New Phytologist

Research

# Unlocking genetic diversity in Colombian cassava landraces for accelerated breeding

**Kehan Zhao[1], Evan Long[2], Francisco Sanchez[3], Erwan Monier[4], Paul Chavarriaga[3] and Grey Monroe[1]**

[1]Department of Plant Sciences, University of California Davis, Davis, CA 95616, USA; [2]Northwest Irrigation and Soils Research Laboratory, United States Department of Agriculture – Agricultural Research Service, Kimberly, ID 83341, USA; [3]Alliance Bioversity International & International Center for Tropical Agriculture – CIAT, Apartado Aéreo, 6713, Cali, Colombia; [4]Department of Land, Air and Water Resources, University of California Davis, Davis, CA 95616, USA

Authors for correspondence:
*Evan Long*
*Email: evan.long@usda.gov*

*Grey Monroe*
*Email: gmonroe@ucdavis.edu*

Received: *30 October 2025*
Accepted: *17 December 2025*

*New Phytologist* (2026)
**doi**: 10.1111/nph.70918

**Key words:** cassava, climate adaptation, coumarins, gene editing, genetic diversity, landraces, loss-of-function mutations, *Manihot esculenta*.

## Summary

- Cassava (*Manihot esculenta*) is a staple crop across the global south, yet modern varieties may have limited genetic diversity due to historical bottlenecks. We investigated the genomic diversity of over 1000 cassava genotypes, incorporating 387 newly sequenced Colombian landraces originating from diverse climates. We hypothesized that landraces retain untapped variation useful for breeding and adaptation.
- Whole-genome sequencing was used to characterize landraces and breeding lines. We assessed genetic differentiation across geography and climate and analyzed the distribution of loss-of-function (LoF) mutations to identify potential targets for gene editing.
- Landraces maintained high and novel dimensions of genetic diversity compared to breeding lines from Asia and Africa. Differentiation among landraces reflected both demography and climate of origin. LoF analyses indicated purging of deleterious alleles through inbreeding, but LoF alleles were retained in genes enriched for coumarin biosynthesis and plant immunity, suggesting selection for postharvest quality and disease resistance. Climate-associated loci were explored for their adaptive potential.
- Cassava landraces represent a critical reservoir of genetic diversity. This study establishes a foundation for leveraging landrace variation to accelerate cassava improvement through gene editing and targeted breeding.

## Introduction

Cassava (*Manihot esculenta Crantz*), a versatile and resilient monoecious root crop, provides a major caloric source for over 500 million people world-wide, especially in the global south (Parmar *et al.*, 2017; Ferguson *et al.*, 2019). The importance of cassava in global agriculture is largely due to its remarkable productivity with minimal inputs and ability to thrive under marginal conditions, such as in water-limited environments (Parmar *et al.*, 2017). However, this adaptability faces new environmental challenges including extreme weather events and plant diseases (Food and Agriculture Organization, 2019). While cassava is a staple in regions of the world with the fastest human population growth, recent projections have predicted a possible drop in cassava yield of up to 10–20% in the next 50–100 yr (Zhu *et al.*, 2023). These underscore the necessity to fortify the resilience of essential crops like cassava for future food security.

Cassava was domesticated in or near the Amazon Basin between 5 and 10 thousand years ago, and has since become a major crop in the tropical regions of South America, Africa, and Asia (Parmar *et al.*, 2017; Ferguson *et al.*, 2019). In recent years, wild germplasm and domesticated accessions not used for intensive breeding have been seen as a valuable resource to access adaptive genetic diversity (Kashyap *et al.*, 2022). Crop landraces, or traditional cultivated varieties, can be valuable sources of beneficial traits and alleles, often due to adaptation reflecting geographical variation. Leveraging landraces or wild relatives to identify genetic variation responsible for environmental adaptation has been successful in multiple crops and other plant species including wheat (Lopes *et al.*, 2015), maize (Janzen *et al.*, 2022), *Brassicaceae* (Turner *et al.*, 2010), and sorghum (Lasky *et al.*, 2015).

Here, we characterized the genomic diversity of Colombian cassava landraces maintained in the living genebank collection at the Alliance Bioversity and the International Center for Tropical Agriculture (CIAT). This genebank contains thousands of cassava clones, many of which were sampled and preserved from locations in Colombia and the rest of South America (Ferguson *et al.*, 2019). These accessions originated from locations that represent diverse environments, especially across the dramatic elevation gradient of Colombia, and thus may also contain alleles responsible for local climate adaptation. As a first look at this genomic variation, we aim to provide foundational resources to discover loci informative for accelerated breeding.

The advent of genome engineering with technologies such as CRISPR-Cas9 presents an opportunity to match the pace of crop improvement with that of the challenges imposed by environmental change (Sedeek *et al.*, 2019). This technology has already proved useful in cassava by inducing resistance to cassava brown streak disease, without relying on transgenics (Gomez *et al.*, 2019). While mapping studies can find markers for certain traits, it remains difficult to reliably find causative mutations that can be the target of genome engineering (Haque *et al.*, 2018). To identify practical targets for gene editing, it is essential to consider the functional characterization of causative mutations. Loss-of-function (LoF) mutations, or gene knockouts, are classified as mutations predicted to disrupt or remove the function of a protein. These include premature stop codons, reading frame shifts, large deletions, or entire gene loss. Naturally occurring LoF mutations are important in the evolution of many plants, contributing to domestication, crop improvement, and adaptation to the environment (Orr, 2005; Monroe *et al.*, 2016, 2018, 2020; Murray, 2020; Xu & Guo, 2020; Klim *et al.*, 2024; Lee *et al.*, 2024). Identifying beneficial LoF alleles in crops can thus inform next-generation breeding efforts for rapid improvement. Indeed, gene knockouts have already led to major agricultural advances. The Green Revolution's dramatic yield increases were driven by LoF mutations in *GA20ox2* in rice and other crops (Spielmeyer *et al.*, 2002). In wheat, knockout of *GRAIN WIDTH2* improves rust resistance and grain weight (Sestili *et al.*, 2019; Liu *et al.*, 2024), while in cassava, knocking out *CYP79D1* and *CYP79D2*, which are key genes in cyanide synthesis, has reduced cyanide levels by 92% in bitter varieties (Jørgensen *et al.*, 2005).

However, detecting causal gene loss using statistical approaches such as Genome-Wide Association Study (GWAS) can be challenging as allelic heterogeneity diminishes the power of statistical testing (Monroe *et al.*, 2021). To address this, LoF burden tests aggregate multiple rare or independent LoF variants within a gene into a single score, enhancing the ability to detect associations between gene disruption and phenotypic variation (Povysil *et al.*, 2019; Spence *et al.*, 2024). In this study, we sequenced and analyzed 387 cassava landrace genomes, together with previous whole-genome sequencing of wild relatives and improved breeding lines. We tested for genetic-environment associations at multiple levels, including loss-of-function alleles. This approach identifies multiple candidate genes linked to climate adaptation and offers new directions for breeding strategies in cassava.

## Materials and Methods

### Landrace sampling

We selected 387 unique Cassava (*Manihot esculenta Crantz*) landraces based on their environmental parameters across Colombia to best represent the geographical diversity within the country. We also sequenced an additional 33 technical duplicates of some of the chosen accessions to verify and ensure sequencing and variant calling accuracy. Tissue cultures of those selected accessions were sent to the laboratory for DNA extraction from the CIAT genebank.

### DNA isolation and sequencing

Young leaf tissue from each accession was sampled to extract genomic DNA using the DNeasy plant mini kit (QIAGEN). DNA samples were sequenced with DNBSEQ™ at BGI America, yielding a total of *c.* $2.7 \times 10^{10}$ 150-bp paired-end reads and an average depth of coverage at 25.20× for each sample.

### Variant calling

We downloaded previously published Cassava whole-genome sequencing data from (Ramu *et al.*, 2017), (Kistler *et al.*, 2025) (excluding herbarium and archaeological samples), (Hu *et al.*, 2021), (Bredeson *et al.*, 2016), and (Wang *et al.*, 2014), and conducted *de novo* variant calling along with our sequencing data of Colombian cassava landraces and close relatives. Raw reads were trimmed using TRIMMOMATIC (v.0.39) to control read quality. The clean reads were mapped to the reference genome v8 (https://phytozome-next.jgi.doe.gov/info/Mesculenta_v8_1) using BWA-MEM (v.0.7.17-r1188). The mapped reads were then sorted and duplicates were removed by SAMTOOLS (v.1.13). The reads were realigned and the variants were called for each accession using DEEPVARIANT (v.1.5.0 (Poplin *et al.*, 2018)). Subsequently, gVCF merging and joint variant calling were performed using GLNEXUS (v.1.4.1 (Yun *et al.*, 2021)). To determine the effects of variants on the genome, we used SNPEFF (v.5.1d (Cingolani *et al.*, 2012)) to build a database from the reference genome (v8) and annotated the VCF.

