scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A T2T reference genome and haplotype-resolved assembly of cultivar 'Guozhixian' for future apricot breeding

Qiuping Zhang[1,3]✉, Tsan-Yu Chiu[2,3], Ning Liu[1], Yuping Zhang[1], Chunhai Chen[2], Ming Xu[1], Shuo Liu[1], Yujun Zhang[1], Xiaoxue Ma[1] & Weisheng Liu[1]✉

Apricots (*Prunus armeniaca* L.) are valued for their flavor and appearance. Recent advancements have enabled the first Telomere-to-Telomere (T2T) genome assembly for the hybrid cultivar *P. armeniaca* L. 'Guozhixian (GZX)', derived from 'Sungold' (Western) and 'Chuanzhihong' (Eastern). The high-quality T2T genome assembly is 253.5 Mb in length with an N50 contig size of 31 Mb. Its completeness, assessed using BUSCO, is 98.9%. This assembly surpasses its parent genomes, filling all gaps-unlike 'Chuanzhihong' (262 gaps) and 'Sungold' (172 gaps)-and fully resolves telomeres and centromeres. The haplotype-resolved genomes provide a comprehensive genetic blueprint, unlocking deeper insights into apricot biology. This resource will enhance the study of inherited traits and facilitate precision breeding strategies to develop superior cultivars. The T2T genome marks a major milestone in apricot genomics, overcoming previous challenges posed by the genome's size and complexity.

Background & Summary

Apricot (*Prunus armeniaca* L.) is an economically important fruit crop, valued for its attractive skin of golden yellow and juicy and aromatic fruit flesh[1]. They are cultivated for their fruits and kernels in the Mediterranean region, the Middle East, Caucasus, Central Asia, and China with a yield of around 4.1 million tons annually[2]. They originated in the upper reaches of Yellow River of China and the earliest cultivation evidence can be dated back to Xia Dynasty (about 2000 B.C.)[3]. Contrary to long-standing beliefs, the current study reveals that apricot domestication events occurred independently in China, Europe, and Central Asia[4]. Genomic analysis revealed that selection affected only a small proportion of the apricot genome, a pattern consistent with what is typically observed in perennial crops[5]. Notably, the region under selection differ between European and Chinese cultivated apricots, suggesting that similar phenotypic traits may have arisen through distinct genetic pathways than convergent evolution[5].

The field of apricot breeding and crop improvement has witnessed significant advancements in recent years, driven by the integration of traditional breeding methods with modern genomics and biotechnology[6]. This has not only accelerated the development of new cultivars with improved traits but has also contributed to the overall sustainability and resilience of apricot agriculture. While annual crops often exhibit pronounced genetic variations due to domestication, perennial fruit trees like apricots, with their extended juvenile phases and high outcrossing rates, pose unique challenges[7]. Nevertheless, human selection has also shaped traits related to reproduction (self-incompatible), fruit quality, growing traits, and stress response in apricots[8–11]. However, most genome assemblies in Rosaceae (except Peach Genome V2.0 and cold-tolerant apple pollen-tissue cultured line sequencing) are constructed with plant materials with heterozygous genetic backgrounds (www.rosaceae.org)[12]. This results in "chimeric" genome data, where only one set of parental genetic information is mixed due to the diploid nature of these plants. A well assembled genome is emerging as new tool in genomic research and brings new opportunities to gain a deeper understanding of these regions.

To address the current storage challenges faced by domestic apricot varieties in China, we proposed a novel breeding approach, integrating the desirable traits of Western varieties 'Sungold' (attractive appearance, firm

[1]Liaoning Institute of Pomology, Yingkou, 115009, China. [2]BGI Genomics, Shenzhen, 518083, China. [3]These authors contributed equally: Qiuping Zhang, Tsan-Yu Chiu. ✉e-mail: lbzhangqiuping2@163.com; wsliulaas@163.com

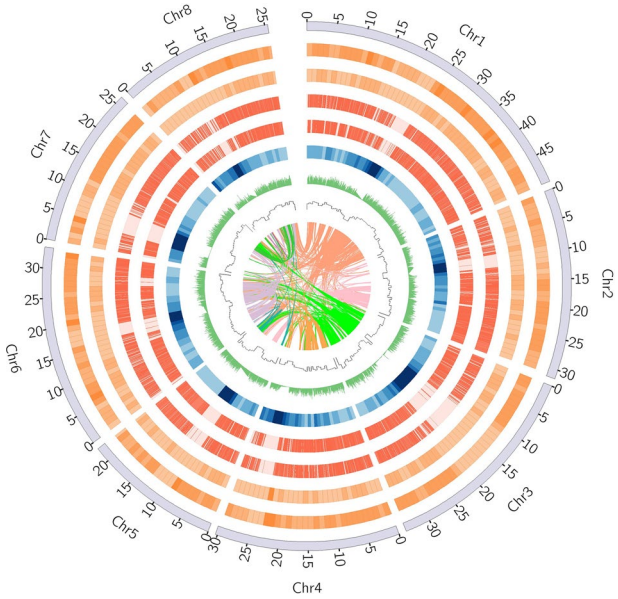

Fig. 1 Circos plot of the GZX genome. From inside to outside, I: collinear regions within the GZX genome; II: gene density in nonoverlapping 1 Mb windows (histograms); III: the median TPM values of genes across 12 transcriptome samples (lines); IV: percent coverage of repetitive sequences in nonoverlapping 1 Mb windows (heat maps); V: the homologous regions between GZX genome and Chuanzhihong genome. White indicates the absence of homologous regions between the two genomes (heat maps); VI: the homologous regions between GZX genome and Sungold genome. White indicates the absence of homologous regions between the two genomes (heat maps); VII the SNP density between GZX genome and Chuanzhihong genome (heat maps); VIII: the SNP density between GZX genome and Sungold genome (heat maps); IX: 8 chromosomes. Lengths are shown in Mb.

