBS-LRR proteins (Table S4). These classes represent central components of plant immunity and have been widely associated with pathogen resistance and tolerance to abiotic stresses³⁷.

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^{38, 39} (Table S4). 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 analyses⁴⁰, which reported PR2 proteins co-localized within a previously reported QTL interval previously on the same chromosome, suggesting that this genomic region represents a candidate locus that may harbor multiple components of the defense response, although functional assays are required to confirm this relationship.

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, Table S6, S7, S8), is consistent with previous functional studies of *T. cacao* resistance to *P. palmivora*³⁶, where these chromosomes were also identified as carrying genes containing NBS, LRR, and RLK-type kinase domains, supporting their potential involvement 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. Whether these uncharacterized genes represent lineage-specific candidates in *T. cacao* or conserved genes across the *Theobroma* genus remains unknown and will require comparative genomic analyses.

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 are theoretically predicted to generate truncated, extended, or structurally altered proteins and hypothetically may compromise gene function. 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 SNPs⁴¹.

Functional enrichment indicated that a subset of 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 infection⁴².

Variants in genes within the phenylpropanoid pathway may influence resistance by altering lignin biosynthesis and the accumulation of antimicrobial phenolic compounds, including flavonoids thereby affecting structural and chemical barriers associated with plant immunity. Increased lignification may reinforce cell walls and hinder pathogen penetration, whereas phenolic compounds can contribute to chemical defense through antimicrobial activity. These defense mechanisms may contribute to limiting pathogen penetration and colonization during *Phytophthora* infections⁴³. Similar mechanisms may also contribute to restricting the establishment of *Moniliophthora* spp., although the contribution of phenylpropanoid-associated variation to resistance in this pathosystem remains less well characterized.

In addition to defense-related pathways, functional enrichment also identified broad biological processes related to general metabolism, signaling, and transport (Fig. 3d). While these broad categories represent systemic housekeeping and baseline cellular functions, they constitute the physiological framework required to support active immune responses. Within the localization and transport categories, specific defense-related functional evidence emerged, six genes showed high-impact mutations: Tc02v2_g026000, Tc05v2_g004320, Tc09v2_g011380, Tc05v2_g005990, Tc06v2_g003200, and Tc02v2_g001040. The last gene, 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 impact on protein function that may influence 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 may inhibit its activity, destabilizing its structure and potentially compromising its catalytic function⁴⁴. 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 resistance⁴⁵. The gene Tc05v2_g005990 encodes a protein containing a nodulin transporter domain (Fig. 4d) and represents another candidate gene associated with transport processes that may have potential implications in immunity. This type of domain has previously been linked to plant defense responses, including resistance of banana against *Xanthomonas campestris* pv. *musacearum*⁴⁶.

In addition, genes associated with response to stimulus, small molecule metabolic processes, and response to oxidative stress also harbored high-impact mutations providing localized functional evidence within wider metabolic networks. 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 responses⁴⁷. 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.^{48,49}.

On the other hand, within the response to oxidative stress category, a biological process closely linked to early immune signaling 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 responses⁵⁰. In poplar, for example, the gene PdePRX12 modulates resistance by regulating hydrogen peroxide (H₂O₂) levels after infection by *Botryosphaeria dothidea* and *Alternaria alternata*⁵¹. The SNPs identified in these genes may influence defense signaling and the transport of diverse molecules, including defense proteins, phytohormones and metabolites, potentially altering their biological function and defense-related signaling and transport processes involved in pathogen responses.

Some of these genes were co-localized with previously reported QTL intervals identified in studies of *Phytophthora* spp. in *T. cacao*. 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 proteins participate in the responses to different pathogens, with NADH oxidoreductases contributing to ROS production, while aquaporins act as regulators of immune signaling by facilitating H₂O₂ transport across membranes, thereby contributing to MAPK activation, defense gene induction, and PTI in plants^{10,11,52}. Recent studies further emphasize that ROS-mediated defense responses do not function in isolation but instead depend on a tightly regulated crosstalk with reactive nitrogen species (RNS) and downstream hormonal signaling networks, which collectively modulate plant immunity and stress resilience⁵³. Therefore, the co-localization of genes encoding an NADH oxidoreductase and an aquaporin within a previously reported QTL interval identified in studies of resistance to *Phytophthora* spp., together with the identification of candidate peroxidases associated with oxidative stress, suggests a possible contribution of comprehensive redox-regulated immune signaling to defense responses in *T. cacao*.