### Data filtering and quality control

We employed a combination of novel and standard filtering methods to obtain high-confidence variants. Among our genotyped individuals, we obtained 33 technical replicates (i.e. 33 individuals whose DNA was sampled, extracted, and sequenced in duplicate). We used the sites with discordance between the two replicate genotype calls as a training set to train a gradient boost model (R package XGboost (Chen & Guestrin, 2016)) to detect erroneous variant sites. In brief, a set of variant site characteristics (percent missing genotypes, Quality score, minor allele frequency, number of alleles, and depth mean and variance) was used to predict erroneous sites. The percent missing genotypes was the most significant predictor of discordant sites between technical replicates. This was done in a leave-one-out method for each chromosome, and the predicted erroneous sites were removed from the available variants. The variants were further filtered using more standard methods using VCFTOOLS (v.0.1.16), sites with Quality value above 30 (--minQ 30), Genotype Quality value above 20 (--minGQ 20), and Minor Allele Frequency (MAF) above 0.05 (--maf 0.05) were kept for downstream analysis as rare variants (MAF < 0.05) are both more prone to spurious associations and less likely to contribute meaningfully to adaptive or beneficial variation, and thus were not the focus of our study. Variants were then phased using BEAGLE V5 (Browning & Browning, 2007) with the ne = 1000.

### Assessing genetic diversity and population structure

Principal component analysis (PCA) was performed on the filtered VCF using PLINK (v.1.9). Genetic assignment analysis was conducted using the filtered SNPs with the ADMIXTURE program (v.1.3.0 (Alexander *et al.*, 2009)). The sequencing data from Hu *et al.* (2021) had an average depth of *c.* 8.39×, while our newly sequenced Colombian landraces were generated at a depth of 25×. Accessions from Kistler *et al.* (2025) had an average depth of 21.31×, but ranged widely from 0.01× to 88.01×. These substantial differences in sequencing depth may significantly affect variant calling, data filtering, and downstream analyses. Therefore, accessions from Hu *et al.* (2021) and Kistler *et al.* (2025) were included only in the PCA or ADMIXTURE analysis to provide a global context and were not used in subsequent analyses of linkage disequilibrium (LD), nucleotide diversity ($\pi$), and fixation index ($F_{ST}$) comparisons.

To estimate nucleotide diversity unbiased by missing data in the VCF, all-site VCFs (VCF including invariant sites) of 18 chromosomes were produced using GATK HaplotypeCaller, GenomicsDBImport, and GenotypeGVCFs (v.4.5.0.0). $F_{ST}$ and $\pi$ between South American cassava, African cassava, and wild relatives were then calculated using Pixy (1.2.10.beta2 (Korunes & Samuk, 2021)) after basic filtering (--max-missing 0.8/--min-meanDP 20/--max-meanDP 500). LD was also calculated using PLINK (v.1.9) on a window size of 100 kb.

### Environmental variables and phenotypes

The geographic origins of the sampled cassava landraces were obtained from the CIAT genebank database, and then environmental variables were collected using the latitude and longitude of the selected samples (Supporting Information Table S1). Bioclim variables (Fick & Hijmans, 2017) were extracted using the R package 'RASTER' v.3.6.26, with a tile resolution of 2.5. Elevation values were extracted using the R package 'ELEVATR' v.0.99.0. Composite environmental phenotypes were generated using principal components analysis on all environmental variables corresponding to EnvPC1-5. Phenotypic data were also previously collected and reported by the CIAT genebank (Fukuda & Guevara, 1998; Ferguson *et al.*, 2020), provided as metadata for each accession (Table S1).

### Annotating loss-of-function mutations

We used mutation effect and protein structure prediction to evaluate LoF mutations and mutation impact prediction to infer potential LoF mutations. We used the primary transcripts annotated from the v.8 cassava genome assembly (Nordberg *et al.*, 2014) for all gene analyses. The program SnpEff (Cingolani *et al.*, 2012) was used to annotate mutation effects from our genotype variants. All mutations classified as 'HIGH' effect were putative LoF mutations, including mutations such as frameshift insertions and deleterious gain of a premature stop codon, loss of the start codon, and splice interruption. Although SnpEff evaluates variants affecting canonical splice donor and acceptor sites, alternative splicing isoforms were not incorporated into our LoF categorization. Post-translational modifications were also not considered, as a comprehensive dataset for cassava is currently unavailable.

To improve the accuracy of functional predictions based solely on sequence, we additionally incorporated protein structure prediction in our LoF pipeline using ESM-fold (Lin *et al.*, 2023). We used two values of amino acid structure prediction to measure their characteristics, the confidence or 'predicted local distance difference test' score (pLDDT) and the relative available surface area (rASA). These protein characteristics were measured across all proteins using a github pipeline https://github.com/em255/PopulationPDBStats (Long & Monroe, 2025). These protein structures often have 5′ and 3′ tails that are highly disordered and predicted with low confidence, where we see an enrichment for mutations. We annotated the 5′ tail of a protein as the region starting from the first amino acid until the first confidently folded amino acid occurs within an ordered region of the protein (pLDDT > 70 & rASA < 0.5). The 3′ tail was annotated in a similar manner, starting from the 3′ end. Using these disordered tail annotations, we chose to disregard certain LoF mutations that we estimated not to impact overall protein structure. All mutations in the disordered 3′ tail were ignored. Premature stop mutations and start codon losses were disregarded in the 5′ tail if another in-frame start codon was present within the region. While this method may discount some functional importance of these disordered regions, it allows for the enrichment of likely high-impact LoF mutations. Finally, using phased genotype information multiple LoF mutations were collapsed into a single functional status for each gene with either 0, 1 (heterozygous), or 2 functional copies of the gene. Importantly, rare alleles (MAF < 0.05) were included in this functional genotyping.

### Genome-wide association

Genome-wide association analyses were performed using TASSEL (Bradbury *et al.*, 2007). For genotype–environment and genotype–phenotype associations, we used the TASSEL General Linear Model framework with either: (1) the full SNP dataset; (2) the uncollapsed LoF genotype matrix; or (3) the collapsed LoF burden matrix as predictors. The first five genetic principal components were included as fixed covariates to account for population structure. For LoF-based analyses, we applied a minor LoF allele frequency threshold of 5%, analogous to a MAF filter, to exclude extremely rare variants.

To assess how controlling for different levels of genetic structure affects the detection of associations, we conducted a separate series of linear-model analyses in R. In this analysis, LoF status was regressed on environmental variables while varying the number of included genetic PCs. This procedure was used solely to evaluate the impact of population structure adjustment on association signal, not to identify significant associations for interpretation. Significant associations were determined using false discovery rate (FDR)–corrected p-values. Specifically, FDR correction was applied using the p.adjust() function in R, and signals with FDR-corrected $P < 0.05$ are considered significant.

### Redundancy analysis

We performed redundancy (RDA) and partial redundancy analysis (pRDA) following published tutorials (Capblancq & Forester, 2021). We applied this method to our genotype and environmental data to evaluate the genetic variance explained by environmental predictors. Environmental predictors were standardized to ensure comparability and then used to estimate the total variance explained of the genetic relationships. We then used pRDA to partition the variance explained by climate, neutral genetic structure, and geography. This involved creating models with population allele frequencies as the response variable and sets of bioclimate variables, genetic structure proxies, and geographical coordinates as explanatory variables. This approach allowed us to decompose the contributions of different factors to genetic variation and assess their independent and combined effects.

Loci with *P*-values below a stringent threshold were identified as candidate adaptive outliers. Results were visualized using an RDA biplot and a Manhattan plot, showing the correlation between genetic variation and environmental predictors, and highlighting significant outliers. This approach allowed us to identify loci potentially under selection due to environmental factors while accounting for population structure, providing insights into the genetic basis of local adaptation.

### Gene ontology

To investigate the biological significance of LoF mutations, we identified putative orthologs of cassava genes in *Arabidopsis thaliana* (TAIR10) using BLASTP, retaining matches with an *E*-value < 1e-10 and bit score > 100. For each cassava gene containing at least one predicted LoF allele, we mapped its Arabidopsis ortholog and used the associated Gene Ontology (GO) annotations to perform enrichment analysis, focusing on the Biological Process (BP) category. Enrichment was conducted using the TopGO package in R, applying the 'weight01' algorithm and Fisher's exact test. Only GO terms that remained significant after Bonferroni correction for multiple testing (adjusted $P < 0.05$) were considered enriched. To assess lineage-specific patterns, we performed the analysis separately for wild relatives, traditional landraces, and modern breeding lines.

### Testing for selection and inbreeding effects on LoF accumulation

To evaluate whether deleterious LoF mutations are purged by negative selection under inbreeding, we tested for an association between LoF burden and genomic inbreeding coefficients in cassava landraces. In this study, we use the term inbreeding to refer to realized inbreeding coefficients estimated from genome-wide genotype data, rather than deliberate selfing or controlled inbreeding practices. Although cassava is predominantly outcrossing, realized inbreeding can arise through clonal propagation combined with occasional sexual reproduction, bottlenecks during domestication and local lineage maintenance, or the use of related parents in modern breeding programs. For each individual landrace, we quantified the total number of LoF events, defined as the number of genes with at least one high-confidence LoF allele. In parallel, we estimated the inbreeding coefficient (F) using VCFtools, based on genome-wide biallelic SNP variation. We then assessed the relationship between inbreeding and LoF accumulation by ranking landraces according to their F values and their total number of LoF genes. A negative correlation between ranked F and LoF counts would support the hypothesis that inbreeding facilitates purging of deleterious variants through increased homozygosity and stronger exposure to selection.

To investigate whether certain LoF mutations may be beneficial in specific functional contexts, we repeated this analysis while restricting LoF counts to genes annotated with specific GO BP categories. This allowed us to identify cases where LoF accumulation is positively associated with inbreeding, which may indicate adaptive or tolerated LoF in particular functional pathways.