Genome feature	GZX_Primary	GZX_HapA	GZX_HapB	Sungold (♂)	Chuanzhihong (♀)
Assem Chr #	8	8	8	8	8
Contig #	8	8	8	180	270
Gap #	0	0	0	172	262
Contig N50 (bp)	31,483,207	29,602,923	30,921,754	7,134,962	1,018,044
Contig N90 (bp)	25,539,148	23,904,160	24,263,261	1,919,751	313,287
Scaffold N50 (bp)	31,483,207	29,602,923	30,921,754	26,110,760	25,125,992
Scaffold N90 (bp)	25,539,148	23,904,160	24,263,261	21,140,521	18,857,615
Gene #	30,543	28,461	28,689	32,636	30,436
LAI	19.25	20.71	19.54	18.59	19.55

Table 1. Summary of the *Prunus armeniaca* L. 'GZX' genome assembly data. Note: Assem Chr #: Number of assembled chromosomes; Contig #: Contig Number; Gene #: Number of genes; LAI: LTR (long terminal repeats) Assembly Index.

flesh) with those of Chinese local varieties 'Chuanzhihong' (strong aroma, nice flavor, broad adaptability)[13], leading to the successful development of the new apricot cultivar *P. armeniaca* L. 'GZX'[14], combing the characteristics of strong flavor, attractive appearance, firm flesh and wide adaptability. Therefore, unraveling the structural variations in the genomes of *P. armeniaca* L. 'GZX' and its parental sources provides valuable genetic insights for future apricot breeding.

A gap-free reference genome named 'GZX' was achieved, containing 8 chromosomes with a total length of 253,575,856 bp (Fig. 1 and Table 1). This high-quality GZX genome resolved two different haplotypes with 238,646,170 bp (GZX_HapA) and 248,939,259 bp (GZX_HapB) respectively and each assembly contained a total of 8 contigs (Fig. 2 and Table 1). Compared to its parent genomes, the newly assembled GZX genome exhibits substantial advancements in completeness and contiguity. The GZX_Primary, GZX_HapA, and GZX_HapB assemblies are highly contiguous, with each chromosome represented by a single contig. In contrast, the 'Sungold' and 'Chuanzhihong' assemblies display higher fragmentation, as indicated by the increased number of contigs per chromosome (Table 1).

The 'Chuanzhihong' genome has 262 gaps acrossing 8 linkage groups, suggesting unresolved regions likely caused by repetitive or complex sequences (Table 1). Similarly, the 'Sungold' genome contains 172 gaps (Table 1). In contrast, the GZX genome has successfully resolved all gaps, resulting in a fully continuous assembly without

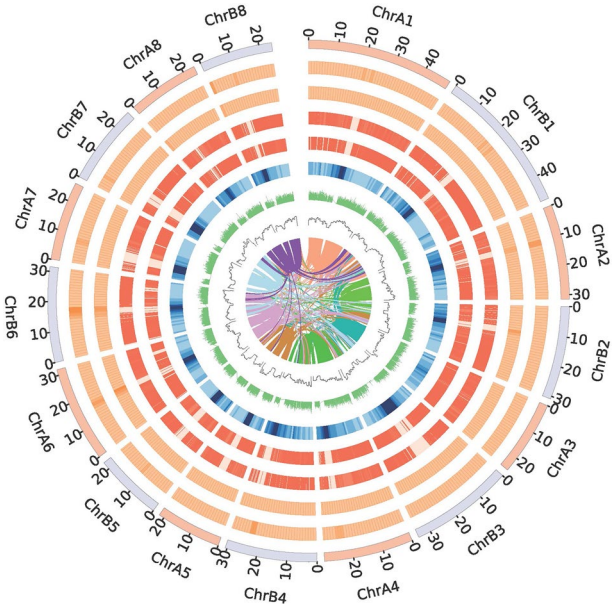

Fig. 2 Circos plot of two haplotypes. From inside to outside, I: collinear regions between haplotype A and haplotype B; II: gene density in nonoverlapping 1 Mb windows (histograms); III: the median TPM values of genes across 12 transcriptome samples (lines); IV: percent coverage of repetitive sequences in nonoverlapping 1 Mb windows (heat maps); V: the homologous regions between haplotypes and Chuanzhihong genome. White indicates the absence of homologous regions between the two genomes (heat maps); VI: the homologous regions between haplotypes and Sungold genome. White indicates the absence of homologous regions between the two genomes (heat maps); VII: the SNP density between haplotype A and haplotype B (heat maps); VIII: the SNP density between haplotypes and GZX (heat maps); IX: 8 chromosomes. Lengths are shown in Mb.

unresolved regions. This represents a significant improvement over both 'Chuanzhihong' and 'Sungold', underscoring the enhanced quality and completeness of the GZX genome. These advancements provide a more reliable and comprehensive genomic resource for future research.

