

Haplotype-resolved regulation of secondary metabolism in cannabis

Sen Cai^{1,†} , Zhiyuan Zhang^{1,†}, Xiang Li¹, Yucheng Zhao¹, Chengwei Lin¹, Manru Huang¹, Mingwei Shen¹, Deyong Ao¹, Yaxing Li¹, Shijun You¹ , Nirit Bernstein², Kailei Tang³ and Yuanyuan Liu^{1,*} 

¹Haixia Institute of Science and Technology, State Key Laboratory of Agricultural and Forestry Biosecurity, College of Life Sciences, Instrumental Analysis Center, Fujian Agriculture and Forestry University, Fuzhou 350002, China

²Institute of Soil Water and Environmental Sciences, Volcani Institute, 68 HaMaccabim Road, Rishon LeZion 7505101, Israel

³School of Agriculture, Yunnan University, Chengong District, Kunming 650500, China

*Corresponding author. E-mail: yyliu@fafu.edu.cn

[†]These authors contributed equally to this work.

Abstract

Cannabis sativa produces a rich array of specialized metabolites, including cannabinoids, terpenoids, and flavonoids, which are of immense therapeutic interest. However, the genetic and regulatory complexity arising from its highly heterozygous genome has obscured a complete understanding of their biosynthesis. Here, we present a fully phased, chromosome-level genome assembly of the wild diploid cannabis accession CSLZ. Comparative genomics and transcriptomics reveal that extensive haplotype-specific variation is a major driver of metabolic diversification within this accession. We find that structural variants and transposable element insertions differentially shape the allelic architecture of key biosynthetic genes, leading to divergent expression and function between haplotypes. This is exemplified by haplotype-specific loss of germacrene D synthase activity, copy number variation in oxidosqualene cyclases, and silencing of a flavonoid glycosyltransferase allele via promoter motif loss. Furthermore, allele-specific expression contributes to the spatiotemporal regulation of metabolic pathways, from cuticle formation in leaves to cannabinoid production in glandular trichomes. Our haplotype-resolved resource uncovers the pervasive role of allelic divergence in generating the specialized metabolome of wild cannabis accession CSLZ, providing a foundation for genomics-guided breeding and metabolic engineering.

Introduction

Secondary metabolites are pivotal for plant adaptation, defense, and mediating ecological interactions [1]. In the globally significant species *Cannabis sativa* L., these compounds including cannabinoids, terpenoids, and flavonoids define critical agronomic and therapeutic traits, such as chemotype, aroma, and medicinal value [2]. However, elucidating the genetic regulation of these complex pathways is challenging due to the plant's high heterozygosity and polygenic control. Traditional genomic studies reliant on a single, collapsed reference genome fail to capture the distinct contributions of homologous chromosomes, thereby obscuring haplotype-specific variation that are fundamental to a comprehensive understanding of gene expression and trait heritability.

This challenge is not unique to cannabis but is common among diverse plant lineages, including highly heterozygous diploids and complex polyploids. Recent advances in haplotype-resolved genome assemblies have transformed our ability to dissect allele-specific regulation, uncovering allele-specific expression (ASE), structural variation, and regulatory asymmetry that collectively

shape plant traits. For example, allele-aware assemblies identified ~19% of genes in diploid, hybrid-origin apple [3] displaying ASE during fruit development and diploid potato demonstrating 16.6% of allelic genes exhibiting differential expression and 30.8% differential methylation [4]. In gymnosperms such as *Pinus densiflora*, haplotype-level variation illuminated structural rearrangements and transcription factor expansions driving evolutionary adaptation [5]. Haplotype-resolved assemblies for perennial crops such as lemon (*Citrus limon*) and avocado (*Persea americana*) have clarified the genetic basis of fruit and essential oil traits [6, 7]. The octoploid cultivated strawberry [8] and the tetraploid potato cultivar 'Otava' [9] further highlight how phasing massive structural rearrangements in polyploids is key to understanding domestication traits such as fruit color and firmness. In the tea plant (*Camellia sinensis*), a haplotype-resolved genome uncovered extensive structural variation and widespread ASE associated with domestication traits, including dwarfing through selection on brassinosteroid metabolism genes (CsBAS1 and CsDWF4), as well as cold tolerance and herbivore defense [10].

Building on these insights, haplotype-aware genomic frameworks have also advanced our understanding of secondary

Received: 1 November 2025. Accepted: 9 April 2026. Published: 28 April 2026. Corrected and Typeset: 1 September 2026

© The Author(s) 2026. Published by Oxford University Press on behalf of the Nanjing Agricultural University.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

metabolite biosynthesis, revealing how structural and regulatory divergence between alleles shapes specialized metabolic pathways. For instance, the haplotype-resolved genome of the rubber tree (*Hevea brasiliensis*) revealed that more than 30% of annotated genes exhibiting ASE, indicating an imbalanced genetic contribution to rubber biosynthesis [11]. In perennial plants such as mulberry, resolving haplotypes was crucial for identifying regulatory variation, including promoter indels and missense mutations, that govern anthocyanin accumulation in fruits [12]. In the diploid medicinal plant *Artemisia annua*, a haplotype-aware assembly revealed a positive correlation between artemisinin content and the copy number of the amorpho-4,11-diene synthase (ADS) genes, revealing that gene family expansion is a key factor governing artemisinin yield [13]. In the medicinal autotetraploid crop rhubarb (*Rheum officinale*), the haplotype-resolved genome uncovered numerous allele-differentially-expressed genes, often due to transposable element (TE) insertions in regulatory regions, revealing an imbalanced genetic contribution to the anthraquinone biosynthesis pathway [14]. In vegetatively propagated *Zingiber officinale* (cultivated ginger), haplotype resolution uncovered large inversions linked to infertility, while ASE of key genes such as *C4H* in the phenylpropanoid pathway regulates gingerol biosynthesis [15]. In *C. sinensis*, ASE was further detected in *F3'H* and *CsXDH*, key genes for catechin and caffeine biosynthesis, respectively [10].

Recent progress in cannabis genomics led to a comprehensive *Cannabis* pangenome comprising 181 new and 12 previously published genomes [16]. These resources highlight genomic and structural diversity across hemp, marijuana, and feral or wild accessions, as well as TE activity shaping both the sex chromosomes and the cannabinoid synthase gene clusters. The pangenome revealed that wild Asian accessions maintain distinct genomic resources that differ markedly from domesticated fiber or drug-type varieties, providing valuable allelic and regulatory diversity for understanding secondary metabolism and adaptation. Nevertheless, critical questions remain regarding the haplotype-specific regulatory features that drive allelic imbalance. Despite the growing availability of pangenomes and haplotype-resolved assemblies, the extent to which structural variants (SVs), TEs, and promoter divergence contribute to ASE and metabolic specialization in cannabis remains largely unexplored. In this study, we present a chromosome-scale, haplotype-resolved genome of a wild cannabis accession Linzhi (CSLZ) from Tibet and integrate multi-omics analyses to dissect haplotype-specific regulation of secondary metabolite biosynthesis, including cuticle, terpenoids, cannabinoids, flavonoids, and lignin. Our findings provide new insights into the genetic and regulatory basis of metabolite diversity, trichome development, and adaptation in cannabis, highlighting the critical role of allelic divergence in shaping key agronomic and pharmacological traits.