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 addition, this gene has previously been reported as differentially expressed during *Phytophthora megakarya* infection in *T. cacao*, where reduced transcript accumulation was observed 48 h after inoculation (Table S21)^{43,54}. Although this expression pattern does not establish a causal role in resistance, it provides independent transcriptomic evidence supporting the potential involvement of this locus in cacao–pathogen interactions. 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*^{55,56}.

The gene Tc08v2_g004590, located on chromosome 8, was identified within the QTL interval BROphy-2CHR8. 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 HR⁵⁷. The high variability of high and moderate impact variants detected in these regions is consistent with previous studies that reported QTL intervals on these chromosomes in relation to disease-resistance traits, suggesting that these genomic regions represent promising targets for the future characterization of alleles potentially involved in cacao–pathogen interactions^{36,40}.

Our nucleotide variation and Tajima's D analysis performed on genes previously predicted to be located in coding regions revealed a heterogeneous distribution of genomic windows exhibiting atypical Tajima's D values across the ten chromosomes. Chromosome 6 presented eight genes located within genomic windows exhibiting

putative evidence of balancing selection, associated with diverse biological processes, including metabolic processes, transport, and response to stimuli, and harboring missense variants.

Among these genes, Tc06v2_g000370 and Tc06v2_g003200 were of particular interest because they were involved in biological processes associated with plant defense and were co-localized with the previously reported QTL intervals BARq1BPPctCHR6 and AKAprrICS100CHR62, which were identified in studies of resistance to *Phytophthora* spp. in *T. cacao*³⁶. Although the co-localization of these genes with previously reported QTL intervals is consistent with a potential contribution of pathogen-mediated balancing selection, the relative influence of demographic history, population structure, and other environmental selective pressures on the observed Tajima's D patterns remains uncertain.

On chromosome 7, two genes were identified within genomic windows exhibiting putative signals of positive selection, and nine genes were located within genomic windows exhibiting putative signatures of balancing selection. Among these, Tc07v2_g001630 and Tc07v2_g008920, both located within genomic windows exhibiting putative evidence of balancing selection, were associated with transport and metabolism related biological processes. Chromosomes 2, 3, 5, and 9 not only harbored the highest numbers of high-impact SNPs, but also contained variants located within genomic windows exhibiting putative signatures of positive and balancing selection.

On chromosomes 2 and 5 (Fig. 4a-d), four genes located within genomic windows exhibiting putative signatures of 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, contributing to comparatively higher levels of genetic diversity in these genomic regions. The occurrence of genomic windows exhibiting putative signals of balancing selection on chromosomes 2 and 5 is of interest because both chromosomes have previously been associated with quantitative resistance to witches' broom and frosty pod diseases (Brown *et al.*, 2005;⁴⁰.

On the other hand, genes identified within genomic windows exhibiting putative signatures of positive selection, including four genes identified on chromosomes 2, 3, 9, and 10, showed fewer missense variants (Fig. 4e-h). These patterns may reflect differences in allele frequency distributions among genomic regions; however, Tajima's D alone is insufficient to distinguish between selection and alternative demographic processes, and these results should therefore be interpreted as preliminary.

Limitations and future perspectives

Although this study provides a comprehensive assessment of genetic variation within candidate resistance-associated genes in *T. cacao*, several limitations should be considered. First, all candidate genes, SNPs, and predicted functional effects were identified exclusively through bioinformatic approaches and have not been experimentally validated; therefore, the biological effects of variants predicted by VEP and SIFT remain hypothetical and may not necessarily result in complete loss of protein function, as alternative transcripts or compensatory mechanisms could mitigate their effects. In addition, resistance-gene annotation relied partly on homology-based annotation using *A. thaliana*. Although complementary domain-based analyses using DRAGO3, InterProScan, and HMMER provided additional support, functional assignments derived from homologous genes should be interpreted with caution and lineage-specific resistance genes lacking detectable homologs in *A. thaliana* may have been underrepresented. Furthermore, the broad genomic intervals encompassed by the analyzed QTLs limit the precise identification of specific causal variants.