Methods

Plant material and sequencing. The young leaves from *Prunus armeniaca* 'GZX' were collected from the National Germplasm Repository for Plums and Apricot (Xiongyue, Liaoning, China). The leaves were carefully harvested for genomic DNA extraction, employing a modified cetyltri-methylammonium bromide (CTAB) method[15]. The quality and quantity assessment of the isolated DNA were carried out using a NanoDrop 2000 (Thermo Scientific, CA, USA) and a Qubit 2.0 Fluorometer (Life Technologies, CA, USA). After purification with the Qiagen genomic kit (Qiagen, 13343), approximately 5 μg of *Prunus armeniaca* 'GZX' DNA was used to construct short DNA insert size (~350 bp) libraries with the MGIEasy Universal DNA Library Prep Kit. Additionally, 20 kb PacBio HiFi sequencing libraries were generated using the SMRTbell Prep Kit 2.0. The short libraries underwent sequencing on an MGI-T7 platform with 150 bp paired-end reads. The parental genomes were obtained from the Genome Database for Rosaceae (GDR; https://www.rosaceae.org/). Specifically, the Western cultivar 'Sungold' corresponds to *Prunus armeniaca* Sungold Whole Genome v1.0, and the Eastern cultivar 'Chuanzhihong' corresponds to *Prunus armeniaca* Genome v1.0.

For the construction of ultra-long Nanopore libraries, approximately 10 μg of gDNA with a size of about 100 kb was selected using the SageHLS HMW library system (Sage Science, USA). This DNA was utilized for library construction with the ONT 1D Sequencing Kit (SQK-LSK109). Sequencing of PacBio HiFi and ultra-long ONT libraries was performed on the PacBio Sequel II platform and Nanopore PromethION sequence, respectively. Subreads were generated and processed using the CCS algorithm of SMRTLink (v8.0.0).

***De novo* genome assembly.** In order to assemble a high-quality genome of *Prunus armeniaca* L. 'GZX', different sequencing reads including HIFI reads, ONT Ultra long reads, Hi-C reads and DNBseq reads were generated. A total of 34 Gb (~ 135x coverage) HiFi reads using the Pacbio sequel II platform with N50 of 13.781 Kb were generated. A contig genome was assembled using HIFI reads with Hifiasm v0.19.5 with default parameters[16]. Further, the Hi-C data were aligned to this contig genome with BWA software (v0.7.12)[17] and the alignment information was processed by Juicer (v1.5.6)[18]. Last, the contig sequences were arranged into chromosome sequences using the 3D DNA (v180922)[19], and a chromosomal-level genome was achieved. We also assembled another chromosomal-level genome using ONT Ultra-long reads. A total of 68.09 Gb (~272x coverage) using Oxford Nanopore Technology (ONT) ultra-long reads with N50 of 60.431 Kb were obtained. The initial contig genome was assembled by using Nextdenovo (v2.5.0)[20] and the chromosomal-level assembly was done with Hi-C reads using the same procedures as described with the contig genome based on HIFI reads. Comparing the N50 of these two assemblies, the chromosomal-level genome based on HIFI reads shows higher continuity. Thus, we

built the final T2T genome based on assemblies built with the HIFI reads and used the ONT ultra-long reads to close the gap and correct the wrong assemblies using JUICEBOX Assembly Tools (v2.15.07)[21] and LR_gapcloser[22] with the parameter '-r 2 -s nanopore -m 1000000 -v 10000'. We generated two high-contiguity haplotype-resolved assemblies (GZX_HapA and GZX_HapB) using PacBio HiFi reads via Hifiasm, which inherently performs phasing during assembly.

In order to grade a gap-free, telomere-to-telomere and structurally stable primary assembly version (GZX_primary), we prioritized the structures that showed stronger Hi-C interaction signals and greater alignment consistency with ultra-long ONT reads for inclusion in the primary assembly, the phased contig set with the highest overall continuity (contig N50 = 31.5 Mb), completeness (BUSCO score: 98.9%), and assembly quality (QV score = 39.7, LAI = 19.25). When conflicting structural configurations were present between haplotypes, the resolution was guided by the Hi-C interaction maps and manual inspection in Juicebox Assembly Tool. Furthermore, alignment with parental genomes ('Sungold' and 'Chuanzhihong') helped identify chimeric or misassembled regions, allowing correction via gap closure and contig reordering using long-read data and Hi-C scaffolding.

Repeat annotation. Repetitive elements were identified through a combination of homology-based and *de novo* methods using default parameters. Homology-based sequence alignment method relies on the RepBase database[23]. The RepeatMasker[24] and RepeatProteinMask are used to identify sequences similar to known repetitive sequences and classify them accordingly. *De novo* prediction method involves the use of sequence or structural features of repetitive sequences or transposons themselves to construct algorithms or software for sequence recognition. The RepeatModeler[25] and LTRharvest[26] are used to establish a *de novo* repeat sequence library, followed by prediction using RepeatMasker[24]. Additionally, Tandem Repeats Finder[27] is used to identify tandem repeat sequences in the genome. The advantage of *de novo* prediction method lies in its ability to predict based on the structural features of transposon elements themselves, independent of existing transposon databases, thus enabling the discovery of unknown transposon elements.

