Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A draft genome assembly and resequencing analysis of Chinese cherry (Cerasus pseudocerasus) reveal structural variants associated with fruit traits

Liu Zhuqin

lzq048@163.com

Ningbo Academy of Agricultural Science

Wang Xueping

Nanjing Agricultural University

Zhao Xiuhua

Ningbo Academy of Agricultural Science

Gu Chao

Nanjing Agricultural University

Research Article

Keywords: Chinese cherry, draft genome, population genetic relationships, deletion structural variants, fruit traits

Posted Date: August 28th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4793503/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Version of Record: A version of this preprint was published at Plant Growth Regulation on November 22nd, 2024. See the published version at https://doi.org/10.1007/s10725-024-01252-9.

Abstract

Chinese cherry (*Cerasus pseudocerasus*) is a significant fruit crop that enjoys widespread cultivation in China. Here, we report a draft autotetraploid genome assembly of the Chinese cherry cultivar 'Huangguo', characterized by a size of 340.99 Mb and comprised of 261,760 scaffold sequences. We further obtained resequencing data of 8 Chinese cherry varieties at an average sequencing depth of about 104× per individual. Population structure analysis revealed that the 8 varieties were partitioned into two distinct groups, and the G matrix analysis found that the closest genetic background was shared between the 'Changbing' and 'Duanbing' varieties. In addition, we have creatively established a workflow for transforming heterozygous genotypes from diploid format to tetraploid format by conducting secondary genotyping in deletion structural variations, and through this process, structural variants related to Chinese cherry fruit peel color or size were initially screened out. In summary, this study provides valuable resources for population genetic relationships and will promote functional genomics studies in Chinese cherry and other crops.

1 Introduction

Chinese cherry (*Cerasus pseudocerasus*) belonging to the Rosaceae family under the genus *Cerasus*, is an economically significant fruit crop with a cultivation history of over 3000 years (Liu et al. 2024; Liu et al. 2008). It originates from the southwestern region of China, and a recent study has suggested the Longmenshan Fault zone, a center of biological and geological activity in southwest China, is the specific origin of cultivated Chinese cherry (Zhang et al. 2016; Cao 2018; Zhang et al. 2021). Chinese cherry is characterized by excellent broad adaptability, high disease resistance, and early flowering and ripening (Wang et al. 2022c; Yü 1979). However, as a perennial woody fruit crop, the juvenile phase of Chinese cherry is moderately long, from three to six years, which has lagged behind the breeding efficiency (Khan and Korban 2022; Yü 1979). Therefore, it is extremely necessary to accelerate the breeding process using genomics-based breeding strategies (Shirasawa et al. 2017; Raina et al. 2023).

In recent years, whole-genome resequencing technology has widely realized genetic evolution analysis and functional gene prediction by sequencing the genomes of individuals with reference genome and detecting different variants, including single nucleotide polymorphisms (SNPs) sites, insertion and deletion variants (InDels), and structural variants (SVs) (Song et al. 2023). For example, the resequencing of 161 peach accessions in northwestern China, reveals that these peach landraces are derivatives of peaches from eastern and southern China (Li et al. 2024). Liu et al. (2023) conducted resequencing analysis on four typical tea landraces from Hunan province, delineating that these landraces can be classified into two distinct groups. In addition, they identified genes associated with selection processes in these tea landraces (Liu et al. 2023). Furthermore, the genetic variation and population structure characteristics between wild and cultivated varieties in crops such as tomato (Li et al. 2023), orange (Huang et al. 2023), and sweet cherry (Xanthopoulou et al. 2020) have been investigated by this strategy.

Structural variation plays an essential role in plants and has been widely demonstrated to directly impact phenotype (Gabur et al. 2019; Mahmoud et al. 2019; Guo et al. 2020). Additionally, in plant population analysis, the different genotypes of the same SV have been reported to be highly correlated with agronomic traits. For example, the phenotype of red flesh color around the stone was found to be strongly associated with the genotype of a 487 - bp deletion in peach (Guo et al. 2020). However, prevalent second-generation sequencing tools for detecting structure variations, such as BreakDancer (Fan et al. 2014), Delly (Rausch et al. 2012), and Lumpy (Layer et al. 2014), are primarily designed to record SVs in diploid data formats. The genotypes of identified SVs are typically denoted as 0/0, 0/1, and 1/1, representing homozygous reference (RR), heterozygous (RA), and homozygous alternate (AA) genotypes, respectively. For homozygous variants (0/0 and 1/1), there is no difference between diploid and tetraploid data. However, for heterozygous variants (0/1), there are three possible types in tetraploid data: 0001, 0101, and 0111. Hence, for the tetraploid structural variant dataset, the 0/1 genotype requires a secondary genotyping procedure. To be specific, the detection of deletions is more comprehensive in comparison to other events in NGS data (Meng et al. 2023). Therefore, only the deletions structural variants (DELs) were explored in this paper.

In the previous studies, we have determined the tetraploid characteristics of Chinese cherry cultivars by flow cytometry (Gu et al. 2010; 2013; 2014). In this study, we constructed and annotated the *C. pseudocerasus* reference genome for the cultivar ‘Huangguo’, and furtherly identified the autotetraploid level of it. Moreover, using the ‘Huangguo’ genome as a reference genome, we conducted whole-genome resequencing analysis on 8 major Chinese cherry cultivars to elucidate their genetic evolution relationships by high-quality SNP variations and explore key trait candidate sites based on deletion structural variants by a new method. Our research provides valuable resources for genetic improvement, enhancing our understanding of the genetics and evolution of domestic Chinese cherry, and will facilitate the process of molecular breeding in Chinese cherry.

2 Materials and methods

2.1 Sample collection

The young leaves of ‘Huangguo’ and ‘Duanbing’ were obtained from Binghuai orchard, Jiangning District, Nanjing, Jiangsu Province, China. And ‘Changbing’, ‘HXxinzhong’, ‘HXyaoduan’, ‘Yongzhuyihao’, ‘HXyaohuang’, ‘Heizhenzhu’, and ‘Hongfei’ cultivars were collected from Ningbo Academy of Agricultural Science, Zhejiang Province, China. The ‘Huangguo’ cultivar was utilized for *de novo* genome sequencing and assembly, and the others were resequenced. All samples were frozen in liquid nitrogen and stored in the freezer at minus 80 degrees Celsius.

2.2 Genome sequencing

Genomic DNA was extracted with a plant DNA Kit (OMEGA) according to the manufacturer’s instructions. Then, the qualified DNA sample was ultrasonically interrupted with the Covaris M220 instrument to 300-500bp, and Illumina library was constructed by TruSeq™ Nano DNA Sample Prep Kit. Finally, 150 bp

paired-end sequencing of the constructed library was performed on the Illumina NovaSeq 6000 platform (Illumina, USA). Approximately 63 Gb of raw data from Huangguo was obtained by 150 bp paired-end read length.

2.3 Illumina data quality control and genome survey

Illumina raw reads of Huangguo were firstly processed for quality by Trimmomatic (v0.39) (Bolger et al. 2014) software with parameters: 'ILLUMINACLIP:adapters.fa:2:30:10 SLIDINGWINDOW:4:15 MINLEN:75'. Specifically, low-quality reads were filtered out based on the following principles: reads with equal or greater than 10% unidentified nucleotides (N) were removed; the adapter sequence in reads or reads sequencing quality score less than Q20 or non-AGCT bases contained at the 5 ' end were trimmed; reads less than 75bp in length after the above quality pruning were discarded. After filtering, the clean data was used to estimate genome size and heterozygosity by k-mer frequency analysis with Jellyfish (v2.2.6) (Marçais and Kingsford 2011), kmc (3.2.1) (http://sun.aei.polsl.pl/kmc), GenomeScope 2.0 (1.0.0) (Ranallo-Benavidez et al. 2020), and Smudgeplot (0.2.3) (Ranallo-Benavidez et al. 2020).

2.4 Genome assembly and assessment

For genome assembly, the optimized sequences were assembled by employing the ABySS (Simpson et al. 2009) assembly software with various K-mer parameters to achieve the most favorable assembly outcomes. Subsequently, the GapCloser (Luo et al. 2012) software was utilized to conduct local gap filling and base correction on the assembled results. To assess the completeness and quality of assembly, clean reads were aligned to reference genome using BWA (0.7.12-r1039) (Li 2013). Subsequently, SAMtools (v1.14) (Li et al. 2009) was used to evaluate sequencing depth and coverage. Benchmarking Universal Single-Copy Orthologs (BUSCO) (v5.6.1) (Manni et al. 2021) was applied to determine genome completeness based on the eukaryota_odb database.

