scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Haplotype-resolved genome of *Citrus* × *sinensis* 'Pera IAC', the most widely cultivated sweet orange in Brazil

Monica Neli Alves[1], Leny Calvez[1], Mateus de Almeida Santos[1], Vitor Trinca[2], Mauro de Medeiros Oliveira[2], Jesus Aparecido Ferro[2], Eliane Cristina Locali[1], Patrick Ollitrault[1], Nelson Arno Wulff[1] & Alessandro M. Varani[2]✉

A chromosome-scale, haplotype-resolved genome assembly is reported for *Citrus* × *sinensis* 'Pera IAC', the most widely cultivated sweet orange in Brazil and a major contributor to global orange juice production. Leveraging PacBio Revio HiFi and Arima Hi-C sequencing, both parental haplotypes were reconstructed into nine chromosome-scale assemblies each (2n = 18), with telomeric repeats detected at multiple chromosome ends, capturing contributions of its ancestral progenitors, *C. maxima* (pummelo) and *C. reticulata* (mandarin). The assembly spans ~620 Mb, with 99.9% BUSCO completeness, and provides comprehensive structural and functional annotation (58,171 predicted gene models and 67,796 CDS including isoforms; ~51% transposable elements). Complete chloroplast and mitochondrial circular genomes were also assembled, expanding the genomic resources for citrus. Technical validation included BUSCO, LTR Assembly Index, Inspector, and Merqury analyses, all supporting the accuracy and completeness of the assembly. Synteny analyses and genetic map anchoring further confirmed the structural integrity and chromosome phasing. All datasets are publicly available, providing a robust reference for allele-aware analyses, comparative genomics, and the development of improved sweet orange cultivars.

Background & Summary

Sweet orange (*Citrus* × *sinensis* (L.) Osbeck) is among the most widely cultivated fruit crops worldwide, and Brazil remains its leading producer[1,2]. Within the São Paulo–Triângulo Mineiro Citrus Belt (Fig. 1), Pera sweet orange accounts for roughly 35–36% of bearing trees, where 'Pera IAC' is the most cultivated clone among this cultivar, surpassing 'Valencia' (~31%) and 'Hamlin/Westin/Rubi' (~16%), due to its high juice yield, balanced flavour, and stable mid-season maturity[3,4]. This cultivar plays a strategic role in the Brazilian juice industry, as its harvest period bridges early cultivars such as 'Hamlin' and late cultivars such as 'Valencia', contributing to a more continuous supply of fruit for industrial processing. In addition, 'Pera IAC' is highly valued for its not from concentrate (NFC) orange juice production due to its desirable sensory qualities and consistent industrial performance.

Because ~80% of Brazil's orange harvest is processed for juice and the country supplies ~75% of the frozen-concentrate orange juice traded globally[4,5], genomic resources for 'Pera IAC' have national and international importance, providing essential support for breeding programs. Moreover, 'Pera IAC' is known for requiring cross-protection with mild strains of *Citrus tristeza virus* (CTV) to prevent severe stem-pitting symptoms, a management strategy that has enabled the cultivar to become the most widely planted sweet orange in Brazil[6]. Together, these agronomic and biological characteristics make 'Pera IAC' an especially relevant genetic background for genomic studies aimed at improving disease management, fruit quality, and citrus productivity.

[1]Fundo de Defesa da Citricultura, 14807-040, Araraquara, Brazil. [2]Departamento de Biotecnologia Agropecuária e Ambiental, Universidade Estadual Paulista (UNESP), Faculdade de Ciências Agrárias e Veterinárias, 14884-900, Jaboticabal, SP, Brazil. ✉e-mail: alessandro.varani@unesp.br

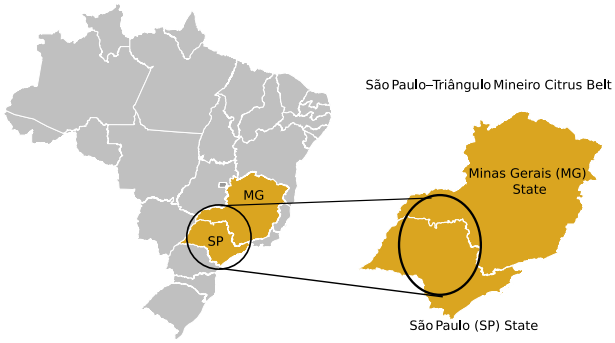

Fig. 1 Approximate representation of the "*São Paulo–Triângulo Mineiro*" Citrus Belt in Brazil. Map of Brazil highlighting the "*São Paulo–Triângulo Mineiro*" Citrus Belt, covering approximately 394,918 hectares (≈975,800 acres) of commercial sweet-orange groves.

Commercial sweet oranges, including 'Pera IAC', originated from an ancient genome admixture of two ancestral species, *C. reticulata* Blanco (mandarins) and *C. maxima* (Burm.) Merr. (pummelos)[7]. Their propagation by nucellar apomixis and grafting practices has maintained a narrow genetic base and high heterozygosity of this horticultural group[8]. This genetic uniformity contributes to susceptibility to major diseases such as huanglongbing (HLB) and citrus canker[9,10]. At the same time, breeding is further limited by apomixis, a long juvenile phase, and the accumulation of hidden[11] and deleterious[12] mutations. The first sweet-orange draft captured only ~87% of the genome and collapsed haplotypes[8]. More recently, long-read whole-genome sequencing (WGS) technologies provided three haplotype-level sweet orange genome assemblies[10–12]. Although dozens of citrus genomes are now available and a genus-wide pangenome initiative is underway[13], a haplotype-resolved reference for 'Pera IAC' has been missing. This absence is particularly critical given that no South American cultivar of comparable importance—naturalized to Brazil's tropical environment—has yet been sequenced, despite 'Pera IAC' dominating national juice production and contributing substantially to the global orange-juice market.

Here, a chromosome-scale, haplotype-phased genome assembly of *Citrus* × *sinensis* 'Pera IAC' was generated using PacBio Revio HiFi reads and Arima Hi-C sequencing. The assembly resolves both parental haplotypes, one mostly originating from *C. reticulata* and the other from *C. maxima*, into nine chromosome-scale sequences each, with telomeric repeats detected in all chromosomes. It faithfully captures the cultivar's high heterozygosity, resolves both gene-rich and repetitive regions, and documents haplotype-specific structural variants (Fig. 2). This dataset provides a robust framework for allele-level analyses and molecular marker development in sweet-orange breeding, while filling a key gap in citrus pangenome resources.

Methods

Plant material, DNA isolation, and PacBio Sequel II sequencing. Fully expanded young leaves from a healthy *Citrus* × *sinensis* 'Pera IAC' tree maintained at Fundecitrus (Araraquara, Brazil; 21.8166° S, 48.1822° W), originally selected and maintained at "*Centro de Citrultura Sylvio Moreira, Cordeirópolis, São Paulo, Brazil*" were harvested, kept for 24 h in the dark to deplete starch, flash-frozen in liquid nitrogen, and shipped on dry ice to the Arizona Genomics Institute (AGI), Tucson, USA. High-molecular-weight (HMW) genomic DNA was extracted with a modified cetyl-trimethyl-ammonium bromide (CTAB) protocol[14] that includes additional PVP (2%, w/v) and RNase A steps to remove polyphenols and RNA. DNA concentration and purity were measured with the Qubit dsDNA High-Sensitivity assay (Thermo Fisher Scientific) and a NanoDrop ND-1000 spectrophotometer; fragment size and integrity were confirmed on a Femto Pulse System (Agilent) and by 0.75% pulsed-field gel electrophoresis. HMW DNA was mechanically sheared to 15–25 kb with a Covaris g-TUBE, purified with AMPure PB beads (ratio 0.45 ×), and size-selected on a SageELF to enrich fragments >15 kb. Library construction utilized the SMRTbell Express Template Prep Kit v3.0 (Pacific Biosciences). One SMRT Cell 25 M was run on a PacBio Sequel II instrument using the HiFi Circular Consensus Sequencing (CCS) protocol, resulting in Q30 HiFi reads with an average insert length of ~15 kb (Table 1).

Arima Hi-C library construction and Illumina sequencing. Proximity-ligation libraries were prepared from the same flash-frozen leaf tissue using the Arima-HiC + kit (Arima Genomics, USA), which employs a four-enzyme restriction cocktail to promote uniform chromatin fragmentation. Procedures followed the manufacturer's guidelines: in-nucleus formaldehyde cross-linking, enzymatic digestion, fill-in with biotin-dATP, proximity ligation, reversal of cross-links, and DNA purification. Dual-indexed libraries were constructed using the KAPA HyperPrep kit (Roche, Switzerland) and size-selected with SPRIselect beads (Beckman Coulter, USA) to a range of 350–800 bp. After eight cycles of PCR amplification, libraries were sequenced on an Illumina NovaSeq X Plus platform (Illumina, USA; S4 flow cell) to obtain 2 × 150 bp paired-end reads suitable for Hi-C scaffolding (Table 1).

