nature communications

Article

https://doi.org/10.1038/s41467-026-68463-8

The *Camellia sinensis* var. *sinensis* cv. Fuding Dabaicha genome unveils structural variation-driven metabolic innovation

Received: 5 March 2025

Accepted: 7 January 2026

Published online: 14 January 2026

Check for updates

Weiyi Zhang[1,6], Xiaohui Jiang[1,2,6], Shijie Luo[1], Arslan Tariq[3], Jan Buchmann[3], Dawei Gao[1], Xiaoliang Zhang[1], Alisdair R. Fernie[4], Björn Usadel[3,5] & Weiwei Wen[1]

Tea plants possess a highly heterozygous genome and produce diverse beneficial metabolites, yet the genomic basis of its metabolic diversity remains fully elusive. Here, we show that single-cell sequencing of 107 sperm cells, combined with PacBio HiFi and ONT ultra-long sequencing, enables an accurate haplotype-resolved genome assembly of Fuding Dabaicha (FDDB). Structural variations (SVs) between the two haplotypes comprise 23.8% of the genome and strongly influence crossover patterns. Using these phased genomes, we establish a half-sib-based QTL mapping platform and identify a *Gypsy* LTR insertion in the promoter of *CsDFRb*, which associated with the increasing of *p*-coumaroylquinate levels in young leaves. Moreover, mGWAS reveals 2649 additional loci associated with 2837 metabolites when using the FDDB genome rather than the 'Tieguanyin' reference genome for variant calling. Functional validation of *CsC3H* and *CsST2Ac* confirms their role in determining chlorogenic acid and sulfated metabolite levels in tea plant. In addition, we observe that allelic heterogeneity at *CsST2Ac* affects sulfated metabolite abundance. These phased genomes illuminate how SVs drive metabolic diversity and offer valuable resources for tea breeding.

Tea (*Camellia sinensis*), one of the most significant beverage plants globally, is featured by remarkable morphological, metabolic, and genetic diversity[1–3]. The tea plant genome has a large size (~3 Gb) and is characterized by high complexity and heterozygosity, underscoring the necessity of assembling high-quality genomes for dissecting some important agricultural traits, particularly the desirable metabolic traits.

Many chromosome-level genome assemblies have been achieved for tea, including those for elite cultivars and wild tea trees[4–9]. Additionally, two pan-genome studies of tea have assembled 22 and 11 genomes, respectively, revealing a high diversity within the species[10,11]. However, most of these genomes were assembled by PacBio Continuous Long Read (CLR) sequencing, and therefore have relatively low contiguity and base quality. Recently, high-quality assemblies have been achieved for many plant species using PacBio HiFi or ONT sequencing and have become a popular way of plant genome assembly[12,13]. However, tea is of high self-incompatibility and heterozygosity, and, therefore, haplotype information is crucial for functional genomics research in tea. Without haplotype genomes, significant genomic information may be lost in even consensus genomes reaching near complete levels[14,15].

Currently, the methods commonly used for genome phasing are generally based on three principles: (i) reads-based methods, such as

[1]National Key Laboratory for Germplasm Innovation & Utilization of Horticultural Crops, Hubei Hongshan Laboratory, College of Horticulture and Forestry Sciences, Huazhong Agricultural University, Wuhan, Hubei, China. [2]College of Food and Pharmaceutical Engineering (Guangxi Liupao Tea Modern Industry College), Wuzhou University, Wuzhou, Guangxi, China. [3]Institute for Biological Data Science, Heinrich Heine University, Düsseldorf, Germany. [4]Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm, Germany. [5]Institute of Bio- and Geosciences, IBG-4: Bioinformatics, CEPLAS, Forschungszentrum Jülich, Jülich, Germany. [6]These authors contributed equally: Weiyi Zhang, Xiaohui Jiang. e-mail: wwwen@mail.hzau.edu.cn

PacBio Hifi or ONT reads[16] or linked reads[17]; (ii) employing Hi-C data to assemble haplotype contigs, as demonstrated by software such as ALLHiC[18], HapHiC[19], and hifiasm[20]; and (iii) utilizing parentage information for trio-binning assembly to generate parent-specific *k*-mers for phasing[21]. Another method known as Gamete-binning involves leveraging the haploid nature of gamete cells to construct a genetic map and then achieve chromosomal-level phasing and haplotype genome assembly[22]. Many plant genomes have been phased using these methods. For example, the lychee genome was phased using reads-based approaches[23], while the sugarcane[18], kiwifruit[14], and rose[24] genomes were phased using Hi-C data. The pear genome was phased using trio-binning[19], and the apricot[22] and potato[25] genomes were phased using gamete-binning. In the case of tea plant, two haplotype-resolved genomes of the elite oolong tea cultivars, Tieguanyin (TGY) and Huangdan, were phased and assembled using Hi-C data with ALLHiC algorithm[4,9]. However, the phasing accuracy and overall quality of tea plant genomes still require improvement. Recently, a phasing pipeline that utilizes sperm cells from tea plants has been developed[26,27]. Combining phased SNPs with long-read sequencing has the potential to significantly enhance the quality of haplotype-resolved genomes for tea plants.

Fuding Dabaicha (FDDB; *C. sinensis* var. *sinensis*) is an elite tea cultivar widely grown in China for over 100 years. Parentage analysis has revealed that FDDB is one of the most frequently utilized parents in tea breeding[5,26,28]. Recently, we performed a pilot single sperm cell sequencing experiments for phasing of chromosomes and analysis of ASE genes in FDDB[26,29]. However, these experiments only phased the single nucleotide polymorphisms (SNPs) of FDDB, and it remains challenging to fully dissect the genetic mechanisms underlying allelic variations in FDDB due to the lack of a comprehensive haplotype genome. Therefore, it is particularly important to assemble haplotype genomes in tea plants, particularly in some elite accessions such as FDDB, for identifying superior haplotypes to facilitate tea breeding.

Plants produce structurally diverse metabolites, which play essential roles in plant growth, development, and stress responses, and are also important resources for human nutrition, medicine, and bioenergy[30]. Dissection of metabolic variations and the underlying genetic basis is critical for plant improvement through breeding[31]. Two primary methods are often used to detect metabolic quantitative trait loci (mQTL) in plants, namely linkage analysis using bi- or multi-parental populations and metabolite-based genome-wide association study (mGWAS) using natural populations. To date, there have been several studies of mQTL mapping in tea plants. For instance, a previous study identified two *O*-methyltransferases that play a role in the biosynthesis of *O*-methylated catechins through linkage analysis[32]. Another study revealed that *CsF3′5′H* and *CsANR* contribute to higher levels of catechins in tea plants through mGWAS[5,33]. Furthermore, a total of 14,022 mQTL were mapped through comprehensive large-scale targeted metabolite profiling of 2837 metabolites in the first and third leaves of 215 diverse tea accessions coupled with GWAS[34]. However, due to the lack of haplotype-resolved genomes, it remains challenging for linkage analysis in bi- or multi-parental populations to pinpoint critical mutations in candidate genes. Additionally, the outcomes of QTL detection are significantly influenced by the quality of the reference genome[35–37]. Different tea accessions have high genomic diversity and significant structural variations (SVs)[10]. Hence, high-quality haplotype-resolved genomes have overwhelming advantages in genetic studies and exploration of how genetic variations between different genomes affect the mGWAS results.

In this work, we first use PacBio HiFi, ultra-long ONT, single sperm sequencing, and combine DNA sequencing of offspring population to develop a pipeline for the assembly of accurate haplotype-resolved genomes of an elite tea cultivar, FDDB. Then, based on the assembled haplotype-resolved genomes, we investigate SVs and their formation mechanisms as well as the allele-specific expression (ASE) patterns in nine different tissues and an FDDB offspring population. We further establish an effective QTL mapping pipeline coupled with ASE analysis platform and reveal associations of 2837 metabolites from fresh tea leaves with allelic variations in the offspring population. Finally, we carry out mGWAS of these 2837 metabolites with the assembled genome, and compare the results with those obtained with the TGY genome to elucidate how the high-quality and haplotype-resolved genome assembled in this study improves mQTL detection in tea plants, and verify the candidate genes that are involved in chlorogenate biosynthesis and sulfation. The high-quality single-cell phasing based haplotype-resolved genomes unlock hidden SVs and SV-metabolism nexus, greatly enhance our understanding of metabolic innovation in tea plants.

Results

Assembly of consensus genome for FDDB

We developed a sperm cell and long read-based pipeline to assemble the haplotype genomes of FDDB, with the first step of our assembly strategy being the generation of a consensus genome (Supplementary Fig. 1 and Supplementary Method 1). The young leaves of FDDB were harvested and sequenced using a combination of PacBio and ONT platform. A total of 87 Gb (~29×) and 96 Gb (~32×) PacBio HiFi reads, and ultra-long ONT reads (N50 100 Kb) were produced, respectively. The HiFi and ONT reads were then combined to generate the preliminary genome assembly using hifiasm[20]. Compared with using HiFi reads alone, the incorporation of ONT reads could greatly improve the genome assembly quality. For example, contig N50 increased from 93.0 Mb to 175.7 Mb while the contig number decreased from 638 to 400 (Supplementary Table 1). After comparison with published tea organelle sequences (NCBI accession numbers: MH019307 and MK574877), 49 contigs of the 400 contigs were matched with organelle sequences and removed for further analysis (Supplementary Data 1).

To scaffold contigs to pseudomolecules, we constructed a genetic map based on 107 single-sperm sequencing data of FDDB. Sperm cell reads were mapped to the preliminary genome assembly, resulting in the detection of 13,262,473 high-quality SNPs and 1342 genetic bins. Interestingly, there was a significant increase in the number of SNPs while an obvious decrease in that of bins compared with using DASZ genome as the reference[26], probably due to the higher assembly quality of FDDB than DASZ or large-scale SVs between the two genomes. Except for two genetic bins, all genetic bins were anchored to 15 linkage groups. Each double crossover (CO) event within a contig was manually checked, and no obvious assembly errors were detected. Four short contigs were overlapped with large contigs were considered redundant contigs and removed (Supplementary Table 2). Finally, 28 contigs with a total length of 3,090,354,494 bp (95.83% of preliminary genome assembly) were anchored according to the genetic map. It is worth noting that only 13 chromosomal gaps remained (Supplementary Fig. 2).

Combination of results from different assembly software may help improve the continuity of the draft genome[38]. Hence, we constructed two more genome assemblies with verkko[39] and necat[40] to fill the gaps on chromosomes. It seemed that the assembly results from both verkko and necat were not as good as those from hifiasm, as their contig N50 was 11.5 Mb and 3.6 Mb, respectively (Supplementary Table 3). The verkko and necat assemblies filled six and four chromosomal gaps, respectively (Supplementary Figs. 3 and 4). Several potential assembly errors at the end of contigs were detected and corrected. For instance, a redundant fragment was identified at the side of gap 1 on chromosome 1, which was corrected by a contig from necat (Supplementary Fig. 3). Following gap filling and correction, only three gaps remained in the final consensus genome, all of which were located in highly repetitive regions, making them difficult to be filled

https://doi.org/10.1038/s41467-026-68463-8

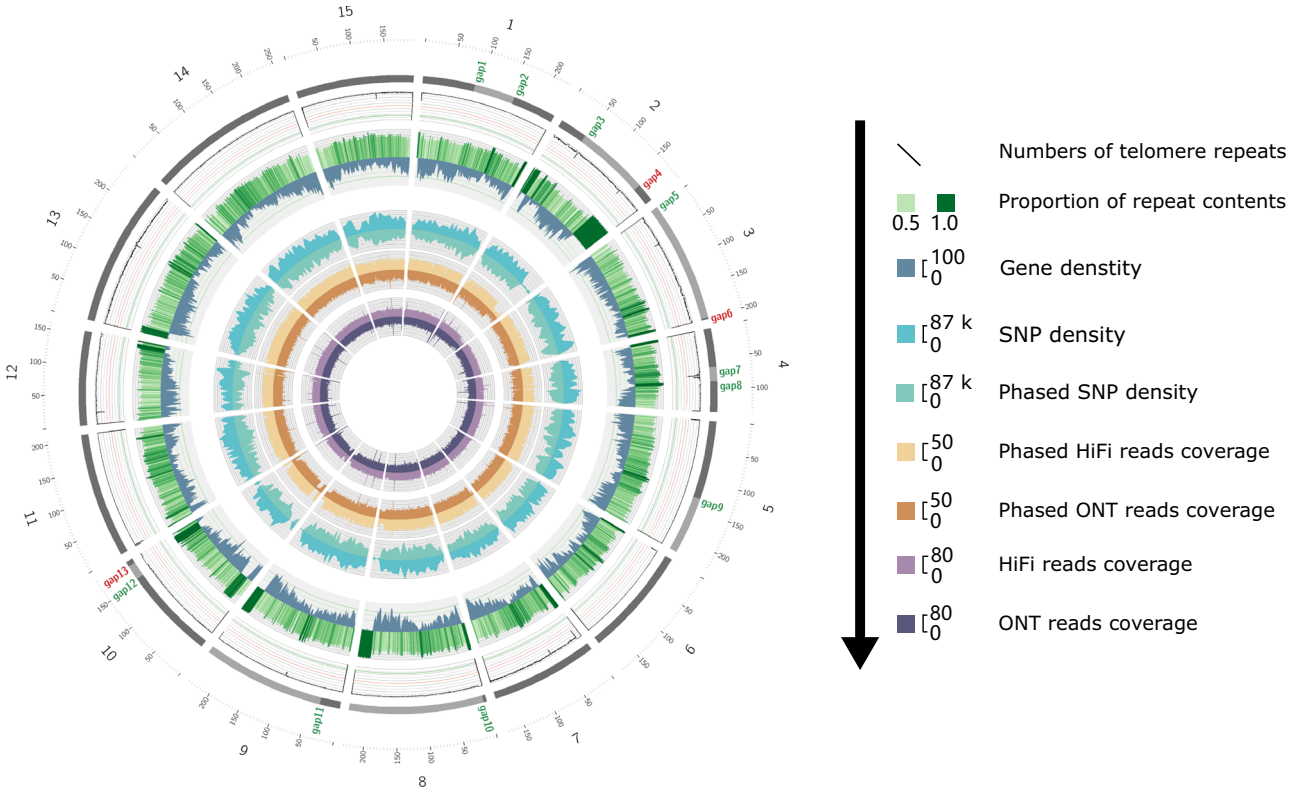

Fig. 1 | Circos plot of genome assembly and annotation of FDDB. Switches between light grey and dark grey on the chromosomes indicate gaps in the preliminary assembly. The green text indicates gaps filled by verkko or necat assembly, while red text represents gaps that remain unfilled. From the outside to the inside of circus plot are the numbers of telomere repeats, proportion of repeat contents, gene density, SNP density, phased SNP density, phased Hifi reads coverage, phased ONT reads coverage, coverage of HiFi and ONT reads, respectively. Source data are provided as a Source Data file.