2.5 Genome annotation

We used the combination methods of *de novo* prediction and homologous protein alignment to annotate genes. Trained with reference genomes, *de novo* gene prediction was performed using AUGUSTUS (v3.2.3) (Stanke et al. 2006). The gene coding region and intron region were identified by precise alignment of the protein sequence and the reference genome DNA sequence with Genewise (v2.4.1) (Birney et al. 2004). Finally, the data from the above steps were comprehensively integrated with EVidenceModeler (v1.1.1) (Haas et al. 2008).

For repeat sequence, Tandem repeats finder (TRF) (Benson 1999) and RepeatMasker (http://www.repeatmasker.org) were used to identify tandem repeats sequence and interspersed repeats sequence respectively. Gene function annotation was predicted by aligning protein sequences to five databases (Nr, http://www.ncbi.nlm.nih.gov/, Swiss-Prot, http://www.ebi.ac.uk/uniprot, eggNOG, http://eggnogdb.embl.de/, KEGG, http://www.genome.jp/kegg/, and GO, http://geneontology.org/) with BLAST (2.7.1) search (E-value no more than 1e-5) (Altschul et al. 1990). To ensure its biological significance, only one optimal comparison result was retained for each query gene per database.

2.6 resequencing and data quality control

A total of 8 accessions were resequenced, namely 'Changbing', 'Duanbing', 'HXxinzhong', 'HXyaoduan', 'Yongzhuyihao', 'HXyaohuang', 'Heizhenzhu', and 'Hongfei'. The library construction method, sequencing platform, and raw data quality control process were the same as above (Illumina NovaSeq sequencing in 2.2 and data quality control in 2.3). After filtering, we generated a total of 418.11Gb clean reads which were used for further analysis.

2.7 Read mapping and SNP/Indel calling

The clean paired-end reads were mapped to the 'Huangguo' reference genome using BWA (0.7.12-r1039) (Li 2013) with parameters: 'mem -t 20 -M'. PCR duplicates were removed by Picard-tools (https://github.com/broadinstitute/picard). Then, the mapped reads were sorted and indexed using SAMtools (v1.14) (Li et al. 2009). SNP and Indel calling were performed by the GATK (Genome Analysis Toolkit, 4.1.2.0) package with parameters:'HaplotypeCaller -stand-call-conf 50 --max-genotype-count 500'. To filter the potential false positive SNPs or Indels, GATK (4.1.2.0) (McKenna et al. 2010) and VCFtools (v0.1.16) (Danecek et al. 2011) were utilized with the following criteria: biallelic sites were maintained, read depth > = 4, variations quality score > = 20, and QD > = 2.0 or FS < = 60.0 or MQ > = 40.0 or MQRankSum >= -12.5 or ReadPosRankSum >= -8.

2.8 Annotation of genetic variants

The ANNOVAR package (v2019-10-24) (Wang et al. 2010) was utilized with default parameters to effectively annotate SNP and Indel variants based on the reference genome. For SNPs, variant sites were classified into seven distinct categories, namely exonic regions, intronic regions, splicing regions, upstream regions, downstream regions, upstream and downstream regions, and intergenic regions. Furthermore, SNPs within the coding sequence region were grouped into four types: nonsynonymous SNV, synonymous SNV, stopgain SNV, and stoploss SNV. For Indels, in addition to the above seven categories, variant sites also occurred in UTR5 (5' untranslated region). Indels in coding sequence region were further categorized into frameshift deletion/insertion, nonframeshift deletion/insertion, stopgain SNV, and stoploss SNV.

2.9 Phylogenetic tree, principal component analysis, and population structure analysis

To clarify the genetic relationship of 8 individuals, the vcf file of SNPs was firstly filtered using criteria of MAF > = 0.05 and a missing rate of no more than 90%. Subsequently, a neighbor-joining (NJ) tree was constructed through the following steps: i. conversion of the filtered snp.vcf file to phy format using vcf2phylip (Ortiz 2019). ii. construction of an NJ evolutionary tree with a bootstrap value set to 1000 in MEGA7 (Kumar et al. 2016). The principal component analysis (PCA) was conducted by GCTA (1.93.2) (Yang et al. 2011) software and plotted in the first three significant components using the ggplot2 (Hadley, 2016) R package. And ADMIXTURE (1.3.0) (Alexander et al. 2009) was applied to infer the population structure among 8 samples. To determine the optimal K value, number of population groups

(K) was calculated from 1 to 8, and cross-validation (CV) error values were extracted and visualized. The genetic relationship (G) matrix analysis was also employed by GCTA (1.93.2) (Yang et al. 2011) software.

2.10 Structure variation (SV) calling

We chose LUMPY (v0.3.1) (Layer et al. 2014) to call SV. Firstly, the assembly alignments file obtained above in 2.7 were used to extract split reads and discordant reads by Samtools (v1.14) (Li et al. 2009) and extractSplitReads_BwaMem module in LUMPY (v0.3.1) (Layer et al. 2014). Secondly, SV calling was performed separately for each sample to obtain individual VCF files by lumpyexpress module in LUMPY (v0.3.1) (Layer et al. 2014). In addition, SVTyper (v0.7.1, https://github.com/hall-lab/svtyper) was used to call genotypes on output VCF files with a Bayesian maximum likelihood algorithm. Finally, merge VCF files by SURVIVOR (v1.0.7) software with paraments '1000 2 1 1 0 30' (Jeffares et al. 2017). The results of SVs were counted in 4 types: deletions (DEL), duplications (DUP), inversions (INV), and translocations (BND). The SnpEff software (v5.2c) (Cingolani et al. 2012) was used for annotating SVs.

2.11 Secondary genotyping of deletion heterozygous variants in SVs

The workflow includes four steps:

1. Selecting the surrounding region size for heterozygous DEL variation in each individual file by a python script named '01select_region_size.py'. We try our best to ensure that there is no other variation in the surrounding interval of the deletion heterozygous variant (in general, set to 100bp). If it is unavoidable, we will take the surrounding area of 10bp for subsequent analysis (but this probability is only about 0.1% in our data set).
2. Extracting the average read depth of this variation (V_{rd}) and its surrounding area (S_{rd}). Firstly, use the command 'genomecov -ibam' of bedtools (Quinlan and Hall 2010) to count the number of reads at each position for each bam file. And then a python script named '02extract_average_rd.py' is devised to meticulously compute the V_{rd} and S_{rd}.
3. Calculating the radio (V_{rd}/S_{rd}) and cleaning data. A python script, denoted as '03calculate_to_clean.py', facilitates the execution process. It involves calculating the ratio (V_{rd}/S_{rd}), collecting basic information on the ratio data of each file, removing outliers and calculating the proportion of outliers, and separately creating box plots for the raw data and the data after outliers have been removed.
4. Determining the secondary genotype and modifying it into the population vcf file. In theory, the secondary typing of a heterozygous genotype should be as follows: if the radio (V_{rd}/S_{rd}) is equal to 0.5, set to 0101 genotype, if the ratio is greater than 0.5, set to 0001 genotype, and if the ratio less than 0.5, set to 0111 genotype. However, in fact, due to the limitation of sequencing technology, a threshold range is required for the ratio judgment, which we set as 0.5 ± 0.15, that is, if the radio (V_{rd}/S_{rd}) is greater than or equal to 0.35 and less than or equal to 0.65, the genotype is set as 0101; if the ratio is greater than 0.65, it is set as 0001; and if the ratio is less than 0.35, it is set as 0111.

Given that the SURVIVOR (Jeffares et al. 2017) software retains the basic information of each individual site when merging each individual file, we can extract the 0/1 genotype site of DEL type variations from the merged file and restore it to each individual file, thereby determining the secondary subtyping type of the heterozygous genotype in the merged file. A python script, named '04genotype_to_merge.py', can be utilized to implement this process.

After completing the workflow, candidate sites can be manually identified by the similar trend between the phenotype and variant genotype.

2.12 Fruit weight and color assessment for eight individuals

For the analysis of fruit basic traits, the fruits of all 8 cultivars were collected. Six fruits were randomly selected from three trees for each accession. A HunterLab ColorFlex EZ (CFEZ) (Konica Minolta, Inc., Tokyo, Japan) was used to measure the fruit peel color according to the CIE system. The positive a* values indicate red and purple, and the negative values correspond to green and blue. Each sample was measured for three biological replicates. The weight of the fruit was quantified utilizing a QUINTIX213-1CN electronic balance (Sartorius, Germany).