Genome size and heterozygosity survey. Genome size and heterozygosity were estimated from k-mer frequency distributions. K-mers of length 21 were extracted and counted using Jellyfish v2.3[15], and the resulting histogram was analyzed with GenomeScope2[16] to infer genome size and heterozygosity rates. Additional estimates were generated using the *jcvi* genome build script[17] (Fig. 2a).

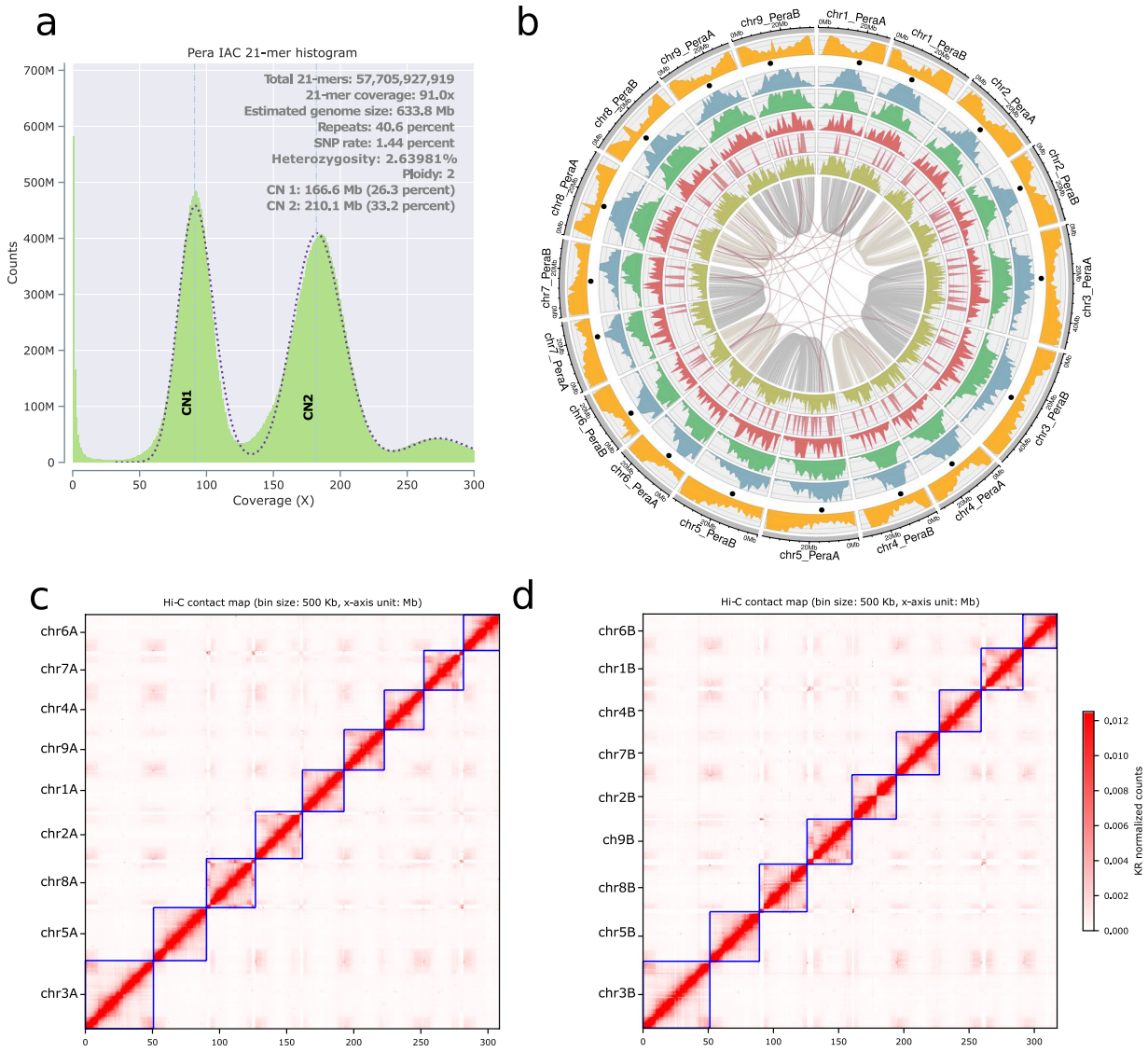

Fig. 2 Haplotype-phased genome of *Citrus × sinensis* 'Pera IAC'. (**a**) Genome survey based on 21-mer frequency analysis. (**b**) Circular representation of the genome. Tracks from outer to inner show: chromosomes (gray), gene density (yellow), predicted centromeric regions (black dots), LTR Gypsy retrotransposons (blue), LTR Copia retrotransposons (green), LARD and TRIM elements (red), DNA transposons (orange), and inter-chromosomal collinearity links. (**c,d**) Hi-C contact maps supporting chromosome scaffolding of hapA and hapB after parental assignment, with hapA being essentially derived from *C. reticulata* and hapB from *C. maxima*. Panels b–d therefore represent the fully phased genome, organized according to its ancestral progenitors.

Genome assembly and primary phasing. PacBio HiFi reads were assembled with *hifiasm* v0.20.0-r639[18] in Hi-C–aware mode (–h1–h2). To improve phasing and enable telomere detection during assembly, the following additional parameters were applied: --n-weight 7 --n-perturb 30000 --f-perturb 0.3 --telo-m CCCTAAA, specifying the canonical plant telomeric motif (TTTAGGG)n. The program produced two haplotype-resolved sets of primary contigs (HapA and HapB). At this stage, the haplotypes were not yet assigned to their ancestral progenitors (*C. maxima* and *C. reticulata*); this assignment was performed in subsequent analyses (see below). Remarkably, even before Hi-C scaffolding, the primary hifiasm assemblies already contained ultra-long contigs spanning nearly entire chromosomes (Table 1).

Hi-C read pairs were then mapped separately to each haplotype assembly (hapA and hapB) using the Arima mapping pipeline v0.3[19], which sequentially wraps bwa-mem2[20], samtools[21], and pairtools[22] with Arima-recommended presets. Only read pairs flagged as valid Hi-C after duplicate removal were retained for scaffolding. Chromosome scaffolding was performed in parallel with two independent tools for cross-validation: YAHS v1.2.2[23] and HapHiC v1.0.7[24], both with default parameters. Contact maps were visualized with the HapHiC plotting utility, and only concordant scaffolding decisions between YAHS and HapHiC were retained in the final build (Table 1).

Genome sequencing statistics			
Number of HiFi reads	4,612,052		
N50 HiFi reads (bp)	14,676		
Average Phred value (Q)	38.5		
GC content of the HiFi reads (%)	36.28		
HiC sequencing			
Total sequences (2×150bp)	129,881,808		
Average Phred value (Q)	38		
GC content of the HiC reads (%)	37.03		
Number of HiC mapped reads pairs	54,819,788 (42.2% of the total)		
Intra-chromosomal pairs ≤ 1 kb	25,905,892 (47.3%)		
Intra-chromosomal pairs ≥ 10 kb	22,547,730 (41.1%)		
Intra-chromosomal pairs ≥ 15 kb	21,776,693 (39.7%)		
Intra-chromosomal pairs ≥ 20 kb	21,193,253 (38.7%)		
Inter-chromosomal pairs	9,441,621 (17.2%)		
Primary assembly (near-chromosome scale, with haplotigs and redundant sequences removed)			
	hapA	**hapB**	
Assembly length (bp)	307,880,319	320,644,964	
Scaffold number	13	12	
Scaffold N50 (bp)	26,087,353	31,878,633	
Scaffold L50 (bp)	5	4	
Longest scaffold (bp)	50,705,368	50,768,643	
Number of scaffolds > 100 Kb	13	12	
Number of scaffolds > 1 M	13	12	
Number of scaffolds > 10 Mb	11	10	
Genome assembly statistics (after Hi-C correction and scaffolding)			
	hapA + hapB	**hapA (C. reticulata)**	**hapB (C. maxima)**
Genome length (bp)	621,264,921	306,441,580	314,143,499
Contig number	20	9	9
Scaffold N50 (bp)	34,005,357	34,577,998	34,005,357
Scaffold L50	8	4	4
GC content of the genome (%)	35.29	35.17	35.39
Assembly gaps (%)	2 (199 Ns)	1 (100 Ns)	1 (99 Ns)
Chromosomes	18 (1n = 9) + plastid and mitochondrial genomes	9	9
Predicted centromeres	18	9	9
Predicted telomeres	25	1 left; 6 right; 2 both ends	2 left; 2 right; 5 both ends

Table 1. Sequencing and assembly statistics for the *Citrus* × *sinensis* 'Pera IAC' genome. Summary of PacBio HiFi and Arima Hi-C sequencing output and assembly metrics for the haplotype resolved genome of *C.* × *sinensis* 'Pera IAC'. Values are provided for the combined assembly (hapA + hapB) as well as for each parental haplotype individually. Assembly statistics include contiguity, GC content, scaffold metrics, gaps, and predicted chromosomal features.