Result

Genome assembly and haplotype chromosome construction

We first performed karyotype analysis to determine the chromosomal composition and ploidy level of the wild cannabis accession CSLZ. The results revealed that CSLZ possesses 10

pairs of chromosomes ($2n=20$), confirming its diploid status and providing a cytogenetic foundation for subsequent genome assembly (Fig. 1A). To estimate the physical genome size, we performed flow cytometry analysis using *Oryza sativa* ssp. *japonica* as an internal reference, which indicated a genome size of ~777.75 Mb (Fig. S1). We generated a total of 274.39Gb of raw sequences, including 71.04 Gb (94.72×) PacBio CLR reads, 29.25 Gb (39X) PacBio CCS reads, and 70.10 Gb (93×) Illumina paired-end reads with 104 Gb (138.67×) Hi-C data, to construct a chromosome-scale haplotype-resolved assembly. We first assembled haploid genomes and then combined PacBio CCS reads with HindIII Hi-C data to phase contigs, ultimately producing a fully phased, chromosome-level assembly (Fig. 1B). After manual curation of contig order and orientation using Hi-C contact maps, 99.1% of both haplotypes were anchored to 10 chromosomes, with total lengths of 739.64 and 739.45 Mb, respectively (Fig. 1C and Fig. S2), covering ~95.1% of the flow cytometry-estimated genome size. The final assemblies of HapA and HapB totaled 1.48 Gb, with contig N50 values of 22.8 and 17.3 Mb, respectively (Fig. 1D).

The assembly quality was rigorously assessed using multiple complementary metrics. Merqury analysis yielded consensus Quality Value (QV) exceeding 65 for both haplotypes (HapA: 66.03; HapB: 65.60), indicating highly accurate base-level assembly. K-mer completeness was estimated to be >98%, demonstrating that the vast majority of the reliable genomic information was successfully captured. Taxonomic classification using Kraken2 with the Standard-8 database confirmed the absence of viral or nonplant contamination in the final chromosome assemblies (Table S1). In addition, the long terminal repeat (LTR) Assembly Index scores were 20.8 for HapA and 22.2 for HapB, reflecting high continuity and gold-standard assembly quality.

To ensure a comprehensive repeat landscape prior to gene annotation, we integrated multiple complementary approaches, combining standard *de novo* pipelines (RepeatModeler and The Extensive *de novo* TE Annotator) with targeted identification using MITE-Hunter and Tandem Repeats Finder (TRF). Overall, repetitive elements accounted for 70.68% and 72.59% of HapA and HapB, respectively (Table S2). Following repeat masking, we predicted 32 213 and 32 037 protein-coding genes in HapA and HapB, respectively, with both assemblies achieving 98.9% completeness according to BUSCO assessment (Fig. 1D). Compared with previously published cannabis genomes, the CSLZ assembly exhibits higher contiguity and more complete gene representation, underscoring its superior quality and reliability for downstream analyses (Fig. S3).

Evolutionary insights and gene family expansion in cannabis

Using CSLZ as the representative genome for *C. sativa*, we performed comparative genomics analyses with nine representative plant species spanning key evolutionary nodes: *O. sativa*, *Solanum lycopersicum*, *Vitis vinifera*, *Arabidopsis thaliana*, *Morus notabilis*, *Ficus macrocarpa*, *Parasponia andersonii*, *Trema orientale*, and *Humulus lupulus*. A total of 606 single-copy orthologous genes were identified and used to construct a phylogenetic tree (Fig. 2A, Table S3). The results showed that *C. sativa* had the closest evolutionary relationship with *H. lupulus*, with an estimated divergence

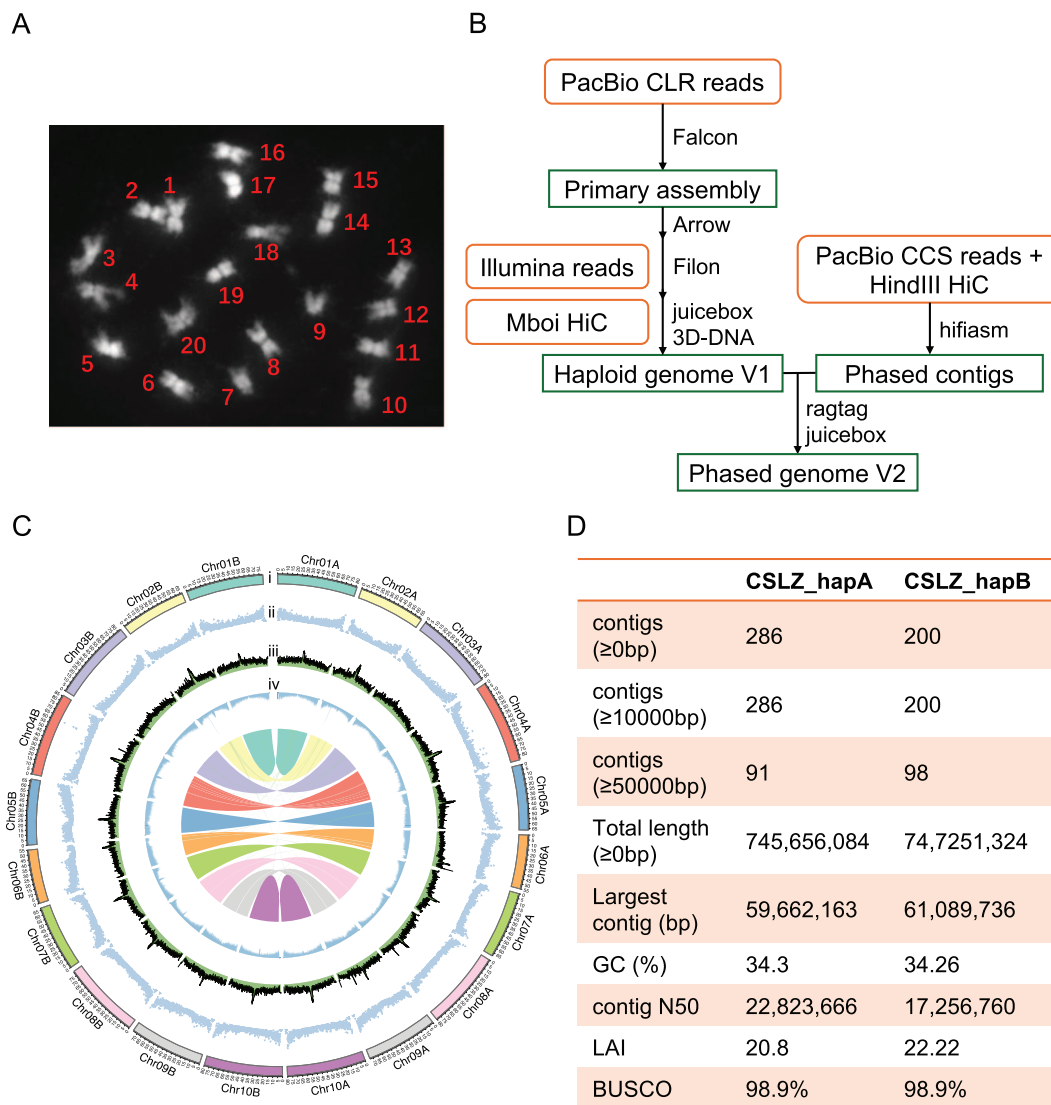


Figure 1 Sequencing samples and genome assembly results. (A) Karyotype of *C. sativa* Linzhi. (B) Workflow of genome assembly. (C) Characteristics of the CSLZ genome. Genomic landscape of the 20 assembled chromosomes. Tracks are displayed from the outer to inner rings as follows: (i) chromosome length (Mb); (ii) gene density; (iii) GC content; and (iv) TE density. All densities were calculated in 100 kb windows. The innermost arcs represent synteny blocks. (D) Summary of genome assembly.