2.13 Data availability

The genome assembly and gene annotation data are available at Figshare: Zhuqin, Liu; Xueping, Wang; Xiuhua, Zhao; Chao, Gu (2024). Genome sequencing of Chinese cherry (*Cerasus pseudocerasus*) reveals structural variants associated with fruit traits. figshare. Dataset. https://doi.org/10.6084/m9.figshare.25856074.v1. The raw reads of resequensing data generated in this study have been deposited in the CNCB genome sequence archive (GSA) with the accession number PRJCA026162 (https://ngdc.cncb.ac.cn/gsa/s/WY5h5b9E).

3 Results

3.1 Genome assembly and annotation

To generate the genome assembly, we sequenced a total of 63.0 Gb of Illumina short reads with about 137 x coverage (Table S1). After filtering, a total of 61.3 Gb of clean data was retained for further analysis (Table S1). To gain insight into the characteristics of *C. pseudocerasu* genome, a genome survey analysis was conducted employing the K-mer method (Marçais and Kingsford 2011; Ranallo-Benavidez et al. 2020; Simpson et al. 2009), the results represented that it was an autotetraploid genome (the AAAB genome structure rate was 70%), with estimated genome size, heterozygosity rate, and repetition rate of 459.67 Mb, 1.61%, and 1.75%, respectively (Fig. 1, Table S2). The final assembly size of 'Huangguo' genome was 340.99Mb, consisting of 261,760 scaffolds. Additionally, the scaffold N50 was 1.56Kb and the GC content rate was determined to be 37.58% (Table 1). The quality and completeness of 'Huangguo' genome were assessed through BUSCO (Manni et al. 2021) in eukaryota_odb10 database. BUSCO analysis revealed that the genome completeness was 50.6% (Table S3).

For genome annotation, a total of 51,264 coding genes were predicted in ‘Huangguo’ genome with a mean coding sequence length of 572 bp (Table 1, Table S4). Moreover, the genome contained approximately 158.78 Mb (46.57%) of repetitive sequences, comprising 143.1 Mb of dispersed repeats and 15.68 Mb of tandem repeats (Table 1, Table S5). In genome function annotation, 92.67% of genes were annotated by the NR, GO, COG, KEGG, or Swiss-Prot databases (Table S6).

(A) ‘Huangguo’ Genome K-mer frequency distribution map. (B) The smudge plot analysis in *C. pseudocerasu* genome.

Table 1
Summary of ‘Huangguo’ genome assembly

Genomic feature	Huangguo
Estimated genome size (Mb)	459.67
Assembled genome size (Mb)	340.99
Total number of scaffolds	261760
Scaffold N50 (Kb)	1.56
Complete BUSCOs (%)	50.6
Number of genes	51,264
Percentage of repeat sequences (%)	46.57

3.2 Genome resequencing and SNPs/Indels calling

To clarify the evolutionary relationships of 8 cultivated Chinese cherries (‘Changbing’, ‘Duanbing’, ‘HXxinzhong’, ‘HXyaoduan’, ‘Yongzhuyihao’, ‘HXyaohuang’, ‘Heizhenzhu’, and ‘Hongfei’), we generated a total of 388.27 Gb of raw Illumina sequence data with an average of 48.53 Gb per individual (Table S7). Approximately 375.99 Gb of high-quality sequencing reads were aligned to the ‘Huangguo’ genome, yielding a mapping rate between 99.19% and 99.39% (Table S7, Table S8). Moreover, sequencing coverage depths varied from 78 to 121 x (Table S8). Consequently, we called 27,169,381 SNPs and 3,733,525 Indels for eight Chinese cherry cultivars (Fig. 2A, Table S9). Among the 8 cultivars, ‘Yongzhuyihao’ has the fewest SNP/Indel variant sites, while ‘Changbing’ has the most (Fig. 2A, Table S9). Additionally, Type C: G change to T: A and type T: A change to C: G constituted the majority of the SNP mutation types (Fig. 2B).

In all identified SNPs, a total of 78.29% (21.3 million) were located in the intergenic regions, while only 5.6% (1.5 million) were situated within the exonic regions (Fig. 2C, Table S10). For these SNPs in coding regions, we detected 727,248 synonymous SNPs and 778,697 non-synonymous SNPs, resulting in a non-synonymous/synonymous ratio of 1.07 (Table S10). Similarly, concerning all identified Indels, 80.63%

(3.0 million) were located in the intergenic regions, and only 1.2% (0.04 million) were detected within the exonic regions (Fig. 2D, Table S11).

(A) Statistics on the various numbers of SNPs and InDels in eight samples. (B) Distribution of SNP mutation type in eight samples. The x-axis delineates the type of SNP mutation, and the y-axis quantifies the number of SNPs. (C) Annotation distribution of SNPs variation. Different colors show the SNPs distribution of intergenic region, upstream, downstream, exon, intron, and splice. (D) Annotation distribution of InDels variation. Different colors show the InDels distribution of intergenic region, upstream, downstream, exon, intron, splice, and UTR5.

3.3 Phylogenetics and population structure

In order to explore the genetic relatedness among eight domestic Chinese cherry varieties, we constructed an unrooted neighbor-joining phylogenetic tree based on the filtered SNP data. The result indicates that the 8 individuals were divided into two major subclades. The first major subclade consists of ‘Changbing’, ‘Duanbing’, and ‘HXxinzhong’, with ‘Changbing’ and ‘Duanbing’ clustering on an internal node, suggesting that these two individuals have the closest evolutionary relationship. Similarly, the second major subclade includes ‘Yongzhuyihao’, ‘HXyaohuang’, ‘Heizhenzhu’, ‘Hongfei’, and ‘HXyaoduan’, with ‘HXyaohuang’ and ‘Yongzhuyihao’ clustering on an internal node, inferring that these two individuals have the closest genetic relationship within this branch (Fig. 3A). To further estimate different ancestral proportions in 8 individuals, population structure analysis was conducted by assuming a given number (*K*) of ancestral populations. When $K = 2$, the CV (Cross validation) error value was minimized (Fig. 3B), and the 8 cherry accessions were divided into two groups, which was consistent with the subclades of the phylogenetic tree. Additionally, ‘HXxinzhong’ and ‘HXyaohuang’ were shared ancestry between these two groups. When $K = 4$, there was a division or shared ancestry between each individual except ‘Changbing’ and ‘Duanbing’ (Fig. 3C). Further combined with the analysis of the G matrix, the obtained G values for the ‘Changbing’ and ‘Duanbing’ cultivars reached up to 0.85 (Fig. 3E), indicating a high degree of homology or closely related breeding histories between them. Moreover, The Principal Component Analysis (PCA) also yielded similar results in the phylogenetic tree among the individuals. (Fig. 3D).

(A) Unrooted neighbor-joining tree constructed with a bootstrap value of 1000 between individuals. (B) Cross-validation error results from different K values. (C) ADMIXTRUE analysis among the accessions (K = 2–4). (D) The principal component plot of 8 individuals. (E) Genetic relationship (G) matrix analysis between pairs of individuals, where the larger G value indicates the closer relationship between the samples.

3.4 SVs identification and annotation

A total of 2,829,651 SVs were identified in 8 domestic Chinese cherry varieties. Among them, the individual ‘Duanbing’ had the fewest variant sites with 259,237, while ‘Yongzhuyihao’ had the highest number of variant sites with 407,093, which was inconsistent with the variation trend for SNPs and InDels. In terms of structure variation types, there were 2,700,528 CTX types, 102,302 DEL types, 16,879 INV types, and 9,942 DUP types in all 8 samples. The proportion of CTX (95.44%) is extremely high

among the different types of structure variations, potentially attributable to the limitation of genome scaffold assembly. Therefore, we took DEL, INV, and DUP types in 8 samples for further analysis.

After merged by SURVIVOR (Jeffares et al. 2017), there were 16,878 DEL types, 1,883 DUP types, and 753 INV types were found at least in two individuals. We also analyzed the distribution of these SVs throughout the genome and determined that 27.5% of SVs coincide with genic regions, including CDS, introns, and promoters (Table S12).

3.5 Secondary genotyping of heterozygous variants in SVs

As is well-known, there are three strategies to detect structural variation: discordant read pair, split reads, and read depth (Alkan et al. 2011). But for the DEL type, no matter which detection method is used, there is no doubt that the read depth of the variation region is always lower than that at wild-type sites (Cai et al. 2019). Therefore, the core idea of the secondary typing of the DEL variant type is to determine whether the heterozygous genotype (0/1) is 0001, 0101, or 0111 according to the ratio of the reads depth at the heterozygous mutation locus (that is, the number of reads supporting reference) to the reads depth within a specified interval size (such as 100bp) on the both sides of the variant without any variation.

