scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Near T2T haplotype-resolved genomes of cacao (*Theobroma cacao*) variety CCN51

Huawei Tan[1], Yaxin Lou[1], Yuchen Yan[2], Lyndel W. Meinhardt[3], Bryan Bailey[3], Osman Gutierrez[4], Sunchung Park[3], Stephen P. Cohen[3], Dapeng Zhang[3] ✉ & Yanbin Yin[2] ✉

As a cornerstone of the global chocolate industry, cacao cultivation supports millions of livelihoods. The modern hybrid variety CCN51 is the most widely planted cultivar in Latin America, valued for its exceptional yield, broad disease resistance, and superior adaptability. Here, we present a high-quality, haplotype-resolved, near Telomere-to-Telomere (T2T) genome assembly for CCN51, generated using PacBio HiFi long-reads and Hi-C data. The assembly spans 414.0 Mb and 417.7 Mb for the two haplotypes and exhibits a high k-mer completeness of 99.24%. Our comprehensive annotation predicted 22,941 and 22,948 protein-coding genes in the two haplotypes, respectively. Furthermore, our improved analysis of transposable elements (TEs) revealed a high TE content of 62.36% in CCN51, with a classification rate of 59%, representing a significant improvement over previous reports. This comprehensive genomic resource provides a vital foundation for deciphering the genetic basis of CCN51's exceptional agronomic traits. The genomic data generated from this work will directly support genomics-assisted breeding, facilitating the development of high-yield, high-quality cacao varieties with enhanced resilience to environmental and biological stresses.

Background & Summary

Cacao (*Theobroma cacao*) is grown extensively as the sole source of cocoa butter and powder for the global confectionery industry. Its production primarily occurs on small-scale farms in developing countries across Africa, Asia, and Latin America[1]. Cacao farming predominantly takes place on small-scale farms located in tropical developing nations. It is estimated that five to six million small holder farmers are engaged in cacao farming worldwide, supporting the livelihoods of approximately 40–50 million people[2]. The global production of dry cacao beans was 4.75 million metric tons in 2020 (The International Cocoa Organization, 2020), and the global chocolate industry had an expected retail market value of USD 189.9 billion by 2026[3]. Despite its economic significance, cacao production remains largely within low-input, low-output systems. Most cacao farmers use minimal technology and financial resources, resembling subsistence farming[4]. Challenges such as declining soil fertility, increasing pest and disease damage, and rising labor costs limit cacao sustainability. As a result, rather than increasing yields, the solution to meet growing demand has often been the expansion of production into new regions[1].

Cacao breeding began in the 1940s, with early varieties selected from open-pollinated families[5]. In the 1950s, the discovery of heterosis in cacao led to the development of bi-parental hybrid varieties selected from F1 families[5]. F1 hybrid clone became the primary breeding method around the world and numerous hybrid clones have been created since then[5]. Among them, CCN51 [Colección Castro Naranjal 51] stands out as one of the most successful cacao varieties. Developed in Ecuador in the 1960s and released publicly in 1984, the popularity of CCN51 has surged rapidly, particularly in Latin America due to its high productivity, disease resistance, and adaptability to various environmental conditions[6,7]. Additionally, it is a precocious variety, producing large pods and beans after two years of transplanting. The beans contain a high butter content (54%), which is particularly beneficial for the cacao butter industry[8]. This variety produced an average of 1,300–1,800 kg dry bean per hector

[1]School of Life Sciences, Nanjing University, Nanjing, Jiangsu, 210023, P. R. China. [2]Nebraska Food for Health Center, Department of Food Science and Technology, University of Nebraska-Lincoln, Lincoln, NE, USA. [3]U.S. Department of Agriculture, Sustainable Perennial Crops Laboratory, Beltsville, MD, 2005, USA. [4]U.S. Department of Agriculture, Subtropical Horticulture Research Station, Miami, FL, USA. ✉e-mail: dapeng.zhang@usda.gov; yyin@unl.edu

Metrics	Value
Number of HiFi subreads	1,706,361
Total HiFi data (bp)	24,096,216,362
HiFi data Depth (×)*	57.65
N50 of HiFi subreads (bp)	14,215
Mean length of HiFi subreads (bp)	14,121
Max length of HiFi subreads (bp)	44,154
Number of Hi-C paired-end reads	91,806,025
Total Hi-C data (bp)	27,541,807,500
Hi-C data Depth (×)*	65.89
Hi-C Q20 reads ratio	96.40%

Table 1. Statistics of PacBio HiFi and Hi-C sequencing data for CCN51. *Depth was calculated under the estimate of a genome size of 418 Mb.

in Latin America, which is higher than other economically relevant clones in the region[7]. Under the full-sun cultivation system, dry bean yield of 2,890 kg/ha was reported for CCN51[9]. In addition, resistance to black pod disease caused by *Phytophthora palmivora* and *P. citrophthora* and to witches' broom (WB) (*Moniliophthora perniciosa*) have been reported[10–12]. It is also being used as common source of resistance to Frosty Pod Rot disease, caused by *Moniliophthora roreri* in Latin America (Bailey *et al*., 2018). Moreover, CCN51 has a particularly high content of cacao butterfat with commercial possibilities in different markets[8]. As a result, it has become a key variety for smallholder cacao growers seeking to revitalize their farms, especially in Latin America.

Parentage analysis confirms that CCN51 is a cross between IMC67 and ICS95[13]. IMC67 is a wild population from Iquitos, Peru[14], while ICS95 is a Trinitario variety resulting from the natural hybridization of Criollo and Amelonado[15,16]. CCN51 is self-compatible, a rare trait among modern hybrid cacao varieties. It also produces a larger number of pollinated flowers compared to Nacional varieties. Furthermore, its pollen remains viable for approximately 24 hours, much longer than the 3 to 7 hours typical of Nacional types, providing a significant pollination advantage[17]. Additionally, CCN51 has a low pod index, a crucial yield component for cacao[13].

Despite its favorable traits, CCN51 faces criticism due to its flavor quality. It is often regarded as "bulk cacao" lacking the floral, nutty, and fruity notes found in premium fine-flavor cultivars. Sensory panels describe its taste as having strong acidity, bitterness, and astringency, along with undesirable green flavors. Because CCN51 is typically seen as a low-quality bulk cacao variety, its increasing adoption has sparked controversy. A particular concern is that its widespread use could lead to the decline of traditional varieties, such as Nacional, especially in Ecuador and other Latin American countries, which is both the center of origin and the center of diversity for *T. cacao*[18,19].