Comparative analysis of repetitive elements across the genomes revealed that the GZX assemblies (Primary, HapA, and HapB) consistently exhibit higher proportions of repetitive elements compared to their parent genomes, Sungold and Chuanzhihong. The *de novo* repeat model identified the largest fraction of repetitive elements in all genomes, with GZX_Primary showing the highest proportion at 53.83%. Similarly, Tandem Repeats Finder detected significantly more repeats in the GZX assemblies (13.54–15.95%) compared to Sungold (4.18%) and Chuanzhihong (5.02%). In contrast, Repeatmasker[24] and Proteinmask repeats showed relatively consistent proportions across all genomes, ranging from 11.04–12.28% and 5.45–6.09%, respectively. These findings provide valuable insights into the structural and evolutionary characteristics of the genomes.

A detailed breakdown of transposable elements (TEs) further highlighted distinct patterns across the genomes (Table 2). Long Terminal Repeat (LTR) elements were the most abundant TEs, ranging from 35.15% in Chuanzhihong to 45.22% in GZX_Primary. DNA transposons were the second most prevalent, with proportions ranging from 4.06% in GZX_Primary to 7.44% in Chuanzhihong (Tables 2, 3). LINE and SINE elements were less abundant, contributing 0.46–1.14% and 0.01–0.07% to the genomes, respectively. Unclassified or novel repeats also represented a notable fraction, ranging from 7.34% in GZX_HapB to 8.26% in Chuanzhihong (Table 3).

Overall, the GZX assemblies displayed higher proportions of TEs compared to Sungold and Chuanzhihong. This trend was particularly evident in LTR and DNA elements, which were more abundant in the GZX genomes. The higher TE content in the GZX assemblies may reflect differences in genome complexity, evolutionary history, or assembly quality. These findings underscore the significant role of TEs in shaping genome structure and provide critical insights into the genomic diversity and dynamics of these assemblies.

Using the seven-base telomere repeat sequence ('AAACCCT') as a sequence query, we identified all 16 telomeres of the GZX genome (Fig. 3). Plant centromeres are usually composed of tandem repeat monomers (or satellite DNA), with thoroughly investigating the architecture of centromeres is crucial for gaining insight into genome stability and the process of cell division. We employed Tandem Repeats Finder[27] and HMMER[28] to identify all centromeres across each chromosome of the GZX genome. The length of these putative centromeres ranged from 2,496,729 bp to 7,701,153 bp.

Gene prediction and annotation. Homolog and transcriptome alignment algorithms were used to predict genes. For transcriptome prediction, RNA seq data from 'Yinxiangbai' were assembled from a hisat2-stringtie pipeline[29] and mapped onto the GZX genome. Then we use featureCounts[30] to compute the count of reads aligned to gene regions and subsequently determine the gene expression levels using TPM calculation formula[31]. TPM computation method involves dividing the read count by the length of each gene (measured in kilobases) to derive reads per kilobase (RPK). The sum of all RPK values in the sample is then divided by 1,000,000 to yield the 'per million' scaling factor. Proportions of genes that could be functionally annotated and transcriptionally detected in GZX_Primary, GZX_HapA and GZX_HapB (Fig. 4J).

Homolog protein comparison was conducted using GeMoMa[32]. The PASA[33] annotation was performed to combine gene model evidence from these two methods and obtain a final gene set. A total of 30,543 protein-coding gene models were fully annotated (Table 4). The protein-coding genes were functionally annotated based on seven databases including NR, Swissprot, KEGG, KOG, TrEMBL, Interpro and GO. Specifically, 88% of the genes in the gene sets of the T2T genome and the two haplotype genomes were successfully functionally annotated in the examined databases. Approximately 40.20% of the genes (corresponding to 12,847, 12,277, and 12,219 genes in the T2T genome and the two haplotype genomes, respectively) were co-annotated across the five major databases: Swissprot, KEGG, KOG, TrEMBL, and Interpro (Fig. 4F–H). The average length of a coding sequence (CDS) is 986 bp. An average exon length is 236 bp, with a mean of four exons per gene. The

GZX_Primary	Type	Repbase TEs % in genome	TE protiens % in genome	De novo % in genome	Combined TEs % in genome
	DNA	4.06	0.62	3.24	6.24
	LINE	0.46	0.18	0.58	0.95
	SINE	0.01	0.00	0.03	0.04
	LTR	7.05	4.79	44.68	45.22
	Other	0.00	0.00	0.00	0.00
	Unknown	0.00	0.00	7.38	7.38
	Total	11.42	5.59	53.80	54.37
GZX_HapA	DNA	4.05	0.65	2.89	6.20
	LINE	0.44	0.19	0.36	0.71
	SINE	0.01	0.00	0.03	0.04
	LTR	6.92	4.62	42.05	42.58
	Other	0.00	0.00	0.00	0.00
	Unknown	0.00	0.00	7.62	7.62
	Total	11.25	5.47	50.75	51.42
GZX_HapB	DNA	3.99	0.64	3.17	6.17
	LINE	0.46	0.27	0.53	0.93
	SINE	0.01	0.00	0.03	0.04
	LTR	6.75	4.54	43.55	44.13
	Other	0.00	0.00	0.00	0.00
	Unknown	0.00	0.00	7.34	7.34
	Total	11.04	5.45	52.72	53.41

Table 2. Overview of repetitive sequence composition in the genome of *Prunus armeniaca* L. 'GZX' and of its parents. Note: This statistical table does not contain Tandem Repeats, some elements may partly include another element domain. *Combined: the non-redundant consensus of all repeat prediction/classification methods employed. LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements; LTR, long terminal repeat.