time of 18 (13.2–21.8) million years ago (Mya; Fig. 2A). *P. andersonii* and *T. orientale* divergence in ~10.2 (5.1–15.7) Mya (Fig. 2A). Compared to the most recent common ancestor of *H. lupulus* and *C. sativa*, 686 orthogroups were expanded and 1745 orthogroups contracted in cannabis (Fig. 2A). Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis revealed that significantly expanded orthogroups were associated with metabolic pathways, including tryptophan metabolism, sesquiterpenoid and triterpenoid biosynthesis, and cyanoamino acid metabolism (Fig. 2C, Table S4). The cyanoamino acid metabolism pathway is a key component of plant defense, contributing to cyanogenic glycoside biosynthesis that deters herbivores and pathogens while influencing defense signaling and broader metabolic regulation [17]. Gene family clustering analysis indicated that 8085 gene families were shared among all 10 species (Table S3), and 4095 genes were unique to *C. sativa* (Fig. 2B, Table S5). KEGG enrichment analysis showed that cannabis-specific genes were primarily enriched in phenylpropanoid biosynthesis, ribosome

biogenesis, and starch and sucrose metabolism (Fig. 2C, Table S6). Together, the expanded and cannabis-specific gene families were predominantly associated with the biosynthesis of specialized metabolites such as terpenoids and phenylpropanoids, as well as defense-related pathways including cyanoamino acid metabolism. This coordinated expansion may help explain why the wild accession CSLZ is enriched in medicinal compounds and exhibits strong environmental adaptability.

Whole-genome duplication (WGD) events during cannabis evolution were examined by calculating the density distribution of the synonymous substitution rate (K_s), and two significant peaks were identified (Fig. 2D). The first K_s peak at ~1.98 represented the ancient γ triplication event. The second K_s peak appears at ~0.05, which may represent recent gene expansion. Analysis of LTR expansion revealed that both Copia and Gypsy underwent rapid expansion ~0.1 Aya, suggesting that recent genetic expansion in cannabis may have been driven by intense LTR transposon bursts (Fig. 2E).

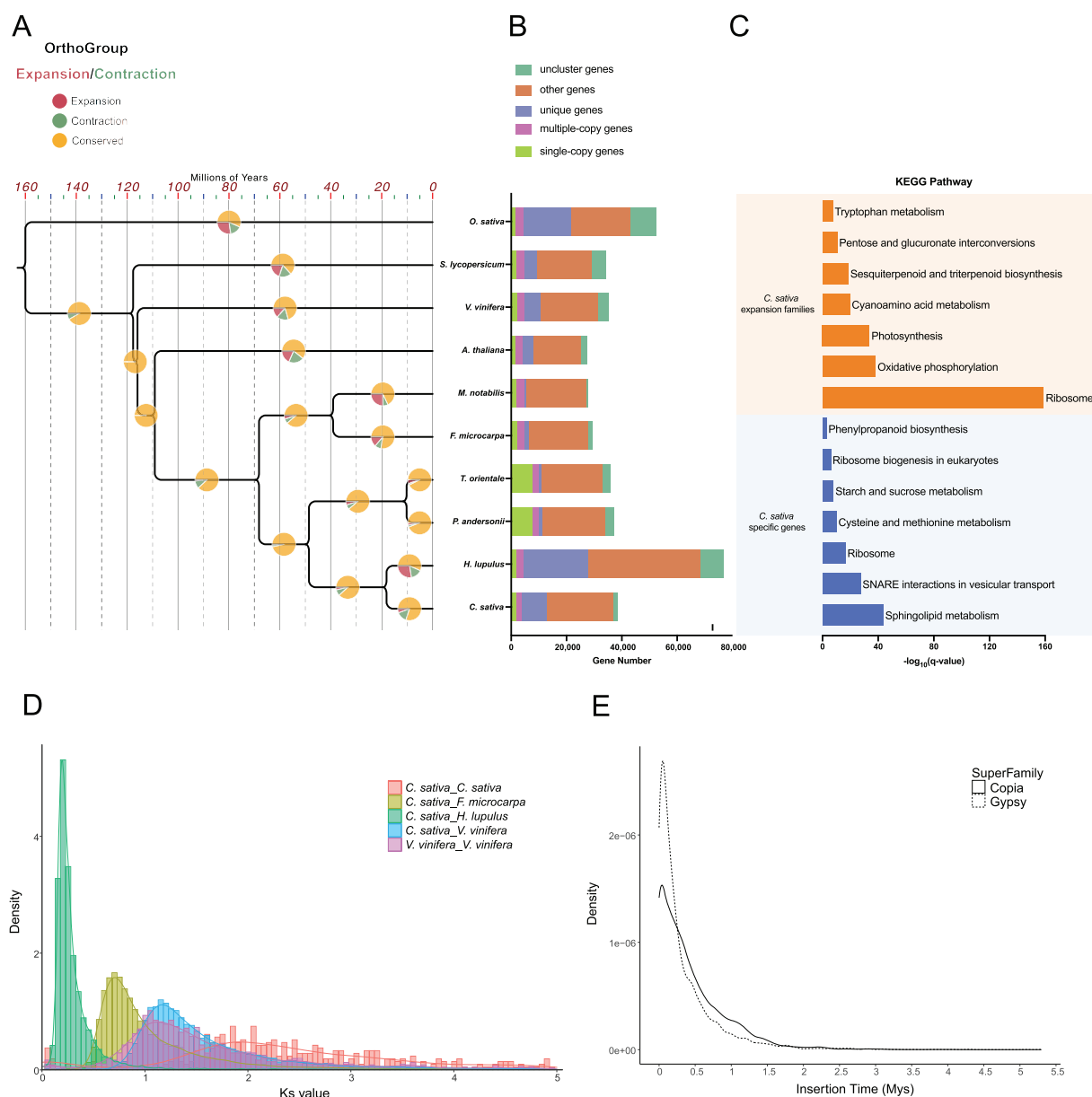


Figure 2 Comparative genomics and TE expansion. (A-B) The phylogenetic tree was constructed based on 21 single-copy genes of *O. sativa*, *S. lycopersicum*, *V. vinifera*, *A. thaliana*, *M. notabilis*, *F. macrocarpa*, *P. andersonii*, *T. orientale*, *H. lupulus*, and *C. sativa*. Orthogroup dynamics, including expansions, contractions, and conservative families, are indicated at each node as defined in the internal figure legend. (C) KEGG enrichment analysis of cannabis expanded gene families (up) and species-specific gene families (down). (D) Ks distributions for paralogs of cannabis and orthologs between cannabis and other species. (E) Insertion time of LTR estimated by sequence divergence of Copia and Gypsy.

Comparative genomics reveals conserved architecture and lineage-specific innovation in cannabis

To investigate evolutionary divergence among cannabis lineages, we perform whole-genome synteny analysis between CSLZ and other representative cannabis varieties. Overall, chromosomes exhibited clear one-to-one correspondence across varieties, indicating conserved genome collinearity. Nevertheless, numerous structural rearrangements were detected, reflecting substantial intraspecific genomic plasticity (Fig. 3A). Gene ontology (GO) enrichment of 7461 conserved syntenic block genes revealed

significant enrichment in terms related to ‘nucleobase-containing small molecule metabolic process’, ‘response to salicylic acid’, and ‘response to oomycetes’, suggesting evolutionary conservation of metabolic and defense functions among cannabis genomes (Fig. 3B, Table S7).