In response to concerns about "mono-cropping" of CCN51 and the perceived low quality of cacao beans and decrease of on-farm genetic diversity in Latin America, cacao breeders have focused on developing new varieties using CCN51 as a parent. The high productivity, disease tolerance, and low fertilizer requirements of CCN51 have been incorporated into a series of new varieties, while breeders have also sought to improve flavor profiles through hybridization with fine-flavor Nacional varieties native to Ecuador. Recently, clones such as EETP880 and EETP801 have been released in Ecuador, leveraging the beneficial traits of CCN51 while enhancing flavor quality[6,7,20].

The cacao reference genomes have provided valuable resources for developing advanced breeding tools aimed at improving yield, stress tolerance, and quality attributes in cacao[21–25]. However, the development of these initial genomes has primarily focused on geographically distinct landraces[26–28], wild accessions[29–31], and one specialized vascular streak dieback-susceptible progeny tree[24]. Crucially, few assemblies specifically represent modern cacao varieties, particularly the elite hybrid varieties. In this study, we present a high-quality, *de novo* genome assembly for one of the most significant modern cacao varieties, CCN51 (Table 1). We achieved a complete chromosome-scale, haplotype-resolved genome assembly using PacBio HiFi long reads and Dovetail Hi-C technology. This high-quality genome will be highly useful to improve our understanding of the superior characteristics of CCN51, such as high productivity, earliness, adaptability, diseases resistance, pollen longevity and self-compatibility. The genomic data developed in this research will aid in genomics-assisted cacao breeding, contributing to an effective response to changing biotic and abiotic threats and facilitating the development of new cacao varieties with high yield and superior quality attributes.

Materials and Methods

Plant material and sequencing. The commercial *T. cacao* clone 'CCN51' was collected from the greenhouse of USDA-ARS Sustainable Perennial Crops Laboratory (SPCL), Beltsville, Maryland. Approximately 0.1 g of leaf tissue was harvested for DNA extraction. DNA was subsequently extracted and purified using the phenol: chloroform: isoamyl alcohol (25: 4: 1) method[32].

For PacBio HiFi sequencing, a single library was constructed using the DNA Template Prep Kit 3.0 and DNA/Polymerase Binding Kit. The DNA sample was sequenced on a PacBio Sequel II 8 M SMRT cell, yielding 1.7 Mb reads. This resulted in 24.1 gigabases (Gb) of data for CCN51, providing an average depth of approximately 54.65 × (Table 1).

To enable chromosome-level scaffolding, the Omni-C technique from Dovetail Genomics was used for chromatin conformation capture[33]. The protocol began by fixing chromatin *in situ* using formaldehyde, followed

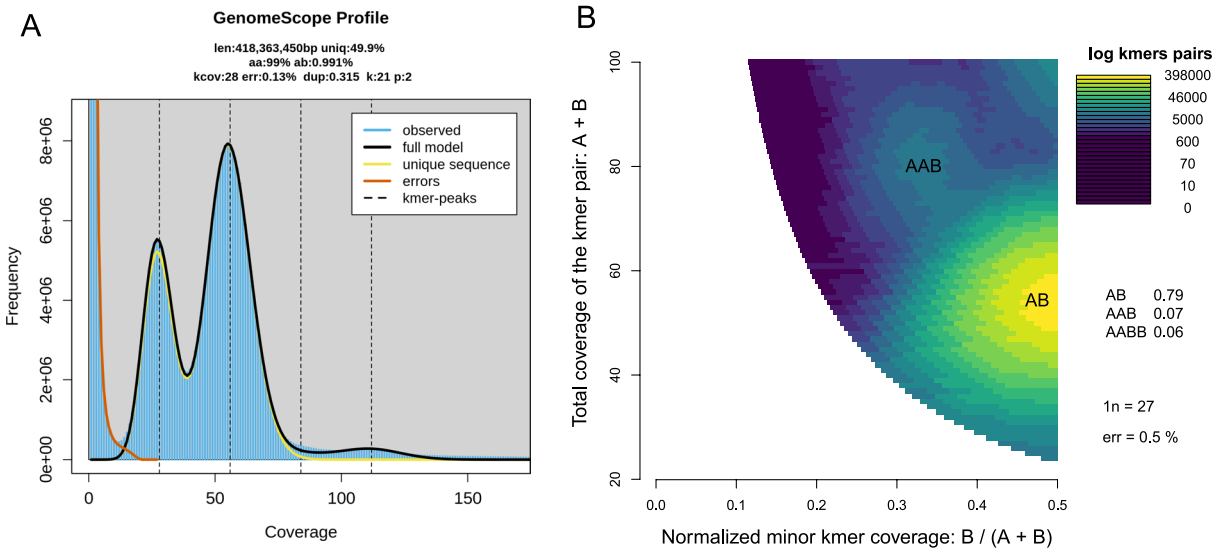

Fig. 1 Estimation of CCN51 genome size by k-mer analysis. (**A**) The figure showed the frequency of 21 k-mers. A total of ~24 G k-mers were analyzed and the major (2^{nd}) peak of k-mer depth is 56. Genome size was estimated as (total k-mer number)/(the volume peak), which is 418 Mb. Two peaks were shown suggesting this is a diploid genome, with a heterozygosity rate at 0.991%. (**B**) Smudgeplot of CCN51 HiFi reads.

by extraction. The fixed chromatin was then digested with DNase I, and the resulting fragments were repaired and ligated to a biotinylated bridge adapter. Subsequent proximity ligation of adapter-containing ends was performed. After reversing the crosslinks and purifying the DNA, biotin was removed from any unincorporated fragments. Sequencing libraries were prepared using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing fragments were isolated using streptavidin beads and amplified via PCR. The libraries were then sequenced on an Illumina HiSeqX system. Then, a total of 91.81 million (Mb) Hi-C reads, representing 27.54 Gb of data (~65.89 × depth), were obtained (Table 1).

***De novo* assembly of cacao CCN51.** To conduct a genome survey of CCN51, we performed k-mer counting (k = 21) with Jellyfish (v2.3.0)[34] to estimate genome characteristics using Genomescope (v2.0)[35]. This analysis predicted the CCN51 genome size to be 418.36 Mb, with a repeat ratio of 49.9% and a heterozygosity rate of 0.991% (Fig. 1A). To further analyze the genome's ploidy and complexity based on k-mer counts, we used Smudgeplot (v0.4)[35], which confirmed that CCN51 has a diploid genome (Fig. 1B).

Prior to genome assembly, PacBio HiFi reads were subjected to a quality control and filtering pipeline. Kraken2[36], with the pluspfp (v20240904) database, was used to identify and remove a small number (22) of reads classified as bacterial or human contaminants. Subsequently, to assemble the organellar genomes, 56,605 HiFi reads were aligned to publicly available plant mitochondrial and chloroplast sequences from the NCBI database using Minimap2[37], and then assembled using Flye (v2.9.5) with the --meta parameter[38].