Sungold	Type	Repbase TEs % in genome	TE protiens % in genome	De novo % in genome	Combined TEs % in genome
	DNA	4.36	0.60	3.47	6.78
	LINE	0.46	0.12	0.46	0.81
	SINE	0.01	0.00	0.05	0.07
	LTR	7.30	4.87	35.29	35.92
	Other	0.00	0.00	0.00	0.00
	Unknown	0.00	0.00	8.00	8.00
	Total	11.96	5.59	44.88	45.59
Chuanzhihong	DNA	4.43	0.72	4.34	7.44
	LINE	0.48	0.29	0.76	1.14
	SINE	0.01	0.00	0.03	0.04
	LTR	7.52	5.08	35.15	35.80
	Other	0.00	0.00	0.00	0.00
	Unknown	0.00	0.00	8.26	8.26
	Total	12.28	6.09	45.98	46.77

Table 3. Overview of repetitive sequence composition in the genome of *Prunus armeniaca* L. 'Sungold' and 'Chuanzhihong'. Note: This statistical table does not contain Tandem Repeats, some elements may partly include another element domain. *Combined: the non-redundant consensus of all repeat prediction/classification methods employed. LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements; LTR, long terminal repeat.

predicted gene was evaluated based on the embryophyte database using the BUSCO software[34], which delivered a score of completeness of 92.5% (Fig. 4I).

Chromosomal synteny analysis. To reveal the genomic differences contributing to the formation of variety characteristics, we align the genome of 'GZX', 'Sungold', and 'Chuanzhihong' using MUMmer4[35] with default parameters (Fig. 3). The results indicated that the 'GZX' and 'Sungold' showed much higher collinearity with a