Refined haplotype phasing and inference of phylogenomic structure in the 'Pera IAC' genome. Although *hifiasm*, in combination with Hi-C data from the same *C.* × *sinensis* 'Pera IAC', produced two haplotype-resolved assemblies (hapA and hapB), this approach does not unambiguously assign parental origin to each chromosome. Because Hi-C data were not derived from the parental lines (i.e., no trio binning), the phasing might still contain a mixture of genomic segments from both ancestral species (*C. maxima* and *C. reticulata*). To clarify parental assignment and ancestral genomic composition of *C.* × *sinensis* 'Pera IAC', the mosaic contributions of *C. maxima* (pummelo) and *C. reticulata* (mandarin) were assessed. For this purpose, a curated set of 323,899 *C. maxima* and 555,790 *C. reticulata* diagnostic SNPs (DSNPs) previously identified in the *C. reticulata* 'Cleopatra mandarin' reference genome (CITRE)[25] and annotated for their ancestral origin[26], was employed (Table S1).

For each DSNP, 50-bp flanking probes upstream and downstream were extracted and mapped to the pre-phased haplotype-A and haplotype-B assemblies using ScaffHunter's LocOnRef[27] with bwa-mem2[20]. Probe coordinates were offset-corrected, and only uniquely mapped, strand-consistent probes were retained. The

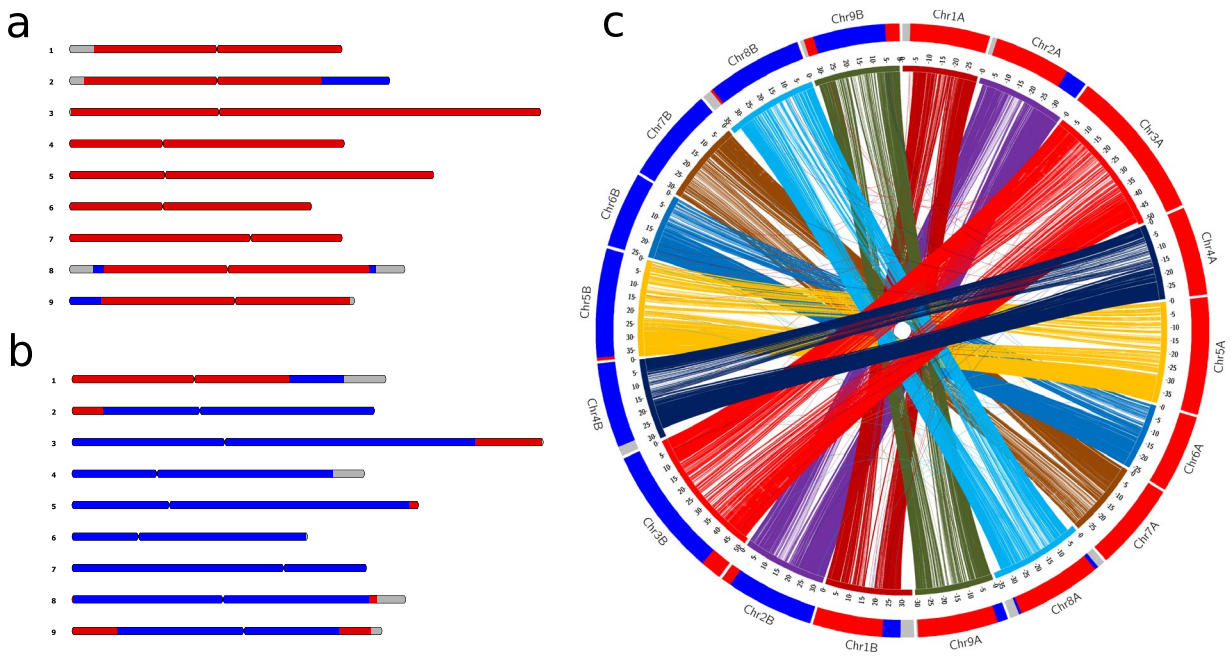

Fig. 3 Phylogenomic ancestry structure of the *Citrus* × *sinensis* 'Pera IAC' genome. (**a,b**) Distribution of diagnostic SNPs (DSNPs) from *C. reticulata* (mandarin) and *C. maxima* (pummelo) and along hapA and hapB assemblies, displayed as ancestry tracks on ideograms (detailed mapping results are shown in Supplementary Figures S1, S2). (**c**) Circos plot illustrating the chromosome-scale mosaic architecture. In all panels, red indicates *C. reticulata* ancestry, blue indicates *C. maxima* ancestry, and gray represents undefined ancestry. Together, these representations document the phased parental composition of the 'Pera IAC' genome following assignment with diagnostic SNPs.

corresponding *C.* × *sinensis* 'Pera IAC' alleles were then compared with inferred ancestral states (Table S2). Collectively, these steps enabled fine-resolution phasing of both haplotypes, confirming their suitability for downstream ancestry inference and structural analyses (Supplementary Figures S1, S2).

The validated SNP set was used to estimate allele frequencies (pummelo, mandarin, or undefined) along each chromosome and to derive ancestry tracks. Circos plots[28] were generated within Galaxy[29], and ideograms were rendered using the Gemo platform (https://gemo.southgreen.fr) (Fig. 3). Centromere positions were approximated using two complementary approaches: (i) gene-density profiling in 500-kb windows based on the official annotation, and (ii) MareyMap[30] plots contrasting physical coordinates with genetic positions of SNP markers from the consensus map[31] (Fig. 3).

Gene density (see Annotation procedures) was profiled along each chromosome, revealing gene-poor regions that typically mark centromeric and pericentromeric domains. Although the consensus genetic map does not represent recombination from a single population, alignment of genetic and physical coordinates with MareyMap highlighted low-recombination plateaus coincident with these domains. Patterns were visualized in Circos alongside gene-density histograms, providing cross-validation for the inferred centromere positions (Table S3, Supplementary Figure S3).

This phylogenomic inference consolidated the parental assignment of haplotype A to *C. reticulata* and haplotype B to *C. maxima*, while documenting the chromosome-scale mosaic structure of the 'Pera IAC' genome.

Collinearity assignment of phased haplotypes using consensus map markers. To assess the structural accuracy of the phased assemblies, SNP markers from the *Citrus* consensus genetic map[31] were anchored onto both phased hapA and hapB (Table S4). This *Citrus* consensus genetic map resulted from the high synteny and collinearity among five ancestral *Citrus* species and three interspecific hybrids, including *C. maxima*, *C. reticulata* representatives, and one *C. maxima x C. reticulata* hybrid. This consensus map is highly syntenic and collinear with recent high-quality genome assemblies of numerous *Citrus* species[31]. A total of 17,718 probes of 201 bases originally designed on the *C. clementina* v1.0 reference genome[32], were anchored in each haplotype (100 bases each side of the SNPs mapped on the consensus map) using the LocOnRef tool from the Scaffhunter toolbox[27] with bwa-mem2[20]. Genomic coordinates of uniquely aligned probes were retained, and to reduce redundancy, only one SNP per gene of the 'Pera IAC' assemblies was selected. Marker positions in the assemblies were then compared with their genetic positions in the consensus map to assess genome-wide collinearity. Spearman's rank correlation coefficients were calculated for each chromosome, and Circos plots[28] were generated through the Galaxy platform[29] to visualize synteny patterns (Fig. 4a,b, Tables S5–S7).