Second, population-genomic inferences based on Tajima's D should also be interpreted as exploratory, as they were derived from a heterogeneous set of publicly available genomes representing diverse genetic backgrounds. Without explicit correction for population structure or demographic history, observed deviations in Tajima's D may reflect a combination of selection, demography, admixture, or sampling effects rather than true adaptive

selection alone. Likewise, no direct genotype–phenotype association analyses were performed, limiting causal inference between genetic variation and resistance traits.

Finally, genomic data alone do not capture dynamic transcriptional responses during pathogen infection, restricting the biological interpretation of candidate genes. Future studies integrating population-aware analyses, genotype–phenotype association studies, functional validation, and transcriptomic approaches, including single-cell and spatial transcriptomics, will be essential to confirm the biological relevance of the candidate genes and evaluate their potential application in marker-assisted breeding programs.

Conclusions

Our results identified a repertoire of genes potentially associated with resistance in *T. cacao*, characterized by classical defense-related functional domains. The integration of functional annotation, predicted SNP effects, and co-localization with previously described QTLs provides complementary bioinformatic evidence supporting the prioritization of candidate genes, involved in processes related to MAPK signaling, transport, oxidative metabolism, and pathogen recognition. In addition, the occurrence of high- and moderate-impact variants in these genes, together with putative evidence of positive and balancing selection inferred from Tajima's D analyses, highlights genomic regions that may have contributed to the diversification of defense-related functions and warrant further investigation. Overall, these findings provide an exploratory framework for future functional genetic, and population-genomic studies. Experimental validation will be required to confirm the biological relevance of the prioritized genes and variants and to assess their potential utility in breeding programs aimed at developing disease-resistant cacao cultivars.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-026-65298-7>.

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).

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.

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.

Data availability All data supporting the findings of this study are available within the paper, its Supplementary Information, and the public repository: <https://github.com/gnramliz/cacao-genomic-variability>.