## Results

### Genetic diversity and population structure

To investigate the genetic variation of Cassava landraces in Colombia, we sampled and sequenced 387 unique cassava accessions by whole-genome sequencing (WGS). The Colombian cassava landraces in our dataset span a wide climatic range, from lowland tropical regions with high temperatures and precipitation to cooler mid-elevation zones with more moderate rainfall. This environmental heterogeneity captures much of the climatic variation across cassava-growing regions of Colombia (Table S1). On average, 25.20× sequencing coverage was generated for each accession. We then combined our data with previously published cassava WGS data, including 174 accessions from (Ramu *et al.*, 2017), 158 accessions from (Kistler *et al.*, 2025), 338 accessions from (Hu *et al.*, 2021), 59 accessions from (Bredeson *et al.*, 2016) and 2 accessions from (Wang *et al.*, 2014), and conducted *de novo* variant calling (Table S1). To ensure consistency and minimize batch effects, all samples were processed using the same quality control, filtering, and variant calling pipeline (see the Materials and Methods section).

To dissect the population structure underlying cassava samples, we performed PCA of selected high-confidence variants (see the Materials and Methods section), which reveals substantial, previously untapped genetic diversity among Colombian cassava landraces (Fig. 1). In our diversity panel of *M. esculenta* and using the selected MAF threshold, we identified a total of 29.23% genetic variance captured in the first two principal components (PCs). South American accessions displayed greater genetic differentiation compared to cultivated cassava from Africa and Asia, although some overlap was observed, suggesting a genetic bottleneck associated with the introduction of cassava to these regions (Fig. 1b). In Colombia, elevation is a major determinant of climate and notably, PC1 of genotypic diversity shows a strong correlation with elevation and temperature (Figs 2b, S1). The genetic differentiation along PC1 may reflect local adaptation or environmental selection associated with climate. Highland and

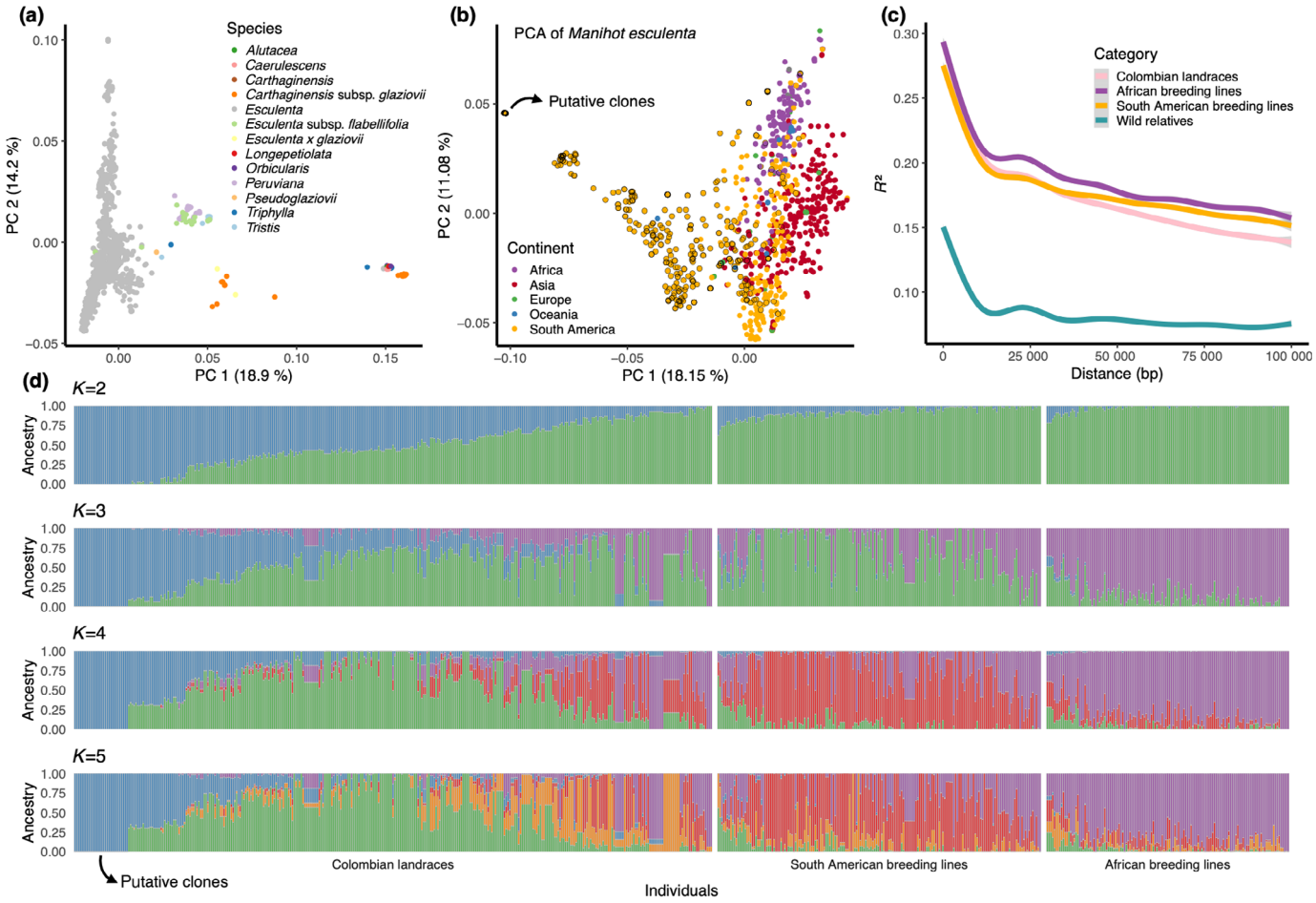


**Fig. 1** Population structure, genetic diversity, and linkage disequilibrium decay of cassava accessions and close relatives. (a) Principal component analysis (PCA) of the first two dimensions of genotype data from all accessions, including close relatives. Data points are colored by species. (b) PCA of the first two dimensions of genotype data from *Manihot esculenta* accessions, with data points colored by geographical location (continent). Colombian landrace accessions sequenced in this study are highlighted in black outline. (c) Linkage disequilibrium (LD) decay in Colombian landraces, African breeding lines, South American breeding lines, and wild relatives. (d) ADMIXTURE plots of Colombian landraces, African breeding lines, and South American breeding lines. Cassava accessions progressively separate into minor subpopulations as K increases. Accessions are arranged in the order of PC1.

lowland populations harbor distinct allelic compositions, possibly tied to important traits such as drought and heat tolerance along with demographic history. Later, we test the efficacy of genotype–environment analyses to disentangle the joint effects of demography and selection on alleles associated with the climate of origin.

LD decay patterns are generally similar across populations, with LD decreasing sharply within the first *c.* 20 kb in all groups (Fig. 1c). Although differences in sequencing methods and depths among previous studies may introduce bias, Colombian landraces consistently exhibit lower overall LD across distances compared to African and South American breeding lines (Fig. 1c), indicating more abundant historical recombination events in the landraces or greater underlying genetic diversity. As expected, LD is lowest in the wild relatives, consistent with their high genetic diversity. Genetic assignment analysis using ADMIXTURE reveals a more complex genetic structure in Colombian landraces. At $K = 5$, these landraces comprise five distinct ancestral components, each predominant in different subsets of accessions. By contrast, African breeding lines are largely dominated by a single ancestral group, which again implies a strong bottleneck effect and their potential demographic origin (Fig. 1d). Because the ADMIXTURE cross-validation error decreased gradually without identifying a clear optimal $K$, we also present results for $K = 6$–10 (Fig. S2a). These higher $K$ values largely partition existing clusters and do not alter the major ancestry patterns (Fig. S2b).

The identity-by-state (IBS) matrix derived from the VCF reveals that both Asian and African breeding lines have a higher genetic similarity to lowland (< 500 Meters Above Sea Level, MASL) South American accessions than to highland (> 500 MASL) accessions (Fig. S1a). These observations suggest that the African and Asian breeding accessions likely originated from genotypes with closer genetic relatedness to modern landraces found in the lowland regions of South America. The IBS analysis between domesticated cassava and other species or subspecies indicates a close relationship with *flabellifolia*, *peruviana*, and *tristis* (Fig. S1b), consistent with findings from the previous phylogenetic study (Simon *et al.*, 2022).