Table 1 | Summary of the FDDB assembly

Assembly	FDDB consensus	FDDB haplotype A[b]	FDDB haplotype B[b]	TGY consensus	TGY haplotype A	TGY haplotype B
Contig N50 (Mb)	208.99	208.85	199.17	1.94	0.22	0.22
Scaffold N50 (Mb)	208.99	208.85	199.17	213.47	208.55	198.56
Total number of Gaps	3	3	2	3,536	22,805	22,296
Identified telomeres	26	27	24	/	/	/
Length of assembly (Gb)	3.08	3.02	2.99	3.06	3.06	2.93
Genome BUSCO (%)[a]	99.0	99.0	99.1	93.7	84.8	83.2
Proteome BUSCO (%)[a]	98.8	99.0	98.8	92.4	85.0	82.4
QV	51.19	61.44	62.25	/	/	/
Error rate	7.60×10^{-6}	7.17×10^{-7}	5.96×10^{-7}	/	/	/
Total number of genes	50,496	49,388	49,076	51,384	29,792	22,828
Total number of gene functional annotated	35,959	35,172	35,109	34,911		
Percent of repeat elements (%)	79.64	79.34	79.27	78.2	74.28	74.19
LAI	19.57	25.85	26.17	10.17	/	/
K-mer completeness	77.02	76.16	76.04	/	/	/
Combined *k*-mer completeness	96.80			/		

[a]BUSCO was estimated by data base "embryophyta_odb10" and the total number of BUSCOs is 1614.
[b]Genome evaluation was based on V2 of FDDB haplotype genomes.

(Fig. 1). We used telomere repeats (5′-CCCTAAA-3′) to search for telomeres along each chromosome. As a result, 26 out of the 30 telomeres were detected, and telomeres at one side of chromosomes 2, 10, 12, and 13 were not detected (Fig. 1). In summary, we achieved a high-quality consensus genome of FDDB (3.08 Gb; Table 1). A further comparison with other published tea genomes demonstrated that the FDDB genome assembled in this study has an obviously higher contig N50 than all other tea genomes, indicating assembly of a high-quality genome for FDDB (Supplementary Data 2 and Table 1).

Genome annotation and quality assessment

A total of 79.64% of FDDB genome was identified as repetitive sequences, which is similar to the case for other tea accessions (Table 1)[10]. *Gypsy* played a dominant role, accounting for 38.25% of the

genome, followed by CACTA DNA transposon, which accounted for 5.14% of the genome (Supplementary Data 3). A total of 50,496 protein-coding genes were identified, and 35,959 of them had functional annotations in GO, KEGG, or Pfam databases (Table 1).

We then investigated gene and transposable element (TE) distribution along chromosomes of the FDDB genome. Interestingly, several TE-enriched regions (>10 Mb) were identified, which contained nearly no genes and tended to be located at the end of the chromosomes. We used coverage of HiFi and ONT reads to assess the accuracy of the assembly correction. Most genomic regions showed a uniform distribution of sequencing depth, although some high-coverage regions remained in the FDDB genome, suggesting potential assembly errors that may require further improvement (Fig. 1). To further evaluate assembly in TE-enriched regions, we analyzed the sequencing depth of HiFi and ONT reads with mapping quality (mapQ) > 30. These regions were generally well covered, with an average depth around 15 for both HiFi and ONT (Supplementary Fig. 5). Although some low-coverage areas were observed, especially on chromosome 10, the overall pattern supports the correctness of the assembly in TE-rich regions (Supplementary Fig. 5). The largest TE-enriched region was found on chromosome 2 (chr2-2), spanning an interval with a length over 40 Mb (Fig. 1). We further classified TEs in these regions and revealed that *Gypsy* played a dominant role in chr2-2 (chr2: 152–198 Mb), chr8-1 (chr8: 204–230 Mb), Chr9-1 (chr9: 218–237 Mb) and chr10-1 (chr10: 13–32 Mb), while *Mutator* and *TcMar* were enriched in chr2-1 (chr2: 4–16 Mb) and chr10-2 (chr10: 165–181 Mb), respectively (Supplementary Fig. 6a). We further determined the LTR insertion time for the four *Gypsy* enrichment regions. The results indicated that *Gypsy* expansion in three of these four regions (chr2-2, chr8-1, and chr9-1) occurred ~0.15 million years ago (Mya), which is slightly earlier than the LTR burst in tea plants (Supplementary Fig. 6b). However, *Gypsy* expansion on chromosome 10 occurred ~2.5 Mya, and was more ancient than that on other chromosomes (Supplementary Fig. 6b). *Gypsy* burst was unevenly distributed across the genome and may form some hotspots in some specific chromosomal positions in tea plants. However, these TE hotspots were not observed in other tea genomes (Supplementary Fig. 7), demonstrating high quality of the FDDB genome.

We further aligned the Illumina short reads of FDDB to the final assembly to verify the assembly quality. The results showed that 99.31% of the reads were mapped, covering 99.99% of the FDDB genome, which indicated a high accuracy of the genome sequences. According to Benchmarking Universal Single Copy Orthologs analysis (BUSCO), 1607 of 1614 core plant genes were detected, and 1598 (99.0%) of them were complete genes (Table 1), implying a high integrity of the genic sequences. The proteome BUSCO score was 98.8 %, indicating a high completeness of the annotated gene set (Table 1). LTR Assembly Index (LAI), which can be used to evaluate the completeness of repetitive regions by estimating the percentage of intact LTR retroelements, was 19.57 for the FDDB genome, suggesting a high level of completeness of the repetitive sequences, which can be assigned to reference class (Table 1). A total of 25,961 homozygous SNPs were detected with a base error rate of 8.42×10^{-6}. A *k*-mer based method was also employed to determine the base quality and completeness of the genome. The consensus quality (QV) and error rate were calculated to be 51.19 and 7.60×10^{-6}, respectively, which are similar to those of T2T genome assemblies for other species such as maize[41] and kiwi fruit[14]. However, the *k*-mer completeness of the FDDB genome was estimated to be 77.02 (Table 1), indicating that considerable genomic resources may have been lost without a haplotype genome.

Haplotype genome assembly of FDDB

The above results have indicated the necessity of assembling haplotype-resolved genomes to rescue the completeness of the consensus genome of FDDB. Single sperm genome sequencing data provide a great genetic resource for accurately phasing the two haplotypes of FDDB genome. We re-mapped 107 sperm cell data to the consensus genome of FDDB and obtained a total of 15,818,536 highly accurate phased SNPs, which accounted for about 63.8% of the total heterozygous SNPs in the FDDB genome. We further integrated the DNA sequencing data of seven FDDB offspring to phase the remaining heterozygous SNPs in the FDDB genome (Supplementary Figs. 1, 8, and 9; Supplementary Method 1). By using this pipeline, a total of 20,586,760 SNPs were phased, among which 14,126,398 SNPs were common SNPs phased by using sperm cells (Supplementary Table 4). Only 0.012% (1,762) of the SNPs were inconsistent between offspring and sperm cells, indicating a high phasing accuracy of this pipeline (Supplementary Table 4). We then merged the SNPs phased from the two datasets, resulting in the phasing of a total of 22,277,136 SNPs, which accounted for 89.80% of the total SNPs (Supplementary Table 4), and there was no obvious difference in density between all SNPs and phased SNPs (Fig. 1). We then assigned the HiFi and ONT reads to the two haplotypes by using these phased SNPs, and these phased reads covered 99.95% of the consensus genome of FDDB, suggesting that a high-completeness haplotype genome can be generated using these phased SNPs (Fig. 1 and Supplementary Table 5).

Additional PacBio HiFi (reaching 197 Gb in total) and ultra-long ONT data (reaching 165 Gb in total) were incorporated to further improve contiguity and gap filling. We also phased the short reads of FDDB and generated haplotype-specific *k*-mers for the two haplotypes, respectively, which were thus subsequently combined with 197 Gb HiFi reads (~65×) and 165 Gb ultra-long ONT reads (~55×) to conduct trio binning assembly using hifiasm and verkko2, respectively (Supplementary Fig. 1 and Supplementary Data 4). The raw assembly of haplotype A and B generated by hifiasm was scaffolded according to the consensus genome of FDDB and the remaining gaps were filled using assembly from verkko2, resulting in the anchoring of a total of 18 and 17 contigs of haplotype A and B to pseudomolecules, respectively (Fig. 2a, b and Table 1). The contig N50 of the two haplotype genomes reached 208.85 Mb and 199.17 Mb, respectively (Fig. 2a, b and Table 1). Both haplotype A and B genomes showed a high QV (61.44 and 62.25, respectively) and low error rate (7.17×10^{-7} and 5.96×10^{-7}, respectively), suggesting that the haplotype genomes have a high base accuracy (Table 1). Moreover, 27 and 24 telomeres were detected in haplotype A and B genomes, respectively, revealing their high completeness (Fig. 2a, b and Table 1). A total of 49,388 and 49,076 protein-coding genes were identified in haplotype A and B genomes, respectively. The complete BUSCO of haplotype A and B genomes reached 99.0% (1598 out of 1614) and 99.1% (1599 out of 1614), respectively. The corresponding proteome BUSCO scores were 99.0% and 98.8%, indicating accurate and comprehensive reconstruction of the genic region and annotations for both haplotypes (Table 1). Both the genomic and proteomic BUSCO scores of FDDB were higher than those of the haplotype assembly of TGY[4], indicating the high-quality haplotype assembly of FDDB (Table 1). In addition, haplotype A and B genomes had 79.34% and 79.27% repetitive contents, respectively (Table 1). The *k*-mer completeness of haplotype A and B genomes was 76.16 and 76.04, respectively, but when the two genomes were merged, the completeness rose to 96.80 (Table 1), underscoring the effectiveness of haplotype genome assembly in enhancing the overall genome completeness. The LAI values were 25.85 and 26.17 for haplotype A and B genomes, respectively, indicating that both of them could be assigned to the golden level[42] (Table 1).

Due to the lack of parental data of FDDB, we employed *k*-mer generated from single sperm sequencing to evaluate haplotype phasing. The blob plot shows that most of the contigs were partitioned to the expected haplotypes (Fig. 2c). Specifically, haplotype A and haplotype B markers were identified in their respective genomes, with minimal contaminating markers from each other (Fig. 2c). The switch