The workflow includes four steps and is described in materials and methods in detail. Since these sites have been identified as heterozygous genetype (0/1) of DEL variations by Lumpy (Layer et al. 2014) software algorithms, we employ boxplots to examine the distribution pattern and remove outliers within the radio data that we acquired in 8 individuals (Table 2, Fig. 4). In the 8 sample boxplots, the upper quartile, median, and lower quartile distributions were 0.762–0.773, 0.631–0.658, and 0.476–0.510, respectively. And the proportion of outliers in each file ranges from 2.29–3.06% (Table 2).

Table 2 The statistics of the original radio data in 8 individuals.

Sample	Count	Mean	Std	Min	25%	50%	75%	Max	Count_outliers
Changbing	7950	0.653	0.313	0.005	0.497	0.656	0.773	8.498	217(2.73%)
Duanbing	5666	0.639	0.321	0.004	0.476	0.631	0.762	11.251	164(2.89%)
Heizhenzhu	6238	0.658	0.307	0.004	0.510	0.658	0.771	9.572	180(2.89%)
Hongfei	6279	0.651	0.303	0.003	0.494	0.650	0.771	9.188	169(2.69%)
HXxinzhong	6185	0.644	0.352	0.005	0.483	0.638	0.766	11.092	155(2.51%)
HXyaoduan	7437	0.644	0.329	0.005	0.486	0.640	0.765	10.392	224(3.01%)
HXyaohuang	5987	0.658	0.370	0.006	0.495	0.648	0.768	12.990	183(3.06%)
Yongzhuyihao	4750	0.646	0.286	0.077	0.494	0.639	0.763	7.539	109(2.29%)

Count: total number of heterozygous genotypes in DEL variation. Std: standard deviation. 25%: the lower quartile. 50%: the same as the median. 75%: the upper quartile. Count_outliers: the number and proportion of outliers filtered by the boxplot.

(A) Boxplots of the raw ratio data in 8 individuals. (B) Boxplot of the ratio excluding outliers in 8 individuals. The y-axis indicates the radio of V_{rd}/S_{rd}. V_{rd}: the average read depth of the variation; S_{rd}: the

average reads depth of its surrounding area.

Subsequently, we use a threshold range to determine secondary typing on each individual file, and replace the heterozygous loci of each DEL type in the merged file based on the results with the secondary typing type. After obtaining the population vcf file modified into secondary typing, we selected the genotypes that exhibit a similar trend based on the phenotype data of fruit peel color from 8 samples. In 8 individuals, the ‘Changbing’ variety showed the reddest fruit peel color, with an average of 26.736 in a* value, which was more than twice that of the yellow sample ‘HXyaohuang’ (10.042) (Fig. 5A). It indicates that the genotype of fruit peel color in the two samples was significantly different. Given that it is uncertain whether the variant exerts a dominant or recessive effect on the phenotype, it can only be extrapolated that if one individual exhibits a genotype of 1111, the corresponding individual must invariably display a genotype of 0000, and the other samples show an increasing or decreasing trend of dominance. Based on this, 2 loci correlated with fruit peel color have been discerned within 10 scaffold locations (Fig. 5C-D, Table S13). For the phenotypic data of fruit weights, we utilized the same strategy and identified a locus associated with fruit weights from 11 scaffold positions (Fig. 5E, Table S14).

(A) Comparison of fruit peel color in 8 samples. The y-axis indicates the a* value in fruit peel color, the higher the value, the redder the peel color. (B) Comparison of fruit weight in 8 samples. ∗∗∗∗ P-value < 0.0001, t-test. The sample is arranged in ascending order of values. (C) The reads depth of the candidate scaffold site (‘scaffold7622’) related to fruit peel color in 8 samples. (D) The reads depth of the candidate scaffold site (‘scaffold112378’) related to fruit peel color in 8 samples. (E) The reads depth of the candidate scaffold site (‘scaffold45003’) related to fruit weight in 8 samples. The red dashed line represents the DEL-type SV interval in 8 samples.

4 Discussion

Chinese cherry is a notable economic significance fruit crop that is extensively cultivated in China. Here, we generated a scaffold-level genome for Chinese cherry cultivar ’Huangguo’ and performed resequencing on eight cultivars. Additionally, we proposed a novel methodology to convert heterozygous genotypes from diploid format to tetraploid format by secondary genotyping in deletion structural variants, which explores the association between the variant genotype and Chinese cherry phenotypic attributes including fruit peel color and fruit size. These data reveal meaningful genetic variations and infer the evolutionary history of the species.

In nature, Chinese cherry populations exhibit a stable ploidy level of tetraploidy, yet there has been a persistent dearth of evidence as to whether they are autotetraploid or allopolyploid (Wang et al. 2018). However, this issue has been recently addressed. Wang et.al (2023) utilized Chromosomal Karyotype and GISH analysis, suggesting that Chinese cherry likely originated from autotetraploidy. The phylogenetic and comparative genomic analyses conducted on the four haplotypes within the *Cerasus pseudocerasus* genome indicated that the Chinese cherry is a stable, random-pairing autotetraploid species (Jiu et al. 2024). In our work, the autotetraploid level of it is additionally supported by genome

survey analysis, with an observed autotetraploid rate of up to 70%, which adds another strong evidence to the autotetraploid nature of Chinese cherry.

Autotetraploidy genomes assembly differ from allopolyploids, with the chromosome phasing being extremely challenging, especially when distinguishing highly similar sequences in homologous chromosomes (Zhang et al. 2024a; Wei et al. 2023). And the situation is further complicated by the same high level of repetitive sequences and polyploidy complexity. Nonetheless, owing to the rapid advancement of third-generation sequencing technology and the combination with next-generation sequencing data, the successful assembly of autotetraploid genomes at the chromosome level had been extensively reported in the last three years, including sugarcane (*Saccharum spontaneum*) (Zhang et al. 2022), a potato cultivar (*Solanum tuberosum*, 'Qingshu No.9') (Wang et al. 2022a), wax apple (*Syzygium samarangense* (Blume) Merr. and Perry) (Wei et al. 2023), rhubarb (*Rheum officinale*) (Zhang et al. 2024b), tetraploid strawberry (*Fragaria moupinensis*) (Qiao et al. 2024), hardy kiwifruit (*Actinidia arguta*) (Zhang et al. 2024a), and Chinese cherry (*Cerasus pseudocerasus*) (Jiu et al. 2024). Therefore, it is our limitation to only use Ilumina sequencing data, we finally obtained a draft genome of 340.99 Mb for Chinese cherry, which covered 74.2% of the estimated genome size of about 459.67Mb.

Compared with the other cherry species in the *Cerasus* genu, the genome of the Chinese cherry displays a repetitive rate of 46.57%, which is closely aligned with the flowering cherry (*Prunus yedoensis*, nudiflora) (Baek et al. 2018) at 47.2%. In contrast, the repetitive rate in the genome of the sour cherry (*Prunus cerasus*, Montmorency) (Goeckeritz et al. 2023) is higher at 48.5%, while the sweet cherry (*Prunus avium*, Tieton) (Wang et al. 2020) showcases the highest repetitive rate at 59.40%. Wang et al. (2022b) shown that the closest phylogenetic relationship between Chinese cherry and flowering cherry (*Prunus yedoensis*) by assessing the similarity of repetitive sequences in different genomes. Hence, it is possible that the proportion of repetitive sequences in Chinese cherry and flowering cherry (*Prunus yedoensis*) may also exhibit similarity.

Chinese cherry originated from Longmenshan Fault zone in China, and spread widely along the Yangtze River basin, showing remarkable adaptation to a variety of ecological environments (Zhang et al. 2021; Wang et al. 2022c). Previous research used traditional SSR marker methods to analyze the genetic diversity and population structure of wild and cultivated cherries (Zhang et al. 2016), but the evolutionary relationships within cultivated Chinese cherry populations are not clear. In our study, we selected 8 representative cultivated varieties for whole-genome resequencing analysis. Based on the results of phylogenetic tree analysis, PCA, and population structure, when K = 2, 8 cultivated Chinese cherry varieties were divided into two groups, suggesting they derive from different ancestors. Moreover, taking genetic relationship matrix analysis into account, in group I, 'Changbing' and 'Duanbing' have the closest genetic background with a genetic distance of 0.85, and in group II, 'Yongzhuyihao' and 'HXyaohuang' have a closer genetic relationship with a genetic distance of 0.61. These findings will serve as a valuable resource for plant breeders and assist in Chinese cherry breeding programs.

Heterozygous and homozygous genotypes of the same structural variants have different effects on the same phenotype, which was confirmed by SV-GWAS structural variation analysis (Guo et al. 2020). However, the heterozygous genotypes of tetraploid are more complex than those of diploid data, with three types: 0001, 0101, and 0111. Therefore, based on the SV calling software LUMPY (Layer et al. 2014), we established a workflow for secondary typing of the deletion heterozygous structural variants and examined related loci that may affect the fruit peel color or fruit size of Chinese cherry. However, the potential contribution of deletion structural variants to related phenotypes still needs to be further analyzed.