Orthogroup clustering across nine cannabis genomes identified lineage-specific gene families, with the number of variety-specific orthogroups ranging from 23 orthogroups (638 genes) in Pink Pepper to 499 orthogroups (2827 genes) in Finola (Fig. 3C). GO enrichment of CSLZ-specific genes highlighted overrepresentation of categories associated with ribosomal small subunit assembly and ATP synthase complex assembly (Fig. 3D). These processes support enhanced protein synthesis and energy

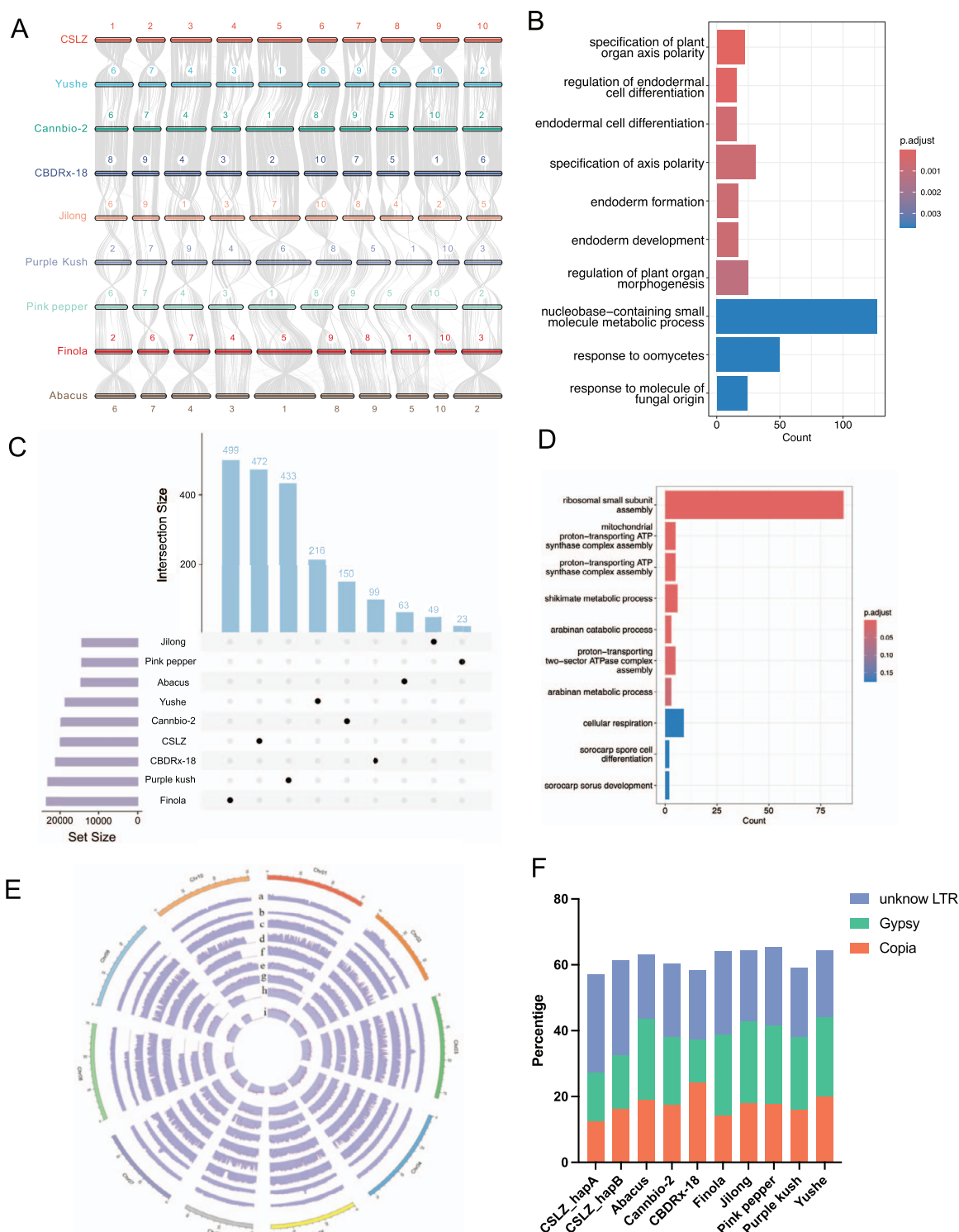


Figure 3 Conservation and specificity among cannabis varieties. (A) Collinearity map of nine cannabis genomes. (B) GO enrichment of genes within conserved regions among cannabis varieties. (C) Unique gene families across cannabis genomes. (D) GO enrichment of unique gene families in CSLZ varieties. (E) Distribution of TEs in the genomes of different cannabis varieties. Tracks are displayed from the outer to inner rings as follows: (a) Linzhij; (b) Abacus; (c) Cannbio-2; (d) CBDRx-18; (e) Jilong; (f) Pink Pepper; (g) Yushe; (h) Finola; (i) Purple Kush. (F) Percentage of different types of TE in the genomes of different cannabis varieties.

production, implying that CSLZ might have evolved elevated translational and metabolic adaptability. This pattern suggests that, as a wild accession, CSLZ follows an evolutionary trajectory emphasizing both basal cellular function and metabolic flexibility.

We further compared TE content and composition among the nine genomes. Finola harbored the largest TE complement, spanning 711.6 Mb and accounting for 70.5% of its genome (Fig. 3E). Given that LTR retrotransposons comprise the majority

of TE sequences, we quantified the relative proportions of Copia, Gypsy, and unclassified LTRs across varieties. In both haplotypes of CSLZ, Copia elements represented the smallest fraction (12.37% and 16.28%, respectively), similar to proportions observed in other cannabis genomes (Fig. 3F, Table S8).

Transposable element insertions shape allelic regulation of cuticle biosynthesis

To systematically assess the impact of TEs on gene expression, we classified genes harboring TEs within their genomic regions as TE-associated genes (TE-genes) and the remainder as noTE-genes. Transcriptome sequencing of CSLZ glandular trichomes, female flowers, xylem, phloem, roots, and leaves revealed that noTE-genes consistently exhibited significantly higher expression levels than TE-genes across all tissues (Fig. 4A), indicating a pervasive suppressive effect of TEs on transcriptional activity. GO enrichment analysis of TE-genes showed significant enrichment in pathways related to lipid, terpene, wax, and fatty acid biosynthesis (Fig. 4B, Table S9), while KEGG analysis further highlighted ‘steroid hormone biosynthesis’, ‘sesquiterpenoid and triterpenoid biosynthesis’, and ‘cutin, suberin, and wax biosynthesis’ (Fig. 4C, Table S10). Given that cutin and wax constitute essential components of the leaf cuticle, which mediates protection against desiccation and environmental stress, we focused on leaf transcriptomes to elucidate the molecular basis of cuticle formation.

By comparing gene expression between leaves and other tissues across two haplotypes, and integrating ASE in leaves, we identified 1100 leaf-specific genes exhibiting haplotype-biased expression (Fig. 4D). GO enrichment analysis indicated significant overrepresentation of terms related to secondary metabolic processes, particularly the cutin biosynthetic process (Fig. 4E, Table S11). Seven of these genes encoded enzymes involved in cutin biosynthesis, including three lipid transfer proteins (LTPs), two glycerol-3-phosphate acyltransferases (GPATs), and two HXXXD-type acyltransferase family proteins (Fig. 4F). Notably, LTP genes showed markedly higher expression in haplotype A (Fig. 4F). Structural annotation identified an LTR insertion within the intron of *LTP4-like_hapB* gene, which was absent from *LTP4-like_hapA* (Fig. 4G). These observations suggest that the LTR insertion in *LTP4-like_hapB* may drive the observed allelic expression divergence, linking TE insertion to regulatory variation in cuticle-associated genes.