https://doi.org/10.1038/s41467-026-68463-8

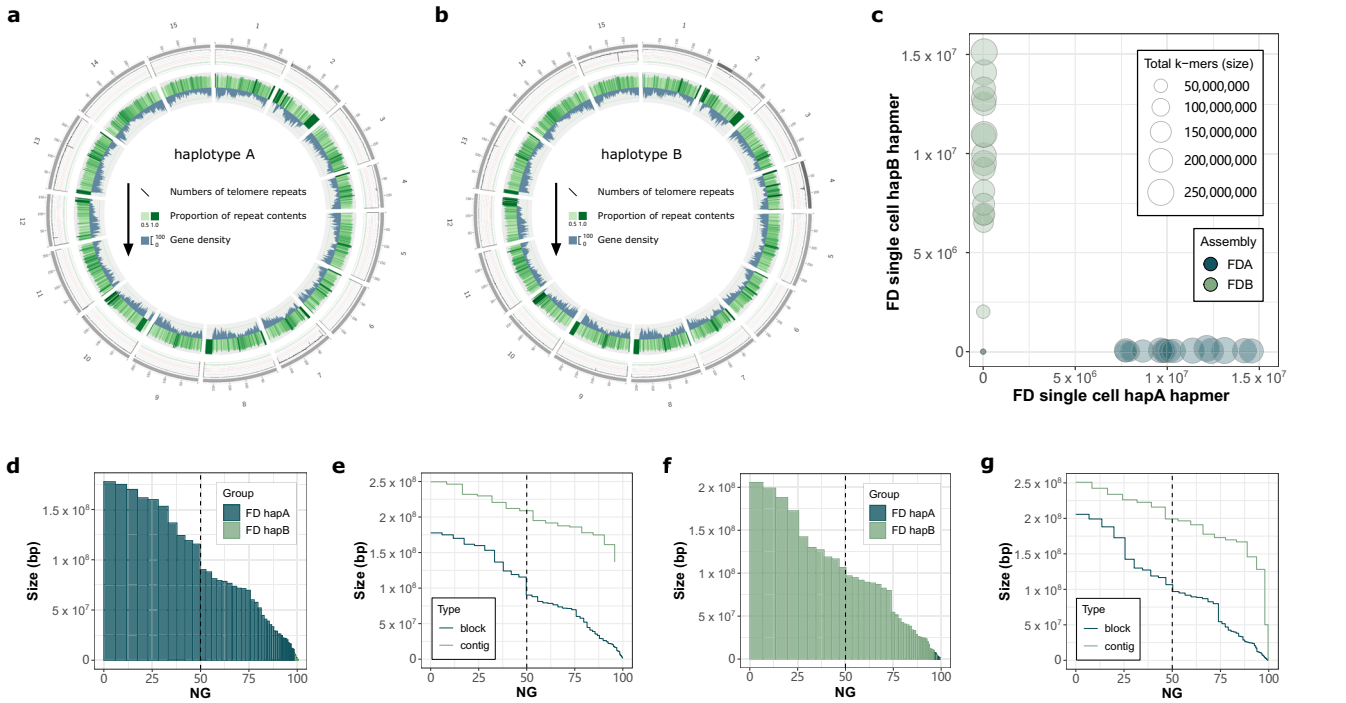

Fig. 2 | Haplotype genome assembly of FDDB. a, **b** Circos plot of genome assembly and annotation of haplotype A and B of FDDB (V2). Switches between light grey and dark grey on the chromosomes indicate gaps in the haplotype assembly. From the outside to the inside of circos plot are telomere repeats, proportion of repeat contents, and gene density, respectively. Hap-mer blob plot of haplotype assembly (V2). Blue and green blobs indicate contigs of hapotype A and B, respectively. Blob size represents total *k*-mer size for each contig, and each blob is plotted according to the number of contained haplotype A (*x* values) and B (*y* values) hap-mers. **d**, **f** NG plots of the phase blocks of haplotype A and B assembly, respectively (V2). Haplotype blocks are sorted by size. *X*-axis is the percentage of the genome size covered by phase blocks, and *y*-axis indicates the length for each haplotype block (bp). Blue and green represent blocks from haplotype A and B, respectively. **e**, **g** Comparison of NG plot between phased block and contig length for haplotype A and B genomes, respectively (V2), showing relatively high continuity of phased assembly. *X*-axis is the percentage of the genome size covered by phase blocks and contigs, while *y*-axis is the length for phase blocks and contigs. Length of the phased block is shorter than that of the contig due to the switches between the two haplotypes. Source data are provided as a Source Data file.

error rate was estimated to be 0.38% and 0.42% for haplotype A and B, respectively, suggesting a high accuracy of haplotype phasing. We then calculated the phased blocks for each haplotype genome by using a 500-Kb window with at most 1000 switches. The phased blocks were sorted by size, and only a few small blocks were derived from incorrect haplotypes (Fig. 2d, f), with the N50s of phased blocks reaching 90.19 Mb and 106.62 Mb for haplotype A and B, respectively, which demonstrated good performance in haplotype phasing (Fig. 2e, g).

Structural variation and crossover (CO) landscape between two haplotypes

Different from traditional diploid genome assemblies, which combine both haplotypes into a single synthetic consensus sequence, individual haplotype genome assemblies enable the investigation of haplotype diversity. When the two haplotype genomes were aligned, large-scale rearrangements between them were observed (Fig. 3a). A total of 297 inversions, 6,424 duplications, 98,865 deletions (>50 bp), and 102,871 insertions (>50 bp) were detected between the two haplotype genomes, which spanned ~143.1 Mb, 67.3 Mb, 247.8 Mb, and 257.1 Mb, respectively (Fig. 3b and Supplementary Data 5). SVs accounted for ~23.8% of the genome (~3 Gb), which is twice as many as the genomic differences between different tea accessions[10], indicating the high quality genome enhance SV detection.

To further evaluate the impact of assembly quality on comparative analyses, we compared the two haplotypes of FDDB with the two haplotypes of TGY using SyRI (Supplementary Fig. 10). The syntenic alignment length between FDDB and TGY haplotypes ranged from 1.07 to 1.17 Gb, substantially shorter than the syntenic regions observed between the two haplotypes of FDDB (2.43–2.44 Gb). Consistently, the total size of insertions and deletions between FDDB and TGY haplotypes (11.2–12.9 Mb) was markedly smaller than that detected between the two FDDB haplotypes (223.7–226.1 Mb; Supplementary Data 5). Similar patterns were also observed when comparing FDDB with Huangdan (HD), another elite oolong tea cultivar (Supplementary Data 5 and Supplementary Fig. 11). Specifically, the syntenic regions between FDDB and HD were notably smaller than those between the two FDDB haplotypes, and the total size of deletions and insertions between FDDB and HD (12.1–14.3 Mb) was much smaller than that between the two FDDB haplotypes (Supplementary Data 5). These observations highlight that high-quality genome assemblies are essential for reliable detection of PAVs, CNVs, and other structural variants.

To investigating the mechanism of SV formation in tea genome, we analyzed the sequences as well as the break points of SVs. A total of 100,942 SVs (~48.4% of total SVs) overlapped within a single transposable element (TE), and among them, 56,276 (~27% of total SVs) of SVs covered at least 90% of a single TE, indicating that they may be originated from TE insertion events (Fig. 3c). TE classification analysis demonstrated that *Gypsy* (Class I) and *Harbinger* miniature inverted-repeat transposable elements (MITE; Class II) represent the two predominant TE categories in TE insertion events (Supplementary Fig. 12). We further characterized the TEs with intact structures in these TE insertion events and identified 12,959 SVs (99.5 Mb) that could be associated with these structurally intact TEs. TE insertion time analysis using intact TEs releveled those distinct patterns of insertion history, with a single recent burst peak of Class I TEs (~0.025 MYA) and two burst peaks for the Class II TEs (~0.01 and ~0.39 MYA, respectively), suggesting recent TE mobilization persists in the tea genome, while

https://doi.org/10.1038/s41467-026-68463-8

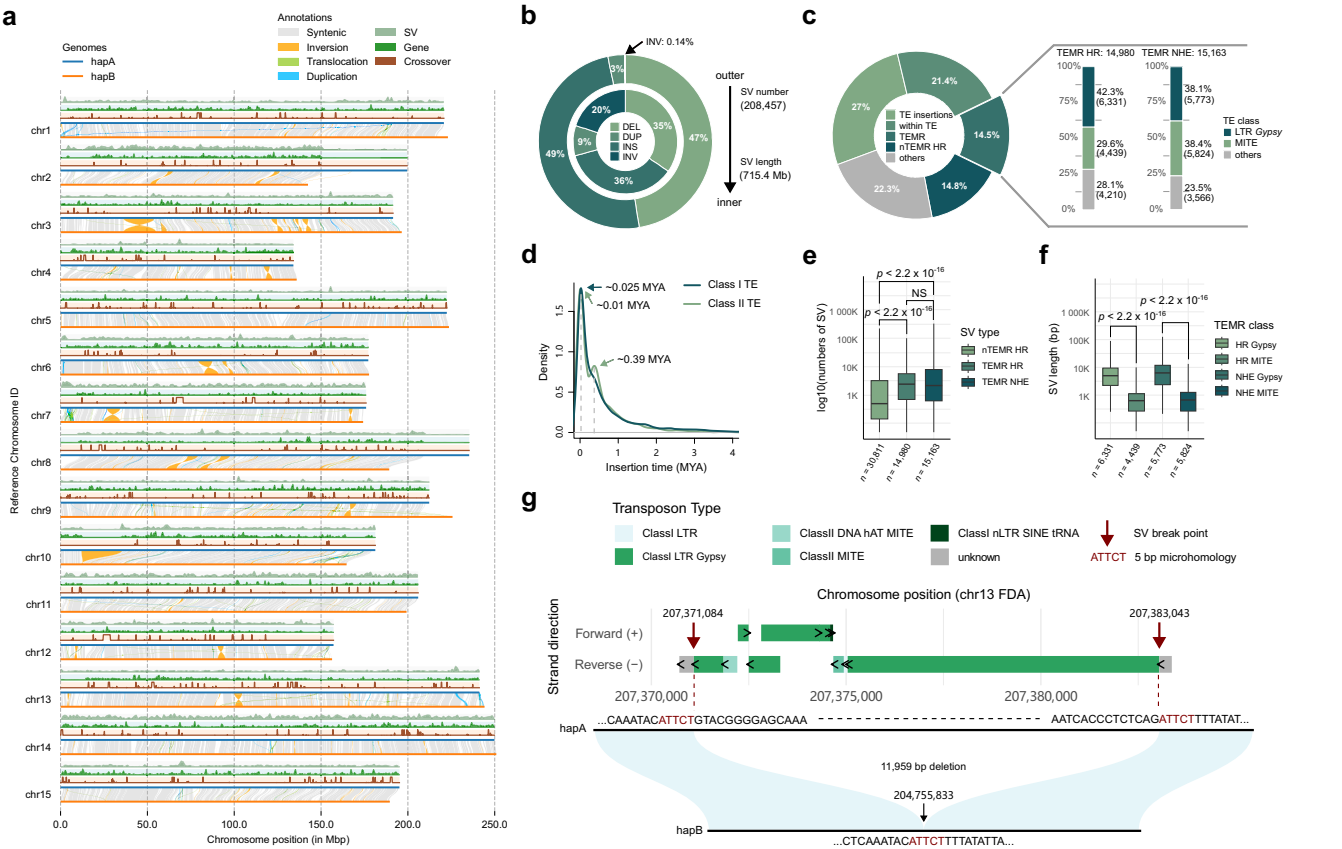

Fig. 3 | Crossover landscape and SVs between the two haplotype genomes. **a** Crossover landscape of FDDB. The chromosomes of haplotype A and B are shown in blue and orange lines, respectively. Grey, orange, green, and blue lines between two haplotype chromosomes indicate syntenic, inversion, translocation, and duplication alignments, respectively. Light green, green and dark red lines above the chromosomes represent SV density, gene density plot, and distribution of COs, respectively. **b** Proportion of the major four types of SVs. The outer circle and inner circle represent the percentage of SV numbers and SV length of insertion (INS), deletion (DEL), inversion (INV) and duplication (DUP), respectively. **c** Summary of the origins of SVs. The percentage in the plot indicated the SV numbers. **d** The density of estimate Class I and Class II TE insertion time. **e** Boxplot of SV length generated by TEMR and nTEMR HR. The boxes represent 75% and 25% quartiles; the central line represents the median. *P*-value is calculated using the normal approximation by Wilcoxon test (two-sided; $n = 30,811$, $n = 14,980$, and $n = 15,163$ for nTEMR HR, TEMR HR, and TEMR NHE, respectively). **f** SV length of Gypsy and MITE mediated TERMs. The boxes represent 75% and 25% quartiles; the central line represents the median. *P*-value is calculated using the normal approximation by Wilcoxon test (two-sided; $n = 6331$, $n = 4439$, $n = 5773$, and $n = 5824$ for HR Gypsy, HR MITE, NHE Gypsy, and NHE MITE, respectively). **g** A 11,959 bp deletion TEMR between two Gypsy repeats. Top panel: TE distribution in haplotype A genome of FDDB. The triangle represents the direction for each TE. Bottom panel: Alignment at the break points of deletion. "ATTCT" represents the 5 bp microhomology. These analyses were based on V1 of FDDB haplotype genomes. Source data are provided as a Source Data file.

ancient insertions/deletions may be maintained population-level polymorphisms (Fig. 3d).

Homologous transposable elements (TEs) can act as substrates for ectopic DNA repair, leading to SVs through TE-mediated rearrangements (TEMRs). This mechanism has been established as a significant contributor to SV formation in both humans and plants[43,44]. We examined the flanking sequences of SV breakpoints (±100 bp) and identified that approximately 14.5% (30,143) of SVs may have originated from TEMRs (Fig. 3c). To assess the significance of TEMRs in tea plant genomes, we shuffled the SVs 10,000 times and compared the observed versus expected TEMR events. The number of observed TEMRs was nearly three times higher than the expected, indicating that TEMRs play a crucial role in SV formation in tea plants ($p < 0.0001$; Supplementary Fig. 13a). We categorized TEMRs into homologous recombination (HR) and non-homologous end joining (NHE) events, based on whether the breakpoints contained microhomology sequences (≥3 bp). A total of 14,980 and 15,163 of TEMRs were classified into HR and NHE categories, respectively (Fig. 3c). In alignment with the patterns observed in TE insertion events, *Gypsy* and MITEs were identified as the two predominant categories of TEMR events in tea plants (Fig. 3c). The lengths of SVs associated with *Gypsy*-mediated TEMR events, including both NHE and HR, were significantly greater than those associated with MITE-mediated TEMR events (Fig. 3f; Wilcoxon test; $p < 2.2 \times 10^{-16}$). In addition, we investigated HR events (microhomology ≥3 bp) that occurred without homologous TEs at the breakpoints (nTEMR HR) in the tea genome, identifying 30,811 such events, accounting for about 14.8% of the total SVs (Fig. 3c). A permutation test (10,000 iterations) revealed that the observed number of HR events was about 1.4-fold and 4-fold higher than the expected values for 3- and 4-bp microhomology, respectively, further confirming that HR is a key mechanism for SV formation in tea plants ($p < 0.0001$; Supplementary Fig. 13c, d). Although the number of nTEMR HR events was approximately twice that of TEMR HR events, the length of SVs mediated by TEMR HR was significantly greater than those mediated by nTEMR HR (Fig. 3e; Wilcoxon test; $p < 2.2 \times 10^{-16}$). Two representative examples of TEMR HR and nTEMR HR events are shown in Fig. 3g and Supplementary Fig. 14, respectively.