5 Conclusion

Here, a scaffold-level autotetraploid genome assembly of Chinese cherry was generated. The total length of the assembled sequences was 340.99 Mb, with the repetitive sequence proportion of 46.57%. Population structure analysis revealed that the 8 Chinese cherry cultivars were divided into two groups, the first group included ‘Changbing’, ‘Duanbing’, and ‘HXxinzhong’ cultivars, and the second group included ‘Yongzhuyihao’, ‘HXyaohuang’, ‘Heizhenzhu’, ‘Hongfei’, and ‘HXyaoduan’. And the G matrix analysis furtherly confirmed the closest genetic relationship between ‘Changbing’ and ‘Duanbing’. Moreover, a secondary typing workflow designed for transforming heterozygous genotype from diploid format to tetraploid format in deletion structural variants was built and applied to tentatively screen out candidate sites that link to fruit peel color or fruit size traits in Chinese cherry cultivars. Our results advance our understanding of the genetic relationships in domestic Chinese cherries and provide a new idea for analyzing heterozygous genotypes of deletion structural variants in tetraploid data.

Declarations

Acknowledgements

This work was supported by Ningbo Science and Technology Innovation 2025 major projects 2019B10024 and the Open Project of Ningbo Key Laboratory of Characteristic Horticultural Crops in Quality Adjustment and Resistance Breeding.

Author’s contributions

Z.L and C.G. designed the project. Z.L. conducted sample collection and genome sequencing and assembly with assistance from X.Z. X.W. performed resequencing analysis and drafted the manuscript. Z.L and C.G. revised the manuscript. The authors read and approved the final manuscript.

Competing interests

No conflict of interest is declared.

References

1. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9):1655–1664. 10.1101/gr.094052.109
2. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12(5):363–376. 10.1038/nrg2958
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. 10.1016/s0022-2836(05)80360-2
4. Baek S, Choi K, Kim GB, Yu HJ, Cho A, Jang H, Kim C, Kim HJ, Chang KS, Kim JH, Mun JH (2018) Draft genome sequence of wild *Prunus yedoensis* reveals massive inter-specific hybridization between sympatric flowering cherries. Genome Biol 19(1):127. 10.1186/s13059-018-1497-y
5. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580. 10.1093/nar/27.2.573
6. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. Genome Res 14(5):988–995. 10.1101/gr.1865504
7. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. 10.1093/bioinformatics/btu170
8. Cai L, Wu Y, Gao J (2019) DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. BMC Bioinformatics 20(1):665. 10.1186/s12859-019-3299-y
9. Cao SY (2018) Local Varieties of Chinese Cherry. China Forestry, Beijing
10. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6(2):80–92. 10.4161/fly.19695
11. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27(15):2156–2158. 10.1093/bioinformatics/btr330
12. Fan X, Abbott TE, Larson D, Chen K (2014) BreakDancer: Identification of Genomic Structural Variation from Paired-End Read Mapping. Curr Protoc Bioinf 45. 15.16.11-11
13. Gabur I, Chawla HS, Snowdon RJ, Parkin IAP (2019) Connecting genome structural variation with complex traits in crop plants. Theor Appl Genet 132(3):733–750. 10.1007/s00122-018-3233-0
14. Goeckeritz CZ, Rhoades KE, Childs KL, Iezzoni AF, VanBuren R, Hollender CA (2023) Genome of tetraploid sour cherry (*Prunus cerasus* L.) 'Montmorency' identifies three distinct ancestral *Prunus* genomes. Hortic Res 10(7):uhad097. 10.1093/hr/uhad097
15. Gu C, Liu Q-Z, Khan MA, Wu J, Zhang S-L (2014) Hetero-diploid pollen grains that represent self-compatibility are incompatible with non-self receptors in tetraploid Chinese cherry (*Prunus pseudocerasus* Lindl). Tree Genet Genomes 10(3):619–625. 10.1007/s11295-014-0708-2

16. Gu C, Liu QZ, Yang YN, Zhang SJ, Khan MA, Wu J, Zhang SL (2013) Inheritance of hetero-diploid pollen S-haplotype in self-compatible tetraploid Chinese cherry (*Prunus pseudocerasus* Lindl). PLoS ONE 8(4):e61219. 10.1371/journal.pone.0061219
17. Gu C, Zhang S-L, Huang S-X, Heng W, Liu Q-Z, Wu H-Q, Wu J (2010) Identification of S-genotypes in Chinese cherry cultivars (*Prunus pseudocerasus* Lindl). Tree Genet Genomes 6(4):579–590. 10.1007/s11295-010-0273-2
18. Guo J, Cao K, Deng C, Li Y, Zhu G, Fang W, Chen C, Wang X, Wu J, Guan L, Wu S, Guo W, Yao JL, Fei Z, Wang L (2020) An integrated peach genome structural variation map uncovers genes associated with fruit traits. Genome Biol 21(1):258. 10.1186/s13059-020-02169-y
19. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9(1):R7. 10.1186/gb-2008-9-1-r7
20. Huang Y, He J, Xu Y, Zheng W, Wang S, Chen P, Zeng B, Yang S, Jiang X, Liu Z, Wang L, Wang X, Liu S, Lu Z, Liu Z, Yu H, Yue J, Gao J, Zhou X, Long C, Zeng X, Guo YJ, Zhang WF, Xie Z, Li C, Ma Z, Jiao W, Zhang F, Larkin RM, Krueger RR, Smith MW, Ming R, Deng X, Xu Q (2023) Pangenome analysis provides insight into the evolution of the orange subfamily and a key gene for citric acid accumulation in citrus fruits. Nat Genet 55(11):1964–1975. 10.1038/s41588-023-01516-6
21. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ (2017) Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun 8:14061. 10.1038/ncomms14061
22. Jiu S, Lv Z, Liu M, Xu Y, Chen B, Dong X, Zhang X, Cao J, Manzoor MA, Xia M, Li F, Li H, Chen L, Zhang X, Wang S, Dong Y, Zhang C (2024) Haplotype-resolved genome assembly for tetraploid Chinese cherry (Prunus pseudocerasus) offers insights into fruit firmness. Hortic Res 11(7):uhae142. 10.1093/hr/uhae142
23. Khan A, Korban SS (2022) Breeding and genetics of disease resistance in temperate fruit trees: challenges and new opportunities. Theor Appl Genet 135(11):3961–3985
24. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol 33(7):1870–1874. 10.1093/molbev/msw054
25. Layer RM, Chiang C, Quinlan AR, Hall IM (2014) LUMPY: a probabilistic framework for structural variant discovery. Genome Biol 15(6):R84. 10.1186/gb-2014-15-6-r84
26. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Genomics, arXiv
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078–2079. 10.1093/bioinformatics/btp352
28. Li N, He Q, Wang J, Wang B, Zhao J, Huang S, Yang T, Tang Y, Yang S, Aisimutuola P, Xu R, Hu J, Jia C, Ma K, Li Z, Jiang F, Gao J, Lan H, Zhou Y, Zhang X, Huang S, Fei Z, Wang H, Li H, Yu Q (2023) Super-

pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. Nat Genet 55(5):852–860. 10.1038/s41588-023-01340-y

29. Li W, Li Y, Wang X, Zhao G, Zhu G, Cao K, Fang W, Wu J, Ma K, Chen C, Wang L (2024) Genomic analysis provides insights into the westward expansion of domesticated peaches in China. Hortic Plant J 10(2):367–375. 10.1016/j.hpj.2022.07.009
30. Liu CJ, Jin GY, Kong ZC (2008) Archaeobotany—Research on Seeds and Fruits. Science, Beijing
31. Liu Z, Wang H, Zhang J, Chen Q, He W, Zhang Y, Luo Y, Tang H, Wang Y, Wang X (2024) Comparative metabolomics profiling highlights unique color variation and bitter taste formation of Chinese cherry fruits. Food Chem 439:138072. 10.1016/j.foodchem.2023.138072
32. Liu Z, Zhao Y, Yang P, Cheng Y, Huang F, Li S, Yang Y (2023) Population whole-genome resequencing reveals the phylogenetic relationships and population structure of four Hunan typical tea landraces. Beverage Plant Res 3(1):0–0. 10.48130/bpr-2023-0009
33. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. Gigascience 1(1):18. 10.1186/2047-217x-1-18
34. Mahmoud M, Gobet N, Cruz-Davalos DI, Mounier N, Dessimoz C, Sedlazeck FJ (2019) Structural variant calling: the long and the short of it. Genome Biol 20(1):246. 10.1186/s13059-019-1828-7
35. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021) 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(10):4647–4654. 10.1093/molbev/msab199
36. Marçais G, Kingsford C (2011) A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27(6):764–770. 10.1093/bioinformatics/btr011
37. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303. 10.1101/gr.107524.110
38. Meng X, Wang M, Luo M, Sun L, Yan Q, Liu Y (2023) Systematic evaluation of multiple NGS platforms for structural variants detection. J Biol Chem 299(12):105436. 10.1016/j.jbc.2023.105436
39. Ortiz EM (2019) Vcf2phylip v2.0: convert a vcf matrix into several matrix formats for phylogenetic analysis. 10.5281/zenodo.2540861
40. Qiao Q, Cao Q, Zhang R, Wu M, Zheng Y, Xue L, Lei J, Sun H, Liston A, Zhang T (2024) Genomic analyses provide insights into sex differentiation of tetraploid strawberry (*Fragaria moupinensis*). Plant Biotechnol J. 10.1111/pbi.14286
41. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842. 10.1093/bioinformatics/btq033