To further evaluate the structural consistency of the *Citrus* × *sinensis* 'Pera IAC' genome assembly, whole-genome alignments were performed between the phased haplotypes (hapA and hapB) and representative

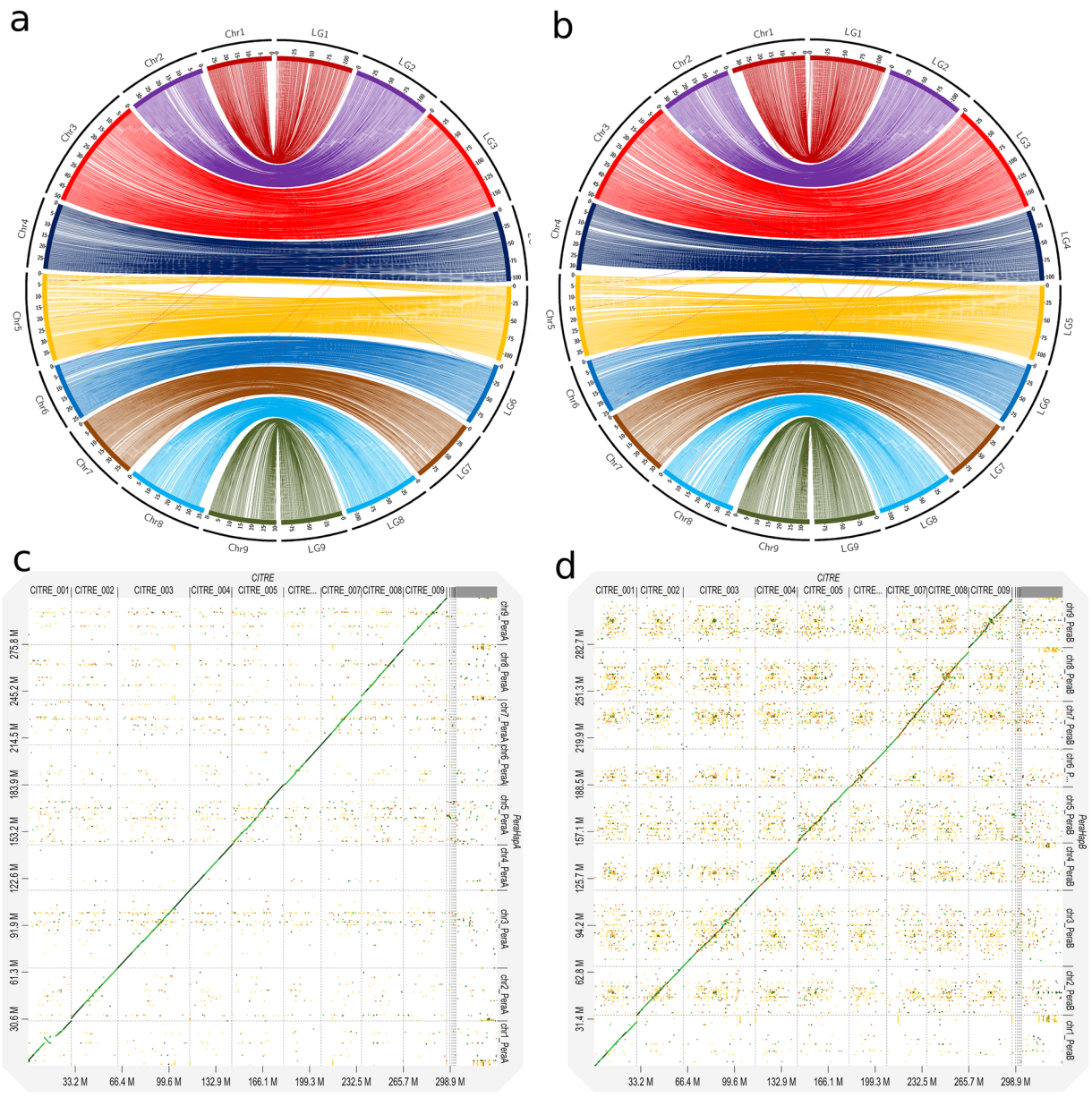

Fig. 4 Anchoring and synteny analyses of the *Citrus* × *sinensis* 'Pera IAC' haplotype assemblies. (**a,b**) Anchoring of SNP probes from the *Citrus* consensus genetic map onto hapA and hapB assemblies, with Spearman's rank correlation coefficients of 0.998 (hapA) and 0.997 (hapB). (**c,d**) Whole-genome alignments of hapA and hapB against the *C. reticulata* 'Cleopatra mandarin' reference genome (CITRE), visualized with D-GENIES dot plots. Both analyses enable the visualization of the genome-wide synteny and large-scale structural conservation, thereby supporting the integrity of the haplotype assemblies.

assemblies of ancestral *Citrus* species, including *C. maxima* 'Shatian'[10], *C. reticulata* ICVN 0110274, *C. medica* 'Corsican' citron ICVN 0100613, and *C. micrantha* ICVN 0101115[26]. Pairwise alignments were generated using D-GENIES[33] with Minimap2[34] (v2.28) to produce genome-wide dot plots and assess chromosome-scale collinearity. Among the currently available haplotype-resolved sweet orange genomes, the 'Valencia' assemblies represent highly contiguous references, including the DVS Valencia haplotypes (GCA_022201045.1 and GCA_022201065.1)[35] and the recently released assemblies ASM4647076v1 and ASM4647077v1. Due to their high contiguity and use in comparative citrus genomics, these assemblies were selected for pairwise structural comparisons with the 'Pera IAC' genome.

Pairwise alignments between the 'Pera IAC' and 'Valencia' haplotypes were performed using Minimap2 with the asm5 preset. Structural variants were identified using SyRI (v1.7.0)[36], which classifies syntenic regions and structural rearrangements, including inversions, duplications, and inter-chromosomal translocations. To ensure consistent visualization, chromosome 9 from the 'Valencia' assembly was reverse-complemented prior to comparison. Structural relationships were visualized using plotsr[37]. To improve representation of large structural

Features			
	Total (hapA + hapB)	**hapA (*C. reticulata*)**	**hapB (*C. maxima*)**
Number of genes	58,171	28,579	29,592
Number of CDSs (including isoforms)	67,796	33,358	34,438
Complete CDS	67,728	33,319	34,409
Start, no stop CDS	29	17	12
Stop, no start CDS	32	19	13
No stop, no start CDS	7	3	4
Mean gene length	3,116	3,126	3,107
Mean CDS length	1,254	1,263	1,246
Mean exons per gene	5	5	5
Mean introns per gene	4	4	4
tRNAs	908	442	466
% of genome covered by genes	29.2	29.2	29.3
% of genome covered by CDS	13.7	13.8	13.7
% of genome covered by TEs	51.56	51.41	51.71
% of genome covered by LTR/Copia	11.90	11.99	11.81
% of genome covered by LTR/Gypsy	15.65	15.98	15.33
% of genome covered by TIR	11.64	11.36	11.91
% of genome covered by Helitrons	1.49	1.50	1.48
% of genome covered by Unknown repeats	6.05	5.70	6.39

Table 2. Annotation features of the *Citrus* × *sinensis* 'Pera IAC' haplotype assemblies. Summary of structural and functional annotation statistics for the combined assembly (hapA + hapB) and for each parental haplotype separately.

events that may be fragmented during whole-genome alignment, adjacent SyRI calls involving the same chromosome pair and variant class were merged when separated by ≤ 100 kb in the reference genome.

High-quality genome annotation. Genome annotation was performed independently for each haplotype assembly (hapA and hapB) in three consecutive stages (Repeat annotation, Structural gene models, and Functional annotation), following best practices for plant genomes[38]. The final annotation resulted in 58,171 predicted gene models and 67,796 CDS (including isoforms) across both haplotypes (Table 2).

Stage 1 – Repeats annotation. Repetitive sequences were identified using EDTA-GUI[39], which wraps EDTA v2.0.1 for *de novo* discovery of plant Transposable Elements (TEs). EDTA-GUI implements domain-based filtering, lineage-level classification with REXdb[40], and refined subtyping of LTR retrotransposons (including TRIMs, LARDs, TR_GAGs, and BARE-2). The resulting TE library was used to soft-mask both haplotype assemblies and exported as GFF3 and FASTA files containing detailed TE annotations. Telomeric and higher-order repeats were detected with quarTeT[41], and putative centromeric satellites were identified using the Centromics pipeline[42]. A descriptive TE annotation is provided in the *Data Overview*.

Stage 2 – Structural gene models and isoforms annotation. Gene prediction integrated multiple sources of evidence. *Ab initio* predictions were generated with BRAKER3 (v3.0.4)[43] in RNA-seq-guided and protein-guided modes, using *C.* × *sinensis* 'Pera' transcript data from CitrusKB[44]. Additional predictions were obtained with GALBA v1.0.7[45] coupled with miniport v0.12[46] TSEBRA v1.1.2[47], GeMoMa v1.9[48], and Exonerate v2.4.0[49]. Each evidence set was benchmarked using BUSCO v6.0.0[50] (embryophyta_odb12), retaining only runs with ≥90% completeness. Transcript structures reconstructed with PASA v2.5.3[51] were merged in EvidenceModeler v1.1.0[52], followed by two refinement rounds to add UTRs, correct exon–intron boundaries, and catalog isoforms.

Candidate loci overlapping TEs or harboring TE domains identified by TEsorter v1.4.1[53] were flagged, and potential false positives were excluded if they met combined criteria [CDS < 250 bp, absence of signal peptides (SignalP 6.0[54]) and transmembrane helices (Phobius[55]), lack of RNA-seq support, and no similarity to UniProt[56] or NCBI-RefSeq.[57] plant proteins]. The final filtered set achieved BUSCO completeness equal to or higher than the assembly itself (99%).

Gene-family organization was assessed with MCScanX[58], classifying duplicates as singleton, tandem, proximal, dispersed, WGD-derived, or transposed. Retrocopies and small-scale duplicates were identified with RetroScan[59] and DupGen_finder[60].