CO plays a pivotal role in generating genetic diversity, thereby ensuring proper chromosome segregation and influencing the inheritance of traits[45]. Single sperm data and high-quality haplotype-resolved genomes provide a great opportunity to investigate how SVs between the haplotype genomes shape the CO patterns in tea plants.

https://doi.org/10.1038/s41467-026-68463-8

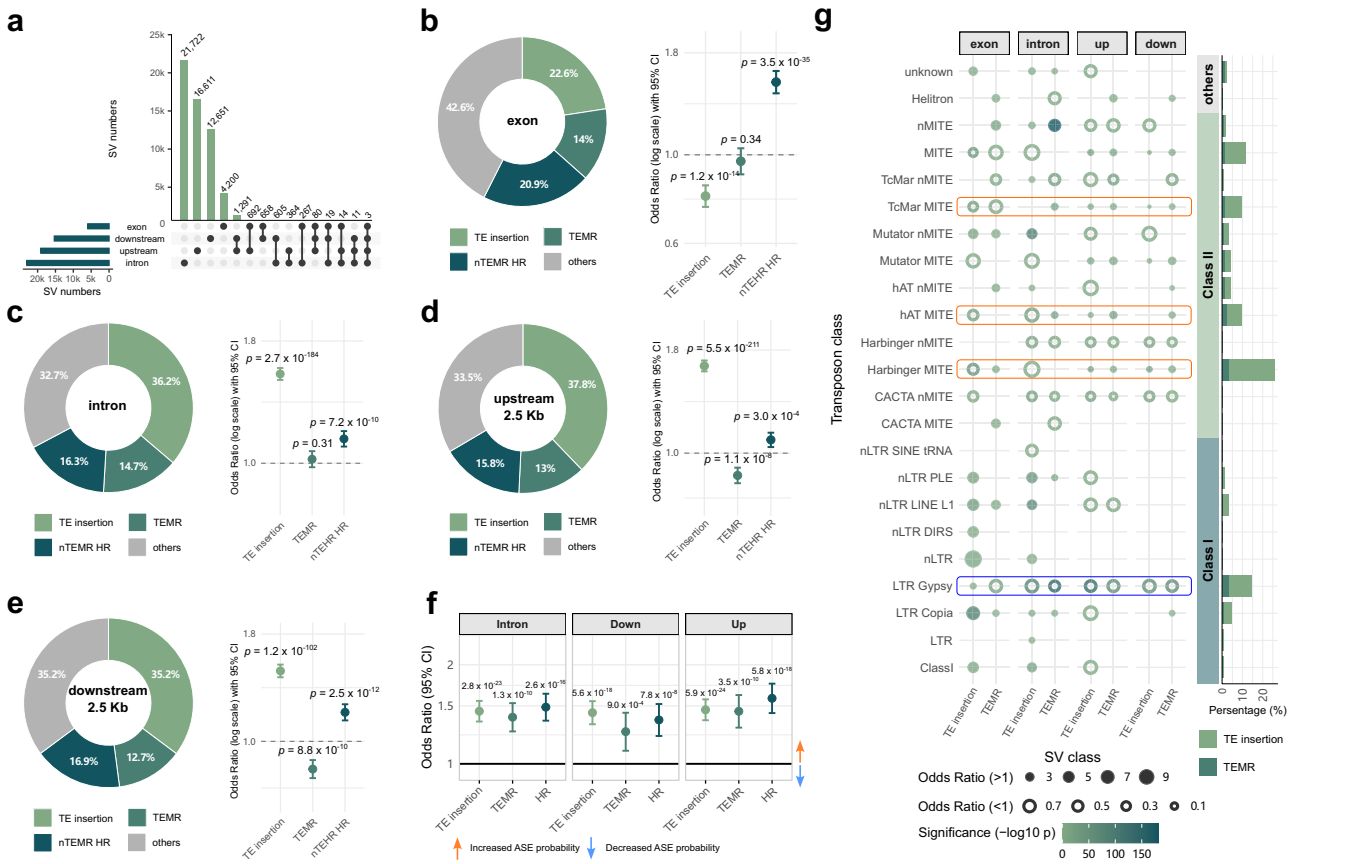

Fig. 4 | SVs induced allelic variations between the two haplotype genomes. **a** Upset plot of SVs overlapped with exon, intron, 2.5 Kb upstream, and 2.5 Kb downstream of gene. **b**–**e** Proportion of different SV types overlapping with exon, intron, upstream, and downstream regions, respectively. The forest plot represents the two-sided Fisher's exact test for each type of SVs. Odds ratio > 1 indicate the enrichment of SVs in the corresponding features. In contrast, odds ration <1 indicate the deplete of SVs in the corresponding features. The centre point shows the odds ratio, and the bar represents the 95% of confidence interval. **b** Exons contained 1339, 831, 1238, and 2525 SVs from TE insertions, TEMRs, nTEMRs, and others, respectively. **c** Introns contained 8335, 3383, 3755, and 7532 SVs from TE insertions, TEMRs, nTEMRs, and others, respectively. **d** In the 2.5-kb upstream regions, the numbers were 7209, 2471, 3005, and 6381. **e** In downstream regions, they were 5397, 1943, 2589, and 5389. The total numbers were 56,276, 30,143, 30,811, and 91,227, respectively. **f** Two-sided fisher exact test of SVs and allele specific expression (ASE) patterns. Odds ratio > 1 suggests the increasing probability of ASE when the SV occurred. In introns, ASE/AEE genes associated with TE insertions, TEMRs, and HR were 1399/3145, 650/1456, and 740/1,562, respectively. In downstream regions, the corresponding numbers were 1048/2330, 336/816, and 495/1,120, and in upstream regions, 1361/3028, 484/1034 and 620/1,227. In total, 5,777 ASE and 17,368 AEE genes were identified. **g** Enrichment patterns of TEs in exon, intron, upstream, and downstream of genes. Left panel: Blob plot of the enrichment of each TE class in the four genic features. Solid circle represents Odds ratio > 1 (enriched), while unfilled circle indicates odds ration <1 (depleted). *P*-value was calculated by two-sided fisher exact test. Right panel: Barplot of percentage of each class of TE. These analyses were based on V1 of FDDB haplotype genomes. Source data are provided as a Source Data file.

Hence, we mapped sperm reads to haplotype A genome to recall the phased SNPs and detect COs. As a result, a total of 1240 COs were identified with an average resolution of 90.9 Kb. Due to the potential presence of various alignment types within CO regions, such as syntenic and translocation alignments, we specifically chose regions where the overlap ratio of CO to alignment exceeded 0.8, resulting in the identification of 846 COs (Supplementary Data 6). Among these COs, 98.7%, 0.83%, 0.35% and 0.12% of them were in syntenic regions, non-assigned regions, highly divergent regions, and duplication regions, respectively. No COs were detected in rearrangement regions (insertion, deletion, translocation, and inversion regions) between the two haplotype genomes (Supplementary Data 6 and Fig. 3a), or in TE enriched regions (Fig. 3a). Therefore, the SVs between the two haplotype genomes greatly shape the CO distribution of tea plants.

SVs contribute to allelic functional imbalance in tea plant genome

SVs have a significantly greater potential to disrupt gene function compared to SNPs. Their larger-scale genomic alterations often interfere with coding sequences, regulatory elements, and chromatin structure, leading to more profound effects on gene activity[46–48]. To investigate potential functional impact of SVs, we identified a total of 5933 exon, 23,005 intron, 19,066 upstream (within 2.5 kb of transcription start sites), and 15,318 downstream (within 2.5 kb of transcription end sites) SVs (Fig. 4a). We examined the proportion of SVs generated by different mechanisms and found that TE insertions and TEMRs exhibited distinct accumulation patterns across genic features. Specifically, TE insertions were significantly depleted in exon regions, while significantly enriched in intron, upstream, and downstream regions (Fisher's exact test; $p < 0.05$; Fig. 4b–e). In contrast, TEMRs showed a significant depletion in upstream, and downstream regions (Fisher's exact test; $p < 0.05$; Fig. 4d–e), a pattern similar to that observed in humans[49], though the enrichment in exonic and intronic regions was not statistically significant (Fig. 4b-c). We further inspected which classes of TEs accumulated most frequently in the genic regions of the tea genome. MITE (61.22%) and *Gypsy* (14.62%) were identified as the two most abundant TE classes in these genic regions (Fig. 4g). Interestingly, except the TEMR events in exon, MITE and *Gypsy* displayed contrasting enrichment patterns across different genic features. MITE was significantly enriched in intron, upstream,

and downstream regions, while it was depleted in exon regions (Fisher's exact test; $p < 0.05$; Fig. 4g). In contrast, *Gypsy* elements were significantly enriched in exon regions but showed significant depleted in other genic features (Fisher's exact test; $p < 0.05$; Fig. 4g). Except for TE-related SVs, we also found that nTEMR HR exhibited significant enrichment across all the four genic features (Fisher's exact test; $p < 0.05$; Fig. 4b–e), indicating that nTEMR HR also plays a potential role in shaping gene function in tea plants. Additionally, the exon of 6852 protein coding genes overlapping with SVs were identified.

High-quality haplotype genomes allow a comprehensive analysis of ASE genes, which play important roles in phenotypic variations[50,51]. Here, we combined the two datasets, including the RNA-seq data of 9 tissues and a half-sib population of FDDB consisting 31 accessions[5,26], to investigate genes showing ASE patterns between the two haplotype genomes. A total of 5439 and 925 ASE genes were identified by 9 tissues and the half-sib population, respectively (Supplementary Figs. 15–17). We found that 50.3% (2783) of ASE genes exhibited tissue-specific expression patterns and a consistent ASE pattern across nine tissues, except for 199 ASE genes with evident differences between the two haplotypes (Supplementary Fig. 17). It is worth noting that 483 ASE genes were detected in both five leaf tissues and half-sib populations, indicating that these genes have relatively stable ASE patterns across different spatial conditions and genetic backgrounds (Supplementary Fig. 16). We further investigated whether the SVs affected the ASE patterns and showed that all of SVs generated from TE insertion, TEMR and nTERM HR significantly enriched in intron, upstream and downstream of ASE genes and no obvious difference was observed among different origins of SVs (Fisher's exact test; $p < 0.05$; Fig. 4f). Additionally, odds ratio of SVs in upstream was slightly higher than SVs in intron and downstream regions, indicating that SVs in the promoter regions had a higher influence in gene expression levels (Fig. 4f).

Effective QTL mapping with FDDB offspring and haplotype genomes

To demonstrate the application of the haplotype genomes assembled in this study, we conducted QTL mapping by using the large-scale metabolomic data (2837 metabolites profiled in two types of leaf tissues) of 31 offspring accessions of FDDB generated in our previous study (Supplementary Method 2)[34]. We first identified 299 genetic bins with an average length of 7.9 Mb from 31 FDDB offspring accessions (Supplementary Data 7) and utilized these bins as markers to identify metabolic QTL from the two leaf tissues. A total of 282 metabolites from young leaf (YL) and 386 metabolites from the third leaf (TL) showed significantly differential accumulation between haplotype A and B in FDDB offspring (ANOVA; $p < 0.001$; Fig. 5a and Supplementary Data 8). We further identified the potential QTL hotspots by counting the number of QTL for each bin. A hotspot was defined as having no fewer than 15 QTL, and the adjacent bins were merged into one QTL hotspot, resulting in the identification of a total of 11 QTL hotspots (q1–q11) for 198 and 274 metabolites from YL and TL, respectively (Fig. 5a and Supplementary Data 9). To gain deeper insights into each hotspot, an MS/MS fragment ion network was constructed with no fewer than four shared fragments in 2 ppm (Fig. 5b). Some QTL with shared fragments were detected within the same hotspot. For example, q3 harbored five unknown metabolites that shared fragments with *p*-coumaroylquinate, indicating that this hotspot may be related to the phenylpropanoid pathway (Fig. 5b).

We observed a significant difference in the accumulation of *p*-coumaroylquinate in bin99 on chromosome 5 in both YL and TL between the two haplotypes (Fig. 5c, d; $p = 7.58 \times 10^{-6}$ and $p = 5.14 \times 10^{-3}$, respectively; ANOVA). Functional analysis identified *CsDFRb* (*FDA05g15866*), which encodes a dihydroflavonol 4-reductase (DFR) and is a key enzyme in the flavonoid pathway, as a candidate gene that influences the *p*-coumaroylquinate content in FDDB offspring. Flavonoids and *p*-coumaroylquinate have *p*-coumaric acid as a common precursor, which serves as a critical intermediate. *CsDFRb* may indirectly influence the *p*-coumaroylquinate content in tea plants. Gene expression analysis revealed that haplotype A of *CsDFRb* exhibited a significantly higher expression value than haplotype B in FDDB offspring (Fig. 5e; $p = 1.19 \times 10^{-7}$; ANOVA). Moreover, haplotype A of *CsDFRb* also displayed consistently higher expression levels than haplotype B across nine different tissues of FDDB, suggesting a stable ASE pattern of *CsDFRb* across different genetic backgrounds and tissues (Fig. 5f). The expression level of *CsDFRb* and the content of *p*-coumaroylquinate showed significant negative correlations in both YL and TL of FDDB offspring (Supplementary Fig. 18a, b; correlation coefficient = −0.68 and −0.51, respectively; $p = 1.5 \times 10^{-4}$ and 0.008, respectively; *t*-test). Sequence alignment revealed that the promoter of *CsDFRb* in haplotype B genome contains an intact *Gypsy* retrotransposon (~9.6 Kb), which was located 1299 bp away from the start codon. The TE insertion was validated using HiFi reads, and the sequencing depth was notably lower in the TE region compared with in the flanking regions, confirming the reliability of the TE insertion (Fig. 5g). Dual-luciferase assay showed that the promoter activity of haplotype A of *CsDFRb* was significantly higher than that of haplotype B (Fig. 5h, i). Transient overexpression of *CsDFRb* in tobacco leaves resulted in a significant decrease in *p*-coumaroylquinate content (Fig. 5j; $p = 0.038$; Supplementary Fig. 19). These results suggested that *Gypsy* insertion decreases the expression level of *CsDFRb* in haplotype B compared with that in haplotype A and further affects the content of *p*-coumaroylquinate in tea plants. Methylation analysis using ONT reads further revealed strong DNA methylation signals within the TE insertion, but not within the *CsDFRb* gene body (Supplementary Fig. 20 and Supplementary Data 10). These findings suggest that epigenetic silencing of the *Gypsy* element may reduce promoter activity in haplotype B, thereby lowering *CsDFRb* expression and indirectly influencing *p*-coumaroylquinate content in tea plants.