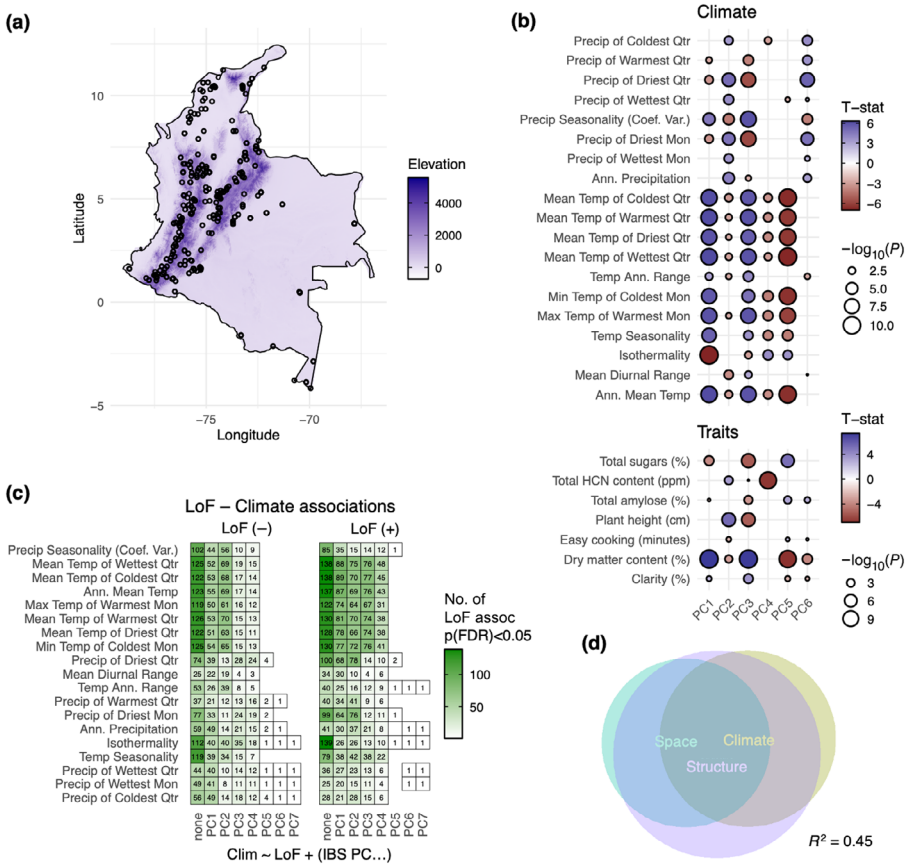


**Fig. 2** Collinearity of environmental adaptation and population structure in Colombian cassava landraces. (a) Geographic locations of Colombian landrace samples analyzed in this study. (b) Correlation of climatic variables and agronomic traits with the top six genetic principal components (PCs). Circle size indicates the significance level of the correlation, while color represents the direction (positive or negative). (c) The number of significant negative and positive associations between loss-of-function (LoF) variants and climate variables decreases as more PCs are included to control for population structure. Numbers in each box indicate the count of significant associations after false discovery rate (FDR) correction. (d) Partitioning of genetic variance through redundancy analysis (RDA), quantifying the contributions of population structure, environmental variation (climate), and geographic location (space). Together, these factors explain 45% of the observed genetic variation.

Interestingly, a small subset of Colombian landraces with the lowest PC1 values cluster tightly together and are clearly separated from the rest of the samples in the principal component space (Fig. 1b). These accessions correspond to individuals that are overwhelmingly composed of a single ADMIXTURE group and show little evidence of genetic mixing with other populations (Fig. 1d). Their pairwise IBS scores are exceptionally high, ranging from 0.9977 to 0.9992, indicating that they are nearly genetically identical. Notably, these accessions are not technical replicates, but rather distinct landrace accessions collected from different geographic regions (Fig. S4). This high degree of genetic similarity across geographically diverse samples could reflect the spread of a clonally propagated or strongly bottlenecked lineage, potentially due to human-mediated dispersal. Such findings also point to the presence of genetically redundant accessions in the genebank, in line with observations from previous studies (Ferguson *et al.*, 2019; Carvajal-Yepes *et al.*, 2023).

Additionally, we carried out an extensive analysis of genetic diversity and differentiation among populations. The wild relatives population displays the highest genetic diversity, with an average genome-wide $\pi_{WR}$ of $1.90 \times 10^{-2}$. The genetic diversity of South American cassava ($\pi_{SA} = 8.91 \times 10^{-3}$), which mostly consists of Colombian landraces, is higher compared to the African cassava population ($\pi_A = 8.22 \times 10^{-3}$), mostly comprising local breeding lines. This result further implies bottleneck effects throughout cassava domestication history. We also observed several genomic windows where $\pi_{WR}/\pi_{SA}$ or $\pi_{WR}/\pi_A$ is significantly higher (Z-score > 3), and those regions are likely subject to selection during domestication (Fig. S5). Similar $\pi_{landrace}/\pi_{improved}$ or $\pi_{wild}/\pi_{landrace}$ peaks have been observed in maize membrane trafficking genes that were selected during maize domestication and improvement (Zheng *et al.*, 2022). The genome-wide average $F_{ST}$ between South American cassava and African cassava is 0.0471, indicating relatively low genetic differentiation between these two cultivated or pre-breeding populations. By contrast, the $F_{ST}$ between South American cassava and wild relatives is higher at 0.134, suggesting a greater genetic divergence. Similarly, the $F_{ST}$ between African cassava and wild relatives is 0.125, also reflecting a significant degree of genetic differentiation.

## Genetic differentiation of Colombian landraces across climate

While the large collection of cassava landraces represents vast untapped genetic diversity, they are also a key resource representing environmental adaptation. These landrace cassava clones originate from locations across vast environments in Colombia (Fig. 2a). Notably, the field sites from which these individuals originate correspond to varied climates. The largest impact of the climatic variation among individuals is driven by elevation distribution around the Andes mountains running through Colombia (Figs 2a, S1). The lowland regions experience higher temperatures and lower precipitation than the highland areas.

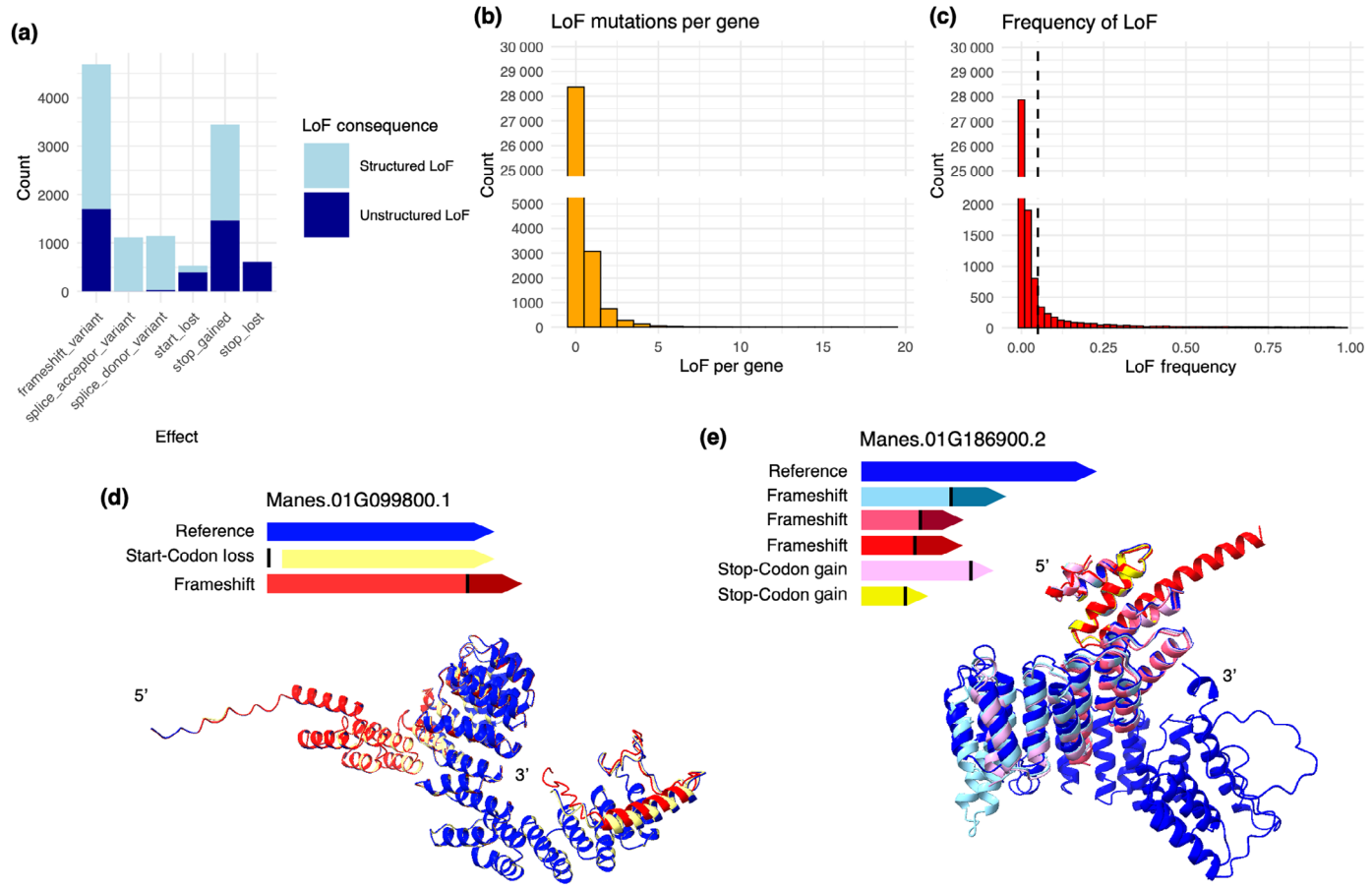


**Fig. 3** Types, distribution, and structural impact of gene loss-of-function (LoF) variants in cassava. (a) Histogram of gene LoF variant types predicted by SnpEff in this study. Frameshift mutations and premature stop codons account for the majority of predicted LoF events. The functional impact of LoF variants is further assessed through protein structure predictions. 'Structured LoF' refers to variants likely occurring within structured protein regions, thus more likely representing true LoF events. (b) Histogram showing the number of independent LoF variants identified per gene. (c) Histogram of LoF variant frequencies per gene across the studied population. The dashed line marks the 5% minor allele frequency (MAF) threshold commonly applied in association studies. (d, e) Illustrative diagrams of LoF variants for two example genes, comparing their predicted protein structures relative to the reference gene model. Black bars indicate the locations of LoF mutations within the gene body. Protein structure models depict the impact of LoF variants, highlighting deviations from the reference structures.