Stage 3 – Functional annotation. Functional annotation followed a hierarchical approach. Homology searches were performed against UniProt and NCBI-RefSeq plants. Transcription factors were annotated with PlantTFDB v5.0[61]; resistance-related genes with NLR-Atlas[62] and PRGdb v4.0[63]; annotation transfer was

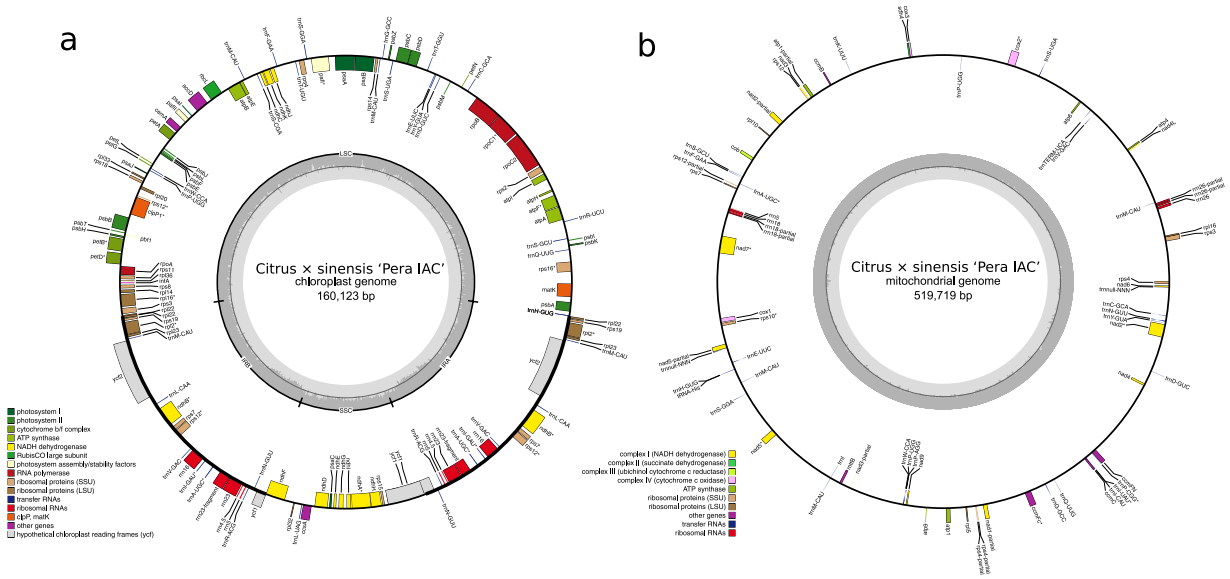

Fig. 5 Organellar genomes of *Citrus* × *sinensis* 'Pera IAC'. (**a**) Chloroplast genome (160,123 bp) shows the canonical quadripartite structure with large single-copy (LSC), small single-copy (SSC), and two inverted repeat (IR) regions. The plastid genome encodes 131 unique genes, including 86 protein-coding genes, eight rRNAs, and 37 tRNAs, corresponding to the standard gene complement for photosynthesis, gene expression, and redox metabolism (Table S9). (**b**) Mitochondrial genome (519,719 bp) with 63 annotated genes, including 40 protein-coding genes, three rRNAs (rrn5, rrn18, rrn26), and 23 distinct tRNAs. The mitochondrial gene set includes components of respiratory complexes I–V, ribosomal subunits, maturases, and cytochrome c biogenesis. Gene models also include fragmented, pseudogenized, trans-spliced, and duplicated loci, consistent with patterns typically observed in angiosperm mitochondrial genomes (Table S10).

performed with eggNOG-mapper v2.1.4[64]. InterProScan v5.61[65] identified domains and motifs, assigning Gene Ontology[66], EC numbers[67], Pfam[68], InterPro[69], CAZy[70], and KEGG[71] classifications. Specialized metabolism clusters were predicted with plantiSMASH v1.0[72]. Blast2GO Basic v6.0[73] was used for GO mapping and annotation, with non-Viridiplantae terms removed. Genome-wide plots were generated with shinyCircos-V2.0[74] (Fig. 2b, Table 2).

Final annotation packaging and interactive visualization. All structural and functional data were consolidated into GFF3 format with AGAT v1.2.0[75], validated for internal consistency, and distributed alongside CDS and protein FASTA files. These outputs integrate repeat, structural, and functional annotations, enabling high-quality queries and visualization. The complete annotation is accessible through an interactive JBrowse2[76] genome browser (https://plantgenomics.ncc.unesp.br/Genomes/Citrus/), which provides user-friendly exploration of TE annotations, RNA-seq evidence, gene models, functional features, and specialized metabolism clusters (plantiSMASH). In addition, the platform features an integrated BLAST[77] search, enabling direct sequence similarity queries against the annotated genome and predicted proteins. Together, these tools offer an accessible, comprehensive, and interactive environment for citrus genomics research.

Organellar genome assembly and annotation. Plastid and mitochondrial genomes were assembled using a workflow adapted from de Abreu *et al.*[78]. Reference databases for organellar sequences were obtained from the NCBI Organelle Genome Resources (https://www.ncbi.nlm.nih.gov/genome/organelle/). HiFi reads were classified against these databases using Kraken2[79], and organelle-assigned reads longer than 13 kb were extracted with the "extract_kraken_reads.py" script from KrakenTools. Filtered reads were assembled with Flye v2.9[80] under default parameters. Plastid and mitochondrial contigs were recovered from the assembly graphs using "get_organelle_from_assembly.py" in GetOrganelle[81], and candidate circular ptDNA and mtDNA molecules were validated by manual inspection in Bandage[82].

Annotation of plastid and mitochondrial genomes was performed with GeSeq[83], using *Arabidopsis thaliana* as reference (NC_000932 for ptDNA and NC_037304 for mtDNA). All predicted features were further manually verified to ensure annotation accuracy. Circular maps were generated with OGDraw[84] (Fig. 5).

Data Records

This study generates and shares multiple genomic datasets for *C.* × *sinensis* 'Pera IAC', including raw sequencing reads, haplotype-resolved nuclear assemblies, organellar genomes, and structural/functional annotations. All resources are deposited in public repositories with persistent identifiers to ensure long-term accessibility and reuse.

- **Raw sequencing data**: PacBio HiFi long reads and Illumina Hi-C paired-end reads are available in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1301450, with accession numbers SRR34861620[85] and SRR34861621[86].
- **Genome assemblies**: The haplotype-resolved assemblies of *Citrus* × *sinensis* 'Pera IAC' are deposited in GenBank under the accessions JBWUIW000000000[87] and JBWUIX000000000[88] for Haplotypes A and B, respectively.
- **Organellar genomes**: The complete plastid (ptDNA) and mitochondrial (mtDNA) genomes are available at GenBank (accessions PX373693[89] and PX373694[90]).
- **Genome annotation**: Structural and functional annotation files, including GFF3 (gene models and TEs), CDS, protein sequences, and TE libraries, are accessible via Zenodo[91] and through the *Pera IAC* public genome browser at: https://plantgenomics.ncc.unesp.br/Genomes/Citrus/.

Each dataset includes accompanying metadata and standardized file formats (FASTQ, FASTA, GFF3) to facilitate reuse.

Data Overview

To provide a descriptive summary of the repetitive landscape in the *C.* × *sinensis* 'Pera IAC' genome, the distribution and diversity of annotated TEs were visualized with the EDTA-GUI (Fig. 6). TEs account for ~51% of the assembly, with LTR retrotransposons representing the predominant fraction. Insertion age profiles indicate a broad spectrum of LTR activity, while phylogenetic reconstruction highlights the diversity of *Gypsy* and *Copia* lineages. These descriptive plots illustrate the composition and structural diversity of the TE dataset, complementing the repeat annotation presented above and offering an overview of key features available to users of the resource.

Technical Validation

To assess the completeness and reliability of the *C.* × *sinensis* 'Pera IAC' genome assembly, three complementary strategies were employed. First, Benchmarking Universal Single-Copy Orthologs (BUSCO v6.0.0) was applied using the *embryophyta_odb12* dataset (n = 2,026 genes), evaluating both the assembled genome and the structural annotation. Second, the LTR Assembly Index (LAI)[92] was calculated to assess assembly contiguity and the resolution of repetitive regions based on the identification and integrity of annotated LTR retrotransposons. Third, alignment-based validation with Inspector[93] and k-mer mapping with Merqury[94] provided independent assessments of consensus accuracy, completeness, and phasing quality (Table 3).

Collectively, these analyses highlight the high quality of the *C.* × *sinensis* 'Pera IAC' assembly. BUSCO recovered nearly all conserved embryophyte genes in both the genome and annotation, indicating near-complete representation of the gene space. The LAI score further places the assembly within the reference-quality range for plant genomes. Inspector confirmed a mapping rate of 99.98% with an estimated QV of 45.5 (error rate $\approx 2.8 \times 10^{-3}$),while Merqury yielded a higher genome-wide QV of 70.5 (error rate $\approx 8.9 \times 10^{-8}$), consistent with the expected difference between alignment-based and k-mer–based validation.