FDDB genome enhances the power of mGWAS

A high-quality reference genome can improve the efficiency of gene mining[35,36]. To ascertain the potential of the FDDB genome in this regard, with a large dataset containing 215 diverse tea accessions with 2837 profiled metabolites, we conducted mGWAS using the consensus FDDB genome as reference. We identified a total of 5125 mQTL associated with 986 metabolites (Supplementary Data 11; $p < 4.04 \times 10^{-6}$). Specifically, 2034 mQTL were linked to 549 metabolites in YL and 3091 mQTL were associated with 708 metabolites in TL (Supplementary Data 11). To explore whether the different reference genome would affect the mGWAS results, we re-performed mGWAS using TGY as the reference genome. A total of 4847 mQTL associated with 985 metabolites were identified using TGY genome. Among these mQTL, 1900 and 2947 were associated with 540 and 687 metabolites in YL and TL, respectively (Supplementary Data 12; $p < 4.14 \times 10^{-6}$). We then used minimap2[52] to align the mQTL region of FDDB genome to that of TGY genome to identify the matched mQTL between them. Surprisingly, when using the FDDB genome, 2,649 of 5,125 mQTL (~51.7%) were newly identified, including 1109 and 1540 mQTL from YL and TL, respectively (Supplementary Data 11). Several mQTL showed stronger QTL signals in Manhattan plot when using the FDDB genome as reference, indicating that it can be used to detect novel mQTL compared with the TGY genome (Supplementary Fig. 21).

To further elucidate how the two reference genomes influence the mGWAS results, we used a sliding window with 1 Mb length and permutation test of 1000 times to identify the mQTL hotspots in FDDB and TGY, respectively ($p < 0.01$; at least seven and nine mQTL for YL and TL, respectively). A total of 76 hotspots were identified using FDDB genome, with 40 and 36 hotspots for YL and TL, respectively (Supplementary Data 13). As a comparison, a total of 71 hotspots were identified using TGY genome, with 38 and 33 for YL and TL, respectively (Supplementary Data 14). And 34 of the 76 (~44.7%) hotspots

https://doi.org/10.1038/s41467-026-68463-8

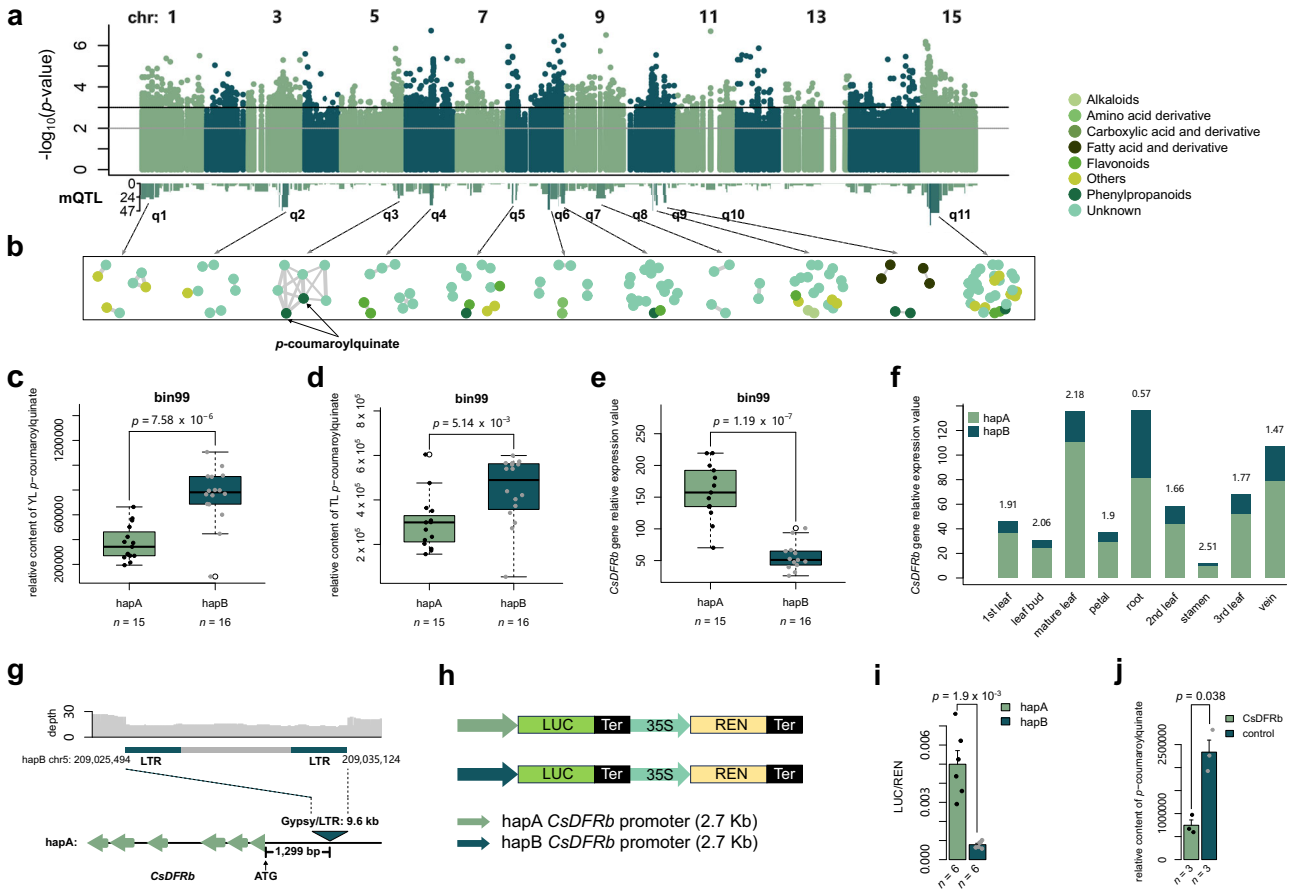

Fig. 5 | mQTL detection in FDDB offspring population. a Manhattan plot of mQTL detection in FDDB offspring population (ANOVA *F*-test; without adjusted). The lower barplot shows the mQTL number distribution along the chromosomes ($p < 0.001$; ANOVA *F*-test; without adjusted). Putative mQTL hotspots are indicated by q1-q11. **b** Isomer network of metabolites in each hotspot. Nodes represent metabolites and edges indicate at least four shared isomers detected within 2 ppm. Colors of nodes indicate classification of metabolites. **c**–**e** Boxplot of *p*-coumaroylquinate contents in YL, TL, and expression level of *CsDFRb* between the two haplotypes of bin99, respectively (ANOVA *F*-test; $n = 15$ and 16 for accessions with haplotype A and B of FDDB, respectively). The central line represents the median; the box edges indicate the first (25%) and third (75%) quartiles, and whiskers show the minimum and maximum values excluding outliers. **f** ASE patterns of *CsDFRb* in nine tissues of FDDB. Numbers above each bar indicate the $\log_2$(fold change) between haplotype A and B. **g** Plot of a gypsy LTR insertion in the upstream of *CsDFRb*. The lower plot shows gene structure of *CsDFRb* and the relative position of Gypsy LTR insertion. The upper plot exhibits the sequencing depth of HiFi reads in the insertion region of Gypsy LTR in haplotype B genome. **h** Diagram of vector construction of dual luciferase assay. **i** Promoter activity between the FDDB haplotype A and B of *CsDFRb* (two-sided *t*-test; $n = 6$). Error bars indicate the standard errors, and the centres indicate the mean values. **j** Transient expression of *CsDFRb* in tobacco leaves. Control indicates the injection of empty vector (two-sided *t*-test; $n = 3$). Error bars indicate the standard errors, and the centres indicate the mean values. These analyses were based on V1 of FDDB haplotype genomes. Source data are provided as a Source Data file.

identified using FDDB were only detected by using the FDDB genome, including 22 and 12 from YL and TL, respectively (Supplementary Data 13). For example, a mQTL hotspot (YL_hotspot_32) on chromosome 13 with 9 metabolites was only identified using FDDB genome, which was not detected using TGY genome (Supplementary Data 13 and Supplementary Fig. 22).

To validate the precision of mQTL identification using the FDDB genome, we examined a mQTL hotspot (TL_hotspot_12) in TL on chr6 of the FDDB genome (Fig. 6a and Supplementary Data 13). Within TL_hotspot_12, a total of nine mQTL were included, which were associated with compounds such as chlorogenic acid (CGA) and its two isomers (CGA (a) and CGA (b)). LocusZoom analysis showed that several SNPs in the exons of two candidate genes were significantly associated with CGA (b) content and formed a strong LD block (Fig. 6b). Functional annotation revealed that *FD06g18840* (*CsC3Ha*) and *FD06g18841* (*CsC3Hb*) encode an ortholog of Arabidopsis *AtC3H*, coumarate 3-hydroxylase, which could catalyze *p*-coumaroylquinate to generate CGA in Arabidopsis (Fig. 6c). Transient expression of *CsC3Ha* obviously increased the contents of five isomers of CGA in tobacco leaves (Fig. 6d and Supplementary Fig. 23). We further transiently expressed *CsC3Ha* in tea leaves, which led to significant increases in the contents of three CGA isomers (paired *t*-test; $p < 0.05$; Fig. 6e and Supplementary Fig. 24). In summary, using the FDDB genome the most likely functional gene influencing the CGA content could be identified within this hotspot in the tea population.

Another noteworthy mQTL hotspot, which was identified as YLhotspot_18 and TLhotspot_15 on chromosome 9 of the FDDB genome, was found in both YL and TL and associated with 15 and 16 metabolites, respectively (Fig. 6f and Supplementary Data 13), including an annotated metabolite eugenolsulfate. Based on the gene functional annotation, three candidate genes, *FD09g28251 (CsST2Aa)*, *FD09g28252* (*CsST2Ab*) and *FD09g28253* (*CsST2Ac*), were selected in the FDDB genome (Fig. 6g and Supplementary Data 15). These three genes belong to the sulfotransferase family, which can catalyze the