Declarations

Competing interest The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the

original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Lahive, F., Hadley, P. & Daymond, A. J. The physiological responses of cacao to the environment and the implications for climate change resilience: a review. *Agron. Sustain. Dev.* **39**, 5. <https://doi.org/10.1007/s13593-018-0552-0> (2019).
2. Kongor, J. E., Owusu, M. & Oduro-Yeboah, C. Cocoa production in the 2020s: challenges and solutions. *CABI Agric. Biosci.* **5**, 102. <https://doi.org/10.1186/s43170-024-00310-6> (2024).
3. Marelli, J. P. et al. Chocolate under threat from old and new cacao diseases. *Phytopathology* **109**, 1331–1343. <https://doi.org/10.1094/PHYTO-12-18-0477-RVW> (2019).
4. Carrero-Gutiérrez, M. L. et al. *Ceratocystis cacaofunesta* is responsible for cocoa crop wilt in Colombia: Morphological and molecular characterization of isolates. *J. Plant Pathol.* **107**, 795–810. <https://doi.org/10.1007/s42161-024-01798-7> (2025).
5. Adeniran, A. O., Onanaye, A. S. & Adeleke, O. J. Optimal control of cocoa black pod disease: a multi-pronged approach. *Franklin Open* **7**, 100100. <https://doi.org/10.1016/j.fraope.2024.100100> (2024).
6. Jibat, M. & Alo, S. Integrated management of black pod (*Phytophthora palmivora*) disease of cocoa through fungicides and cultural practices in southwestern Ethiopia. *Int. J. Food Agric. Nat. Resour.* **4**(3), 43–45. <https://doi.org/10.46676/ij-fanres.v4i3.150> (2023).
7. Pires JL, Pires Melo GR, Yamada MM, Martins ED, Luz N Resistance of cocoa to *Moniliophthora perniciosa*: progenitors and progenies selection. *Revista Ceres* 69:544–552. <https://ojs.ceres.ufv.br/ceres/article/view/8039>. (2025).
8. Sudarma, I. M., Adnyana, I. W. & Darmiati, N. *Trichoderma* sp. as endophytic fungi as a candidate for cocoa pod rot disease control. *GPH-Int. J. Agric. Res.* **6**, 88–96. <https://doi.org/10.5281/zenodo.10441163> (2023).
9. Wijaya, G., Wardani, A. K. & Eris, D. Biocontrol activity of endophytic bacteria from cocoa against *Phytophthora* sp. and *Colletotrichum* sp.. *E-J. Menara Perkebunan* **91**, 72–86. <https://doi.org/10.22302/iribb.jur.mp.v91i1.522> (2023).
10. Yuan, M., Ngou, B. P. M., Ding, P. & Xin, X.-F. PTI–ETI crosstalk: an integrative view of plant immunity. *Curr. Opin. Plant Biol.* **62**, 102030. <https://doi.org/10.1016/j.pbi.2021.102030> (2021).
11. Yuan, M. et al. Pattern-recognition receptors are required for NLR-mediated plant immunity. *Nature* **592**, 105–109. <https://doi.org/10.1038/s41586-021-03316-6> (2021).
12. Wang, D. et al. Suppression of ETI by PTI priming to balance plant growth and defense through an MPK3/MPK6–WRKYs–PP2Cs module. *Mol. Plant* **16**, 903–918. <https://doi.org/10.1016/j.molp.2023.04.004> (2023).
13. Li, Q. et al. Plant genes related to *Phytophthora* pathogen resistance. *Phytopathol. Res.* **6**, 15. <https://doi.org/10.1186/s42483-024-00229-w> (2024).
14. Kulesza, E. et al. The cacao gene atlas: A transcriptome developmental atlas reveals highly tissue-specific and dynamically regulated gene networks in *Theobroma cacao* L.. *BMC Plant Biol.* **24**, 601. <https://doi.org/10.1186/s12870-024-05171-9> (2024).
15. Nieves-Orduña, H. E. et al. Genotyping of cacao (*Theobroma cacao* L.) germplasm resources with SNP markers linked to agronomic traits reveals signs of selection. *Tree Genet. Genomes* **20**, 13. <https://doi.org/10.1007/s11295-024-01646-w> (2024).
16. Fernandes, L. D. S., Correa, F. M., Ingram, K. T., de Almeida, A. A. F. & Royaert, S. QTL mapping and identification of SNP-haplotypes affecting yield components of *Theobroma cacao* L. *Hortic. Res.* **7**, 26. <https://doi.org/10.1038/s41438-020-0250-3> (2020).
17. Osorio-Guarín, J. et al. Genome-wide association study reveals novel candidate genes associated with productivity and disease resistance to *Moniliophthora* spp. in cacao (*Theobroma cacao* L.). *G3 (Bethesda)* **10**, 1713–1725. <https://doi.org/10.1534/g3.120.401153> (2020).
18. Hämälä, T. et al. Genomic structural variants constrain and facilitate adaptation in natural populations of *Theobroma cacao*, the chocolate tree. *Proc. Natl. Acad. Sci. U. S. A.* **118**(35), e2102914118. <https://doi.org/10.1073/pnas.2102914118> (2021).
19. Argout, X., et al. 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> (2023).
20. Andrews, S. FastQC: A quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2010).
21. Li, H. et al. The Sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. <https://doi.org/10.1093/bioinformatics/btp352> (2009).