It is worth noting that Inspector did not detect haplotype switches, as expected for a read-alignment–based validation that primarily captures local structural inconsistencies. By contrast, Merqury reported a switch error rate of 38.4%. However, this is likely an overestimation driven by the elevated heterozygosity of the sweet orange genome (2.64% estimated from k-mer spectra; SNP rate 1.44%). As noted by Rhie *et al.*[94], Merqury can inflate switch error estimates in highly heterozygous plant assemblies, particularly when phasing relies solely on Hi-C data, since k-mer–based evaluation is sensitive to haplotype divergence rather than true large-scale misphasing. This interpretation is supported by the Merqury phased-block blob plot (Supplementary Figure S4), which shows the hapA and hapB sequences separated into large, coherent phased blocks, consistent with effective long-range phasing.

To further validate the assembly, diagnostic SNP probes from the *Citrus* consensus genetic map anchored onto both haplotype assemblies revealed nearly perfect collinearity (Spearman's $\rho = 0.998$ for hapA and 0.997 for hapB). In addition, whole-genome alignments against the *C. reticulata* 'Cleopatra' reference genome confirmed long-range synteny and the absence of major rearrangements. Whole-genome alignments were also performed against additional citrus reference genomes. Comparisons with representative ancestral species (*C. maxima*, *C. reticulata*, *C. medica*, and *C. micrantha*) revealed strong chromosome-scale collinearity across all chromosomes (Supplementary Figure S5), supporting the structural accuracy of the assembly. Pairwise comparisons with the phased 'Valencia' sweet orange genome also demonstrated extensive genome-wide synteny, with most regions aligning along continuous diagonals in whole-genome dot plots (Supplementary Figure S6a). Structural variant analysis using SyRI confirmed that the majority of genomic regions are syntenic between assemblies (Supplementary Figure S6b; Table S8). A previously reported inter-chromosomal rearrangement involving chromosomes 1 and 9 was detected in the 'Valencia' genome[31,35,95,96] but was not observed in the 'Pera IAC' assembly. Thus, these complementary validations demonstrate that the 'Pera IAC' haplotype assemblies preserve accurate chromosome ordering, structural integrity, and reliable parental phasing.

Together, these resources provide an essential reference for the citrus research community. By delivering a chromosome-scale, phased genome of Brazil's most widely cultivated sweet orange, this dataset fills a critical gap in citrus genomics and facilitates a range of downstream applications, including allele-specific analyses of gene function, structural variant discovery, and haplotype-resolved comparative genomics. In addition, the availability of phased assemblies supports the identification of allelic variation underlying agronomically important traits and enables the development of molecular markers for breeding programs aimed at improving yield, fruit quality, and resilience to biotic and abiotic stresses.

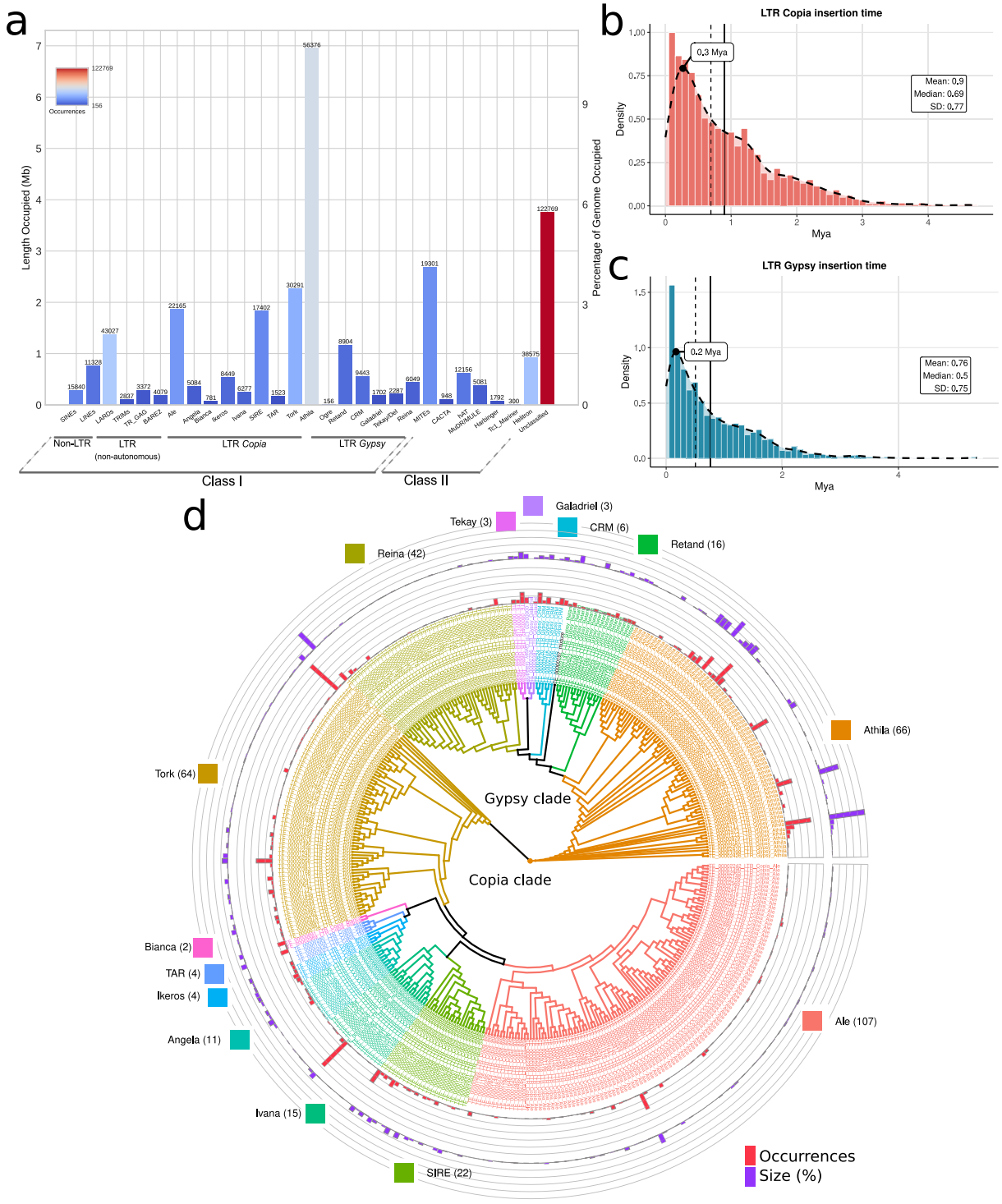

Fig. 6 Overview of transposable elements in the *Citrus* × *sinensis* 'Pera IAC' genome. (**a**) Distribution of TEs grouped by major superfamilies. (**b,c**) Insertion age profiles of LTR retrotransposons estimated from sequence divergence. (**d**) Phylogenetic reconstruction of LTR elements, showing diversity and *Gypsy* and *Copia* lineages abundance. All plots were produced with EDTA-GUI.

Data availability

PacBio HiFi and Illumina Hi-C sequencing reads are available in the NCBI SRA under BioProject PRJNA1301450 (accessions SRR34861620 and SRR34861621)[85,86]. Haplotype-resolved genome assemblies are deposited in GenBank under the accessions JBWUIW000000000[87] and JBWUIX000000000[88] for Haplotypes A and B, respectively. The plastid and mitochondrial genomes are available at GenBank (accessions PX373693 and

Genome BUSCO analysis	
Complete	2,023 (99.9%)
Complete and single copy	26 (1.3%)
Complete and duplicated	1,997 (98.6%)
Fragmented	0 (0%)
Missing	3 (0.1%)
Annotation BUSCO analysis	
Complete	2,016 (99.5%)
Complete and single copy	32 (1.6%)
Complete and duplicated	1,984 (97.9%)
Fragmented	2 (0.1%)
Missing	8 (9.4%)
LTR Assembly Index	27.15
Inspector analysis	
Mapping rate (%)	99.98
Depth	106.9063
Structural error	5
Expansion	3
Collapse	2
Haplotype switch	0
Inversion	0
QV	45.50510
Error rate (E, from QV $= -10\log_{10}E$)	0.0028%
Merqury analysis	
K-mer completeness (%)	99.34
Completeness hapA (%)	100
Completeness hapB (%)	100
QV (genome-wide)	70.52
Switch error rate (%)	38.45
Phased blocks (N)	71
Mean block length (bp)	8,750,178
N50 block length (bp)	33,944,669
Maximum block length (bp)	51,005,916
Error rate (E, from QV $= -10\log_{10}E$)	0.000009%

Table 3. Summary of technical validation metrics for the haplotype-phased genome of *C.* × *sinensis* 'Pera IAC'. BUSCO analyses confirm near-complete representation of conserved embryophyte genes in both the assembly and annotation. The LTR Assembly Index (LAI) indicates reference-quality resolution of repetitive regions. Inspector analysis validates high consensus accuracy and structural integrity, while Merqury analysis confirms near-complete k-mer representation, very high QV scores, and effective haplotype phasing.