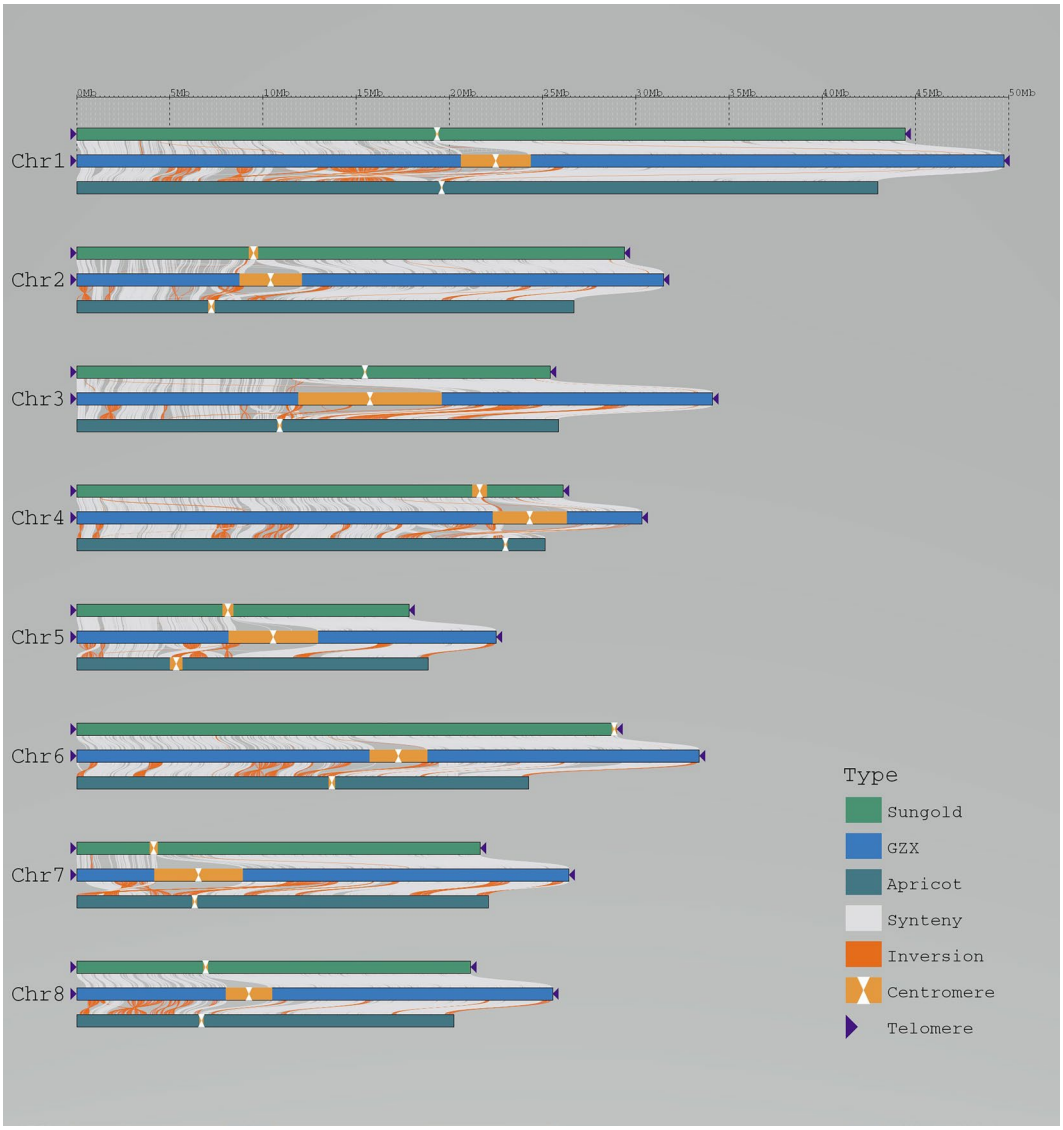

Fig. 3 Structure of T2T and gap-free chromosomes in GZX. All 8 chromosomes of GZX, Sungold, and Chuanzhihong are drawn to scale. Deep purple triangles indicate the presence of telomere sequence repeats. Yellow dumbbell shapes represent the locations and sizes of centromeric regions. Collinearity analysis between GZX, Sungold, and Chuanzhihong with syntenic regions are shown as gray lines, and inversions as orange lines.

total of 633 syntenic regions (~187 Mb). In contrast, there were many more chromosomal inversions between GZX' and 'Chuanzhihong' (Fig. 3).

Data Records

The sequencing data of *Prunus armeniaca* L. 'GZX' has been deposited in the NCBI Sequence Read Archive (SRA) under the project identifier PRJNA1233628 with the accession number SRP569754[36]. Specifically, the Illumina genomic survey sequencing data, PacBio genomic long-read data, Nanopore genomic long-read data, and Hi-C data have been deposited in NCBI SRA with the respective accession numbers SRR32660270, SRR32660273, SRR32660272, and SRR32660271. Additionally, the primary assembly (GZX_Primary), haplotype 1 (GZX_HapA) and haplotype 2 (GZX_HapB) of 'GZX' have been made available in GenBank under the respective BioProjects PRJNA1233628, PRJNA1301539 and PRJNA1301570, with the corresponding accession numbers GCA_053572015.1[37], GCA_053572045.1[38] and GCA_053572545.1[39]. Furthermore, the files of genome assemblies and gene structure annotations have been archived in the Figshare repository[40].

Technical Validation

This high fidelity of the assembly was supported by high mapping rates of 99.83% (HIFI) and 99.97% (ONT). Hi-C reads were mapped to the final version of the assembly using Juicer[18]. In the Hi-C heatmap, there were strong interactive signals of the 8 chromosomes around the diagonal, suggesting that these assemblies had no obvious chromosome assembly errors (Fig. 4A–C). In addition, 98.9% of a genome BUSCO (Benchmarking