Haplotype-specific structural variation drives diversification of terpene biosynthesis

Genomic variation is a major driver of gene innovation and functional diversification. To investigate haplotype-specific SVs, we aligned HapB against HapA using MUMmer and identified SVs with SyRI (Fig. 5A). We detected 32 594 insertion, 116 866 duplicated sequences, 29 055 gap, 1966 inversion, 11 790 relocation, and 22 989 translocation (Fig. 5B, Table S12), indicating extensive structural asymmetry between haplotypes. To further characterize the genomic architecture underlying these variants, we profiled genome-wide sequence complexity, revealing the distribution of low-complexity regions (LCRs) across all chromosomes (Fig. S4).

In cannabis, terpenoids play pivotal ecological roles, mediating plant–insect–microbe interactions and contributing to chemotypic and pharmacological diversity [18]. Sesquiterpenes determine characteristic aroma profiles and synergize with cannabinoids, whereas triterpenes underpin bioactive compound diversity [19]. Germacrene D, a sesquiterpene with antibacterial and anti-inflammatory activity, is produced from farnesyl diphosphate by germacrene-D synthase (GERD; Fig. 5C) [20]. Genomic complexity analysis using AlcoR revealed that the HapB allele resides within a low-complexity region, in contrast to the high-complexity context of the HapA allele (Fig. 5D). To further characterize the structural environment of this locus, we performed a detailed synteny analysis of the local chromosomal region (Figs S5 and S6). We identified a haplotype-specific structural variant at the GERD locus; the HapB allele lacks the 5′ region encoding the first 14 amino acids, likely abolishing enzymatic function (Figs 5E and S7). Moreover, a four-base insertion in the GERD_HapB promoter disrupts a MYC transcription factor-binding motif present in GERD_HapA (Fig. 5F). Transcriptomic profiling across nine female flower developmental stages revealed strong expression of GERD_HapA, whereas GERD_HapB was transcriptionally inactive (Fig. 5F). These findings suggest that haplotype-specific structural and regulatory variations jointly drive allelic expression divergence of a key terpene synthase (TPS).

β -Amyrin, a triterpenoid alcohol that serves as a core precursor for diverse bioactive triterpenes, is synthesized by 2,3-oxidosqualene cyclases (OSCs) that catalyze the cyclization of 2,3-oxidosqualene into α - and β -amyrin (Fig. 5G) [21]. The OSC gene family exhibits haplotype-specific copy number variation in CSLZ, with three copies in HapA and one in HapB (Fig. 5H). Expression profiling across flower development revealed that a single HapA copy (CsLZ_hapA00083) is highly expressed throughout all stages, while the HapB copy remains silent (Fig. 5I). Functional characterization by co-expression of CsLZ_hapA00083 with 3-Hydroxy-3-methylglutaryl CoA reductase (HMGR) in *Nicotiana benthamiana* confirmed β -amyrin production (0.347 μ g/ml; Fig. 5J). Collectively, these results indicate that gene duplication and subsequent expression divergence in HapA created an active, three-copy expansion of OSC, while HapB retained a single, silent copy, revealing haplotype-specific innovation in triterpenoid biosynthesis.

Allelic regulation of flavonoid biosynthesis in glandular trichomes

To examine the contribution of ASE to tissue-specific secondary metabolism, we performed transcriptome sequencing across six CSLZ representative tissues of *C. sativa*: roots, stem xylem, stem phloem, leaves, female flowers (without trichomes), and glandular trichomes. Correlation analysis revealed that global expression profiles were primarily shaped by tissue identity, roots formed a distinct cluster from aerial organs, glandular trichomes clustered closely with female flowers, and leaves grouped with stem phloem (Fig. S8). To identify genes subject to both allelic and spatial regulation, we defined a set of 1252 genes specifically expressed in glandular trichome and differentially expressed between the two haplotypes (Fig. 6A). GO enrichment analysis revealed that these genes were significantly enriched for terms associated with secondary metabolic processes, particularly the biosynthesis of flavonoids and phenylpropanoids (Fig. 6B, Table S13).

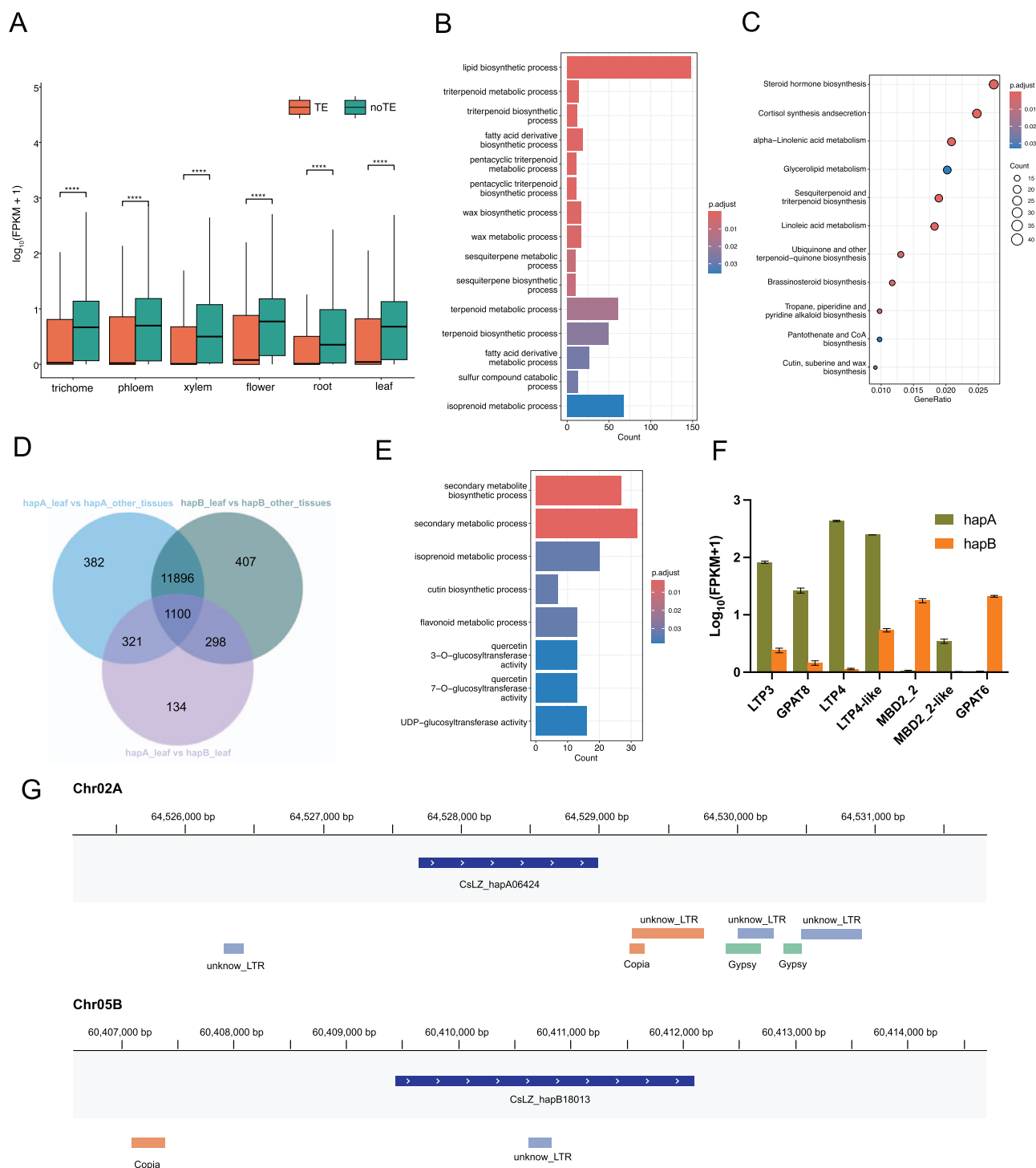


Figure 4 TE-mediated transcriptional suppression and allelic regulation of cuticle biosynthesis (A) Effects of TE insertions on gene expression across different tissues. Statistical significance was determined using the two-sided Mann–Whitney test, **** indicates $P < 0.0001$. (B) GO enrichment analysis of TE-gene. (C) KEGG enrichment analysis of TE-gene. (D) Venn diagrams of the number of different expression genes in the leaf. (E) GO enrichment analysis of leaf tissue-specific genes with allelic differential expression. (F) Expression levels of seven cutin biosynthetic process genes. (G) TE insertion in LTP4-like. LTP4-like_hapA without TE insertion (up) and LTP4-like_hapB with TE insertion (down).