The environments in which these clones are cultivated may reflect adaptive selection histories, as these landraces have been grown in local conditions for hundreds of years. We observed strong correlations between environmental variables and the top genetic PCs. Notably, PC1 and PC3 were positively correlated with temperature, while PC2 showed a positive correlation with precipitation (Fig. 2b). Redundancy analysis (RDA), which assesses the proportion of genotypic variation explained by environmental, spatial, and population structure variables, revealed that these factors together account for a substantial portion of genetic variation ($R^2$ = 0.45). While the contribution of population structure is expected, partial RDA indicates that climate alone explains a significant fraction of this variance (Fig. 2d). The shared variance among climate, space, and population structure underscores the challenge of disentangling the genetic basis of climate adaptation, and can be particularly problematic for GWAS. We further tested for associations between LoF alleles and climate variables, finding that the number of significant associations declined as more genetic PCs were included in the model. When the top 7 PCs were controlled for, few associations remained (Fig. 2c). Nonetheless, these results suggest that strong genetic correlations with environmental variables persist.

## Gene LoF in cassava

In an effort to enrich for impactful genetic variants that may explain climate adaptation, we evaluated the distribution of gene LoF alleles across the cassava landraces. LoF mutations represent a subset of genetic variants that are expected to have large functional effects on gene activity. In total, we found over 11 k variants segregating across the landraces categorized as causing a large disruptive effect by SnpEff (Cingolani *et al.*, 2012). We used protein structure to reduce the false positive rate of classifying LoF, by disregarding mutations that are less likely to impact structured (folded regions with low available surface area), functional regions of the protein (Fig. 3a,d,e). The relative

importance of peripheral, disordered regions of proteins is a subject of current scientific inquiry, but we are choosing to disregard them for this analysis. The LoF mutations are confined to *c.* 6.5 k genes (*c.* 20% of the proteome) with many genes containing multiple segregating variants (Fig. 3b). We collapsed the LoF classification across all genes to a singular functional classification for each gene (Fig. 3d,e) and evaluated the frequency of LoF (Fig. 3c). After filtering out rare, low-frequency functional variation (MAF > 5%), over 2k genes remained.

Cassava is a highly heterozygous crop that experiences severe inbreeding depression (de Freitas *et al.*, 2016; Ramu *et al.*, 2017; Long *et al.*, 2025); therefore, when homozygosity increases, recessive deleterious mutations become exposed and can have pronounced negative effects on fitness. To investigate the relationship between inbreeding and deleterious variation, we calculated the inbreeding coefficient (F) for each individual in our dataset using genome-wide genotype data. We observed a significant negative correlation between the number of LoF alleles per individual and their inbreeding coefficient (Spearman's $\rho = -0.46$, Fig. S6a–c). This suggests that inbred individuals tend to carry fewer LoF alleles, consistent with purging of deleterious mutations through inbreeding.

To investigate the functional role of LoF variants in the evolution and adaptation of cassava, we performed GO analysis on genes retaining LoF alleles (Table S2). Our results reveal that LoF-tolerant genes in cassava are significantly enriched for immune response functions (Fig. 4), consistent with previous findings in Arabidopsis (Zhao *et al.*, 2025). These genes include plausible candidates that may be leveraged to engineer disease resistance in cassava through gene editing. For example, Manes.16G055900 is homologous to Arabidopsis AT2G35110 (*HEM1*), a global translational regulator of plant immunity. Loss of *HEM1* leads to exaggerated cell death that restricts bacterial growth and enhances immunity (Zhou *et al.*, 2023). Manes.08G172100 shares homology with AT1G28380 (*NSL1*) and AT1G29690 (*NSL2*), which participate in the salicylic acid (SA)-mediated programmed cell death pathway in plant immunity (Morita-Yamamuro *et al.*, 2005; Noutoshi *et al.*, 2006; Murakoshi *et al.*, 2024). Mutants with *NSL1* knockout constitutively activate defense responses (Noutoshi *et al.*, 2006). Manes.03G015900, homologous to AT1G51940 (*LYK3*), is involved in suppressing immune responses in the absence of pathogens or following abscisic acid treatment. *LYK3* LoF mutants exhibit increased resistance to *Botrytis cinerea* and *Pectobacterium carotovorum* (Paparella *et al.*, 2014). Manes.08G125011, related to AT5G61900 (*BON1*) and AT1G08860 (*BON3*) in Arabidopsis, may also function in plant immunity, as *BON1* and *BON3* negatively regulate several resistance (R)-like genes (Li *et al.*, 2009), and their knockout in both Arabidopsis and rice has been shown to enhance resistance to bacterial and fungal pathogens (Li *et al.*, 2009; Yang *et al.*, 2017).

In addition to the enrichment for immune response, we observed enrichment for loss of function in genes involved in sideretin biosynthesis and coumarin metabolic process in cultivated, but not wild, cassava (Fig. 4). Coumarins are a class of secondary metabolites that play a central role in iron (Fe)

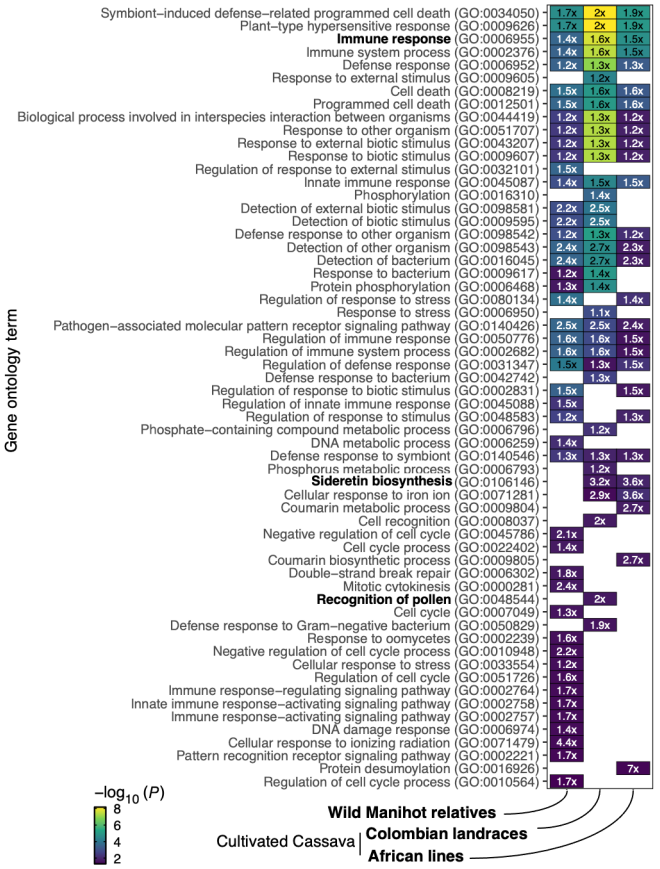


**Fig. 4** Gene ontology (GO) enrichment analysis of loss-of-function (LoF)-tolerant genes in cultivated cassava and wild relatives. Significantly enriched GO terms for LoF-tolerant genes are displayed, ranked by significance. The numbers within each box represent the fold enrichment of the corresponding GO term. Notable biological processes are highlighted in bold for emphasis.

solubilization and uptake, particularly under alkaline soil conditions (Robe *et al.*, 2021). Among them, compounds such as fraxetin and sideretin have been shown to facilitate Fe acquisition, with their efficacy varying depending on soil pH (Rajniak *et al.*, 2018; Paffrath *et al.*, 2024). Beyond their physiological functions, coumarins are also significant for postharvest quality. Scopoletin, the chemical precursor to sideretin, and its glucoside, scopolin, have both been implicated in postharvest physiological deterioration in cassava (Buschmann, 2000; Bayoumi *et al.*, 2008, 2010). Specifically, scopoletin oxidation leads to tissue discoloration, rendering roots unpalatable and reducing their marketability (Liu *et al.*, 2017). We found an enrichment of LoF alleles in genes associated with sideretin biosynthesis, including homologs of *CYP82C2*, *CYP82C3*, and *CYP82C4* from *Arabidopsis thaliana*, which catalyzes the hydroxylation of fraxetin to produce sideretin. Additionally, Manes.07G034800 was identified as a homolog of AT1G55290 (*F6′H2*), AT3G13610 (*F6′H1*), and AT3G12900 (*S8H*). Of these, *F6′H1* is involved in the biosynthesis of scopoletin from feruloyl-CoA, while *S8H* hydroxylates scopoletin to form fraxetin (Rajniak *et al.*, 2018). Whether the accumulation of LoF alleles in coumarin-related

genes reflects selection against undesirable postharvest discoloration or adaptation to varying soil conditions remains unresolved and motivates further investigation.