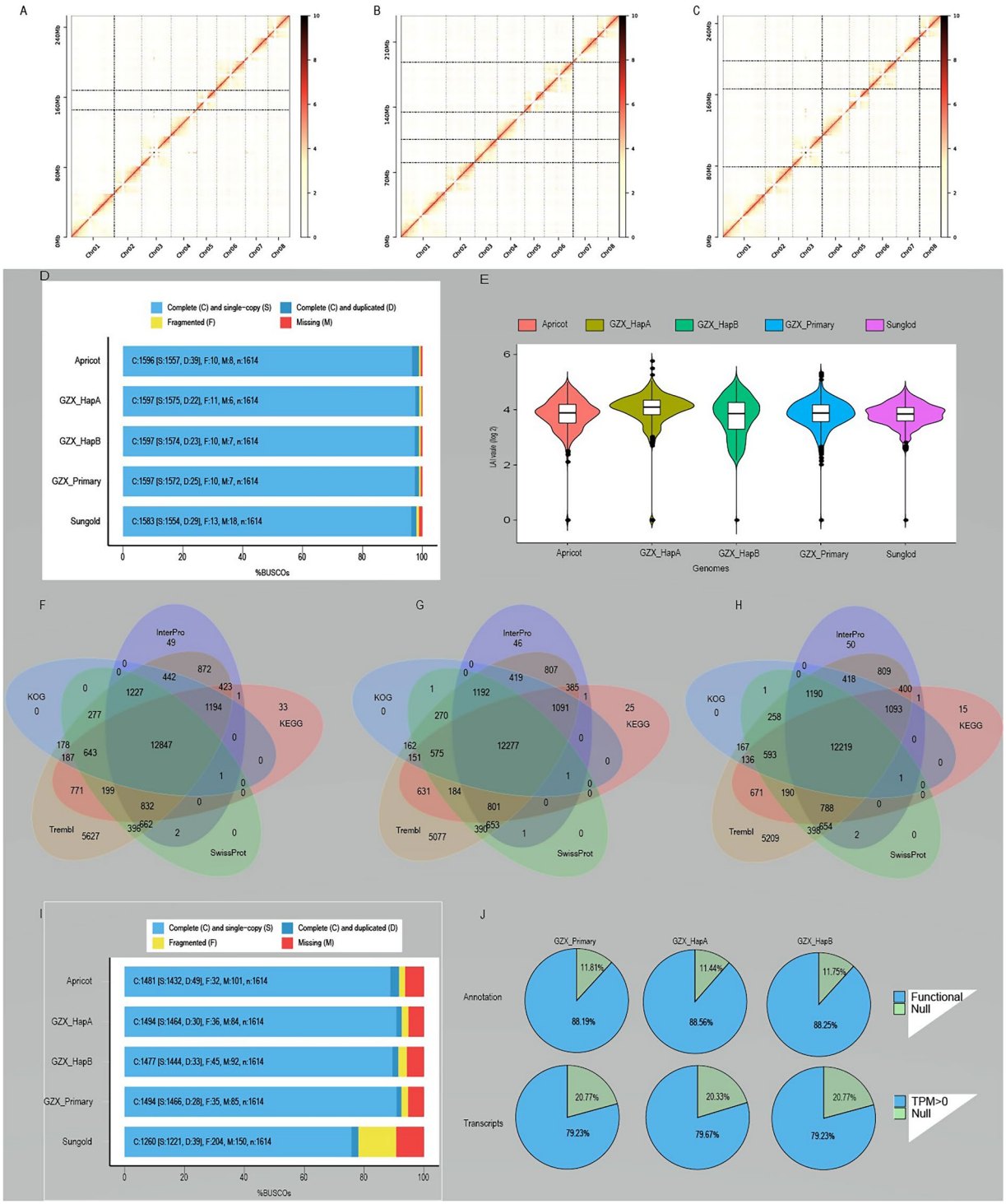

Fig. 4 Quality assessment of the GZX_Primary, GZX_HapA and GZX_HapB genome assembly and protein annotation. (**A**–**C**) Heat map displaying Hi-C interactions of GZX_Primary, GZX_HapA and GZX_HapB pseudomolecules. (**D**) Genomic BUSCO assessments exhibiting proportions classified as Complete and single-copy (S, blue), Complete and duplicated (D, green), Fragmented (F, yellow), and Missing (M, red) categories. (**E**) Comparison of LAI scores among the three assemblies of GZX_Primary, GZX_HapA, GZX_HapB, and sungold, and apricot. (**F**–**H**) Numbers of GZX_Primary, GZX_HapA and GZX_HapB protein-coding genes annotated in the NCBI nr plant, UniProt, GO and Pfam databases are illustrated by a Venn diagram. (**I**) Protein BUSCO assessments exhibiting proportions classified as Complete and single-copy (S, blue), Complete and duplicated (D, green), Fragmented (F, yellow), and Missing (M, red) categories. (**J**) Proportions of genes that could be functionally annotated and transcriptionally detected in GZX_Primary, GZX_HapA and GZX_HapB.

	Gene Number/%	Nr	Swissprot/%	KEGG/%	KOG/%	TrEMBL/%	Interpro/%	GO/%	Overall/%
GZX_Primary	30,543	87.65	55.94	56.09	55.65	87.67	60.74	49.31	88.19
GZX_HapA	28,461	87.98	57.43	56.65	56.71	88.07	62.10	49.89	88.56
GZX_HapB	28,689	87.69	56.80	56.14	56.04	87.81	61.43	49.58	88.25
Sungold	32,636	92.27	61.24	63.96	62.87	92.29	67.03	52.99	92.69
Chuanzhihong	30,436	97.17	64.01	68.90	66.68	97.08	69.87	53.42	97.32

Table 4. Summary of genome-wide annotation of functional genes. Note: Gene Number = Total Number of Predicted Genes; Nr = Non-Redundant Protein Database; Swissprot = high-quality manually annotated protein sequence database; KEGG = Kyoto Encyclopedia of Genes and Genomes; KOG = Eukaryotic Orthologous Groups; TrEMBL = Translated EMBL Nucleotide Sequence Database; Interpro = integrated protein signature database; GO = Gene Ontology; Overall = overall annotation rate. The "/%" suffix for all labels denotes the percentage of sequences annotated against the corresponding database.

Universal Single-Copy Orthologs) reference gene set was detected (Fig. 4D). Merqury (v1.3)[41] was used to evaluate assembly quality value (QV) and the QV value is 39.7. The LTR assembly[42] index (LAI) is 19.25 (Table 1 and Fig. 4E).

We also analyzed the gene expression levels across three GZX genome assemblies: GZX_Primary, GZX_HapA, and GZX_HapB. The total number of genes are identified in each assembly are shown in Fig. 4J. In addition, a significant proportion of these genes show detectable expression (TPM > 0), with GZX_HapA having the highest percentage at 74.48%, followed by GZX_HapB (73.79%) and GZX_Primary (72.19%) (Fig. 4J). This indicates that the majority of genes in all three assemblies are transcriptionally active. All these quality evaluation data demonstrate the high quality and accuracy of this genome assembly.

Data availability

All data supporting this study are publicly available. Raw sequencing reads have been deposited in the NCBI Sequence Read Archive (SRA) under BioProject PRJNA1233628, including Illumina genome survey reads (SRR32660270), PacBio long-read data (SRR32660273), Nanopore long-read data (SRR32660272), and Hi-C reads (SRR32660271). The genome assemblies have been deposited in NCBI GenBank under the Whole Genome Shotgun (WGS) accession numbers JBQWLS000000000, JBQWLQ000000000, and JBQWLR000000000. The genome assemblies and annotation files are available on Figshare database (https://doi.org/10.6084/m9.figshare.29713124).

Code availability

All analyses were conducted using standard bioinformatics tools without custom code. Publicly available protocols were followed, as specified in the methods.