Flavonoids, key derivatives of the phenylpropanoid pathway, serve multifaceted roles in cannabis, contributing to ultraviolet (UV) protection, pigmentation, and defense against pathogens and oxidative stress [22, 23]. To assess allelic contributions to trichome-specific flavonoid biosynthesis, we analyzed ASE

within the flavonoid biosynthetic pathway. Many genes specifically expressed in glandular trichome also exhibited strong ASE bias, suggesting coordinated genetic and spatial regulation (Fig. 6C). For example, the flavonoid 3',5'-hydroxylase (F3'5'H), which catalyzes the hydroxylation of dihydrokaempferol

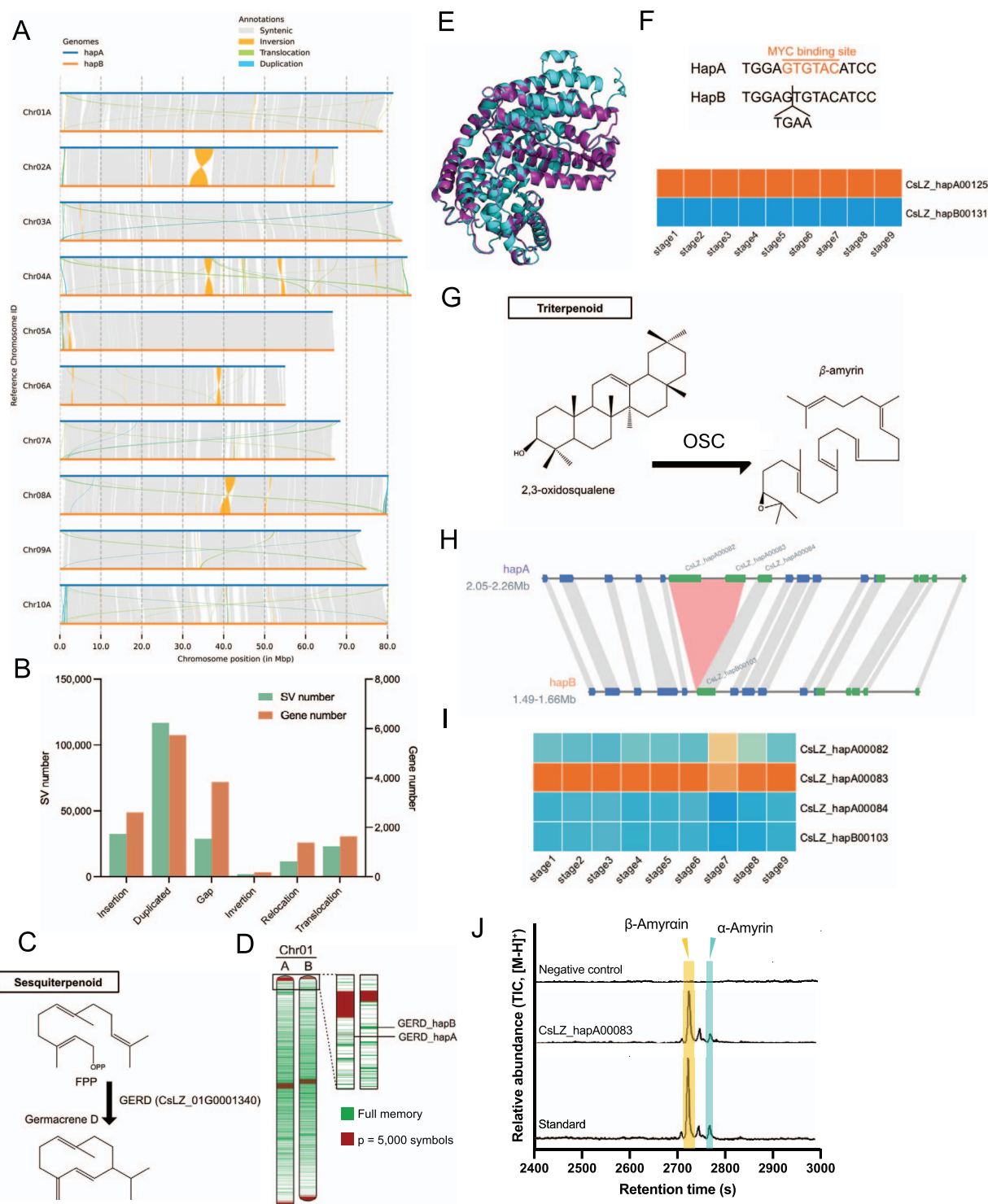


Figure 5 Haplotype-specific structural variation drives allelic divergence in terpene biosynthesis. (A) SVs between HapA and HapB. (B) Total number of variations and overall number of genes containing SVs in the gene region. (C) Structure and catalytic reaction of GERD. (D) Low complexity map of chromosome 1. (E) Multi-sequence comparison between GERD_HapA and GERD_HapB. (F) Differences in GERD expression and promoter region in different haplotypes. (G) Catalytic reaction of OSC. (H) Copy number variation of OSC in different hap. (I) OSC expression heatmap. (J) GC-MS extracted ion chromatogram of m/z 218 in tobacco co-expressing CsLZ_hapA00083 and CSHMGR1.

to dihydromyricetin, showed pronounced haplotype-specific expression, four haplotypes were predominantly expressed in glandular trichomes, one was mainly active in stem xylem and phloem, and one was not transcribed. Both copies of *O*-methyltransferase (MT) were preferentially expressed in glandular

trichomes, with higher expression from the hapB allele. 3-O-glycosyltransferase (3GT), which glycosylates cyanidin to cyanidin 3-O-glucoside, showed marked allelic asymmetry. Among six 3GT haplotypes, most were expressed in glandular trichomes, roots, and leaves, whereas one hapA copy (CsLZ_hapA13739) was