22. Pham M, Tu Y, Lv X 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> (2023).
23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760. <https://doi.org/10.1093/bioinformatics/btp324> (2009).
24. McKenna, A. et al. The Genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. <https://doi.org/10.1101/gr.107524.110> (2010).
25. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330> (2011).
26. Blum, M. et al. InterPro: The protein sequence classification resource in 2025. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkae1082> (2025).
27. García, J. C. et al. PRGdb 4.0: An updated database dedicated to genes involved in plant disease resistance processes. *Nucleic Acids Res.* **50**(1), D1483–D1490. <https://doi.org/10.1093/nar/gkab1087> (2022).
28. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Cancer Res.* **77**, e31–e34. <https://doi.org/10.1158/0008-5472.CAN-17-0337> (2017).
29. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122. <https://doi.org/10.1186/s13059-016-0974-4> (2016).
30. Nei, M. & Li, W. H. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* **76**, 5269–5273 (1979).
31. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**, 437–460. <https://doi.org/10.1093/genetics/105.2.437> (1983).
32. Ge, S. X., Jung, D. & Yao, R. ShinyGO: A graphical gene-set enrichment tool for animals and plants. *Bioinformatics* **36**(8), 2628–2629. <https://doi.org/10.1093/bioinformatics/btz931> (2020).
33. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**(D1), D545–D551. <https://doi.org/10.1093/nar/gkaa970> (2021).
34. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).
35. Wong, C. J. A. et al. A high-density genetic map from a cacao F2 progeny and QTL detection for resistance to witches' broom disease. *Tree Gene. Genom.* **18**, 31. <https://doi.org/10.1007/s11295-022-01562-x> (2022).
36. Mucherino-Muñoz, J. J. et al. Structural and functional genomics of the resistance of cacao to *Phytophthora palmivora*. *Pathogens* **10**, 961. <https://doi.org/10.3390/pathogens10080961> (2021).
37. Kim, H.-J. et al. OsLRR-RLP2 gene regulates immunity to *Magnaporthe oryzae* in japonica rice. *Int. J. Mol. Sci.* **25**, 2216. <https://doi.org/10.3390/ijms25042216> (2024).
38. De Oliveira, I. B. et al. Apoplastomes of contrasting cacao genotypes to witches' broom disease reveal differential accumulation of PR proteins. *Front. Plant Sci.* **15**, 1387153. <https://doi.org/10.3389/fpls.2024.1387153> (2024).
39. Singh, V., Hallan, V. & Pati, P. K. *Withania somnifera* osmotin (WsOsm) confers stress tolerance in tobacco and establishes novel interactions with the defensin protein (WsDF). *Physiol. Plant.* **176**, e14513. <https://doi.org/10.1111/pp1.14513> (2024).
40. Lanaud, C. et al. A meta-QTL analysis of disease resistance traits of *Theobroma cacao* L. *Mol. Breed.* **24**, 361–374. <https://doi.org/10.1007/s11032-009-9297-4> (2009).
41. Guajardo, V. et al. 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. *Sci. Rep.* **10**, 1467. <https://doi.org/10.1038/s41598-020-58271-5> (2020).
42. Baruah, I. K., Ali, S. S., Shao, J., Lary, D. & Bailey, B. A. Changes in gene expression in leaves of cacao genotypes resistant and susceptible to *Phytophthora palmivora* infection. *Front. Plant Sci.* **12**, 780805. <https://doi.org/10.3389/fpls.2021.780805> (2022).
43. Rêgo, A. P. B., Mora-Ocampo, I. Y. & Corrêa, R. X. Interactions of different species of *Phytophthora* with cacao induce genetic, biochemical, and morphological plant alterations. *Microorganisms* **11**(5), 1172. <https://doi.org/10.3390/microorganisms11051172> (2023).
44. Rifat, M. H. et al. Prediction and expression analysis of deleterious nonsynonymous SNPs of *Arabidopsis ACD11* gene by combining computational algorithms and molecular docking. *PLoS Comput. Biol.* **18**, e1009539. <https://doi.org/10.1371/journal.pcbi.1009539> (2022).
45. Dahuja, A. et al. Role of ATP-binding cassette transporters in maintaining plant homeostasis under abiotic and biotic stresses. *Physiol. Plant.* **171**(4), 785–801. <https://doi.org/10.1111/ppl.13302> (2021).
46. Ntui, V. O. et al. Targeted knockout of early nodulin-like 3 (*MusaENODL3*) reveals its function in resistance to *Xanthomonas* wilt disease. *Plant Biotechnol. J.* **22**, 1101–1112. <https://doi.org/10.1111/pbi.14248> (2024).
47. Yan, H. et al. A large-scale screening identifies receptor-like kinases with common features in kinase domains potentially related to disease resistance in planta. *Front. Sci.* <https://doi.org/10.3389/fpls.2024.1503773> (2024).
48. Bethke, G. et al. UDP-glucosyltransferase HvUGT13248 confers type II resistance to *Fusarium graminearum* in barley. *Plant Physiol.* **193**(4), 2691–2710. <https://doi.org/10.1093/plphys/kiad467> (2023).