42. Raina A, Wani MR, Laskar RA, Tomlekova N, Khan S (2023) Advanced Crop Improvement, Volume 1: Theory and Practice. Springer
43. Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020) GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11(1):1432. 10.1038/s41467-020-14998-3
44. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics 28(18):i333–i339. 10.1093/bioinformatics/bts378
45. Shirasawa K, Isuzugawa K, Ikenaga M, Saito Y, Yamamoto T, Hirakawa H, Isobe S (2017) The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. DNA Res 24(5):499–508. 10.1093/dnares/dsx020
46. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123. 10.1101/gr.089532.108
47. Song B, Ning W, Wei D, Jiang M, Zhu K, Wang X, Edwards D, Odeny DA, Cheng S (2023) Plant genome resequencing and population genomics: Current status and future prospects. Molecular Plant
48. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res 34:W435–439. Web Server issue10.1093/nar/gkl200
49. Wang F, Xia Z, Zou M, Zhao L, Jiang S, Zhou Y, Zhang C, Ma Y, Bao Y, Sun H, Wang W, Wang J (2022a) The autotetraploid potato genome provides insights into highly heterozygous species. Plant Biotechnol J 20(10):1996–2005. 10.1111/pbi.13883
50. Wang J, Liu W, Zhu D, Hong P, Zhang S, Xiao S, Tan Y, Chen X, Xu L, Zong X, Zhang L, Wei H, Yuan X, Liu Q (2020) Chromosome-scale genome assembly of sweet cherry (*Prunus avium* L.) cv. Tieton obtained using long-read and Hi-C sequencing. Hortic Res 7(1):122. 10.1038/s41438-020-00343-8
51. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38(16):e164. 10.1093/nar/gkq603
52. Wang L, Wang Y, Zhang J, Feng Y, Chen Q, Liu ZS, Liu CL, He W, Wang H, Yang SF, Zhang Y, Luo Y, Tang HR, Wang XR (2022b) Comparative Analysis of Transposable Elements and the Identification of Candidate Centromeric Elements in the *Prunus* Subgenus *Cerasus* and Its Relatives. Genes (Basel) 13(4). 10.3390/genes13040641
53. Wang Y, Du H-M, Zhang J, Chen T, Chen Q, Tang H-R, Wang X-R (2018) Ploidy level of Chinese cherry (*Cerasus pseudocerasus* Lindl.) and comparative study on karyotypes with four *Cerasus* species. Sci Hort 232:46–51. 10.1016/j.scienta.2017.12.065
54. Wang Y, Hu G-p, Liu Z-S, Zhang J, Ma L, Tian T, Wang H, Chen T, Chen Q, He W (2022c) Phenotyping in flower and main fruit traits of Chinese cherry [*Cerasus pseudocerasus* (Lindl.) G. Don]. Sci Hort 296:110920
55. Wang Y, Li X, Feng Y, Wang J, Zhang J, Liu Z, Wang H, Chen T, He W, Wu Z, Lin Y, Zhang Y, Li M, Chen Q, Zhang Y, Luo Y, Tang H, Wang X (2023) Autotetraploid Origin of Chinese Cherry Revealed by

Chromosomal Karyotype and In Situ Hybridization of Seedling Progenies. Plants (Basel) 12(17). 10.3390/plants12173116

56. Wei X, Chen M, Zhang X, Wang Y, Li L, Xu L, Wang H, Jiang M, Wang C, Zeng L, Xu J (2023) The haplotype-resolved autotetraploid genome assembly provides insights into the genomic evolution and fruit divergence in wax apple (*Syzygium samarangense* (Blume) Merr. and Perry). Hortic Res 10(12):uhad214. 10.1093/hr/uhad214
57. Xanthopoulou A, Manioudaki M, Bazakos C, Kissoudis C, Farsakoglou AM, Karagiannis E, Michailidis M, Polychroniadou C, Zambounis A, Kazantzis K, Tsaftaris A, Madesis P, Aravanopoulos F, Molassiotis A, Ganopoulos I (2020) Whole genome re-sequencing of sweet cherry (*Prunus avium* L.) yields insights into genomic diversity of a fruit species. Hortic Res 7:60. 10.1038/s41438-020-0281-9
58. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1):76–82. 10.1016/j.ajhg.2010.11.011
59. Yü DJ (1979) Classification of Fruit Trees in China. Agricultural, Beijing
60. Zhang F, Wang Y, Lin Y, Wang H, Wu Y, Ren W, Wang L, Yang Y, Zheng P, Wang S, Yue J, Liu Y (2024a) Haplotype-resolved genome assembly provides insights into evolutionary history of the *Actinidia arguta* tetraploid. Mol Hortic 4(1):4. 10.1186/s43897-024-00083-6
61. Zhang H, He Q, Xing L, Wang R, Wang Y, Liu Y, Zhou Q, Li X, Jia Z, Liu Z, Miao Y, Lin T, Li W, Du H (2024b) The haplotype-resolved genome assembly of autotetraploid rhubarb *Rheum officinale* provides insights into its genome evolution and massive accumulation of anthraquinones. Plant Commun 5(1):100677. 10.1016/j.xplc.2023.100677
62. Zhang J, Chen T, Wang J, Chen Q, Luo Y, Zhang Y, Tang H-r, Wang X-r (2016) Genetic diversity and population structure in cherry (*Cerasus pseudocerasus* (Lindl). G. Don) along Longmenshan Fault Zones in China with newly developed SSR markers. Sci Hort 212:11–19. 10.1016/j.scienta.2016.09.033
63. Zhang J, Wang Y, Chen T, Chen Q, Wang L, Liu ZS, Wang H, Xie R, He W, Li M, Liu CL, Yang SF, Li MY, Lin YX, Zhang YT, Zhang Y, Luo Y, Tang HR, Gao LZ, Wang XR (2021) Evolution of Rosaceae Plastomes Highlights Unique *Cerasus* Diversification and Independent Origins of Fruiting Cherry. Front Plant Sci 12:736053. 10.3389/fpls.2021.736053
64. Zhang Q, Qi Y, Pan H, Tang H, Wang G, Hua X, Wang Y, Lin L, Li Z, Li Y, Yu F, Yu Z, Huang Y, Wang T, Ma P, Dou M, Sun Z, Wang Y, Wang H, Zhang X, Yao W, Wang Y, Liu X, Wang M, Wang J, Deng Z, Xu J, Yang Q, Liu Z, Chen B, Zhang M, Ming R, Zhang J (2022) Genomic insights into the recent chromosome reduction of autopolyploid sugarcane *Saccharum spontaneum*. Nat Genet 54(6):885–896. 10.1038/s41588-022-01084-1

Figures

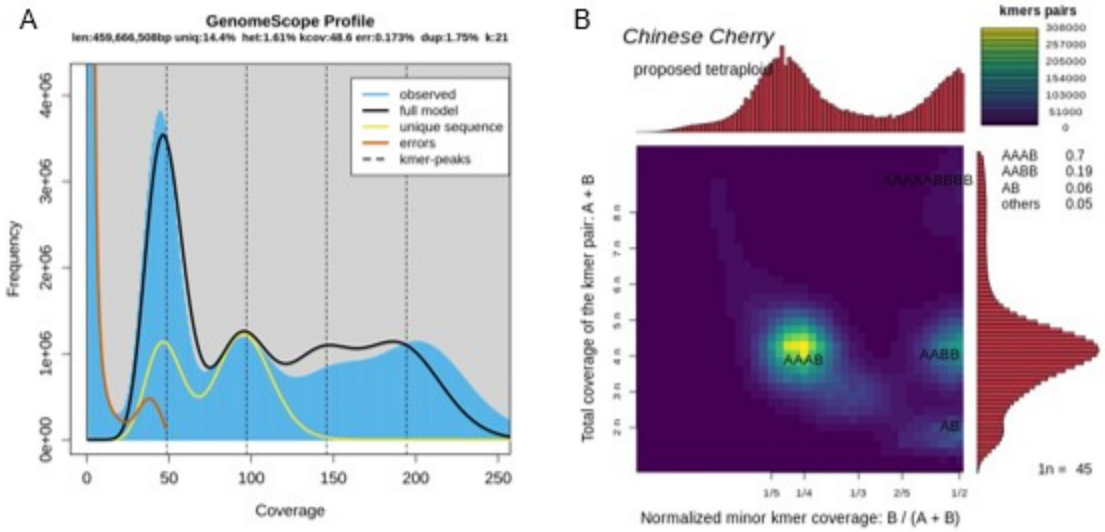


Figure 1

The basic features of *C. pseudocerasu* genome.