https://doi.org/10.1038/s41467-026-68463-8

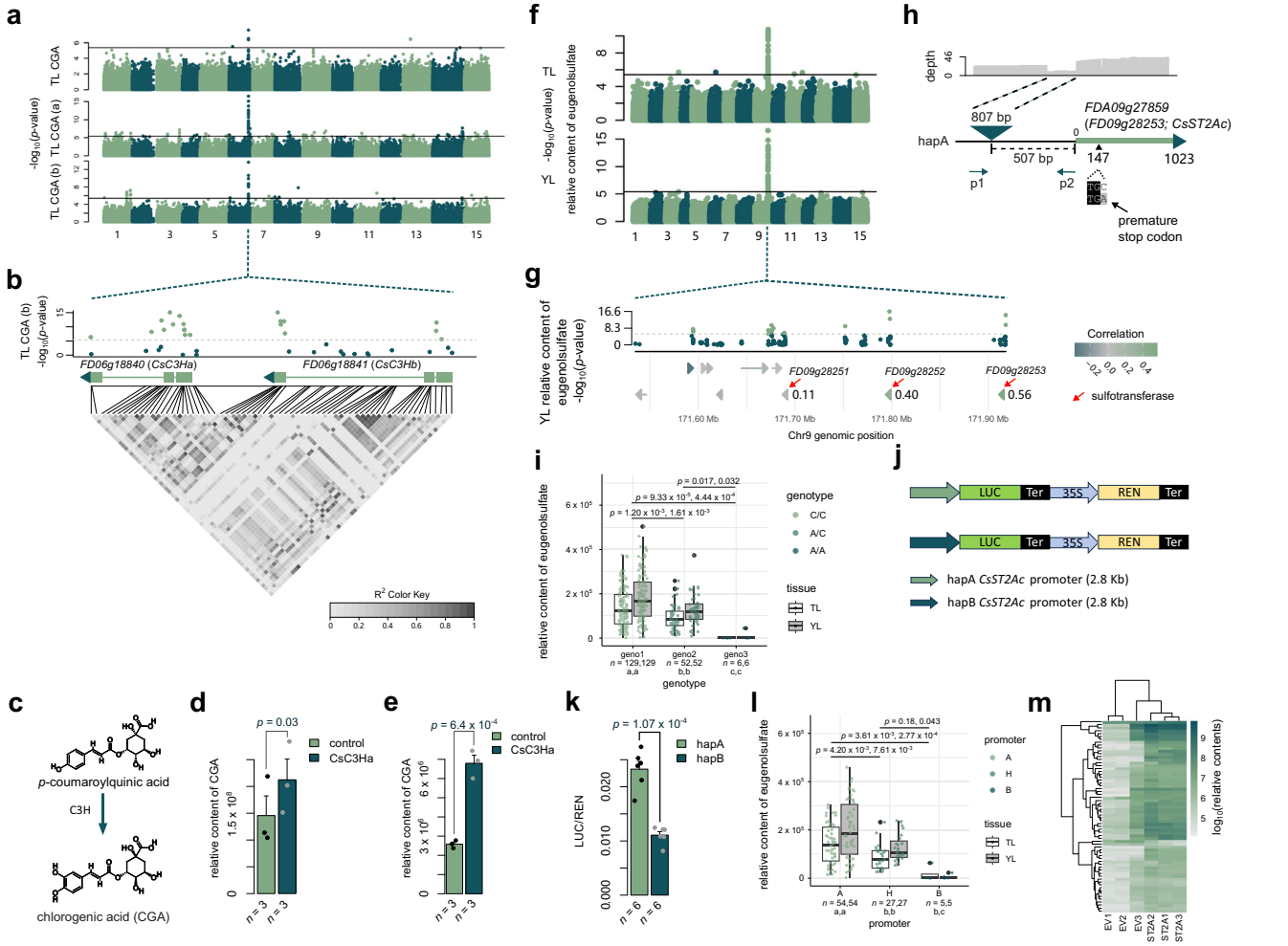

Fig. 6 | Enhancement of mQTL detection through mGWAS by using FDDB genome. a Manhattan plots of CGA and its isomers. **b** Zoomed mGWAS results of CGA in YL. The upper panel shows the Manhattan plot (GEC-adjusted *p*-value; $\alpha = 0.05$). The middle panel indicates candidate gene structure and position. The lower panel shows the LD heatmap of the candidate region. **c** Biosynthetic pathway of chlorogenic acid. **d**, **e** Transient expression of *CsC3Hb* in tobacco and tea leaves, respectively (two-sided paired *t*-test; $n = 3$). Error bars indicate standard errors, and centers indicate mean values. **f** Manhattan plot of eugenolsulfate based on FDDB genome (GEC-adjusted *p*-value; $\alpha = 0.05$). **g** Zoomed mQTL of eugenolsulfate. The upper panel shows the Manhattan plot (GEC-adjusted *p*-value; $\alpha = 0.05$). Three candidate genes identified by FDDB genome are marked with red arrows; gene colors represent correlation coefficients between expression and eugenolsulfate content. **h** Variations of *CsST2Ac* promoter and coding sequence in natural populations. Upper plot shows HiFi read depth mapped to haplotype B genome; P1 and P2 indicate PCR primer positions. **i** Boxplots of eugenolsulfate content affected by premature stop codon of *CsST2Ac* in YL and TL (adjusted *p*-values by Tukey test; $n = 129$, 52, and 6). Central line indicates median; box edges represent first (25%) and third (75%) quartiles, respectively. **j** Vector construction for dual-luciferase assay. **k** Promoter activity of *CsST2Ac* haplotypes in FDDB (two-sided *t*-test; $n = 6$). Error bars are standard errors, and centers indicate mean values. **l** Boxplots of eugenolsulfate content for different *CsST2Ac* promoter genotypes in YL and TL. A, H, B indicate absence, heterozygous, or homozygous insertion (adjusted *p*-values by Tukey test; $n = 52$, 27 and 5). Central line indicates median; box edges represent first (25%) and third (75%) quartiles, respectively. **m** Heatmap of sulfate group metabolites after transient expression of *CsST2Ac* and empty vector. EV1–3 and ST2A1–3 indicate three biological replicates of empty vector and *CsST2Ac* injections, respectively. Analyses are based on FDDB haplotype genome V1. Source data are provided as a Source Data file.

transfer of a sulfate group from a donor molecule to the substrate, and six metabolites mapped to YLhotspot_18 and TLhotspot_15 harbored putative sulfate group (Supplementary Fig. 25). These three genes were thus considered as good candidates in this hotspot. The expression level of *CsST2Ac* showed the highest correlation with the contents of eugenolsulfate among the three genes (correlation coefficient = 0.56; Fig. 6g and Supplementary Data 15). Additionally, sequence analysis revealed that a SNP (C-A) located at 147 bp of *CsST2Ac* coding sequence led to a premature stop codon (Fig. 6h). Genotype analysis showed that accessions harboring this premature stop codon had significantly lower contents of eugenolsulfate in both YL and TL (Fig. 6i; $p < 0.05$; Tukey test). Interestingly, *CsST2Ac* (*FDA09g27859* in haplotype A genome of FDDB) exhibited ASE patterns in eight tissues (except for root) of FDDB, and haplotype A also showed significantly higher expression levels of *CsST2Ac* than haplotype B in the 31 FDDB offspring accessions, indicating that *CsST2Ac* has a relative stable ASE pattern in FDDB ($p = 0.024$; ANOVA *F*-test; Supplementary Fig. 26a, c). Additionally, the accessions harboring haplotype A also had higher eugenolsulfate contents than those harboring haplotype B of *CsST2Ac* (Supplementary Fig. 26b). Promoter analysis showed that an 807 bp insertion was located in 507 bp upstream of ATG of the haplotype B of FDDB (Fig. 6h). Dual luciferase assay indicated that the promoter activity of *CsST2Ac* in haplotype A was significantly higher than that in haplotype B (Fig. 6j, k; $p = 1.07 \times 10^{-4}$; *t*-test). We developed a PCR marker to verify whether the insertion affects the content of eugenolsulfate. The results showed that the accessions containing this insertion had significantly lower levels of eugenolsulfate (Fig. 6l; $p < 0.05$; Tukey test). The main effects of

insertion and premature stop codon of *CsST2Ac* were significant, but no significant interaction effect was detected between them, indicating an additive effect of these two loci on the content of eugenolsulfate ($p = 3.78 \times 10^{-4}$ and 0.011 for premature stop codon and insertion in YL; $p = 5.16 \times 10^{-3}$ and 0.011 for premature stop codon and insertion in TL, respectively; Supplementary Data 16 and Supplementary Fig. 27; ANOVA *F*-test). Further transient overexpression of *CsST2Ac* significantly increased the contents of metabolites with sulfate group in tobacco leaves (Fig. 6m and Supplementary Fig. 28). Finally, we examined the methylation pattern at the *CsST2Ac* locus. Remarkably higher methylation levels were observed in haplotype B than in haplotype A, both within the gene body and at the 807-bp insertion site in the promoter region (Supplementary Fig. 29). This increased methylation is likely responsible for the suppressed expression of *CsST2Ac* observed in haplotype B. Taken together, variations in both the promoter and coding sequence of *CsST2Ac* can affect the contents of metabolites with sulfate group in tea population.

Discussion

In this study, we report a high-quality haplotype-resolved genome by applying sperm cell sequencing in combination with long-read sequencing. Instead of individually assembling the two haplotype genomes using phased reads, as in gamete-binning[22], we utilized phased SNPs to generate haplotype-specific *k*-mers (Supplementary Fig. 1). These *k*-mers were then used to assemble the phased genomes with hifiasm, employing the trio-binning method. This pipeline enabled the use of relatively lower coverage of long reads to achieve a highly continuous haplotype assembly. The phased HiFi and ONT reads covered 99.95% of consensus genome, suggesting a near complete phasing of FDDB genome (Fig. 1 and Supplementary Table 5). Additionally, the haplotype genomes showed high QV and LAI values, and the LAI of both haplotype genomes reached the golden level[42] (Table 1). Hence, we have established an alternative approach for assembling high quality haplotype-resolved genomes without Hi-C data. However, haplotype genome assembly remains challenging in regions with lower SNP densities, such as TE-enriched regions and the telomeric regions of tea plants (Figs. 1 and 2), and a greater depth in ultra-long ONT sequencing combined with Pore-C technology may improve the assembly performance[24,53].

For plant species with high heterozygosity, such as tea plants, a large proportion of genetic information is inevitably lost when assembling a gap-free consensus genome. *K*-mer completeness analysis showed that the consensus tea plant genome may lose ~23% of the genetic information (Table 1), which is higher than the loss rate of genetic information in kiwifruit, apple, Arabidopsis, and humans, indicating a high complexity of the tea plant genome[14,15,54]. We also found that SVs between the two haplotypes account for ~23.8% of the tea plant genome, approximately twice the amount of genomic differences identified in previous tea pan-genome study[10]. This observation indicates that SV diversity is likely to be underestimated without an accurate haplotype-resolved assembly. Moreover, haplotype-resolved genomes have great potential to capture the full extent of genetic variation, thereby complementing existing tea pan-genome resources and pave a way for future functional and evolutionary study of tea plant. The vast majority of these SVs are primarily generated by TE insertions, TEMR, and nTEMR events (Fig. 3b, c). The observations in potato and grapevine suggested that most SVs tend to occur as singletons and are under strong purifying selection[43,55]. In fact, we found that most TE insertion-derived SVs arose relatively recently in the tea genome (<0.025 MYA; Fig. 3d), which indicate their predominance as singletons. Interestingly, Class II TEs exhibited a second burst peak at approximately 0.39 MYA (Fig. 3d), suggesting that some SVs generated during this period may have retained polymorphisms within tea plant populations. This raises the intriguing possibility that these SVs may be subject to balancing selection within tea plant populations. Future research could explore these hypotheses to better understand the evolutionary forces shaping SV diversity in tea plants.

High-quality phased genome assembly not only provides comprehensive insights into chromosomal features but also enhances the effectiveness of detecting QTL associated with important agricultural traits[25,54,56,57]. One frequent application of haplotype genomes is to investigate ASE patterns between two haplotypes. However, most studies of ASE were conducted with different tissue samples emanating from a single individual, which is short of statistical power. Integration of ASE patterns of the offspring population would improve the power to confirm the ASE patterns[58]. Here, we investigated ASE genes in multiple tissues and a FDDB offspring population consisting of 31 accessions. We identified hundreds of genes exhibiting ASE patterns in both different tissues and the offspring population, suggesting their stable ASE patterns across various tissues and genetic backgrounds. These results imply that they are promising candidates for future molecular breeding. Additionally, we also correlated these ASE genes and allelic variations with metabolomic data and revealed the role of *CsDFRb* and its allelic variations in influencing metabolic traits, which exhibits the SV-driven metabolic diversity through *cis*-regulatory rewiring in tea plants. The mapping power can be affected by the limited population size, and the resulting average bin length of 7.9 Mb (Supplementary Data 7) led to low resolution. Therefore, expanding the population size is expected to enhance both mapping precision and gene discovery in the future.

GWAS represents one of the most efficient methods for detecting valuable QTL in plants, and it has been demonstrated that high-quality genomes may help detect novel QTL or refine the QTL regions[25,35,36]. By using the FDDB genome instead of TGY genome as the reference, we identified 2,649 novel mQTL, which accounted for 51.7% of the total mQTL (Supplementary Data 11). Through analysis of two mQTL hotspots, we verified the reliability of the identified mQTL (Fig. 6). The different results obtained by using different genomes may be attributed to two factors. On the one hand, a high-quality genome can correct some dis-assembled regions, thereby reducing mismatch error rates[35]; on the other hand, a substantial number of SVs, including translocations and duplications, exist between these tea plant accessions[10] (Supplementary Figs. 7, 10 and 11). Hence, significantly associated SNPs or candidate genes may be absent in some tea accessions. Both factors can influence read mapping and SNP calling, resulting in divergent GWAS results. Additionally, a high-quality genome enables more comprehensive gene annotations, which can greatly improve the efficacy of candidate gene selection[41]. Another notable finding is that many candidate genes in mGWAS harbor tandem repeats, which necessitates further investigation into the evolutionary and functional diversification of these tandem repeat genes in future studies (Fig. 6)[34].

Diverse metabolites are essential for plant growth and protection against biotic and abiotic stresses, and greatly affect the flavors and health benefits of tea[34]. A comprehensive understanding of the metabolic diversity and its underlying genetic basis of tea plants is essential for ensuring more flavorful and nutritious beverage supply. Although GWAS has shown great capacity to address some important issues regarding the genetic architecture of complex traits, it remains challenging to identify the relevant genes and casual genetic variants. The accurate genome assembled in this study enables efficient dissection of complex metabolic traits. For example, we identified the causal gene *CsST2Ac* and its allelic heterogeneity for sulfate group metabolites by using the assembled genome (Fig. 6f–h). Moreover, haplotype-resolved genomes allowed the identification of an insertion in the promoter of *CsST2Ac*, which affects the expression level of *CsST2Ac* (Fig. 6h–m), and multiple causal variants of *CsST2Ac* were found to have an additive effect on sulfate group metabolites (Supplementary Data 16 and Supplementary Fig. 27). Even though it is believed that analysis of metabolic traits is relatively more straightforward than

analysis of other agronomic traits such as development and resistance, there is still much room for improvement. The high-quality genome, genetic information, and metabolic data in this study, along with the established genetic analysis framework, unlock hidden SVs and SV-metabolism nexus, greatly deepening our understanding of metabolic innovation in tea plants. These resources are valuable for the genetic analysis and improvement of tea quality, as well as for the genetic analysis of other complex traits.

In summary, this study combined long read sequencing and sperm sequencing data to achieve an accurate haplotype-resolved assembly of the tea plant genome, which was used to uncover numerous novel loci associated with metabolic traits and ASE patterns. These findings provide valuable insights into tea plant genetics and metabolic diversity and shed light on future functional genomics and molecular breeding in tea plants.

Methods

Plant materials

Tea accession FDDB was planted in the field experimental station in Huazhong Agricultural University, Wuhan, China. Young leaves for PacBio and ONT sequencing were harvested and immediately frozen by liquid nitrogen after harvest and stored at −80 °C.

Genome assembly of FDDB

For PacBio sequencing, SMRTbell libraries (15 Kb) were constructed and sequenced on PacBio Sequel II system, and consensus reads (Hifi reads) were generated by CCS reads (https://github.com/pacificbiosciences/unanimity) with default parameters. ONT ultra-long libraries (N50 100 Kb) were constructed by using high-molecular-weight gDNA and sequenced on Nanopore PromethION sequencer. ONT reads with meanQ <7 were removed for genome assembly. Preliminary assembly was generated by Hifi and ONT reads using hifiasm[20] with default parameters.