49. Della Gala, V. et al. 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> (2025).
50. Lu, Y. et al. Genome-wide analysis of the class III peroxidase gene family in apple (*Malus domestica*). *PeerJ* **13**, e19741. <https://doi.org/10.7717/peerj.19741> (2025).
51. Cai, G., Zhang, Y., Huang, L. & Wang, N. Uncovering the role of PdePrx12 peroxidase in enhancing disease resistance in poplar trees. *J. Fungi* **9**(4), 410. <https://doi.org/10.3390/jof9040410> (2023).
52. Li, G., Chen, T., Zhang, Z., Li, B. & Tian, S. Roles of aquaporins in plant–pathogen interaction. *Plants* **9**(9), 1134. <https://doi.org/10.3390/plants9091134> (2020).
53. Ali, M. et al. Biosynthesis and multifaceted roles of reactive species in plant defense mechanisms during environmental cues. *Plant Stress* **18**, 101102. <https://doi.org/10.1016/j.stress.2025.101102> (2025).
54. Bailey, B. A. et al. Developmental expression of stress response genes in *Theobroma cacao* leaves and their response to Nep1 treatment and a compatible infection by *Phytophthora megakarya*. *Plant Physiol. Biochem.* **43**(6), 611–622. <https://doi.org/10.1016/j.plaphy.2005.04.006> (2005).
55. Toufiq, N. et al. Improved antifungal activity of barley-derived chitinase I overexpressing a 32 kDa recombinant chitinase in *Escherichia coli*. *Braz. J. Microbiol.* **49**, 414–421. <https://doi.org/10.1016/j.bjm.2017.05.007> (2018).
56. Vaghela, B., Vashi, R., Rajput, K. & Joshi, R. Plant chitinases and their role in plant defense: A comprehensive review. *Enzyme Microb. Technol.* **159**, 110055. <https://doi.org/10.1016/j.enzmictec.2022.110055> (2022).
57. Fernández-Fernández, Á. D., Stael, S. & Van Breusegem, F. Mechanisms controlling plant proteases and their substrates. *Cell Death Differ.* **30**(4), 1047–1058. <https://doi.org/10.1038/s41418-023-01120-5> (2023).
58. Brown, J. S. et al. Mapping QTLs for resistance to frosty pod and black pod diseases and horticultural traits in *Theobroma cacao* L.. *Crop Sci.* **47**, 1851–1858. <https://doi.org/10.2135/cropsci2006.11.0753> (2007).
59. Goodstein, D. M. et al. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **40**(D1), D1178–D1186. <https://doi.org/10.1093/nar/gkr944> (2012).
60. Instituto Brasileiro de Geografia e Estatística (IBGE). Cacao: Produção agropecuária. <https://www.ibge.gov.br/explica/producao-agropecuaria/cacau/br> (2024).
61. International Cocoa Organization (ICCO). Cocoa market and economy. (2024).
62. Krueger, F. TrimGalore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (2019).
63. R Core Team. *R: A language and environment for statistical computing* (Version 4.4.3). R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/> (2025).

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

Authors and Affiliations

Gabriela Nicolle Ramos-Lizardo¹  · Konstantinos G. Alexiou² ·

Jonathan Javier Mucherino-Muñoz³ · María José Aranzana² · Ronan Xavier Corrêa^{1,4} 

✉ Gabriela Nicolle Ramos-Lizardo
nicolleramos94@gmail.com

✉ Ronan Xavier Corrêa
ronanxc@uesc.br

¹ Center for Biotechnology and Genetics, Universidade Estadual de Santa Cruz (UESC), Ilhéus 45662-900, Brazil

² Center for Research in Agricultural Genomics, Building - UAB Campus, Cerdanyola del Vallès, 08193 Barcelona, Spain

³ Institute of Biodiversity and Sustainability (NUPEM), Federal University of Rio de Janeiro (UFRJ), Macaé, Rio de Janeiro, Brazil

⁴ Department of Biological Sciences, State University of Santa Cruz, Ilhéus, Bahia 45662-900, Brazil