Received: 22 April 2025; Accepted: 9 December 2025;
Published online: 17 December 2025

References

1. Mehlenbacher, S. A., Cociu, V. & Hough, F. L. APRICOTS (PRUNUS). *Acta Hortic* 65–110, https://doi.org/10.17660/ActaHortic.1991.290.3 (1991).
2. FAO. *World Production of Apricot and Sloes*. (2019).
3. Zhang, Q. & Liu, W. Advances of the apricot resources collection, evaluation and germplasm enhancement. *Acta Horticulturae Sinica* **45**, 1642–1660 (2018).
4. Liu, S. *et al*. The complex evolutionary history of apricots: Species divergence, gene flow and multiple domestication events. *Mol Ecol* **28**, 5299–5314 (2019).
5. Groppi, A. *et al*. Population genomics of apricots unravels domestication history and adaptive events. *Nat Commun* **12** (2021).
6. Hormaza, J. I., Yamane, H. & Rodrigo, J. Apricot. in *Fruits and Nuts* 171–187 (Springer Berlin Heidelberg, Berlin, Heidelberg). https://doi.org/10.1007/978-3-540-34533-6_7.
7. Smýkal, P., Nelson, M., Berger, J. & Von Wettberg, E. The Impact of Genetic Changes during Crop Domestication. *Agronomy* **8**, 119 (2018).
8. Herrera, S., Lora, J., Hormaza, J. I., Herrero, M. & Rodrigo, J. Optimizing Production in the New Generation of Apricot Cultivars: Self-incompatibility, S-RNase Allele Identification, and Incompatibility Group Assignment. *Front Plant Sci* **9** (2018).
9. Herrera, S., Rodrigo, J., Hormaza, J. & Lora, J. Identification of Self-Incompatibility Alleles by Specific PCR Analysis and S-RNase Sequencing in Apricot. *Int J Mol Sci* **19**, 3612 (2018).
10. Jiang, F. *et al*. The apricot (Prunus armeniaca L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis. *Hortic Res* **6**, 128 (2019).
11. Zhang, L. *et al*. Identification of self-incompatibility (S-) genotypes of Chinese apricot cultivars. *Euphytica* **160**, 241–248 (2008).
12. Jung, S. *et al*. 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Res* **47**, D1137–D1145 (2019).
13. Liu, W. *et al*. Apricot cultivar evolution and breeding program in CHINA. *Acta Hortic* 223–228, https://doi.org/10.17660/ActaHortic.2012.966.35 (2012).
14. Sun, H. Y., Zhang, J. H., Yang, L., Jiang, F. C. & Wang, Y. Z. An overview of apricot breeding in China. *Acta Hortic* 211–216, https://doi.org/10.17660/ActaHortic.2018.1214.37 (2018).
15. Vilanova, S. *et al*. SILEX: a fast and inexpensive high-quality DNA extraction method suitable for multiple sequencing platforms and recalcitrant plant species. *Plant Methods* **16**, 110 (2020).
16. 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).
17. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
18. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).

19. Mercy, G. *et al*. 3D organization of synthetic and scrambled chromosomes. *Science (1979)* 4597 (2017).
20. Hu, J. *et al*. NextDenovo: an efficient error correction and accurate assembly tool for noisy long reads. *Genome Biol* **25**, 107 (2024).
21. Robinson, J. T. *et al*. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst* **6**, 256–258.e1 (2018).
22. Xu, G.-C. *et al*. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8** (2019).
23. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467 (2005).
24. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr Protoc Bioinformatics* **25** (2009).
25. 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).
26. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
27. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
28. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29–W37 (2011).
29. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protocols* **11**, 1650–1667 (2016).
30. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
31. Vera Alvarez, R., Pongor, L. S., Mariño-Ramírez, L. & Landsman, D. TPMCalculator: one-step software to quantify mRNA abundance of genomic features. *Bioinformatics* **35**, 1960–1962 (2019).
32. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. *Methods Mol Biol*. 161–177, https://doi.org/10.1007/978-1-4939-9173-0_9 (2019).
33. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
34. 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. *Mol Biol Evol* **38**, 4647–4654 (2021).
35. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, e1005944 (2018).
36. Zhang, Q. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRP569754 (2025).
37. Zhang, Q. *Prunus armeniaca* L. Guo Zhi Xian (GZX) Genome assembly GZX_Primary_assembly. *GenBank NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_053572015.1 (2025).
38. Zhang, Q., Prunus armeniaca, L. & Guo, Z. Xian (GZX) Genome assembly GZX_HapA. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_053572045.1 (2025).
39. Zhang, Q. Prunus armeniaca L. Guo Zhi Xian (GZX) Genome assembly GZX_HapB. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_053572545.1 (2025).
40. Zhang, Q. The assemblies and annotation files of haplotype and primary genome of *Prunus armeniaca* L. Guo Zhi Xian (GZX). *figshare. Dataset*. https://doi.org/10.6084/m9.figshare.29713124 (2025).
41. 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 (2020).
42. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res* https://doi.org/10.1093/nar/gky730 (2018).

Acknowledgements

This study was funded by the National Natural Science Foundation of China (No. 31972365) and National Key Research and Development Program of China (2019YFD1000600).

Author contributions

Q.Z. and T.-Y.C. designed the study and wrote the manuscript. N.L., Y.Z., M.X., S.L. prepared and collected the plant materials; Y.Z. and X.M. prepared the samples for sequencing. T.-Y.C. and C.C. performed the data analysis; Q.Z. and W.L. supervised and provided valuable suggestions for analysis and the manuscript. All authors reviewed and approved the final version of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.Z. or W.L.