PacBio HiFi reads of the mitochondrial genome were first filtered out. Raw Hi-C data were first processed for quality control using Fastp v0.23.4[39]. During the genome assembly process, the remaining 1,649,734 clean HiFi reads, along with 91,806,025 Hi-C paired-end reads, were then submitted to Hifiasm (v0.24.0-r702)[40] for primary and haplotype assembly. This process yielded an initial assembly consisting of 264 contigs with a total length of 832.91 Mb from two haplotypes. To further refine the assembly, a BLASTn alignment against the NCBI nt database (November 2023 version) was performed. This step identified 7 contigs as potential bacterial, animal rRNA, or mitochondrial contaminants, which were subsequently removed.

The retained contigs were then anchored into a chromosome-level assembly using the Juicer pipeline[41]. First, Hi-C paired-end reads were aligned to the assembly using Bowtie2[42] to obtain unique mapping and localization information. Next, Juicer was used to filter and retain valid interaction paired reads for scaffolding. Finally, the assembly was manually reviewed and adjusted to correct any obvious placement and orientation errors. This yielded a chromosome-level assembly where 98.07% of the total sequence length was successfully anchored. The final assembly was polished using Racon[43], and chromosome IDs were assigned to maintain consistency with the Matina1–6 and Criollo reference genomes.

The final haplotype-resolved genome assembly for the CCN51 cultivar consists of 10 full-length chromosomes for each haplotype, along with a circular mitochondrial and a circular chloroplast genome. Notably, the scaffold N50 values reached 44.56 Mb and 42.77 Mb for the two genomes. Telomeric sequences were successfully identified for all 10 chromosomes across both haplotypes of the CCN51 genome, confirming the attainment of a near Telomere-to-Telomere (T2T) completion level (Fig. 2A). Compared to the previously haplotype-resolved clone26 assembly[24], this effort successfully obtained all telomeres and significantly reduced the number of sequence gaps. While this represents a major step toward completeness, a small number of complex gaps still

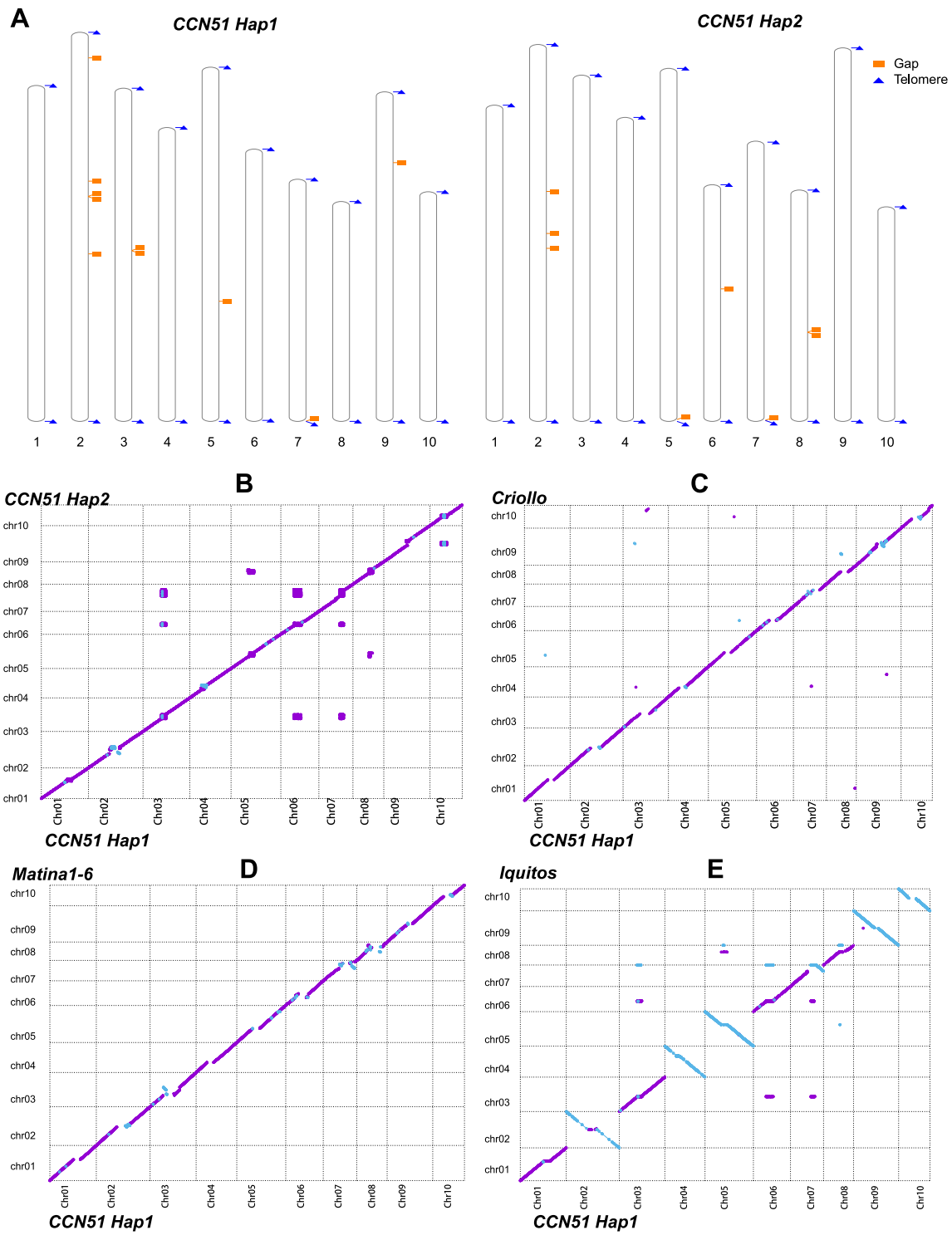

Fig. 2 Two haplotype genomes of CCN51 and comparative genomic synteny with parental cultivars. (**A**) Telomeres are indicated with blue triangles at the two ends of each chromosome. Both haplotype genomes are assembled to the near T2T level with a small number of gaps (orange squares) remained in some chromosomes. (**B**) Synteny between the genomes of CCN51 Hap1 (x-axis) and CCN51 Hap2 (y-axis). (**C**) Synteny between the genomes of CCN51 Hap1 (x-axis) and Criollo (y-axis). (**D**) Synteny between the genomes of CCN51 Hap1 (x-axis) and Amelonado Matina1-6 (y-axis). (**E**) Synteny between the genomes of CCN51 Hap1 (x-axis) and Iquitos IMC67 (y-axis). Synteny plots were generated using Mummer, specifically utilizing the nucmer alignment tool with the filtering parameters: -i 90 (minimum identity of 90%) and -l 20000 (minimum match length of 20 kb).