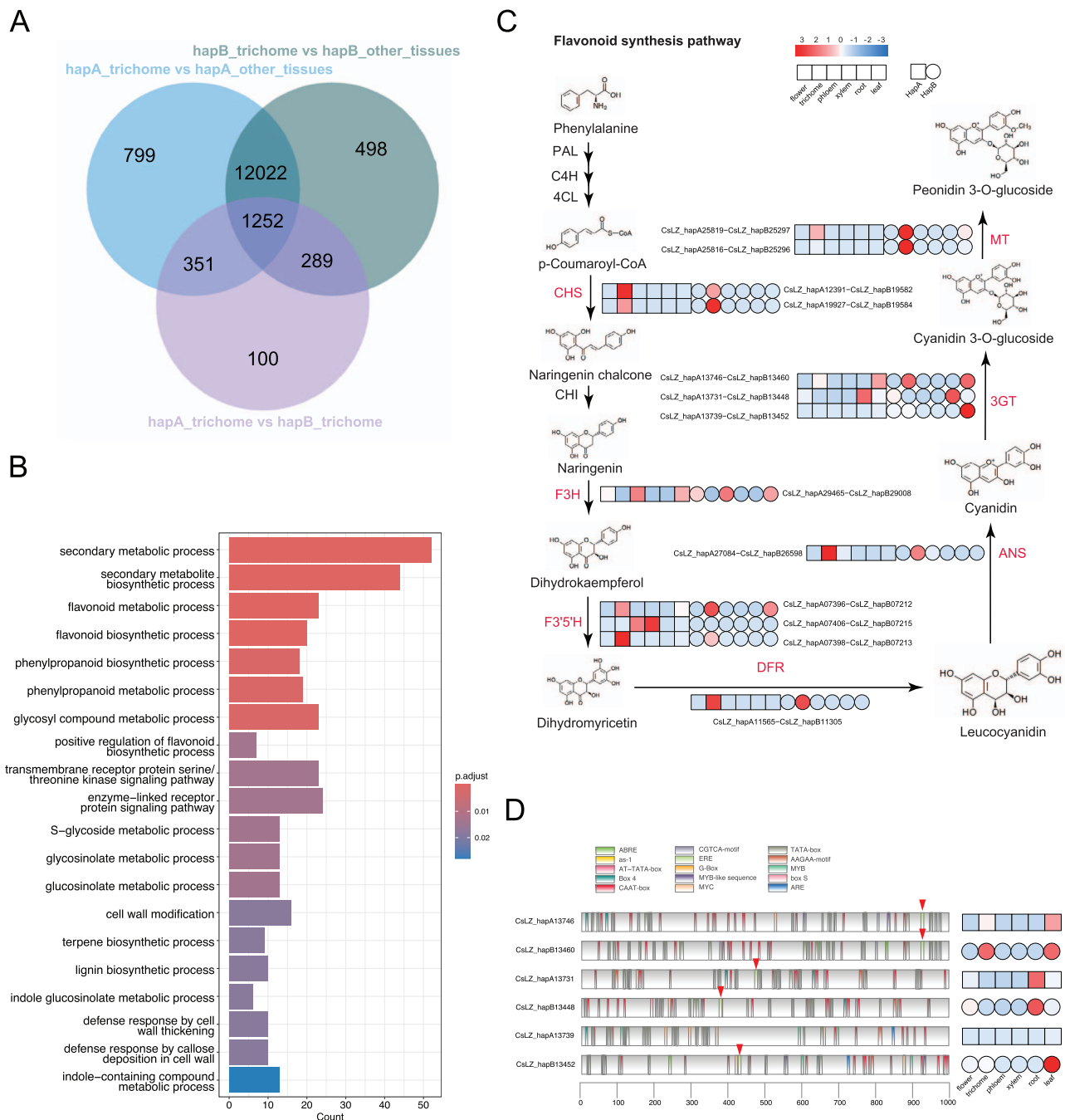


Figure 6 Haplotype-dependent transcriptional regulation of flavonoid biosynthesis in glandular trichomes. (A) Venn diagrams of the number of different expression genes in the glandular trichome. (B) GO enrichment analysis of glandular trichome tissue-specific genes with allelic differential expression. (C) Heatmap of flavonoid synthesis pathway gene expression in different tissues. (D) Transcription factor binding sites and gene expression levels of six 3GT haplotypes. The triangle represents the ERE motif.

transcriptionally inactive across all tissues. Comparative promoter analysis revealed that the non-expressed CslZ_hapA13739 promoter lacked an ethylene-responsive element motif conserved in transcriptionally active 3GT promoters of other species (Fig. 6D), such as *Vaccinium corymbosum* and *V. vinifera* [24, 25]. The loss of this regulatory motif likely contributes to transcriptional silencing of CslZ_hapA13739, suggesting that promoter motif variation mediates allelic regulation of flavonoid biosynthesis and contributes to tissue-specific metabolic diversification in cannabis.

Haplotype-resolved dynamics of gene expression during female flower development

The female cannabis flower is characterized by a perianth enclosing an ovule, protective bracts that host the highest density of glandular trichomes, and two stigmas. From week 1 to week 9 of flowering, calyx size increases from 1 to 6 mm, accompanied by a marked rise in trichome density [26, 27]. To investigate the contribution of ASE to female flower development and metabolism,

we performed transcriptome sequencing across nine flowering stages of the CSLZ accession. Principal component analysis (PCA) revealed that developmental stage accounted for the largest proportion of transcriptional variance (PC1, 40.72%), while haplotype identity contributed a secondary but distinct axis (PC2, 28.61%; Fig. 7A). This indicates that both developmental progression and genetic architecture shape the global transcriptional landscape during cannabis female flower development.

We identified 972 differentially expressed alleles that were responsive to both factors, showing variation between haplotypes at the same developmental stage and across stages within the same haplotype. (Fig. 7B). GO enrichment of this intersecting set revealed overrepresentation of genes involved in secondary metabolic processes, including phenylpropanoid, triterpenoid, flavonoid, and lignin biosynthesis (Fig. 7C, Table S14). Within lignin biosynthetic pathway, phenylalanine ammonia-lyase (PAL) was predominantly expressed from HapB during early and middle developmental stages, while hydroxycinnamoyl-CoA shikimate/quinate hydroxycinnamoyltransferase (HCT), caffeoyl-CoA MT, cinnamyl alcohol dehydrogenase and caffeic acid MT genes displayed haplotype-biased expression across all the stages (Fig. 7D). Promoter analysis of the non-expressed allele HCT2_HapA (CSLZ_hapA19103) revealed deletion of a Box 4 light-responsive *cis*-element due to a haplotype-specific base substitution (Fig. S9), likely causing the observed ASE.

Allelic regulation of trichome development and cannabinoid biosynthesis

In *C. sativa*, glandular trichomes are the primary sites for the biosynthesis and storage of pharmaceutically valuable secondary metabolites, including cannabinoids and terpenes, which collectively determine the plant's medicinal potency. To dissect the allele-specific regulatory basis of trichome development and metabolism, we conducted allele-resolved RNA sequencing across nine developmental stages of glandular trichomes. PCA of the transcriptome data revealed that haplotype identity was the major determinant of expression variance, driving separation along the diagonal of the first principal component (Fig. 8A).

Examination of cannabis homologues of known glandular trichome regulators in tomato uncovered dynamic expression patterns during trichome ontogeny (Fig. 8B, Table S15). Among these, the negative regulator TRIPTYCHON (TRY) and the positive regulator EXPANSIN 12 (EXPA12) exhibited striking haplotype-specific transcriptional bias (Fig. 8B). *TRY_HapB* was strongly expressed, whereas *TRY_HapA* expression was nearly undetectable; a similar trend was observed for *EXPA12*, with *EXPA12_HapB* highly expressed and *EXPA12_HapA* transcriptionally silent (Fig. 8B).

Structural comparison of the TRY locus revealed extensive haplotype divergence. In tomato, TRY encodes a 94-amino-acid protein, whereas the actively transcribed cannabis TRY_HapB produces a 78-amino-acid protein and the inactive TRY_HapA encodes a 239-amino-acid protein (Fig. 8C). Predicted tertiary structures of the two cannabis alleles are distinct, with TRY_HapB more closely resembling the tomato counterpart (Fig. 8C), suggesting that TRY_HapB may have acquired lineage-specific functions in regulating glandular trichome density and spatial

patterning. *EXPA12* belongs to the expansin family, which is broadly conserved among dicotyledonous plants and mediates cell wall loosening by disrupting hydrogen bonds between polysaccharides, thereby facilitating cell expansion under acidic conditions [28]. Canonical expansin protein comprises both an expansin C-terminal domain and a lytic transglycolase domain. Actively transcribed EXPA12_HapB retains both domains, consistent with full functional potential, whereas the non-expressed EXPA12_HapA lacks the C-terminal domain required for full activity (Fig. 8D and Fig. S10). This combination of structural truncation and haplotype-specific expression indicates functional divergence between alleles, which may influence the morphological and developmental specialization of cannabis glandular trichomes.