Details of genome assembly method and gap filling procedures were descripted in Supplementary Method 1. Briefly, single sperm cell sequencing data were obtained from our previous study, which can be downloaded from the National Genomic Center database under the accession PRJCA002764[26]. The single sperm cell sequencing data were mapped to preliminary genome assembly by bwa[59] with default parameters. SNPs were called and filtered by GATK[60] (SNP filtration parameters: QD < 2.0||FS > 60.0||MQ < 40.0 || MQRankSum < −12.5 || ReadPosRankSum < −8.0). SNP phasing and construction of genetic maps were performed following our previous studies[26,29]. Briefly, SNPs were filtered and initially phased by R package Hapi[61]. SNPs without COs were merged into genetic bins. By adopting these bins as markers, genetic maps were constructed using MSTmap[62]. Contigs from preliminary assembly were manually anchored to pseudomolecules based on the constructed genetic map. Small contigs inserted into large contigs according to the genetic map were considered as haplotigs during diploid assembly and were removed for further analysis.

Two more assemblies of FDDB were generated for chromosomal gap filling. Specifically, verkko[39] (ONT + Hifi) and Necat[40] (ONT) were used for genome assembly with default parameters, respectively. Contigs of these two assemblies were aligned to pseudomolecules by minimap2[52] and mummer4[63], respectively. Gaps were filled based on alignment between different assemblies. Each gap filling result was manually checked, and alignment plots were generated using in-house R script.

Genome annotation and evaluation

Repeatmasker[64] was used to scan repeat elements by Repbase, and a de novo repetitive database was constructed by RepeatModeler[65]. Repeat sequences were classified by deepTE[66], and then masked for gene structure annotation by two different methods. First, RNA-seq was performed for nine different tissues of FDDB, including leaf bud, first leaf, second leaf, third leaf, leaf vein, mature leaf, petal, stamen, and root, and gene annotation was conducted by braker pipeline (v3.08)[67]. Second, OrthoDB v12 (https://bioinf.uni-greifswald.de/bioinf/partitioned_odb12/) was used as the protein database to provide additional evidence for gene prediction in the breaker pipeline. Primary transcripts from the two methods were selected by tsebra[68] with the default configuration file, setting intron_support to 0.8 and including the --filter_single_exon_genes parameter. Gene function was annotated by GO, KEGG and Pfam database using GFAP[69]. QV, error rate, and assembly completeness were evaluated by merqury pipeline[15]. LAI score was calculated using LTR_retriver[70]. Complete BUSCO of genome was calculated by BUSCO[71] with embryophyte database. Pairwise LTR divergence was calculated by LTR_retriver[70]. LTR insertion time was calculated with the following equation

$$T = K/2\mathrm{r} \quad (1)$$

where T represents insertion time, K indicates LTR pairwise divergence, and r is the neutral mutation rate (5.62×10^{-9})[8]. Telomere was detected by searching sequence 5′-CCCTAAA-3′ along the chromosome using tidk (https://github.com/tolkit/telomeric-identifier).

SNP phasing of FDDB offspring

Short reads of seven FDDB offspring were used for phasing. The short reads were aligned to the genome by bwa[59]. SNPs were called by GATK[60] with default parameters, and PCR duplications were removed by picard (https://broadinstitute.github.io/picard). A random forest model trained from known FDDB offspring and phased SNPs from single sperm data was used for IBD detection. For each offspring, homozygous SNPs in an IBD region from FDDB hap A or B were considered as phased SNPs of FDDB haplotype A or B, respectively. Phased SNPs were retained based on the following criteria: (i) at least two offspring samples were genotyped without conflict between them, or (ii) at least five offspring samples were genotyped and only one sample had genotype conflict with other samples.

Haplotype genome assembly of FDDB

Short reads of FDDB were obtained from our previous study, which can be downloaded from NCBI under the accession number of PRJNA595898[5]. Short reads were aligned to FDDB genome by bwa[59] and sorted to bam file by samtools[72], and PCR duplications were removed by picard (https://broadinstitute.github.io/picard). Only reads with mapping quality > 40 were kept for phasing. These qualified short reads were then classified into two haplotypes based on the phased SNPs from single sperm data and FDDB offspring. yak (https://github.com/lh3/yak) was used to generate haplotype specific *k*-mers based on the phased short reads.

Trio-binning assembly was performed using hifiasm[20] and verkko2[73] with haplotype *k*-mers, HiFi, and ONT data under default parameters. Contigs generated by hifiasm for each haplotype were anchored to pseudomolecules by RagTag[74] with default parameters. Gaps in the two haplotype assemblies were subsequently closed using the verkko2 assemblies following the same procedure as in the consensus assembly.

To evaluate the haplotype phasing accuracy, short reads from 107 single sperm cells were also separated and merged into haplotype A and B, respectively, according to COs detected for each sperm cell. These reads were used to verify the phased assembly and calculate the switch error rate using the merqury pipeline[15]. For analysis of phase block and switch error rate, at most 1000 switches were allowed in a 500 Kb window.

An updated version of the FDDB genome (V2) was later generated by incorporating additional PacBio HiFi and ONT data to further improve contiguity and gap filling. Because the sequence structure and annotation were nearly identical to those of the original version (V1), all

analyses performed using V1 remain valid and consistent with the V2 genome (Supplementary Data 17).

Identification of SVs between the haplotype genomes

We combined genome alignment and HiFi reads to identify SVs between the haplotype genomes. First, the genomes of two haplotypes of FDDB were aligned by minimap2[52] with '-x asm5 -c --eqx' parameters, and alignments with mapping quality <20 were removed. Structural annotations were performed by SyRI[75] with the follow parameters '-f --cigar'. Insertions and deletions (length > 50 bp) between the two haplotypes were subsequently checked by depth of HiFi reads with minimal length of 2000 bp. Only SVs with significantly different depth greater 5 between SV and flanking regions were kept (*t*-test; $p < 0.0001$). We further used minimap2[52] to align HiFi reads to haplotype A and B genomes, respectively. Sniffles2[76] was performed to detect SVs in each haplotype genome. We only count deletions for each haplotype genome. For filtering the SVs identified by Sniffles2, we extracted the 2 Kb flanking sequence for the two break points of deletions. We paste the 2 Kb flanking sequence into a 4 Kb sequence, and use this 4 Kb sequence as query to map to the other haplotype genome using minimap2. SVs showing alignment length greater than 3 Kb and located in synteny regions (identified by SyRI) between the two haplotypes or the physical distance between the query sequence and mapped position smaller than 5 Mb were kept. Finally, SVs identified by SyRI and Sniffles2 were merged.

Identification of SV formation mechanisms

Intact TEs were identified by EDTA pipeline[77]. SVs derived from TE insertions were defined as those overlapping with a single TE, with coverage > 90%. SVs within TEs were defined as SVs located entirely within a single TE. To identify TEMR events, we extracted ±100 bp flanking sequences at both breakpoints of each SV. If both flanking sequences overlapped with the same class of TEs, the SV was categorized as a TEMR event. HR events were identified by searching for microhomology at the two breakpoints using 100 bp sequences. For microhomology of 3–4 bp, we directly searched at the breakpoints. For microhomology longer than 5 bp, we used blastn[78] to detect microhomology, using the following parameters: -task blastn-short -word_size 4. To assess whether TEMR and HR events were randomly distributed in the tea genome, we used the BEDTools[79] shuffle function to generate random SV datasets. This simulation was repeated 10,000 times to calculate the empirical *p*-value for the permutation test.

CO detection

Raw COs were detected using genetic bins based on the genetic map. The viterbi algorithm in R package HMM was used to screen COs along the chromosome for each sperm cell. Double COs with distance smaller than 500 Kb were filtered out. The CO landscape along the chromosome was plotted by SyRI[75].

ASE analysis

Syntenic gene pairs of the two haplotypes were calculated using JCVI pipeline[29]. Only genes in the syntenic regions between the two haplotypes were kept for ASE analysis. RNA-seq reads from the above-mentioned nine different tissues of FDDB were aligned to FDDB haplotype A genome by hisat2[80]. Alignments were sorted and filtered by samtools[72], and only reads with mapping quality > 40 were kept for ASE analysis. Phased SNPs were used to separate RNA-seq reads into two haplotypes (the same method for phasing short reads of FDDB) and generated haplotype specific bam file by samtools, respectively. Gene expression value (TPM; Transcripts Per Million) was calculated by featureCounts[81] with default parameters. To control the mapping bias between the two haplotype genomes, we substituted the SNPs of haplotype A genomes with those of haplotype B. Subsequently, we re-mapped the RNA-seq reads to compute the TPM values. Only genes with TPM ratios exceeding 0.9 after SNP replacement were retained. Finally, the maximum TPM values for a haplotype greater than 1 and with at least three-fold changes between the two haplotypes were considered as ASEs. Upset plot was generated by R package UpSetR.

In the ASE analysis of the FDDB offspring population, RNA-seq data from each offspring were initially aligned to the FDDB haplotype A genome using hisat2 with default settings. SNP calling for each individual was then carried out using bcftools[82], applying a minimal mapping quality and depth threshold of 40 and 3, respectively. Subsequently, SNP phasing for each offspring was accomplished by comparing their identified SNPs with the phased SNPs of FDDB. This process enabled the generation of haplotype-specific BAM files for each offspring, following the above-mentioned method. Finally, gene expression value was computed by averaging the sequencing depth across the phased SNPs, and then normalized by total read counts. Genes exhibiting ASE in the FDDB offspring population were defined as those displaying a $\log_2$ fold-change greater than one and an ANOVA *p*-value less than 0.01.

Genome-wide association study

The RNA-seq and metabolic data of 215 tea accessions generated from our previous studies were used for GWAS analysis[5,34]. The RNA-seq data were mapped to FDDB genome by hisat2[80] and SNPs were called by GATK[60]. SNPs with a missing rate <80% and an MAF > 0.1 were kept and the missing genotypes were imputated by beagle5[83]. GWAS was performed by EMMAX[84] with default parameters and significant threshold was calculated by GEC[85]. Significant SNPs were merged into raw candidate region with distance <500 Kb or LD > 0.1, and each candidate region had at least two significant SNPs. The final candidate region was determined by an extension of 50 Kb based on the raw candidate region. Candidate genes were selected based on gene functional annotation and correlation between gene expression value and metabolite contents. Hotspots of mQTL were calculated by counting the mQTL number in a 1-Mb sliding window along the chromosome with 1000 times permutation test ($p < 0.01$). ANOVA, fisher exact test, and Tukey test were conducted by R aov, fisher.test, and TukeyHSD function, respectively. The correlation between genes and metabolite contents and the *p*-value of correlation was calculated by R function cor and cor.test, respectively.

Methylation analysis

For methylation analysis, ONT reads were basecalled using Dorado (v0.8.2; https://github.com/nanoporetech/dorado) with the dna_r9.4.1_e8_sup@v3.3 model. The basecalled reads were then mapped to both the collapsed assembly using minimap2[52]. Separately, they were mapped to both haplotype assemblies at the same time (i.e. the whole haplotype resolved genome) to guide reads to the respective haplotype. Subsequently, CpG methylation modifications were identified using Modkit (v0.4.1; https://github.com/nanoporetech/modkit) pileup with the following parameters: --filter-threshold 0.77, --mod-thresholds m:0.99, --motif CG 0, and --combine-strands. Methylation status was determined based on predefined thresholds: CpG sites with a coverage of ≥5 and a base modification frequency of ≥75% were classified as methylated, while sites with a coverage of ≥5 and a base modification frequency of ≤25% were classified as non-methylated. Sites with intermediate modification frequencies (between 25% and 75%) or too low coverage were excluded from further analysis. For a gene-based analysis the identified gene plus a window of 1 kb upstream and downstream was considered, and the average methylation per 100 bp was plotted. Haplotypes that differed by at least 50 percentage points were flagged as likely differentially methylated.

Dual luciferase assay

The dual luciferase assay was performed by Dual Luciferase Reporter Assay Kit (Vazyme). The promoters of *CsDFRb* (2.7 Kb) and *CsST2Ac*

(2.8 Kb) were amplified using STI PCR and cloned into pGreenII 0800 vector to drive the expression of firefly luciferase gene (LUC), respectively. The Renilla luciferase reporter gene (REN), driven by the CaMV 35S promoter in the same vector, was used as the LUC reporter gene and served as an internal reference in each transformation. The vector was transiently expressed in leaves of *Nicotiana benthamiana* through *Agrobacterium*-mediated transformation. Three days after injection, the LUC and REN enzyme activity was measured by a microplate reader of multi-wavelength measurement system (Tecan, Spark, Switzerland). Promoter activity was determined by the ratio between LUC and REN activity. Primers for amplification of promoter sequences are listed in Supplementary Data 18.