Assembly	CCN51 Hap1	CCN51 Hap2	Contamana SCA6*	Iquitos IMC67*	Nanay Pound7*	Criollo*	Amelonado Matina1-6*	clone26 maternal**	clone26 paternal**
HiFi long reads (Gigabases)	24.1		21.4	19.1	20.4	—	—	30.3	
Number of scaffolds	111	88	175	170	1,895	431	711	172	309
Length of Genome (bp)	413,957,861	417,693,076	383,477,156	381,649,652	419,275,778	324,879,930	345,993,675	374,335,369	381,631,648
Length of 10 Chromosomes (bp)	405,542,002	410,217,270	352,058,514	351,354,296	338,407,810	314,189,522	330,456,197	338,223,858	340,140,478
GC (%)	33.64	33.48	33.8	33.8	34.9	32.1	32.5	33.73	33.67
N50 (bp)	44,555,205	42,771,200	39,464,999	39,489,205	34,430,447	36,364,294	34,397,752	38,945,508	39,156,173
L50	5	5	5	5	5	5	5	5	5
N90 (bp)	31,060,405	31,282,783	24,455,724	23,705,257	41,995	21,614,486	21,543,242	25,091,101	1,640,581
L90	9	9	10	10	580	9	9	10	12
k-mer completeness (%)	99.24		86.4	82.9	87.6	92.0	74.5	99.15	
k-mer consensus QV	67.4		61.0	61.6	53.0	38.9	32.4	65.9	
BUSCO genome***	99.1%	99.3%	98.4%	98.8%	98.4%	98.1%	98.3%	98.9%	99.2%
BUSCO proteins	98.7%	98.8%	—	—	—	—	—	—	—

Table 2. Assembly statistics of CCN51 genomes and 6 reference genomes. *The statistics for Contamana SCA6, Iquitos IMC67, Nanay Pound7, Criollo (GCF_000208745.1), and Matina1–6 (GCF_000403535.1) assemblies were obtained from the previous publication[31]. **Genome assemblies for clone26 maternal (GCA_041222385.1) and paternal (GCA_041222375.1) were downloaded from GenBank. Statistics were re-calculated to allow direct comparison with the CCN51 assembly. ***BUSCO scores were calculated with genome sequences as input using the eukaryota_odb10 dataset.

persist within some chromosomes. Future genomic assembly efforts must put emphasis on closing gaps in these remaining challenging regions.

Comparative genomic synteny between the two CCN51 haplotype genomes (Fig. 2B) and other cultivars were performed using MUMmer (v4.0)[44] with default parameters. The synteny plot was generated with the filtering parameters: -i 90 (minimum identity of 90%) and -l 20000 (minimum match length of 20 kb). This collinearity analysis confirmed that CCN51 maintains good collinearity with other cultivars (Fig. 2C–E), providing deeper insights into the evolutionary diversification of these key *Theobroma cacao* varieties.

To assess the assembly's quality and completeness, we performed a k-mer-based analysis using Merqury (v1.4.1)[45]. K-mers shared between the HiFi sequencing reads and the assembly confirmed a high k-mer completeness of 99.24%. We also calculated the assembly's consensus base-pair quality value (QV) based on unique k-mers in the assembly, which resulted in a QV of 67.4. These metrics represent the highest k-mer completeness and QV values reported for any published cacao assemblies to date, underscoring the exceptional quality of our CCN51 genome (Table 2). The completeness of the gene space was further validated using BUSCO (Benchmarking Universal Single-Copy Orthologs)[46]. Against the eudicots_odb10 database, 99.1% and 99.3% of the 2,326 single-copy orthologs were identified as complete in the two CCN51 haplotypes, respectively.

Gene prediction and annotation. Protein-coding genes were predicted using a comprehensive pipeline that integrated evidence from transcriptome sequencing and homologous proteins with *ab initio* predictions. For transcriptome evidence, RNA-seq data from *T. cacao* were mapped to the genome with STAR (v2.7.11b)[47] and assembled into transcripts using Cufflinks[48] and Trinity (v2.15.2)[49]. Trinity was run in genome-guided mode using the following command and parameters for *de novo* assembly: Trinity --max_memory 256 G --genome_guided_bam CCN51.RNA.STAR.bam --genome_guided_max_intron 20000 --CPU 48. These transcripts were used to inform gene structure prediction with the PASA pipeline (https://github.com/PASApipeline/PASApipeline). Concurrently, protein sequences from five annotated cacao varieties (Matina1–6, Criollo, Pound7, SCA6, IMC67) were aligned to the CCN51 genome using Miniprot (v0.13-r248)[50] to generate a homologous protein evidence set. *Ab initio* gene models were also generated using Augustus (v3.2.3)[51] and Braker (v3.0.3)[52], while homologous gene predictions were made using Liftoff[53]. All predicted gene models were then merged and evaluated using EVM (EvidenceModeler, v2.0)[54], which assigned evidence-based quality values to produce the final, high-confidence gene set.

In total, 22,941 and 22,948 protein-coding genes were predicted in the two haplotypes of the CCN51 genome, respectively. These genes were subjected to comprehensive functional annotation using a multi-database approach (Table 3). We performed sequence similarity searches against the UniProtKB database using Diamond (v2.1.8.162)[55]. Additionally, we employed eggNOGmapper (v2.1.12)[56] and InterProScan (v5.72-103.0)[57] to annotate Gene Ontology (GO) terms and identify associations with KEGG[58] and Reactome pathways. The InterProScan analysis integrated protein domain and family predictions from five distinct databases: Pfam[59], SMART[60], Superfamily, Gene3D[61], and CDD[62]. For instance, 22,866 proteins (99.61%) of in CCN Hap1 were successfully annotated in at least one database. The number of proteins annotated by each database/tool was as follows: 6,456 (KEGG), 19,937 (KOG), 17,442 (SwissProt), 22,687 (TrEMBL), 22,156 (InterProScan), and 22,825 (NR). We identified orthologous gene pairs across CCN51, Criollo, Matina1–6, IMC67, Pound7, and SCA6 assemblies using OrthoFinder (v3.0.1)[63]. The tool was run with multiple sequence alignment and phylogenetic

Database	CCN51 Hap1		CCN51 Hap2	
	Number of annotated genes	Percentage of annotated genes (%)	Number of annotated genes	Percentage of annotated genes (%)
All genes	22,941	100.00	22,948	100.00
GO	10,493	45.74	10,550	45.97
KEGG	6,456	28.14	6,502	28.33
KOG	19,937	86.91	19,958	86.97
Swissprot	17,442	76.03	17,465	76.11
TrEMBL	22,687	98.89	22,694	98.89
eggNOG	19,937	86.91	19,958	86.97
InterProScan	22,156	96.58	22,192	96.71
NR	22,812	99.44	22,808	99.39
Total annotated genes	22,853	99.62	22,850	99.57

Table 3. Functional annotation of predicted protein-coding genes in CCN51.

inference enabled: orthofinder -f data1 -t 64 -a 64 -M msa -S diamond -A mafft -T raxml-ng. Using the fold change >2 or <0.5, we identified significant differences in gene family size within CCN51. Specifically, the analysis revealed 147 and 358 gene families that showed expansion and contraction, respectively, in the CCN51 genome compared to the other cultivars.