(A) 'Huangguo'Genome K-mer frequency distribution map. (B) The smudge plot analysis in *C. pseudocerasu* genome.

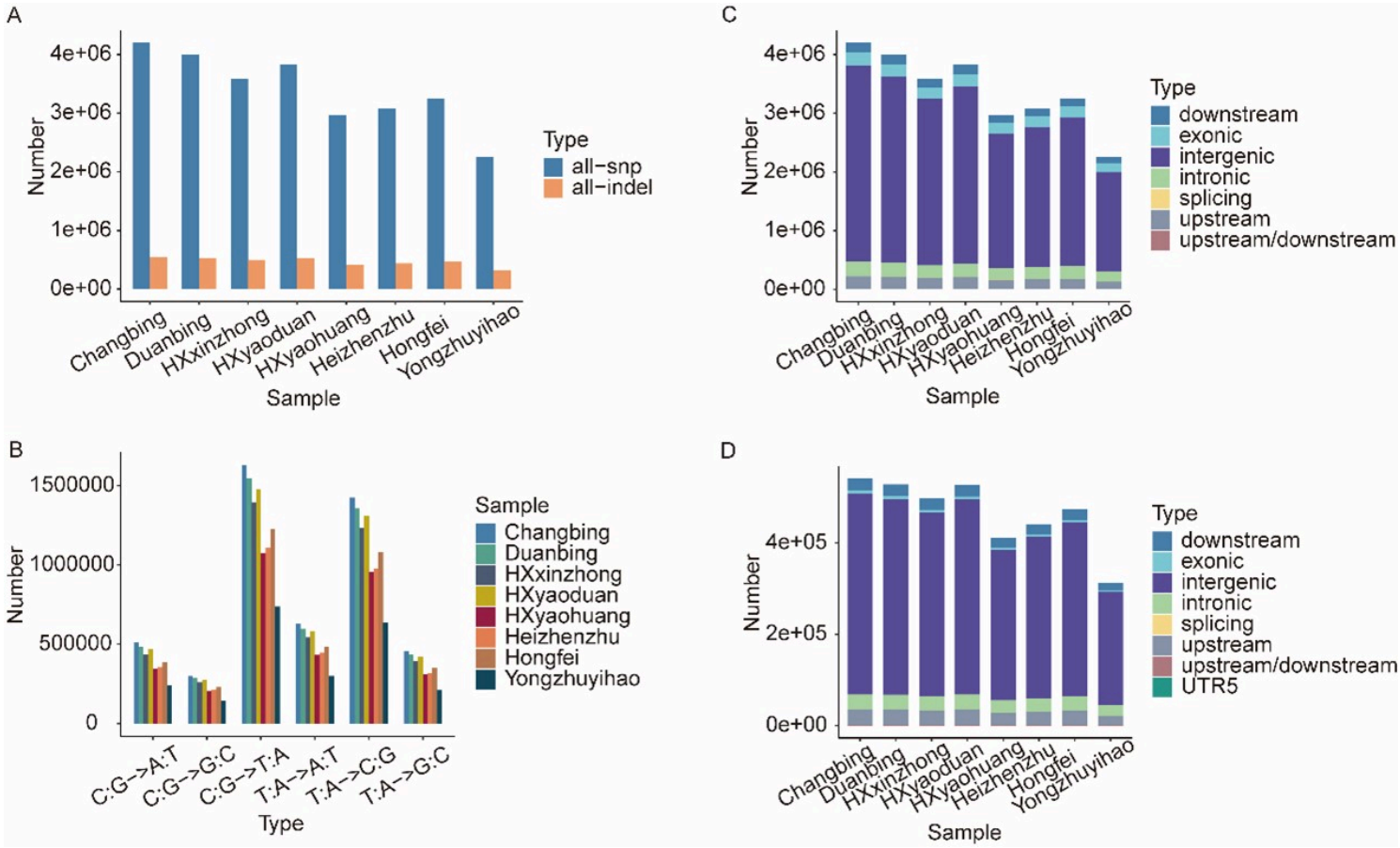


Figure 2

Variation statistic and distribution in 8 accessions.

(A) Statistics on the various numbers of SNPs and InDels in eight samples. (B) Distribution of SNP mutation type in eight samples. The x-axis delineates the type of SNP mutation, and the y-axis quantifies the number of SNPs. (C) Annotation distribution of SNPs variation. Different colors show the SNPs distribution of intergenic region, upstream, downstream, exon, intron, and splice. (D) Annotation distribution of InDels variation. Different colors show the InDels distribution of intergenic region, upstream, downstream, exon, intron, splice, and UTR5.

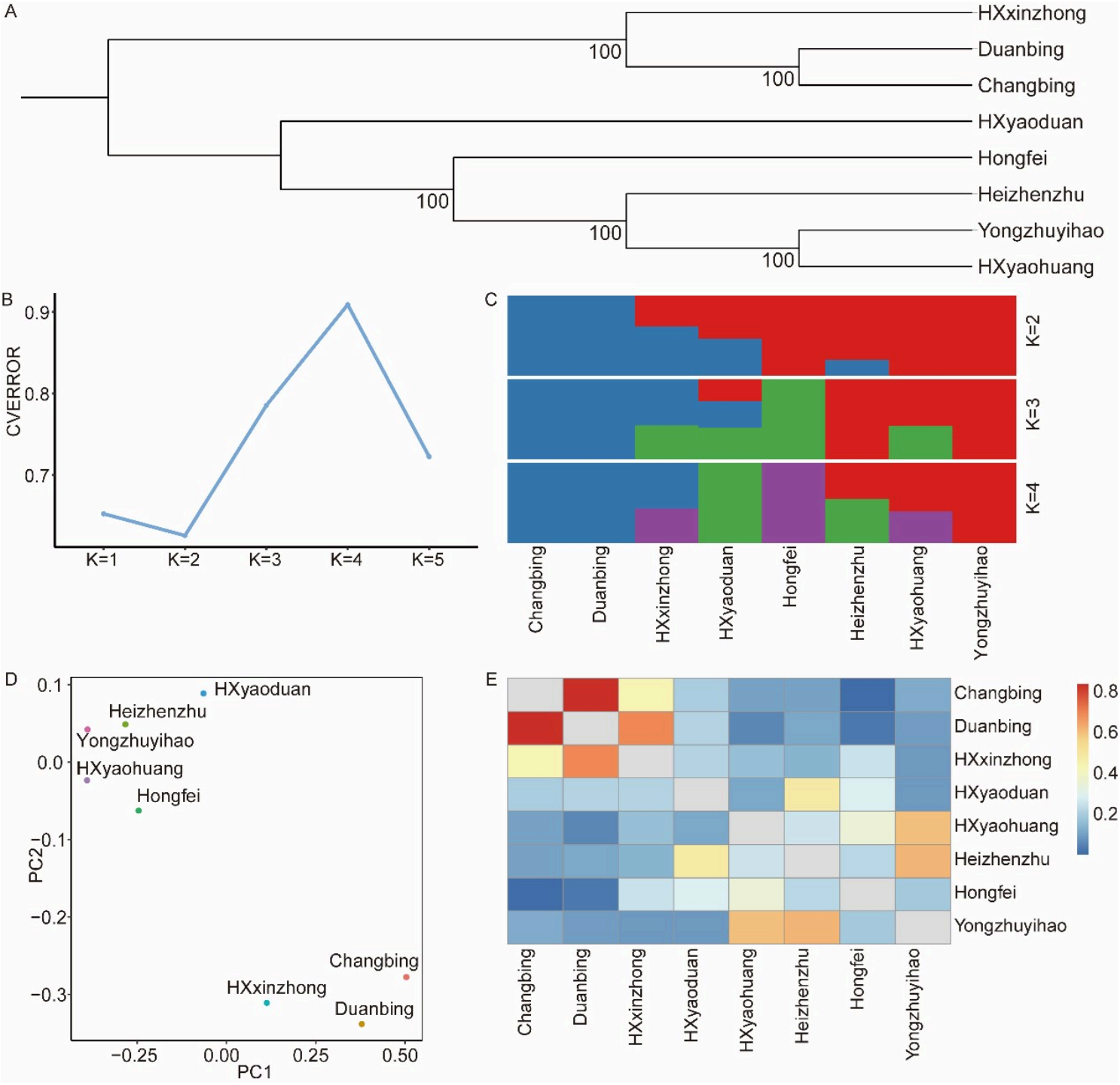


Figure 3

Phylogenetic and population genetic analyses of 8 *Cerasus pseudocerasus* cultivars.