Transient expression assay and metabolite analysis

The coding sequences of *CsDFRb*, *CsC3Ha*, and *CsST2Ac* were amplified from tea leaves (FDDB) and cloned into improvement vector EarleyGate 101 using ClonExpress II One Step Cloning Kit (Vazyme). These genes were transiently expressed in *N. benthamiana* leaves through *Agrobacterium*-mediated transformation. Two days after injection, 200 µMol 3′-phospho-adenosine-5′-phospho-sulfate (PAPS, Merk) was injected in the leaves. Two days later, tobacco leaves were harvested for metabolite profiling according to Qiu et al.[34]. In brief, metabolites were extracted using 1.0 mL of 70% methanol for 12 h at 4 °C. Sample extracts (3 µL) were analyzed by liquid chromatography coupled with Q Exactive Plus mass spectrometry (LC-MS; Thermo Fisher Scientific, California, USA). HPLC parameters included a Waters Atlantis T3 column (260 × 4.6 mm, 5 µm), a solvent system consisting of 0.1% formic acid in water (solution A) and methanol (solution B) at a flow rate of 0.25 mL/min, and a gradient program as follows: 0–1 min, B 2%; 1–10 min, B 2% to 50%; 10–13 min, B 50% to 95%; 13–14 min, B 95%; 14–17 min, B 95% to 2%. Mass spectrogram data were analyzed by Compound Discoverer 3.2 (Thermo Fisher, San Jose, CA, USA). For transient expression in tea leaves, *CsC3Ha* or an empty vector control was introduced into young tea leaves (var. Wuniuzao) via *Agrobacterium*-mediated transformation. Three days post-infiltration, only leaves exhibiting clear fluorescence signals were selected for metabolic analysis, following the same procedures described above. Primers for vector construction are listed in Supplementary Data 18.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw sequencing data and assemblies are available at China National Center for Bioinformation under the bioproject PRJCA023713. The raw reads of Pacbio HiFi, oxford ONT, and RNA-seq of nine tissues of FDDB are available under accession CRA015031 and CRA032018. Short reads sequencing of FDDB are available under accession SRR10696339. Sequencing data of sperm cells are available under accession CRA002708. The consensus genome of FDDB and annotation are available under accession GWHESEG00000000 [https://ngdc.cncb.ac.cn/gwh/Assembly/84476/show]. The genome assembly and annotation of haplotype A and B of FDDB (V1) are available under the accession GWHESEH00000000 [https://ngdc.cncb.ac.cn/gwh/Assembly/84477/show] and GWHESMY00000000 [https://ngdc.cncb.ac.cn/gwh/Assembly/84719/show], respectively. The genome assembly of haplotype A and B of FDDB (V2) are available under the accession GWHGSGX00000000.1 [https://ngdc.cncb.ac.cn/gwh/Assembly/102385/show] and GWHGSGW00000000.1 [https://ngdc.cncb.ac.cn/gwh/Assembly/102384/show], respectively. The annotation of haplotype A and B of FDDB (V2) are available at Figshare [https://doi.org/10.6084/m9.figshare.30484748]. The metabolic data are available at National Omics Data Encyclopedia under accession OEP00006790. Source data are provided with this paper.

Code availability

Scripts and codes used in this study are available at Github [https://github.com/zwycooky/FDhaplotype] and Zenodo [https://doi.org/10.5281/zenodo.17659312].

References

1. Xia, E. H. et al. Tea plant genomics: achievements, challenges and perspectives. *Hortic. Res.* **7**, 7 (2020).
2. Samynathan, R. et al. Recent insights on tea metabolites, their biosynthesis and chemo-preventing effects: a review. *Crit. Rev. Food Sci. Nutr.* **63**, 3130–3149 (2023).
3. Liao, Y., Zhou, X. & Zeng, L. How does tea (*Camellia sinensis*) produce specialized metabolites which determine its unique quality and function: a review. *Crit. Rev. Food Sci. Nutr.* **62**, 3751–3767 (2022).
4. Zhang, X. et al. Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant Camellia sinensis. *Nat. Genet.* **53**, 1250–1259 (2021).
5. Zhang, W. et al. Genome assembly of wild tea tree DASZ reveals pedigree and selection history of tea varieties. *Nat. Commun.* **11**, 3719 (2020).
6. Wang, X. et al. Population sequencing enhances understanding of tea plant evolution. *Nat. Commun.* **11**, 4447 (2020).
7. Xia, E. et al. The reference genome of tea plant and resequencing of 81 diverse accessions provide insights into its genome evolution and adaptation. *Mol. Plant* **13**, 1013–1026 (2020).
8. Zhang, Q.-J. et al. The chromosome-level reference genome of tea tree unveils recent bursts of non-autonomous LTR retrotransposons in driving genome size evolution. *Mol. Plant* **13**, 935–938 (2020).
9. Wang, P. et al. Genetic basis of high aroma and stress tolerance in the oolong tea cultivar genome. *Horticult. Res.* **8**, 107 (2021).
10. Chen, S. et al. Gene mining and genomics-assisted breeding empowered by the pangenome of tea plant Camellia sinensis. *Nat. Plants* **9**, 1986–1999 (2023).
11. Tariq, A. et al. In-depth exploration of the genomic diversity in tea varieties based on a newly constructed pangenome of Camellia sinensis. *Plant J.* **119**, 2096–2115 (2024).
12. Michael, T. P. & VanBuren, R. Building near-complete plant genomes. *Curr. Opin. Plant Biol.* **54**, 26–33 (2020).
13. Gladman, N., Goodwin, S., Chougule, K., McCombie, W. R. & Ware, D. Era of gapless plant genomes: Innovations in sequencing and mapping technologies revolutionize genomics and breeding. *Curr. Opin. Biotechnol.* **79**, 102886 (2023).
14. Han, X. et al. Two haplotype-resolved, gap-free genome assemblies for Actinidia latifolia and Actinidia chinensis shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. *Mol. Plant* **16**, 452–470 (2023).
15. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 1–27 (2020).
16. Patterson, M. et al. WhatsHap: weighted haplotype assembly for future-generation sequencing reads. *Comput. Biol* **22**, 498–509 (2015).
17. Edge, P., Bafna, V. & Bansal, V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812 (2017).
18. Zhang, X., Zhang, S., Zhao, Q., Ming, R. & Tang, H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845 (2019).
19. Zeng, X. et al. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat. Plants* **10,** 1–17 (2024).
20. 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).

21. Koren, S. et al. De novo assembly of haplotype-resolved genomes with trio binning. *Nat. Biotechnol.* **36**, 1174–1182 (2018).
22. Campoy, J. A. et al. Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol.* **21**, 1–20 (2020).
23. Hu, G. et al. Two divergent haplotypes from a highly heterozygous lychee genome suggest independent domestication events for early and late-maturing cultivars. *Nat. Genet.* **54**, 73–83 (2022).
24. Zhang, Z. et al. Haplotype-resolved genome assembly and resequencing provide insights into the origin and breeding of modern rose. *Nat. Plants* **10**, 1–13 (2024).
25. Sun, H. et al. Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat. Genet.* **54**, 342–348 (2022).
26. Zhang, W. et al. A phased genome based on single sperm sequencing reveals crossover pattern and complex relatedness in tea plants. *Plant J.* **105**, 197–208 (2021).
27. Zhang, W. et al. Plant sperm cell sequencing for genome phasing and determination of meiotic crossover points. *Nat. Protoc.* **20**, 1–19 (2024).
28. Tan, L.-Q. et al. Fingerprinting 128 Chinese clonal tea cultivars using SSR markers provides new insights into their pedigree relationships. *Tree Genet. Genomes* **11**, 1–12 (2015).
29. Tang, H. et al. JCVI: a versatile toolkit for comparative genomics analysis. *iMeta* **12**, e211 (2024).
30. Wen, W. et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. *Nat. Commun.* **5**, 3438 (2014).
31. Fernie, A. R. & Tohge, T. The genetics of plant metabolism. *Annu. Rev. Genet.* **51**, 287–310 (2017).
32. Jin, J.-Q. et al. Characterization of two O-methyltransferases involved in the biosynthesis of O-methylated catechins in tea plant. *Nat. Commun.* **14**, 5075 (2023).
33. Kong, W. et al. Pan-transcriptome assembly combined with multiple association analysis provides new insights into the regulatory network of specialized metabolites in the tea plant Camellia sinensis. *Horticult. Res.* **9**, uhac100 (2022).
34. Qiu, H. et al. Depicting the genetic and metabolic panorama of chemical diversity in the tea plant. *Plant Biotechnol. J.* **22**, 1001–1016 (2023).
35. Huang, X. A complete telomere-to-telomere assembly provides new reference genome for rice. *Mol. Plant* **16**, 1370–1372 (2023).
36. Wei, M. et al. Telomere-to-telomere genome assembly of melon (Cucumis melo L. var. inodorus) provides a high-quality reference for meta-QTL analysis of important traits. *Horticult. Res.* **10**, uhad189 (2023).
37. Deng, Y. et al. A telomere-to-telomere gap-free reference genome of watermelon and its mutation library provide important resources for gene discovery and breeding. *Mol. Plant* **15**, 1268–1284 (2022).
38. van Rengs, W. M. et al. A chromosome scale tomato genome built from complementary PacBio and Nanopore sequences alone reveals extensive linkage drag during breeding. *Plant J.* **110**, 572–588 (2022).
39. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **14**, 1474–1482 (2023).
40. Chen, Y. et al. Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nat. Commun.* **12**, 60 (2021).
41. Chen, J. et al. A complete telomere-to-telomere assembly of the maize genome. *Nat. Genet.* **55**, 1–11 (2023).
42. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
43. Cheng, L. et al. Leveraging a phased pangenome for haplotype design of hybrid potato. *Nature* **640**, 1–10 (2025).
44. Collins, R. L. & Talkowski, M. E. Diversity and consequences of structural variation in the human genome. *Nat. Rev. Genet.* **26**, 1–20 (2025).
45. Taagen, E., Bogdanove, A. J. & Sorrells, M. E. Counting on crossovers: controlled recombination for plant breeding. *Trends Plant Sci.* **25**, 455–465 (2020).
46. Zhang, Y. et al. Structural variation reshapes population gene expression and trait variation in 2,105 Brassica napus accessions. *Nat. Genet.* **56**, 1–13 (2024).
47. Long, Y., Wendel, J. F., Zhang, X. & Wang, M. Evolutionary insights into the organization of chromatin structure and landscape of transcriptional regulation in plants. *Trends Plant Sci.* **29**, 638–649 (2024).
48. Wang, X. et al. Genome of Solanum pimpinellifolium provides insights into structural variants during tomato breeding. *Nat. Commun.* **11**, 5817 (2020).
49. Balachandran, P. et al. Transposable element-mediated rearrangements are prevalent in human genomes. *Nat. Commun.* **13**, 7115 (2022).
50. Shao, L. et al. Patterns of genome-wide allele-specific expression in hybrid rice and the implications on the genetic basis of heterosis. *Proc. Natl. Acad. Sci. USA* **116**, 5653–5658 (2019).
51. Albert, E. et al. Allele-specific expression and genetic determinants of transcriptomic variations in response to mild water deficit in tomato. *Plant J.* **96**, 635–650 (2018).
52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
53. Koren, S. et al. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. *Genome Res.* **34**, 1919–1930 (2024).
54. Mansfeld, B. N. et al. A haplotype resolved chromosome-scale assembly of North American wild apple Malus fusca and comparative genomics of the fire blight Mfu10 locus. *Plant J.* **116**, 989–1002 (2023).
55. Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
56. Li, W. et al. Near-gapless and haplotype-resolved apple genomes provide insights into the genetic basis of rootstock-induced dwarfing. *Nat. Genet.* **56**, 1–12 (2024).
57. Zhou, Q. et al. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **52**, 1018–1023 (2020).
58. Fan, J. et al. ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genet.* **16**, e1008786 (2020).
59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
60. McKenna, A. et al. The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
61. Li, R. et al. Inference of chromosome-length haplotypes using genomic data of three or a few more single gametes. *Mol. Biol. Evol.* **37**, 3684–3698 (2020).
62. Wu, Y., Bhat, P. R., Close, T. J. & Lonardi, S. Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* **4**, e1000212 (2008).
63. Marçais, G. et al. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, e1005944 (2018).
64. Chen, N. Using repeat masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **5**, 11–14 (2004).
65. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
66. 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).

67. Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. Whole-genome annotation with BRAKER. *Gene Prediction: Methods and Protocols*, 65-95 (2019).
68. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 1–12 (2021).
69. Xu, D. et al. GFAP: ultrafast and accurate gene functional annotation software for plants. *Plant Physiol.* **193**, 1745–1748 (2023).
70. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
71. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
72. Li, H. et al. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
73. Antipov, D. et al. Verkko2 integrates proximity-ligation data with long-read De Bruijn graphs for efficient telomere-to-telomere genome assembly, phasing, and scaffolding. *Genome Res.* **35**, 1583–1594 (2025).
74. Alonge, M. et al. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol.* **23**, 1–19 (2022).
75. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* **20**, 1–13 (2019).
76. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.* **42**, 1571–1580 (2024).
77. Ou, S. et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20**, 1–18 (2019).
78. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
79. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
80. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
81. 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).
82. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
83. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
84. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
85. Li, M.-X., Yeung, J. M., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant *p*-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum. Genet.* **131**, 747–756 (2012).

Acknowledgements

This study was supported by the National Key R&D Program of China (2022YFF1003103), the National Natural Science Foundation of China (32494781, 32161133017, U23A20213) to W.W., CEPLAS 390686111 to B.U., Deutsche Forschungsgemeinschaft (DFG)-Project number 468870408 to B.U. and A.R.F., and China Postdoctoral Innovation Program (BX20220127) and the National Natural Science Foundation of China (32500478) to W.Z. We thank the Core Facilities in National Key Laboratory for Germplasm Innovation and Utilization of Horticultural Crops for the assistance in metabolite detection.

Author contributions

W.W. designed and managed this research, wrote the paper; W.Z. and X.J. performed experiments, analyzed data, and wrote the paper; S.L. performed RNA-seq analysis, D.G. and X.Z. participated in metabolomic analysis, A.T., J.B., and B.U. participated in genomic analysis; A.R.F. aided in gene functional analysis and editing the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-68463-8.

Correspondence and requests for materials should be addressed to Weiwei Wen.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.