To further assess the completeness of the gene models, we conducted a BUSCO analysis[46]. Against the eudicots_odb10 database, a total of 2,298 complete BUSCOs (98.80%) were identified in the gene models in the CCN51 Hap2 (Table 2), comprising 2,281 single-copy and 17 duplicated orthologs. Only 13 orthologs were fragmented, with just 15 missing from the assembly (Fig. 3A).

We performed tRNA prediction using tRNAscan-SE (v2.0.7)[64] with default parameters, which identified 428 tRNA genes in the Hap1 assembly. The identification of ribosomal RNAs (rRNAs) was conducted using barrnap (v0.9)[65] with the following specific parameters: --kingdom euk –lencutoff 0.8 –evalue 1e-06. For other non-coding RNAs (ncRNAs), sequences were searched and classified against the Rfam database[66]. This analysis resulted in the detection of 132 miRNAs, 817 snoRNAs, 99 snRNAs, and 1,068 other ncRNAs in Hap1 (Table 4).

Repeat annotation and comparisons. To identify transposable elements (TEs) in the CCN51 genome, we constructed a non-redundant, species-specific TE library using a combined *de novo* and homology-based approach: (1) The library was built with the EDTA pipeline (v2.2.1) (parameters: EDTA.pl --genome genome.fa --species others --sensitive 1 --anno 1 --curatedlib SINEbase.fa)[67], integrating tools like MITE-hunter[68] and RepeatModeler[69], and supplemented with elements from SINEbase[70]. (2) To improve the accuracy and breadth of TE classification, we subsequently employed DeepTE (v2022)[71] to classify the remaining unclassified TEs using its pre-trained plant-specific convolutional neural network model. Unclassified Long Terminal Repeat (LTR) retrotransposons were processed separately using the command DeepTE.py -i LTR_unknown.fa -sp P -m_dir Plants_model -fam LTR, while all other remaining TEs were classified with the command DeepTE.py -i Other_unknown.fa -sp P -m_dir Plants_model. Following the DeepTE classification, the newly identified TEs were integrated with the existing RepeatModeler library. (3) The final TE annotation was performed using RepeatMasker (v4.1.2-p1)[72], which was executed through the EDTA pipeline's annotation module (parameters: EDTA.pl --genome genome.fa --step anno --overwrite 1 --anno 1). This three-step process provided a highly comprehensive and accurate final annotation of the transposable elements for CCN51.

Previous studies on *T. cacao* genomes reported repeat content ranging from approximately 52.3% to 61.5%, with a large proportion of these sequences remaining unclassified. Our analysis of the commercial cultivar CCN51 revealed a higher TE content of 62.36%, a proportion exceeding that found in the natural cultivar IMC67 (Table 5, Fig. 4A). A crucial advancement in this study was the use of DeepTE, which enabled the classification of approximately 59% of the TEs, leaving only ~3% unclassified. Among the classified elements, the Gypsy/DIRS1 superfamily was the most abundant, with 96,802 elements accounting for 43.4% of the total TEs in Hap1 (Table 5, Fig. 4A). The remaining ~3% of unclassified TEs are likely attributable to the absence of clear reference information within the pre-trained Plants_model libraries used by both RepeatModeler and DeepTE, highlighting the need for further experimental validation and refined bioinformatics classification in future work.

To explore the evolutionary history and activity of TE families, we calculated the Kimura 2-Parameter (K2P) distance, a proxy for TE age. Our analysis revealed distinct age distributions across different TE families. Overall, LTR elements exhibited the highest sequence identity among all TE classes (Fig. 4B), suggesting recent activity. Further analysis of specific LTR families showed that Copia elements have a recent accumulation peak (K2P distance: 5–25%), indicative of a recent transposition burst. In contrast, Gypsy elements displayed an older, broader distribution with a peak between 25% and 35% (Fig. 4C), suggesting an earlier period of major activity. Similarly, DNA TE families and LINE/L1 and MITE/DTM families all showed older accumulation peaks, clustering around 22–30% and 30%, respectively (Fig. 4D). This analysis demonstrates that TE families within the genome have undergone distinct histories of activity at different evolutionary time points. To reconstruct the accumulation history of intact elements, we further reclassified them using TEsorter (v1.4.6)[73]. This analysis over the past 4.5 million years, the two most prominent LTR families displayed contrasting patterns of activity: LTR/Gypsy elements underwent a significant expansion around 1 million years ago, whereas LTR/Copia elements exhibited a sustained, steady accumulation throughout this entire period (Fig. 4E).