PX373694)[89,90]. Annotation files (GFF3, CDS, proteins, TE libraries) are accessible via Zenodo[91] and the public genome browser https://plantgenomics.ncc.unesp.br/Genomes/Citrus/.

Code availability

All analyses in this study were conducted using publicly available software, with versions and references specified in the Methods. All pipelines and tools are open-source and available under their respective licenses.

Received: 29 September 2025; Accepted: 7 April 2026;
Published online: 14 April 2026

References

1. FAO – Food and Agriculture Organization of the United Nations. *FAOSTAT: Food and Agriculture Data*, https://www.fao.org/faostat/en/#home.
2. IBGE – Instituto Brasileiro de Geografia e Estatística. Produção agrícola municipal: laranja – Brasil. https://www.cnpmf.embrapa.br/Base_de_Dados/index_pdf/dados/brasil/laranja/b1_laranja.pdf (Rio de Janeiro: IBGE, 2023).
3. Fundecitrus. Tree Inventory of the São Paulo and West-Southwest Minas Gerais Citrus Belt – Snapshot of Groves in March 2025. https://www.fundecitrus.com.br/wp-content/uploads/2025/06/Tree-Inventory-and-Orange-Crop-Forecast-2025-2026.pdf (Fundecitrus, Araraquara, SP, Brazil, 2025).
4. USDA – UNITED STATES DEPARTMENT OF AGRICULTURE. Update to Citrus Annual Report for Brazil (Report BR2025-0002). https://apps.fas.usda.gov/newgainapi/api/Report/DownloadReportByFileName?fileName=Update+to+Citrus+Annual+Report+for+Brazil_Brasilia_Brazil_BR2025-0002.pdf (2025).
5. Teixeira, M. Orange juice prices hit all-time high amid bleak production outlook. *Reuters*, https://www.reuters.com/markets/commodities/orange-juice-prices-hit-all-time-high-amid-bleak-production-outlook-2023-10-31/ (2023).

6. Costa, A. S. & Müller, G. W. Tristeza Control by Cross-Protection: A U.S. Brazil Cooperative Success. *Plant Dis.* **64**, 538–541, https://doi.org/10.1094/PD-64-538 (1980).
7. Wu, G. A. *et al*. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.* **32**, 656–662, https://doi.org/10.1038/nbt.2906 (2014).
8. Xu, Q. *et al*. The draft genome of sweet orange (*Citrus sinensis*). *Nat. Genet.* **45**, 59–66, https://doi.org/10.1038/ng.2472 (2013).
9. Wu, G. A. *et al*. Complex history of admixture during citrus domestication revealed by genome analysis. Report No. LBNL Report #: LBNL-7054E, https://escholarship.org/uc/item/3mh4r937 (Lawrence Berkeley National Laboratory, 2014).
10. Liu, S. *et al*. Origin and de novo domestication of sweet orange. *Nat. Genet.* **57**, 754–762, https://doi.org/10.1038/s41588-025-02122-4 (2025).
11. Wang, N. *et al*. Phased genomics reveals hidden somatic mutations and provides insight into fruit development in sweet orange. *Hort. Res.* **11**, https://doi.org/10.1093/hr/uhad268 (2023).
12. Wang, N. *et al*. Structural variation and parallel evolution of apomixis in citrus during domestication and diversification. *National Science Review* **9**, https://doi.org/10.1093/nsr/nwac114 (2022).
13. Nakandala, U., Furtado, A. & Henry, R. J. Citrus genomes: past, present and future. *Hort. Res.* **12**, https://doi.org/10.1093/hr/uhaf033 (2025).
14. Doyle, J. J. & Doyle, J. L. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* **19**, 11–15 (1987).
15. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
16. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
17. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *iMeta* **3**, e211, https://doi.org/10.1002/imt2.211 (2024).
18. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
19. *Arima Genomics' mapping pipeline*, https://github.com/ArimaGenomics/mapping_pipeline. commit: 2e74ea4.
20. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. In *IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 314–324.
21. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, https://doi.org/10.1093/gigascience/giab008 (2021).
22. Open2C. *et al*. Pairtools: From sequencing data to chromosome contacts. *PLOS Computational Biology* **20**, e1012164, https://doi.org/10.1371/journal.pcbi.1012164 (2024).
23. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btac808 (2023).
24. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nature Plants* **10**, 1184–1200, https://doi.org/10.1038/s41477-024-01755-3 (2024).
25. Citrus Genome Hub. *South Green. Citrus × aurantium 'Cleopatra' reference genome (CITRE)*, https://citrus-genome-hub.southgreen.fr/citrus_reticulata.
26. Droc, G. *et al*. A Super-Pangenome for Cultivated Citrus Reveals Evolutive Features During the Allopatric Phase of Their Reticulate Evolution. *Plant Biotechnol. J.*, https://doi.org/10.1111/pbi.70553 (2026).
27. Martin, G. *et al*. Improvement of the banana "*Musa acuminata*" reference sequence using NGS data and semi-automated bioinformatics methods. *BMC Genomics* **17**, 243, https://doi.org/10.1186/s12864-016-2579-4 (2016).
28. Krzywinski, M. *et al*. Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645, https://doi.org/10.1101/gr.092759.109 (2009).
29. Rasche, H. & Hiltemann, S. Galactic Circos: User-friendly Circos plots within the Galaxy platform. *GigaScience* **9**, https://doi.org/10.1093/gigascience/giaa065 (2020).
30. Siberchicot, A., Bessy, A., Guéguen, L. & Marais, G. A. MareyMap Online: A User-Friendly Web Application and Database Service for Estimating Recombination Rates Using Physical and Genetic Maps. *Genome Biol. Evol.* **9**, 2506–2509, https://doi.org/10.1093/gbe/evx178 (2017).
31. Ollitrault, P. *et al*. Comparative genetic mapping and a consensus interspecific genetic map reveal strong synteny and collinearity within the *Citrus* genus. *Frontiers in Plant Science* **15**, https://doi.org/10.3389/fpls.2024.1475965 (2024).
32. Phytozome. *Citrus clementina v1.0 genome*., https://phytozome-next.jgi.doe.gov/info/Cclementina_v1_0.
33. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958, https://doi.org/10.7717/peerj.4958 (2018).
34. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574, https://doi.org/10.1093/bioinformatics/btab705 (2021).
35. Wu, B. *et al*. A chromosome-level phased genome enabling allele-level studies in sweet orange: a case study on citrus Huanglongbing tolerance. *Hort. Res.* **10**, https://doi.org/10.1093/hr/uhac247 (2023).
36. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 277, https://doi.org/10.1186/s13059-019-1911-0 (2019).
37. Goel, M. & Schneeberger, K. plotsr: visualizing structural similarities and rearrangements between multiple genomes. *Bioinformatics* **38**, 2922–2926, https://doi.org/10.1093/bioinformatics/btac196 (2022).
38. Vuruputoor, V. S. *et al*. Welcome to the big leaves: Best practices for improving genome annotation in non-model plant genomes. *Applications in Plant Sciences* **11**, e11533, https://doi.org/10.1002/aps3.11533 (2023).
39. Costa, M. F. S. *et al*. EDTA-GUI: a plant-optimized graphical implementation of the EDTA pipeline enabling lineage-level classification and analysis. *BMC Genomics* **27**, 229, https://doi.org/10.1186/s12864-026-12588-z (2026).
40. Neumann, P., Novak, P., Hostakova, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mob. DNA* **10**, 1, https://doi.org/10.1186/s13100-018-0144-1 (2019).
41. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hort. Res.* **10**, https://doi.org/10.1093/hr/uhad127 (2023).
42. *Centromics*, https://github.com/ShuaiNIEgithub/Centromics.
43. Lars, G. *et al*. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. *bioRxiv*, 2023.2006.2010.544449, https://doi.org/10.1101/2023.06.10.544449 (2023).
44. Ferrasa, A. *et al*. CitrusKB: a comprehensive knowledge base for transcriptome and interactome of *Citrus* spp. infected by *Xanthomonas citri* subsp. *citri* at different infection stages. *Database (Oxford)* **2020**, https://doi.org/10.1093/database/baaa081 (2020).
45. Brůna, T. *et al*. Galba: genome annotation with miniprot and AUGUSTUS. *BMC Bioinformatics* **24**, 327, https://doi.org/10.1186/s12859-023-05449-z (2023).
46. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btad014 (2023).
47. Lars, G., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566, https://doi.org/10.1186/s12859-021-04482-0 (2021).
48. Keilwagen, J., Hartung, F. & Grau, J. in *Gene Prediction: Methods and Protocols* (ed Martin Kollmar) 161–177 (Springer New York, 2019).

49. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, https://doi.org/10.1186/1471-2105-6-31 (2005).
50. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc.* **1**, e323, https://doi.org/10.1002/cpz1.323 (2021).
51. Haas, B. J. *et al*. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
52. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
53. Zhang, R. G. *et al*. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hort. Res*. **9**, https://doi.org/10.1093/hr/uhac017 (2022).
54. Teufel, F. *et al*. SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025, https://doi.org/10.1038/s41587-021-01156-3 (2022).
55. Käll, L., Krogh, A. & Sonnhammer, E. L. L. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *J. Mol. Biol.* **338**, 1027–1036, https://doi.org/10.1016/j.jmb.2004.03.016 (2004).
56. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489, https://doi.org/10.1093/nar/gkaa1100 (2021).
57. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745, https://doi.org/10.1093/nar/gkv1189 (2015).
58. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49–e49, https://doi.org/10.1093/nar/gkr1293 (2012).
59. Wei, Z. *et al*. RetroScan: An Easy-to-Use Pipeline for Retrocopy Annotation and Visualization. *Front. Genet*. **12**, https://doi.org/10.3389/fgene.2021.719204 (2021).
60. Qiao, X. *et al*. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. *Genome Biol.* **20**, 38, https://doi.org/10.1186/s13059-019-1650-2 (2019).
61. Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J. & Gao, G. PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* **48**, D1104–D1113, https://doi.org/10.1093/nar/gkz1020 (2019).
62. Li, X. *et al*. PlantNLRatlas: a comprehensive dataset of full- and partial-length NLR resistance genes across 100 chromosome-level plant genomes. *Frontiers in Plant Science* **14**, 1178069, https://doi.org/10.3389/fpls.2023.1178069 (2023).
63. Calle García, J. *et al*. PRGdb 4.0: an updated database dedicated to genes involved in plant disease resistance process. *Nucleic Acids Res.* **50**, D1483–D1490, https://doi.org/10.1093/nar/gkab1087 (2021).
64. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2019).
65. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics (Oxford, England)* **30**, 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
66. The Gene Ontology Consortium *et al*. The Gene Ontology knowledgebase in 2023. *Genetics* **224**, https://doi.org/10.1093/genetics/iyad031 (2023).
67. McDonald, A. G. & Tipton, K. F. Enzyme nomenclature and classification: the state of the art. *The FEBS Journal* **290**, 2214–2231, https://doi.org/10.1111/febs.16274 (2023).
68. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419, https://doi.org/10.1093/nar/gkaa913 (2020).
69. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354, https://doi.org/10.1093/nar/gkaa977 (2021).
70. Drula, E. *et al*. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res.* **50**, D571–D577, https://doi.org/10.1093/nar/gkab1045 (2021).
71. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res.* **53**, D672–D677, https://doi.org/10.1093/nar/gkae909 (2024).
72. Kautsar, S. A., Suarez Duran, H. G., Blin, K., Osbourn, A. & Medema, M. H. plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.* **45**, W55–W63, https://doi.org/10.1093/nar/gkx305 (2017).
73. Conesa, A. *et al*. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676, https://doi.org/10.1093/bioinformatics/bti610 (2005).
74. Wang, Y. *et al*. shinyCircos-V2.0: Leveraging the creation of Circos plot with enhanced usability and advanced features. *iMeta* **2**, e109, https://doi.org/10.1002/imt2.109 (2023).
75. Dainat, J. *AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. (Version v1.2.0)*, https://github.com/NBISweden/AGAT (2022).
76. Diesh, C. *et al*. JBrowse 2: a modular genome browser with views of synteny and structural variation. *Genome Biol.* **24**, 74, https://doi.org/10.1186/s13059-023-02914-z (2023).
77. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410, https://doi.org/10.1016/S0022-2836(05)80360-2 (1990).
78. de Abreu, V. A. C. *et al*. Comparative analyses of *Theobroma cacao* and *T. grandiflorum* mitogenomes reveal conserved gene content embedded within complex and plastic structures. *Gene* **849**, 146904, https://doi.org/10.1016/j.gene.2022.146904 (2023).
79. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257, https://doi.org/10.1186/s13059-019-1891-0 (2019).
80. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546, https://doi.org/10.1038/s41587-019-0072-8 (2019).
81. Jin, J. J. *et al*. GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. *Genome Biol.* **21**, 241, https://doi.org/10.1186/s13059-020-02154-5 (2020).
82. Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350–3352, https://doi.org/10.1093/bioinformatics/btv383 (2015).
83. Tillich, M. *et al*. GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11, https://doi.org/10.1093/nar/gkx391 (2017).
84. Greiner, S., Lehwark, P. & Bock, R. OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* **47**, W59–W64, https://doi.org/10.1093/nar/gkz238 (2019).
85. *NCBI Sequence Read Archive* http://identifiers.org/nucleotide:SRR34861620 (2026).
86. *NCBI Sequence Read Archive* http://identifiers.org/nucleotide:SRR34861621 (2026).
87. *NCBI GenBank* http://identifiers.org/nucleotide:JBWUIW000000000 (2026).
88. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/nuccore/JBWUIX000000000 (2026).
89. Alves, M. N. *et al*. *Citrus sinensis cultivar PeraRio chloroplast sequence* http://identifiers.org/nucleotide:PX373693.1 (2026).
90. Alves, M. N. *et al*. *Citrus sinensis cultivar PeraRio mitochondrion sequence* http://identifiers.org/nucleotide:PX373694.1 (2026).
91. Alves, M. N. *et al*. *Data from: Haplotype-resolved genome of Citrus × sinensis 'Pera Rio', Brazil's most widely cultivated sweet orange (Version 1)*, https://doi.org/10.5281/zenodo.17014037 (2026).

92. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126–e126, https://doi.org/10.1093/nar/gky730 (2018).
93. Chen, Y., Zhang, Y., Wang, A. Y., Gao, M. & Chong, Z. Accurate long-read de novo assembly evaluation with Inspector. *Genome Biol.* **22**, 312, https://doi.org/10.1186/s13059-021-02527-4 (2021).
94. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
95. Song, S. *et al.* Molecular cytogenetic map visualizes the heterozygotic genome and identifies translocation chromosomes in *Citrus sinensis*. *Journal of Genetics and Genomics* **50**, 410–421, https://doi.org/10.1016/j.jgg.2022.12.003 (2023).
96. Iwamasa, M. in *Proceedings of the First International Citrus Symposium* Vol. 1 *Chromosome aberrations in citrus in relation to sterility and seedlessness* (ed H. D. Chapman) 175–181 (University of California, 1969).

Acknowledgements

The authors thank São Paulo State University (UNESP) and the *Laboratório Multiusuário Centralizado para Sequenciamento de DNA em Larga Escala e Análise de Expressão Gênica* (FAPESP grants #2009/53984-2, #2016/07679-7, and #2023/17106-8) for bioinformatics support. The Applied Research Center in Innovation and Sustainability of Citriculture (CPA-Citrus; FAPESP grant #24/10713-9) is also acknowledged for supporting this study. Fundecitrus (#04 3108) is acknowledged for providing plant material and additional bioinformatics support. The postdoctoral fellowships of Leny Calvez, Mauro de Medeiros Oliveira, and Vitor Trinca are financed by the São Paulo Research Foundation (FAPESP), Brazil (#24/09984-8, #23/04372-1, and #23/10314-4, respectively). NAW has a CNPq fellowship (306365/2024-3).

Author contributions

Conceptualization: Alessandro M. Varani, Monica Neli Alves, Nelson Arno Wulff, and Eliane Cristina Locali Methodology: Vitor Trinca, Leny Calvez, Patrick Ollitrault, Mauro de Medeiros Oliveira, Alessandro M. Varani, and Monica Neli Alves Investigation: Monica Neli Alves, Vitor Trinca, Mauro de Medeiros Oliveira, and Mateus Santos Formal Analysis: Monica Neli Alves, Vitor Trinca, Leny Calvez, Patrick Ollitrault, and Jesus A. Ferro, Mauro de Medeiros Oliveira Supervision: Patrick Ollitrault, Nelson Arno Wulff, and Alessandro M. Varani Resources: Nelson Arno Wulff and Alessandro M. Varani. Writing – Original Draft: Alessandro M. Varani and Monica Neli Alves Writing – Review & Editing: Monica Neli Alves, Leny Calvez, Vitor Trinca, Mauro de Medeiros Oliveira, Jesus A. Ferro, Eliane Cristina Locali, Patrick Ollitrault, Nelson Arno Wulff, and Alessandro M. Varani.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-07229-9.

Correspondence and requests for materials should be addressed to A.M.V.