(A) Unrooted neighbor-joining tree constructed with a bootstrap value of 1000 between individuals. (B) Cross-validation error results from different K values. (C) ADMIXTRUE analysis among the accessions (K = 2 - 4). (D) The principal component plot of 8 individuals. (E) Genetic relationship (G) matrix analysis between pairs of individuals, where the larger G value indicates the closer relationship between the samples.

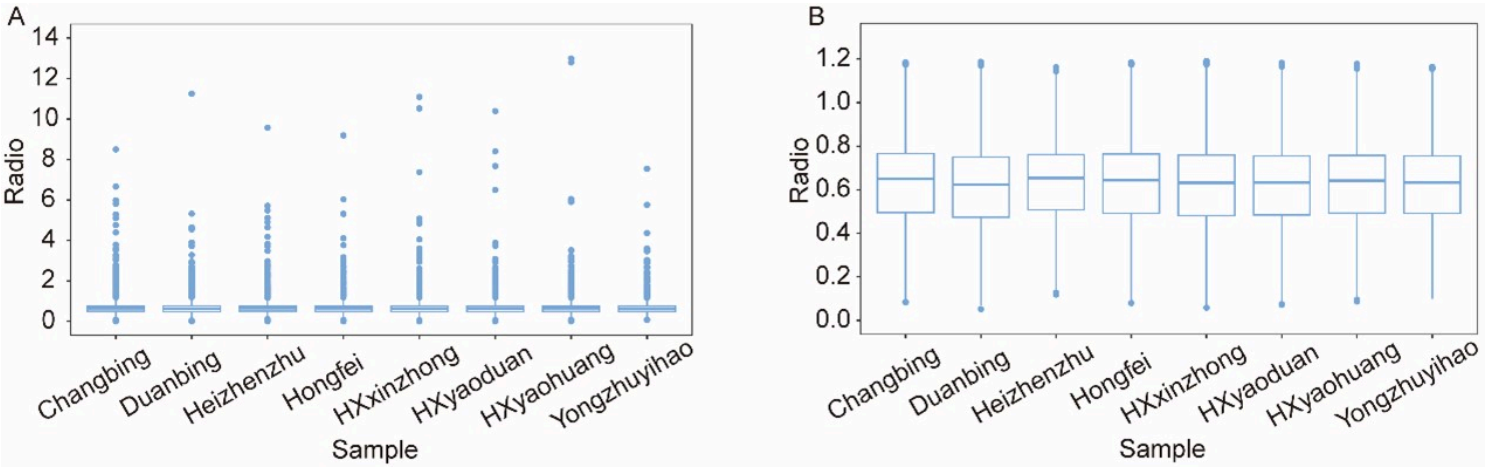


Figure 4

Boxplots of the ratio distribution for the DEL heterozygous genotypes in 8 individuals.

(A) Boxplots of the raw ratio data in 8 individuals. (B) Boxplot of the ratio excluding outliers in 8 individuals. The y-axis indicates the radio of V_{rd}/S_{rd}. V_{rd}: the average read depth of the variation; S_{rd}: the average reads depth of its surrounding area.

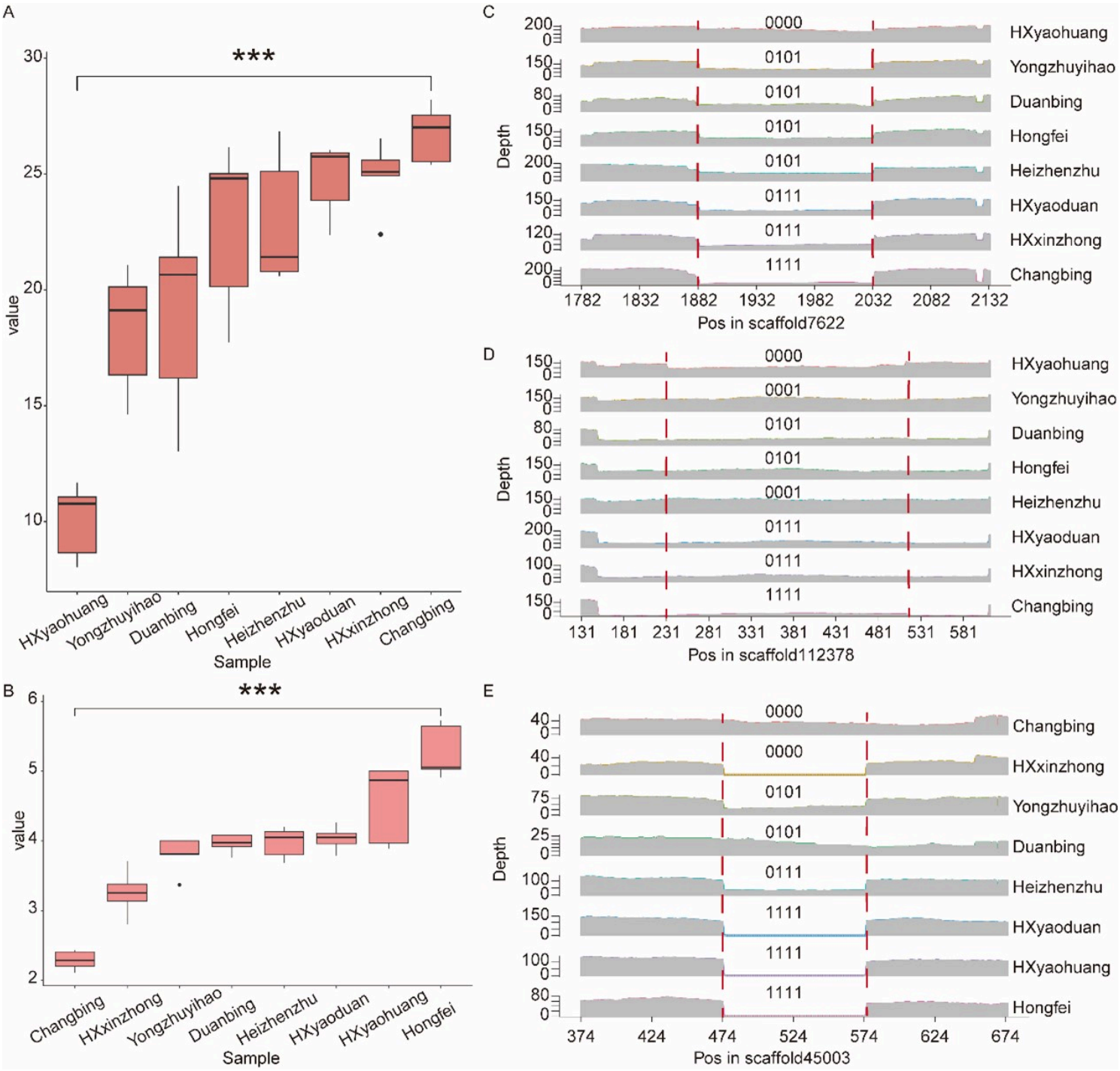


Figure 5

Comparison and candidate scaffold sites of fruit peel color and fruit weight in 8 samples.

(A) Comparison of fruit peel color in 8 samples. The y-axis indicates the a* value in fruit peel color, the higher the value, the redder the peel color. (B) Comparison of fruit weight in 8 samples. ∗∗∗∗ P-value <0.0001, t-test. The sample is arranged in ascending order of values. (C) The reads depth of the candidate scaffold site ('scaffold7622') related to fruit peel color in 8 samples. (D) The reads depth of the candidate scaffold site ('scaffold112378') related to fruit peel color in 8 samples. (E) The reads depth

of the candidate scaffold site (‘scaffold45003’) related to fruit weight in 8 samples. The red dashed line represents the DEL-type SV interval in 8 samples.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytables.xlsx