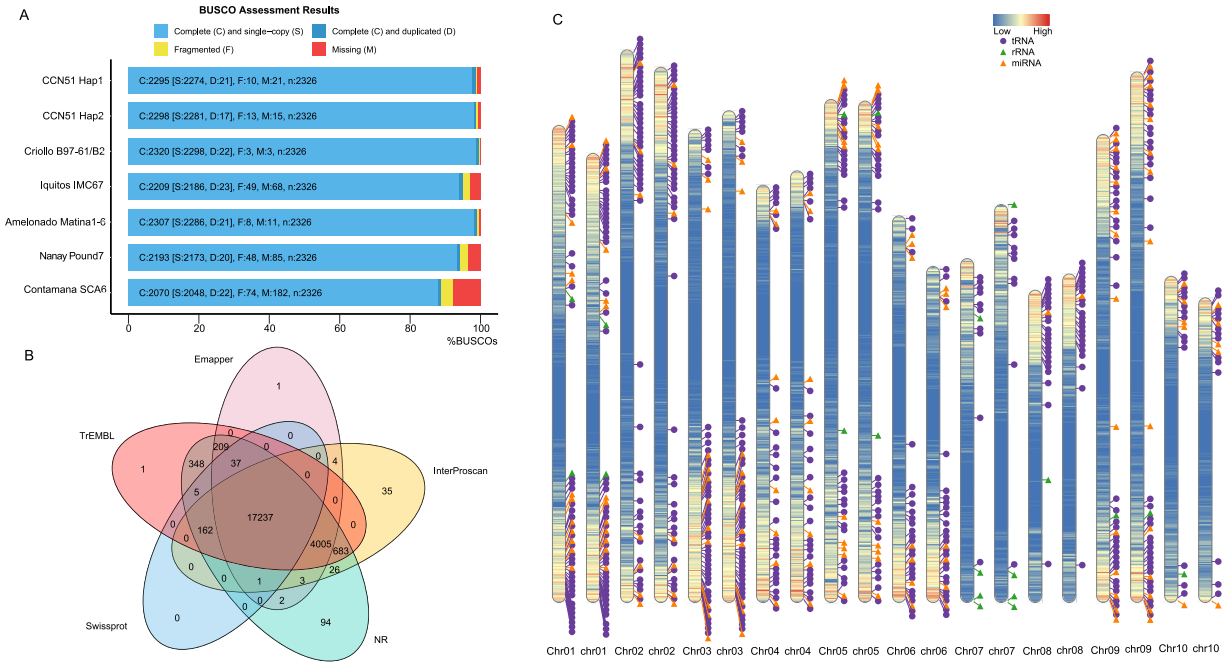

Fig. 3 Gene annotation and assessment. (**A**) BUSCO assessment of CCN51 protein coding genes. (**B**) Annotated protein coding genes in Hap1 by SwissProt, TrEMBL, InterProScan, EggNOG and NR database. (**C**) Protein coding gene density and noncoding RNAs anchored in Hap1 chromosomes.

Type	CCN51 Hap1	CCN51 Hap2
rRNA	1,852	1,142
tRNA	428	426
miRNA	132	132
snoRNA	817	831
snRNA	99	98
Other RNA	1,068	1,080
Total	4,396	3,709

Table 4. Prediction of non-coding RNAs in CCN51.

Class	CCN51 Hap1			CCN51 Hap2		
	No. of elements	Length occupied (bp)	Percentage of sequence	No. of elements	Length occupied (bp)	Percentage of sequence
Retroelements	144,887	215,854,229	52.13%	143,635	220,499,471	52.78%
SINEs	2,466	367,803	0.09%	2,472	371,967	0.09%
LINEs	4,680	3,607,781	0.87%	4,617	3,474,499	0.83%
RTE/Bov-B	213	38,174	0.01%	213	38,114	0.01%
L1/CIN4	4,467	3,569,607	0.86%	4,404	3,436,385	0.82%
LTR elements	137,741	211,878,645	51.17%	136,546	216,653,005	51.86%
Ty1/Copia	34,415	28,297,296	6.83%	34,309	28,278,124	6.77%
Gypsy/DIRS1	96,802	179,805,782	43.42%	95,741	184,629,977	44.20%
DNA transposons	78,526	25,514,116	6.16%	78,751	25,547,765	6.12%
hobo-Activator	1,116	297,740	0.07%	1,119	293,189	0.07%
Tourist/Harbinger	370	342,548	0.08%	352	324,339	0.08%
Rolling-circles	540	301,654	0.07%	523	309,159	0.07%
Unclassified	49,930	12,993,549	3.14%	49,437	12,690,246	3.04%
Total repeats	—	258,213,502	62.36%	—	261,765,353	62.66%

Table 5. Transposable element classes in the CCN51 genome.

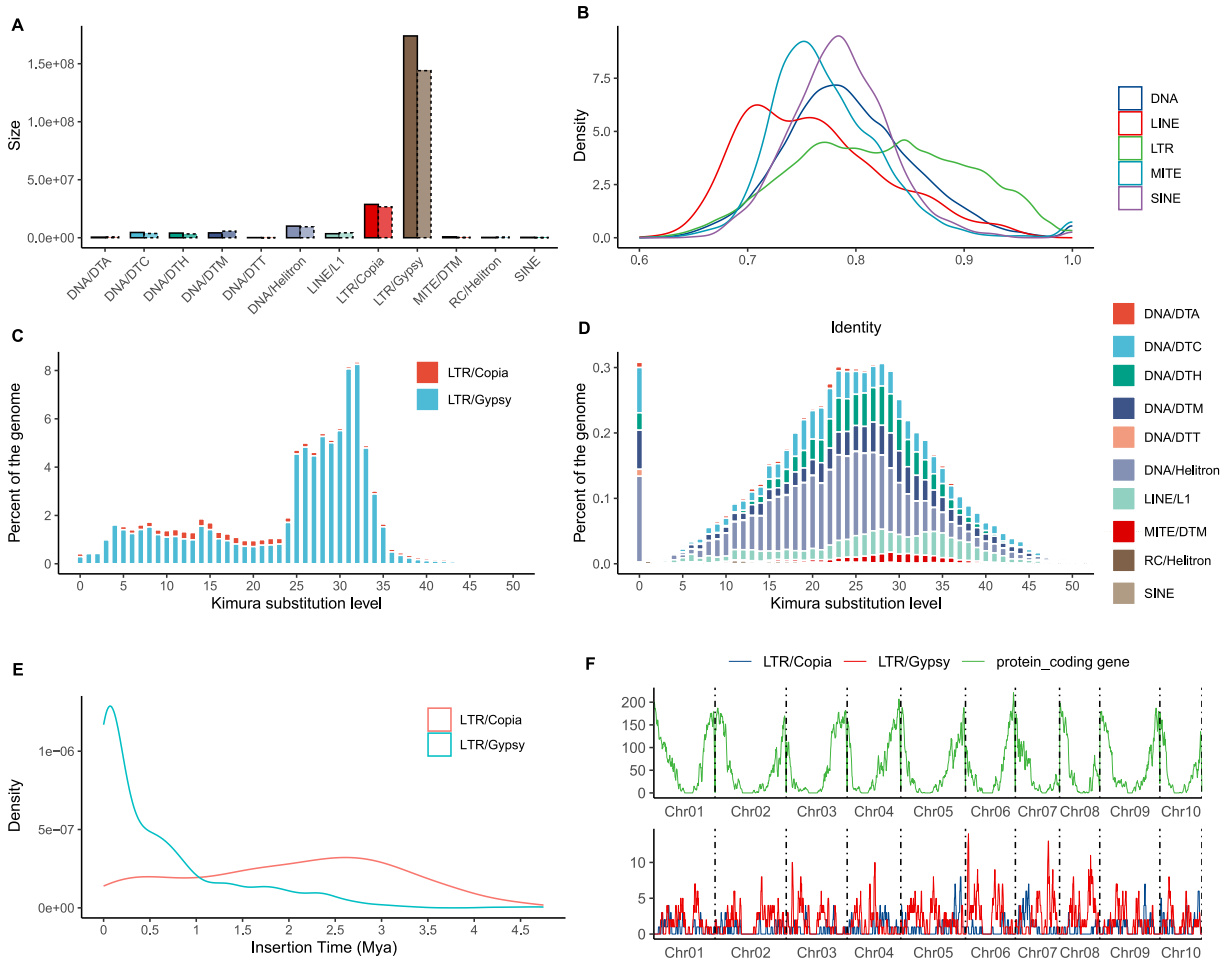

Fig. 4 Transposable elements in CCN51 Hap1 genome. (**A**) The accumulation length of TEs in CCN51 (left) and IMC67 (right). (**B**) The sequence identity within different TE classes. (**C**) The Kimura distance of intact LTR/Copia and LTR/Gypsy. (**D**) The Kimura distance of other TEs. (**E**) The insertion time of intact LTR/Copia and LTR/Gypsy. (**F**) The density of protein coding genes, intact LTR/Copia and LTR/Gypsy in the 10 chromosomes. The analysis uses a sliding window of 1 Mb with a step size of 200 kb to show the distribution of these features.