Moreover, we found that LoF-tolerant genes are enriched for functions related to pollen recognition (Fig. 4). This result is consistent with previous findings that domestication and the shift to clonal propagation in cassava have relaxed selection on sexual reproduction, allowing mutations to accumulate in pollen-associated genes (Long *et al.*, 2025). Interestingly, although there is a genome-wide negative correlation between the number of LoF alleles per individual and their inbreeding coefficient, this correlation is markedly less negative or even positive for genes within significantly enriched GO categories (Fig. S6d). The persistence of LoF alleles in these genes, despite inbreeding, suggests they may be subject to positive selection.

### Genome-environment and genome-phenoptype associations

Toward discovering specific genetic variants responsible for climate adaptation, we performed genotype–environment analyses in the landraces. We used three different classifications of genetic variants for environmental association including all variants, variants predicted to result in gene LoF (Dataset S1), and the consolidated gene LoF burden (Dataset S2).

All 19 available WorldClim bioclimatic variables were tested (Table S3), while we show the results to two primary dimensions of environmental variation: temperature and precipitation (Fig. S7). Genetic associations with annual temperature, an environmental variable of major interest with impending climate change, found a significant locus on chromosome 6 among all variants and LoF variants. This locus, however, was not retained when considering the LoF burden classification that incorporates predicted effects on protein structure. Associations with annual precipitation showed multiple significant variants across the cassava genome. A handful of these variants corresponded to LoF variants and the LoF burden of two specific genes when controlling for the impact on protein structure. These two genes, Manes.09G029200 and Manes.10G053700, show opposite effects in their correlation with high and low precipitation, respectively. In addition to climatic variables, we also tested five additional phenotypes that were previously collected (Fukuda & Guevara, 1998; Ferguson *et al.*, 2020). We found a significant association with high sugar content corresponding to the LoF burden of a gene on chromosome 6. This gene is annotated as a glycosyltransferase, suggesting that its LoF may impact the ability to transport sugars outside of the roots. While high sugar content is generally not a valued trait for cassava production, it is an interesting proof of concept for the utility of LoF association analysis.

We also used redundancy analysis to look for genes responsible for climate adaptation. The RDA method uses collapsed variable space to assess which climate variables explain mutations in the genome, somewhat reversed from a typical genome-wide association analysis. When looking at what genes are most highly explained by climate variables, one gene, Manes.01G186900.2, proved significant (Fig. S8). This gene was among the most significantly associated genes with annual temperature, but did not pass any significance thresholds in the GWAS analysis. The LoF in this gene is also derived from multiple segregating mutations (Fig. 3e), making it a particularly compelling candidate for further investigation. Here, we also emphasize that the interpretation of those results is speculative and based on available functional and association data, though the functions of those genes warrant experimental validation.

## Discussion

In this study, we characterized previously untapped genetic diversity in Colombian cassava landraces. Our analysis of population structure and genetic diversity aligns with the known domestication history of cassava, which originated in South America and was later introduced into Africa and Asia. PCA reveals that African and Asian breeding lines partially overlap with South American accessions, suggesting a genetic connection to the South American gene pool underlying modern cassava cultivars. Modern breeding programs in Africa and Asia were founded largely on introductions from South America, with additional material exchanged among breeding centers over subsequent decades (Vidhi, 2016; Ferguson *et al.*, 2019; Ceballos *et al.*, 2021). We also found that lowland South American landraces are genetically more similar to breeding lines from Africa and Asia. Given that lowland South America is substantially warmer than the highlands, this pre-adaptation to high temperatures may help explain the success of these accessions in African breeding programs. The nucleotide diversity of African cassava is $8.22 \times 10^{-3}$, which is higher than the previously reported nucleotide diversity ($\pi = 3.6 \times 10^{-3}$) by (Ramu *et al.* (2017)). This discrepancy may reflect methodological differences, as nucleotide diversity in the previous study was estimated using VCFtools, which may have limitations under certain conditions (Korunes & Samuk, 2021). Although direct comparisons across studies must be interpreted cautiously due to differences in sequencing platforms, variant calling pipelines, and filtering criteria, the diversity we observe is broadly consistent with values reported for other outcrossing crop species, such as maize landraces ($\pi \approx 6\text{–}10 \times 10^{-3}$) (Tittes *et al.*, 2023). Overall, our results reinforce the view that cassava landraces harbor substantial genetic diversity, a resource that breeding programs can harness for improving traits such as disease resistance, stress tolerance, and nutritional quality.

By sampling the genetic diversity of cassava landraces across the environmental landscapes of Colombia, we aimed to capture potential adaptation to their local environment. This hypothesis that the location in which these landraces were sampled may correspond to environmental adaptation relies upon the assumption that farmers retained those varieties that performed best and were well adapted to their respective regions (Janzen *et al.*, 2022). The correlation of population structure derived from genetic relationships and the climatic variables represents the underlying double-edged sword of this research approach. In contrast to the findings by Kistler *et al.* (2025), which found minimal geographic

population structure in cassava landraces when examined at large spatial scales across the Americas attributed to clonal propagation and human-mediated dispersal, our data reveal pronounced population structure among Colombian landraces. We also found that population structure showed high correlation to many environmental variables, which is consistent with both adaptation and demographic history. Unfortunately, such structure makes it difficult to identify specific adaptive causative mutations. Controlling for population structure in our associations attempts to reduce false positive associations due to structure-environmental correlations; meanwhile, we acknowledge that this is also likely reducing our ability to capture true positive relationships (Lasky *et al.*, 2015; Lotterhos, 2023).

In addition to conventional genotype–environment association using SNP data, we developed a function-based GWAS approach that associates LoF variants and LoF burdens. By collapsing independent LoF alleles into a single allele state, we aim to overcome allelic heterogeneity, which has limited previous GWAS efforts. Moreover, traditional GWAS typically highlights genomic regions rather than pinpointing specific causative mutations, often leaving results difficult to translate into practical applications like gene editing. By contrast, our LoF association tests directly link phenotypes to gene function, enabling the identification of actionable gene-level targets. While population structure complicates the detection of causal variants, we propose that relaxing population structure controls can be justified to uncover candidate genes because of their testability via gene editing, which is highly feasible using the CRISPR-Cas9 system, with multiple studies reporting mutation efficiencies of *c.* 90–100% in successfully transformed plants (Gomez *et al.*, 2019, 2022; Wang *et al.*, 2022; Li *et al.*, 2025; Xiao *et al.*, 2025).

While LoF association tests offer valuable insights, some cautions are warranted. First, relying on a single reference genome may bias LoF detection, missing variation absent from the reference or structural variants (Zhou *et al.*, 2022). Incorporating multiple references or pangenomes can mitigate this. Second, predicted LoF variants do not always cause phenotypic loss due to alternative transcription and/or splicing, compensatory mutations, or genetic redundancy (Singer-Berk *et al.*, 2023). We applied protein structure prediction to reduce false positives, but factors like heterozygosity, dominance, post-translational modifications, and pseudogenes in cassava still complicate interpretation (Rausell *et al.*, 2020). Lastly, LoF variants may be in LD with causal variants, requiring careful analysis and functional validation to confirm causality.

Although our LoF analysis has yielded promising insights into the genetic basis of trait variation and climate adaptation in cassava, further work is needed to fully translate these findings into functional understanding and agronomic application. Experimental validation of candidate LoF genes, such as through CRISPR-based gene editing, will be essential to confirm their roles. Additionally, obtaining agricultural phenotypes in the field under diverse environmental conditions can help bridge the gap between genotype and environment, advancing our ability to harness LoF variation for crop improvement.

## Acknowledgements

We thank members of Monroe Lab for valuable discussions of this work. This work was supported by FFAR grant ICRC20-0000000014. Research was conducted at the University of California – Davis, which is located on land that was the home of the Patwin people for thousands of years.

## Competing interests

None declared.

## Author contributions

KZ and EL contributed to analysis, writing original draft, data curation, and visualization. GM contributed via supervision, project administration, analysis, reviewing/editing, and visualization. PC and FS contributed the methodology and data curation. PC, GM and EM all contributed to the funding acquisition for this project.

## ORCID

Paul Chavarriaga https://orcid.org/0000-0001-7579-3250
Evan Long https://orcid.org/0000-0001-5866-4158
Erwan Monier https://orcid.org/0000-0001-5533-6570
Grey Monroe https://orcid.org/0000-0002-4025-5572
Francisco Sanchez https://orcid.org/0000-0001-5716-8945
Kehan Zhao https://orcid.org/0009-0004-1431-637X

## Data availability

Figures, supplemental data, and code for this research are located at: https://github.com/KehanZhao/ColombianCassavaLandraces. All raw genomic sequencing data generated in this study are publicly available through the NCBI Sequence Read Archive (SRA) under Bioproject accession number PRJNA1228154. Sample metadata for both newly generated and previously published data are provided in Table S1. For access to additional post-processed data formats, please contact the corresponding authors. Loss-of-Function variants are available in Datasets S1 and S2.

## References

**Alexander DH, Novembre J, Lange K. 2009.** Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* **19**: 1655–1664.

**Bayoumi SAL, Rowan MG, Beeching JR, Blagbrough IS. 2008.** Investigation of biosynthetic pathways to hydroxycoumarins during post-harvest physiological deterioration in Cassava roots by using stable isotope labelling. *Chembiochem: A European Journal of Chemical Biology* **9**: 3013–3022.

**Bayoumi SAL, Rowan MG, Beeching JR, Blagbrough IS. 2010.** Constituents and secondary metabolite natural products in fresh and deteriorated cassava roots. *Phytochemistry* **71**: 598–604.

**Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007.** TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635.

**Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, Grimwood J, Schmutz J, Rabbi IY, Egesi C *et al.* 2016.** Sequencing wild and

cultivated cassava and related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology* **34**: 562–570.

**Browning SR, Browning BL. 2007.** Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics* **81**: 1084–1097.

**Buschmann H. 2000.** Accumulation of hydroxycoumarins during post-harvest deterioration of tuberous roots of cassava (*Manihot esculenta* crantz). *Annals of Botany* **86**: 1153–1160.

**Capblancq T, Forester BR. 2021.** Redundancy analysis: A Swiss Army Knife for landscape genomics. *Methods in Ecology and Evolution* **12**: 2298–2309.

**Carvajal-Yepes M, Ospina JA, Aranzales E, Velez-Tobon M, Correa Abondano M, Manrique-Carpintero NC, Wenzl P. 2023.** Identifying genetically redundant accessions in the world's largest cassava collection. *Frontiers in Plant Science* **14**: 1338377.

**Ceballos H, Hershey C, Iglesias C, Zhang X. 2021.** Fifty years of a public cassava breeding program: evolution of breeding objectives, methods, and decision-making processes. *Theoretical and Applied Genetics* **134**: 2335–2353.

**Chen T, Guestrin C. 2016.** XGBoost: a scalable tree boosting system. In: *KDD '16. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery, 785–794.

**Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012.** A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**: 80–92.

**Ferguson M, Fukuda WMG, Guevara CL, Kawuki R. 2020.** Selected morphological and agronomic descriptors for the characterization of cassava.

**Ferguson ME, Shah T, Kulakow P, Ceballos H. 2019.** A global overview of cassava genetic diversity. *PLoS ONE* **14**: e0224763.

**Fick SE, Hijmans RJ. 2017.** WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology: A Journal of the Royal Meteorological Society* **37**: 4302–4315.

**Food and Agriculture Organization. 2019.** *The state of food security and nutrition in the World 2018: building climate resilience for food security and nutrition.* Rome, Italy: Food & Agriculture Organization of the United Nations (FAO).

**de Freitas JPX, da Silva Santos V, de Oliveira EJ. 2016.** Inbreeding depression in cassava for productive traits. *Euphytica* **209**: 137–145.

**Fukuda WMG, Guevara CL. 1998.** *Descritores morfológicos e agronômicos para a caracterização de mandioca (*Manihot esculenta *Crantz)*. Cruz das Almas: Embrapa Mandioca e Fruticultura, 1998.

**Gomez MA, Berkoff KC, Gill BK, Iavarone AT, Lieberman SE, Ma JM, Schultink A, Karavolias NG, Wyman SK, Chauhan RD *et al.* 2022.** CRISPR-Cas9-mediated knockout of CYP79D1 and CYP79D2 in cassava attenuates toxic cyanogen production. *Frontiers in Plant Science* **13**: 1079254.

**Gomez MA, Lin ZD, Moll T, Chauhan RD, Hayden L, Renninger K, Beyene G, Taylor NJ, Carrington JC, Staskawicz BJ *et al.* 2019.** Simultaneous CRISPR/Cas9-mediated editing of cassava eIF4E isoforms nCBP-1 and nCBP-2 reduces cassava brown streak disease symptom severity and incidence. *Plant Biotechnology Journal* **17**: 421–434.

**Haque E, Taniguchi H, Hassan MM, Bhowmik P, Karim MR, Śmiech M, Zhao K, Rahman M, Islam T. 2018.** Application of CRISPR/Cas9 genome editing technology for the improvement of crops cultivated in tropical climates: recent progress, prospects, and challenges. *Frontiers in Plant Science* **9**: 617.

**Hu W, Ji C, Liang Z, Ye J, Ou W, Ding Z, Zhou G, Tie W, Yan Y, Yang J *et al.* 2021.** Resequencing of 388 cassava accessions identifies valuable loci and selection for variation in heterozygosity. *Genome Biology* **22**: 316.

**Janzen GM, Aguilar-Rangel MR, Cíntora-Martínez C, Blöcher-Juárez KA, González-Segovia E, Studer AJ, Runcie DE, Flint-Garcia SA, Rellán-Álvarez R, Sawers RJH *et al.* 2022.** Demonstration of local adaptation in maize landraces by reciprocal transplantation. *Evolutionary Applications* **15**: 817–837.

**Jørgensen K, Bak S, Busk PK, Sørensen C, Olsen CE, Puonti-Kaerlas J, Møller BL. 2005.** Cassava plants with a depleted cyanogenic glucoside content in leaves and tubers. Distribution of cyanogenic glucosides, their site of synthesis and transport, and blockage of the biosynthesis by RNA interference technology. *Plant Physiology* **139**: 363–374.

**Kashyap A, Garg P, Tanwar K, Sharma J, Gupta NC, Ha PTT, Bhattacharya RC, Mason AS, Rao M. 2022.** Strategies for utilization of crop wild relatives in plant breeding programs. *Theoretical and Applied Genetics* **135**: 4151–4167.

**Kistler L, de Oliveira Freitas F, Gutaker RM, Maezumi SY, Ramos-Madrigal J, Simon MF, Mendoza FJM, Drovetski SV, Loiselle H, de Oliveira EJ *et al.* 2025.** Historic manioc genomes illuminate maintenance of diversity under long-lived clonal cultivation. *Science* **387**: eadq0018.

**Klim J, Zielenkiewicz U, Kaczanowski S. 2024.** Loss-of-function mutations are main drivers of adaptations during short-term evolution. *Scientific Reports* **14**: 7128.

**Korunes KL, Samuk K. 2021.** pixy: unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Molecular Ecology Resources* **21**: 1359–1368.

**Lasky JR, Upadhyaya HD, Ramu P, Deshpande S, Hash CT, Bonnette J, Juenger TE, Hyma K, Acharya C, Mitchell SE *et al.* 2015.** Genome-environment associations in sorghum landraces predict adaptive traits. *Science Advances* **1**: e1400218.

**Lee G, Sanderson BJ, Ellis TJ, Dilkes BP, McKay JK, Ågren J, Oakley CG. 2024.** A large-effect fitness trade-off across environments is explained by a single mutation affecting cold acclimation. *Proceedings of the National Academy of Sciences, USA* **121**: e2317461121.

**Li Y, Bao R, Li M, Zeng C, Yang H, Yao Y, Li Y, Wang W, Chen X. 2025.** Improving gene editing of CRISPR/Cas9 using the callus-specific promoter pYCE1 in cassava (*Manihot esculenta* Crantz). *Frontiers in Plant Science* **16**: 1600438.

**Li Y, Pennington BO, Hua J. 2009.** Multiple R-like genes are negatively regulated by BON1 and BON3 in arabidopsis. *Molecular Plant–Microbe Interactions* **22**: 840–848.

**Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y *et al.* 2023.** Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**: 1123–1130.

**Liu S, Liu H, Guo M, Pan Y, Hao C, Hou J, Yan L, Zhang X, Chen X, Li T. 2024.** Knockout of GRAIN WIDTH2 has a dual effect on enhancing leaf rust resistance and increasing grain weight in wheat. *Plant Biotechnology Journal* **22**: 2007–2009.

**Liu S, Zainuddin IM, Vanderschuren H, Doughty J, Beeching JR. 2017.** RNAi inhibition of feruloyl CoA 6′-hydroxylase reduces scopoletin biosynthesis and post-harvest physiological deterioration in cassava (*Manihot esculenta* Crantz) storage roots. *Plant Molecular Biology* **94**: 185–195.

**Long E, Monroe G. 2025.** Protein structure and selection pressure in plants: using mutation to understand the functional importance of protein structure. Research Square.

**Long EM, Stitzer MC, Monier B, Schulz AJ, Romay MC, Robbins KR, Buckler ES. 2025.** Evolutionary signatures of the erosion of sexual reproduction genes in domesticated cassava (*Manihot esculenta*). *G3* **15**: jkae282.

**Lopes MS, El-Basyoni I, Baenziger PS, Singh S, Royo C, Ozbek K, Aktas H, Ozer E, Ozdemir F, Manickavelu A *et al.* 2015.** Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *Journal of Experimental Botany* **66**: 3477–3486.

**Lotterhos KE. 2023.** The paradox of adaptive trait clines with nonclinal patterns in the underlying genes. *Proceedings of the National Academy of Sciences, USA* **120**: e2220313120.

**Monroe JG, Arciniegas JP, Moreno JL, Sánchez F, Sierra S, Valdes S, Torkamaneh D, Chavarriaga P. 2020.** The lowest hanging fruit: Beneficial gene knockouts in past, present, and future crop evolution. *Current Plant Biology* **24**: 100185.

**Monroe JG, McGovern C, Lasky JR, Grogan K, Beck J, McKay JK. 2016.** Adaptation to warmer climates by parallel functional evolution of CBF genes in Arabidopsis thaliana. *Molecular Ecology* **25**: 3632–3644.

**Monroe JG, McKay JK, Weigel D, Flood PJ. 2021.** The population genomics of adaptive loss of function. *Heredity* **126**: 383–395.

**Monroe JG, Powell T, Price N, Mullen JL, Howard A, Evans K, Lovell JT, McKay JK. 2018.** Drought adaptation in *Arabidopsis thaliana* by extensive genetic loss-of-function. *eLife* **7**: XevPb.