When protein-coding genes and the two most abundant LTR types were mapped to the 10 chromosomes, we observed an inverse relationship in their distribution. Gene density was higher in the two chromosome ends lower towards the centromeres. In contrast, the density of Gypsy and Copia LTRs was higher in the centromeric and pericentromeric regions (Fig. 4F).

Data Records

The raw PacBio HiFi reads and HiC reads (in FastQ format) are available in the NCBI SRA database (HiFi: SRX30213285[74] and HiC: SRX30213286[75]) under the project number PRJNA1309843. The genome assemblies (in Fasta format for both DNA and protein sequences) and annotations (in GFF format) are accessible at FigShare[76] and at NCBI (Hap1: PRJNA1355227 & JBSIPM000000000[77] and Hap2: PRJNA1355228 & JBSIPN000000000[78]).

Technical Validation

The quality of the assembled genome was assessed using two independent methods. First, we evaluated assembly completeness at the gene level using BUSCO (v5.8.1), aligned against the eudicots_odb10 gene set (n = 2,326, version Jan-2024). The assembly demonstrated a high BUSCO completeness of 99.0%, with 98.6% single-copy and 0.4% duplicated orthologs, while only 0.3% were fragmented and 0.7% were missing. Additionally, we assessed assembly contiguity and accuracy by calculating the mapping rates of sequencing reads. The mapping rates of original PacBio HiFi and Illumina reads to the assembled genome were 99.22% and 95.47%, respectively.

Data availability

The genomic data are deposited to the NCBI BioProject database under the project number PRJNA1309843.

Code availability

No code was developed for implementing a software.

Received: 25 August 2025; Accepted: 15 December 2025;
Published online: 23 December 2025

References

1. Zhang, D. & Motilal, L. Origin, Dispersal, and Current Global Distribution of Cacao Genetic Diversity. in *Cacao Diseases: A History of Old Enemies and New Encounters* (eds Bailey, B. A. & Meinhardt, L. W.) 3–31, https://doi.org/10.1007/978-3-319-24789-2_1 (Springer International Publishing, Cham, 2016).
2. Kongor, J. E., Owusu, M. & Oduro-Yeboah, C. Cocoa production in the 2020s: challenges and solutions. *CABI Agriculture and Bioscience* **5**, 102 (2024).
3. Voora, V., Bermúdez, S. & Larrea, C. Global Market Report: Cocoa (2019).
4. *Achieving Sustainable Cultivation of Cocoa*. https://doi.org/10.1201/9781351114547 (Burleigh Dodds Science Publishing, London, 2018).
5. Eskes, A. B., CFC, Organization, I. C. & International, B. *Collaborative and Participatory Approaches to Cocoa Variety Improvement: Final Report of the CFC/ICCO/Bioversity Project on "Cocoa Productivity and Quality Improvement: A Participatory Approach" (2004-2010)*. (2011).
6. *Sustainable Cacao Cultivation in Latin America*. https://doi.org/10.4324/9781003381761 (Routledge, London, 2024).
7. Jaimez, R. E. *et al*. Theobroma cacao L. cultivar CCN 51: a comprehensive review on origin, genetics, sensory properties, production dynamics, and physiological aspects. *PeerJ* **10**, e12676 (2022).
8. Criollo Nuñez, J., Ramirez-Toro, C., Bolivar, G., Sandoval A, A. P. & Lozano Tovar, M. D. Effect of microencapsulated inoculum of Pichia kudriavzevii on the fermentation and sensory quality of cacao CCN51 genotype. *Journal of the Science of Food and Agriculture* **103**, 2425–2435 (2023).
9. Benjamin, C. S., Dias, L. A. S., Martins, S. C. V., Aucique-Perez, C. E. & Rosmaninho, L. B. C. Unlocking the potential of cacao yield with full sun cultivation. *Sci Rep* **15**, 4368 (2025).
10. Bahia, R. *et al*. Resistance to Black Pod Disease in a Segregating Cacao Tree Population. *Trop. plant pathol.* **40**, 13–18 (2015).
11. Barreto, M. A. *et al*. Detection of genetic resistance to cocoa black pod disease caused by three Phytophthora species. *Euphytica* **206**, 677–687 (2015).
12. Ribeiro, M. A. Q. *et al*. Rootstock × scion interactions on Theobroma cacao resistance to witches' broom: photosynthetic, nutritional and antioxidant metabolism responses. *Acta Physiol Plant* **38**, 73 (2016).
13. Boza, E. J. *et al*. Genetic Characterization of the Cacao Cultivar CCN 51: Its Impact and Significance on Global Cacao Improvement and Production. *Journal of the American Society for Horticultural Science* **139**, 219–229 (2014).
14. Zhang, D., Mischke, S., Johnson, E. S., Phillips-Mora, W. & Meinhardt, L. Molecular characterization of an international cacao collection using microsatellite markers. *Tree Genetics & Genomes* **5**, 1–10 (2009).
15. Bekele, F. & Phillips-Mora, W. Cacao (Theobroma cacao L.) Breeding. in *Advances in Plant Breeding Strategies: Industrial and Food Crops: Volume 6* (eds Al-Khayri, J. M., Jain, S. M. & Johnson, D. V.) 409–487, https://doi.org/10.1007/978-3-030-23265-8_12 (Springer International Publishing, Cham, 2019).
16. Motamayor, J. C., Risterucci, A. M., Heath, M. & Lanaud, C. Cacao domestication II: progenitor germplasm of the Trinitario cacao cultivar. *Heredity* **91**, 322–330 (2003).
17. García Talledo, B., Bazurto Zambrano, A., García Cruzatty, L. & Zambrano Gavilanes, F. Morphology, viability, and longevity of pollen of National Type and Trinitarian (CCN-51) clones of cocoa (Theobroma cacao L.) on the Coast of Ecuador. *Braz. J. Bot* **42**, 441–448 (2019).
18. Ceccarelli, V. *et al*. Conservation and use of genetic resources of cacao (Theobroma cacao L.) by gene banks and nurseries in six Latin American countries. *Genet Resour Crop Evol* **69**, 1283–1302 (2022).
19. Tscharntke, T. *et al*. Socio-ecological benefits of fine-flavor cacao in its center of origin. *Conservation Letters* **16**, e12936 (2023).
20. De la Peña-Armada, R. *et al*. Unique Composition and Sustainability Aspects of the EETP801 Amazonian Cocoa Cultivar vs. CCN51 and Commercial Cocoas. *Beverages* **11**, 93 (2025).
21. Colonges, K. *et al*. Integration of GWAS, metabolomics, and sensorial analyses to reveal novel metabolic pathways involved in cocoa fruity aroma GWAS of fruity aroma in Theobroma cacao. *Plant Physiology and Biochemistry* **171**, 213–225 (2022).
22. Osorio-Guarín, J. A. *et al*. Genome-Wide Association Study Reveals Novel Candidate Genes Associated with Productivity and Disease Resistance to Moniliophthora spp. in Cacao (Theobroma cacao L.). *G3 Genes|Genomes|Genetics* **10**, 1713–1725 (2020).
23. Romero Navarro, J. A. *et al*. Application of Genome Wide Association and Genomic Prediction for Improvement of Cacao Productivity and Resistance to Black and Frosty Pod Diseases. *Front. Plant Sci*. **8**, (2017).
24. Tobias, P. A. *et al*. Parental assigned chromosomes for cultivated cacao provides insights into genetic architecture underlying resistance to vascular streak dieback. *The Plant Genome* **17**, e20524 (2024).
25. Wickramasuriya, A. M. & Dunwell, J. M. Cacao biotechnology: current status and future prospects. *Plant Biotechnology Journal* **16**, 4–17 (2018).
26. Argout, X. *et al*. The cacao Criollo genome v2.0: an improved version of the genome for genetic and functional genomic studies. *BMC Genomics* **18**, 730 (2017).
27. Argout, X. *et al*. The genome of Theobroma cacao. *Nat Genet* **43**, 101–108 (2011).
28. Motamayor, J. C. *et al*. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biology* **14**, r53 (2013).
29. Hämälä, T. *et al*. Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree. *Proceedings of the National Academy of Sciences* **118**, e2102914118 (2021).
30. Morrissey, J., Stack, J. C., Valls, R. & Motamayor, J. C. Low-cost assembly of a cacao crop genome is able to resolve complex heterozygous bubbles. *Hortic Res* **6**, 44 (2019).
31. Nousias, O. *et al*. Three *de novo* assembled wild cacao genomes from the Upper Amazon. *Sci Data* **11**, 369 (2024).
32. Lee, S. B., Milgroom, M. G. & Taylor, J. W. A rapid, high yield mini-prep method for isolation of total genomic DNA from fungi. *Fungal Genetics Reports* **35**, 23 (1988).
33. Putnam, N. H. *et al*. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
34. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
35. 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 (2020).
36. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**, 257 (2019).
37. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
38. Kolmogorov, M. *et al*. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods* **17**, 1103–1110 (2020).
39. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

40. 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 (2021).
41. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *cels* **3**, 95–98 (2016).
42. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
43. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
44. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology* **14**, e1005944 (2018).
45. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245 (2020).
46. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).
47. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Trapnell, C. *et al*. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).
49. Haas, B. J. *et al*. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494–1512 (2013).
50. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
51. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
52. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777 (2024).
53. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
54. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
55. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
56. Jensen, L. J. *et al*. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Research* **36**, D250–D254 (2008).
57. Blum, M. *et al*. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research* **53**, D444–D456 (2025).
58. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27–30 (2000).
59. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412–D419 (2021).
60. Letunic, I., Khedkar, S. & Bork, P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res* **49**, D458–D460 (2021).
61. Lewis, T. E. *et al*. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res* **46**, D435–D439 (2018).
62. Marchler-Bauer, A. *et al*. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* **39**, D225–D229 (2011).
63. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**, 157 (2015).
64. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **49**, 9077–9096 (2021).
65. Loman, T. A Novel Method for Predicting Ribosomal RNA Genes in Prokaryotic Genomes. http://lup.lub.lu.se/student-papers/record/8914064 (2017).
66. Ontiveros-Palacios, N. *et al*. Rfam 15: RNA families database in 2025. *Nucleic Acids Research* **53**, D258–D267 (2025).
67. Su, W., Ou, S., Hufford, M. B. & Peterson, T. A Tutorial of EDTA: Extensive *De Novo* TE Annotator. in *Plant Transposable Elements: Methods and Protocols* (ed. Cho, J.) 55–67, https://doi.org/10.1007/978-1-0716-1134-0_4 (Springer US, New York, NY, 2021).
68. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* **38**, e199 (2010).
69. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
70. Vassetzky, N. S. & Kramerov, D. A. SINEBase: a database and tool for SINE analysis. *Nucleic Acids Research* **41**, D83–D89 (2013).
71. Yan, H., Bombarely, A. & Li, S. DeepTE: a computational method for *de novo* classification of transposons with convolutional neural network. *Bioinformatics* **36**, 4269–4275 (2020).
72. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **25**, 4.10.1–4.10.14 (2009).
73. Zhang, R.-G. *et al*. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research* **9**, uhac017 (2022).
74. *NCBI Sequence Read Archive:* https://identifiers.org/ncbi/insdc.sra:SRX30213285
75. *NCBI Sequence Read Archive:* https://identifiers.org/ncbi/insdc.sra:SRX30213286
76. *FigShare:* https://doi.org/10.6084/m9.figshare.29410400
77. *NCBI Genome Assembly of Hap1:* https://identifiers.org/ncbi/insdc:JBSIPM000000000
78. *NCBI Genome Assembly of Hap2:* https://identifiers.org/ncbi/insdc:JBSIPN000000000

Acknowledgements

We thank Indrani K. Baruah for sample preparation, and Xuehuan Feng for data curation and valuable discussions during the project. This work was primarily funded by the United States Department of Agriculture (USDA)/Agricultural Research Service (ARS) awards [58-8042-9-089, 58-8042-3-076], and partially by the National Institutes of Health (NIH) R01 award [R01GM140370]. The mention of trade names or commercial products in this publication is for informational purposes only and does not imply endorsement or recommendation by the U.S. Department of Agriculture. The USDA is an equal opportunity provider and employer.

Author contributions

D.Z., L.W.M., B.B. and Y.Y. conceived and designed the project. B.B., O.G., S.P.C., L.W.M., I.B. collected the plant materials and generated the sequencing data. H.T., Y.L. and Y.Yan performed the data analysis. H.T., D.Z., S.P. and O.G. draft the manuscript, H.T., D.Z., Y.Y. finalized the manuscript. All authors contributed and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to D.Z. or Y.Y.

Reprints and permissions information is available at www.nature.com/reprints.

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