**Morita-Yamamuro C, Tsutsui T, Sato M, Yoshioka H, Tamaoki M, Ogawa D, Matsuura H, Yoshihara T, Ikeda A, Uyeda I *et al.* 2005.** The Arabidopsis gene CAD1 controls programmed cell death in the plant immune system and

encodes a protein containing a MACPF domain. *Plant & Cell Physiology* **46**: 902–912.

**Murakoshi Y, Saso Y, Matsumoto M, Yamanaka K, Yotsui I, Sakata Y, Taji T. 2024.** CAD1 contributes to osmotic tolerance in Arabidopsis thaliana by suppressing immune responses under osmotic stress. *Biochemical and Biophysical Research Communications* **717**: 150049.

**Murray AW. 2020.** Can gene-inactivating mutations lead to evolutionary novelty? *Current Biology* **30**: R465–R471.

**Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I. 2014.** The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Research* **42**: D26–D31.

**Noutoshi Y, Kuromori T, Wada T, Hirayama T, Kamiya A, Imura Y, Yasuda M, Nakashita H, Shirasu K, Shinozaki K. 2006.** Loss of Necrotic Spotted Lesions 1 associates with cell death and defense responses in *Arabidopsis thaliana*. *Plant Molecular Biology* **62**: 29–42.

**Orr HA. 2005.** The genetic theory of adaptation: a brief history. *Nature Reviews. Genetics* **6**: 119–127.

**Paffrath V, Tandron Moya YA, Weber G, von Wirén N, Giehl RFH. 2024.** A major role of coumarin-dependent ferric iron reduction in strategy I-type iron acquisition in Arabidopsis. *Plant Cell* **36**: 642–664.

**Paparella C, Savatin DV, Marti L, De Lorenzo G, Ferrari S. 2014.** The Arabidopsis LYSIN MOTIF-CONTAINING RECEPTOR-LIKE KINASE3 regulates the cross talk between immunity and abscisic acid responses. *Plant Physiology* **165**: 262–276.

**Parmar A, Sturm B, Hensel O. 2017.** Crops that feed the world: production and improvement of cassava for food, feed, and industrial uses. *Food Security* **9**: 907–927.

**Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT *et al.* 2018.** A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology* **36**: 983–987.

**Povysil G, Petrovski S, Hostyk J, Aggarwal V, Allen AS, Goldstein DB. 2019.** Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews. Genetics* **20**: 747–759.

**Rajniak J, Giehl RFH, Chang E, Murgia I, von Wirén N, Sattely ES. 2018.** Biosynthesis of redox-active metabolites in response to iron deficiency in plants. *Nature Chemical Biology* **14**: 442–450.

**Ramu P, Esuma W, Kawuki R, Rabbi IY, Egesi C, Bredeson JV, Bart RS, Verma J, Buckler ES, Lu F. 2017.** Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nature Genetics* **49**: 959–963.

**Rausell A, Luo Y, Lopez M, Seeleuthner Y, Rapaport F, Favier A, Stenson PD, Cooper DN, Patin E, Casanova J-L *et al.* 2020.** Common homozygosity for predicted loss-of-function variants reveals both redundant and advantageous effects of dispensable human genes. *Proceedings of the National Academy of Sciences, USA* **117**: 13626–13636.

**Robe K, Izquierdo E, Vignols F, Rouached H, Dubos C. 2021.** The coumarins: Secondary metabolites playing a primary role in plant nutrition and health. *Trends in Plant Science* **26**: 248–259.

**Sedeek KEM, Mahas A, Mahfouz M. 2019.** Plant genome engineering for targeted improvement of crop traits. *Frontiers in Plant Science* **10**: 114.

**Sestili F, Pagliarello R, Zega A, Saletti R, Pucci A, Botticella E, Masci S, Tundo S, Moscetti I, Foti S *et al.* 2019.** Enhancing grain size in durum wheat using RNAi to knockdown GW2 genes. *Theoretical and Applied Genetics* **132**: 419–429.

**Simon MF, Mendoza Flores JM, Liu H-L, Martins MLL, Drovetski SV, Przelomska NAS, Loiselle H, Cavalcanti TB, Inglis PW, Mueller NG *et al.* 2022.** Phylogenomic analysis points to a South American origin of Manihot and illuminates the primary gene pool of cassava. *New Phytologist* **233**: 534–545.

**Singer-Berk M, Gudmundsson S, Baxter S, Seaby EG, England E, Wood JC, Son RG, Watts NA, Karczewski KJ, Harrison SM *et al.* 2023.** Advanced variant classification framework reduces the false positive rate of predicted loss-of-function (pLoF) variants in population sequencing data. *The American Journal of Human Genetics* **110**: 1496–1508.

**Spence JP, Mostafavi H, Ota M, Milind N, Gjorgjieva T, Smith CJ, Simons YB, Sella G, Pritchard JK. 2024.** Specificity, length, and luck: How genes are prioritized by rare and common variant association studies. *bioRxiv*.

**Spielmeyer W, Ellis MH, Chandler PM. 2002.** Semidwarf (*sd-1*), 'green revolution' rice, contains a defective gibberellin 20-oxidase gene. *Proceedings of the National Academy of Sciences, USA* **99**: 9043–9048.

**Tittes S, Lorant A, McGinty S, Holland JB, Sánchez-González J d J, Seetharam A, Tenaillon M, Ross-Ibarra J. 2023.** Not so local: the population genetics of convergent adaptation in maize and teosinte. *eLife* **12**: RP92405.

**Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV. 2010.** Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nature Genetics* **42**: 260–263.

**Vidhi J. 2016.** Cassava: origin, hybridisation and breeding methods. *Biology Discussion*.[WWW document] URL https://www.biologydiscussion.com/vegetable-breeding/cassava-origin-hybridisation-and-breeding-methods-india/68768

**Wang W, Feng B, Xiao J, Xia Z, Zhou X, Li P, Zhang W, Wang Y, Møller BL, Zhang P *et al.* 2014.** Cassava genome from a wild ancestor to cultivated varieties. *Nature Communications* **5**: 5110.

**Wang Y-J, Lu X-H, Zhen X-H, Yang H, Che Y-N, Hou J-Y, Geng M-T, Liu J, Hu X-W, Li R-M *et al.* 2022.** A transformation and genome editing system for cassava cultivar SC8. *Genes* **13**: 1650.

**Xiao L, Cheng D, Ou W, Chen X, Rabbi IY, Wang W, Li K, Yan H. 2025.** Advancements and strategies of genetic improvement in cassava (*Manihot esculenta* Crantz): from conventional to genomic approaches. *Horticulture Research* **12**: uhae341.

**Xu Y-C, Guo Y-L. 2020.** Less is more, natural loss-of-function mutation is a strategy for adaptation. *Plant Communications* **1**: 100103.

**Yang D-L, Shi Z, Bao Y, Yan J, Yang Z, Yu H, Li Y, Gou M, Wang S, Zou B *et al.* 2017.** Calcium pumps and interacting BON1 protein modulate calcium signature, stomatal closure, and plant immunity. *Plant Physiology* **175**: 424–437.

**Yun T, Li H, Chang P-C, Lin MF, Carroll A, McLean CY. 2021.** Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics* **36**: 5582–5589.

**Zhao K, Lensink M, Monroe JG. 2025.** Functional insights into dispensable genes using genome-wide loss-of-function burden tests in Arabidopsis. *bioRxiv*.

**Zheng C, Yu Y, Deng G, Li H, Li F. 2022.** Network and evolutionary analysis reveals candidate genes of membrane trafficking involved in maize seed development and immune response. *Frontiers in Plant Science* **13**: 883961.

**Zhou Y, Niu R, Tang Z, Mou R, Wang Z, Zhu S, Yang H, Ding P, Xu G. 2023.** Plant HEM1 specifies a condensation domain to control immune gene translation. *Nature Plants* **9**: 289–301.

**Zhou Y, Zhang Z, Bao Z, Li H, Lyu Y, Zan Y, Wu Y, Cheng L, Fang Y, Wu K *et al.* 2022.** Graph pangenome captures missing heritability and empowers tomato breeding. *Nature* **606**: 527–534.

**Zhu L, Yi H, Su H, Guikema S, Liu B. 2023.** Impacts of climate change on cassava yield and lifecycle energy and greenhouse gas performance of cassava ethanol systems: an example from Guangxi Province, China. *Journal of Environmental Management* **347**: 119162.

## Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

**Dataset S1** Predicted LoF variants in variant call format.

**Dataset S2** Gene-level LoF burdens in variant call format.

**Fig. S1** Correlation between principal components and geography.

**Fig. S2** ADMIXTURE cross-validation error and ancestry assignments across multiple K values.

**Fig. S3** Identity-by-state (IBS) relationship between cassava accessions and close relatives.

**Fig. S4** Geographic location of accessions with high genetic identity similarity.

**Fig. S5** Nucleotide diversity ratio of wild relatives to cultivated lines.

**Fig. S6** Relationship between loss-of-function (LoF) variation and inbreeding across the Genome.

**Fig. S7** Manhattan plots of genome-wide association tests using all variants, LoF variants and LoF burdens.

**Fig. S8** Genotype–environment association using redundancy analysis (RDA).

**Table S1** Meta information of 1152 accessions included in this study.

**Table S2** Summary table of GO analysis.

**Table S3** Summary table of GWAS and RDA.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.