Given that cannabinoids are the principal bioactive compounds in cannabis, we next profiled the transcriptional dynamics of the cannabinoid biosynthetic pathway during trichome development. Gene expression heatmaps revealed stage-dependent activation of key biosynthetic enzymes and widespread allelic expression divergence (Fig. 8E). 1-Deoxy d-xylulose-5-phosphate synthase and acyl-activating enzyme were preferentially expressed during early stages, whereas most other pathway genes peaked at later stages. Notably, all genes except olivetol synthase (OLS) were expressed from both haplotypes. OLS, which catalyzes the formation of olivetolic acid, a precursor of cannabigerolic acid, showed extreme allelic bias, with OLS-HapB highly expressed in late developmental stages (stage 9), while OLS-HapA remained almost silent. Structural analysis revealed three consecutive TE insertions in the intron of *OLS-HapA* (Fig. 8F), leading to substantial structural divergence relative to OLS-HapB. These TE insertions likely contribute to observed allelic silencing, suggesting how TE-induced SVs can rewire allelic regulation and shape metabolic specialization in cannabis glandular trichomes.

Discussion

The immense therapeutic and ecological significance of specialized metabolites in plants, such as the complex cannabinoid, terpenoid, and flavonoid profiles of *C. sativa*, demands a complete understanding of their regulatory control. Although the key biosynthetic pathway genes have been identified, the genetic and regulatory mechanisms governing their production remain incompletely understood, particularly in the context of the plant's highly heterozygous and repetitive genome. To address this gap, we generated a high-quality, fully phased diploid genome of the wild cannabis accession CSLZ, enabling an allele-aware exploration of genomic evolution and metabolic specialization. Integrative transcriptomic and structural analyses reveal extensive structural and regulatory asymmetry between homologous chromosomes, demonstrating that haplotype-specific variation is a critical, yet underappreciated, driver of metabolic divergence in this wild accession, directly shaping specialized metabolite biosynthesis and tissue-specific accumulation.

Our comparative genomic analyses revealed that *C. sativa* has experienced lineage-specific expansion of gene families involved in secondary metabolism and primary-metabolite support pathways. Expanded orthogroups include those linked to tryptophan, sesquiterpenoid, and triterpenoid biosynthesis, consistent with the amplification of TPS and cytochrome P450 families observed in other metabolite-rich species such as

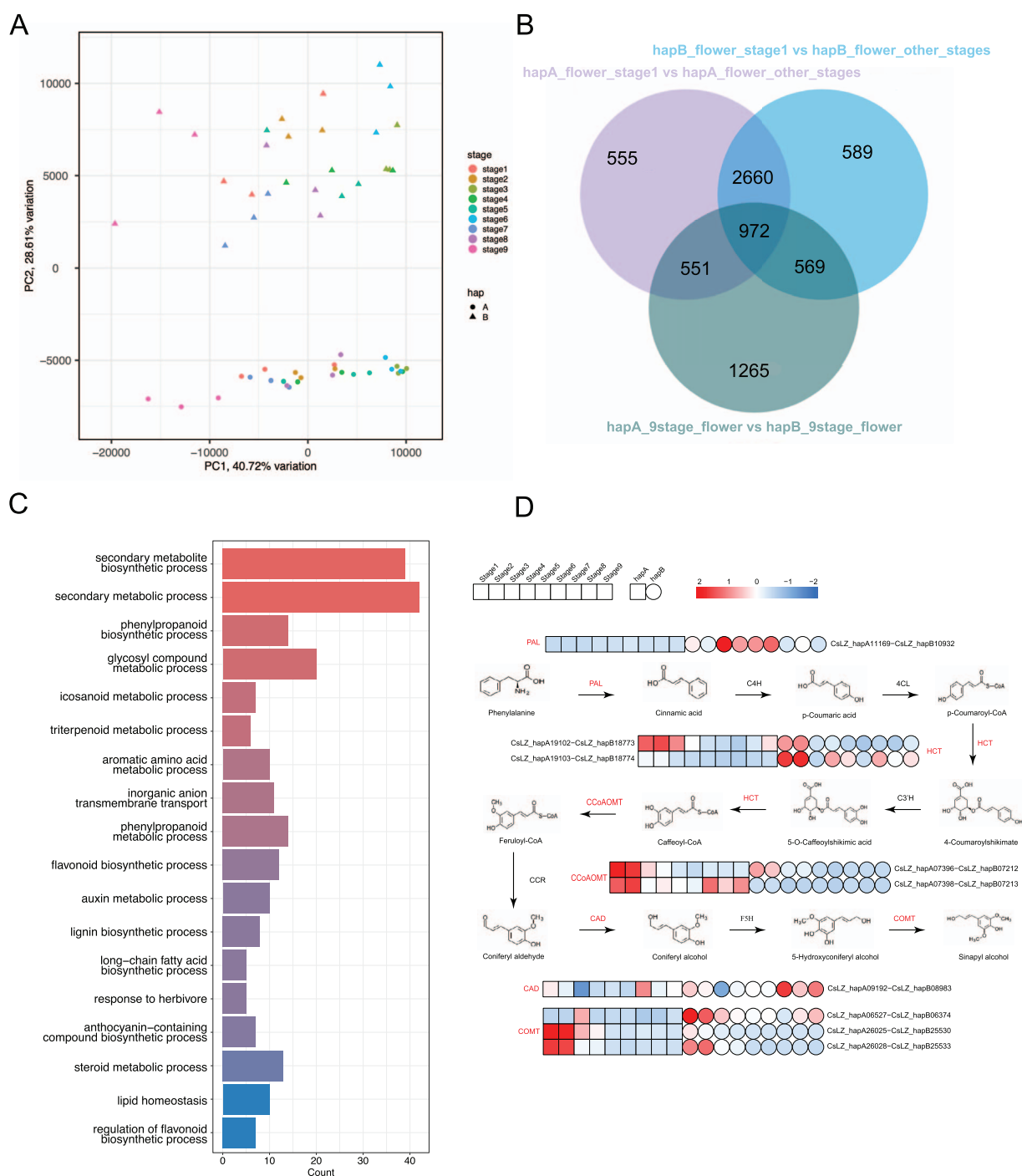


Figure 7 Allele-specific regulation and developmental control of secondary metabolism in female flowers. (A) PCA of the female flower RNA-seq data. (B) Venn diagrams of the number of different expression genes in female flower. (C) GO enrichment analysis of differentially expressed genes with allelic differential expression during female flower development. (D) Heatmap of phenylpropane biosynthesis pathway gene expression during female flower development.

Pinaceae [5] and *R. officinale* [14]. These expansions likely enhance both metabolic capacity and environmental resilience, coupling biosynthetic innovation with stress adaptation. Enrichment of cannabis-specific gene families in phenylpropanoid biosynthesis, ribosome biogenesis, and starch and sucrose metabolism further suggests that elevated primary metabolic flux supports intensified secondary metabolism in specialized tissues such as glandular trichomes.

A central finding of this study is the pervasive role of TEs in shaping allelic expression and functional divergence, particularly within specialized metabolic pathways. Genes harboring TEs within their structure or regulatory regions consistently exhibited significantly reduced expression across all sampled tissues. This global suppressive effect likely arises from multiple mechanisms, including the disruption of gene structure, altered DNA methylation, chromatin remodeling, or the introduction of novel

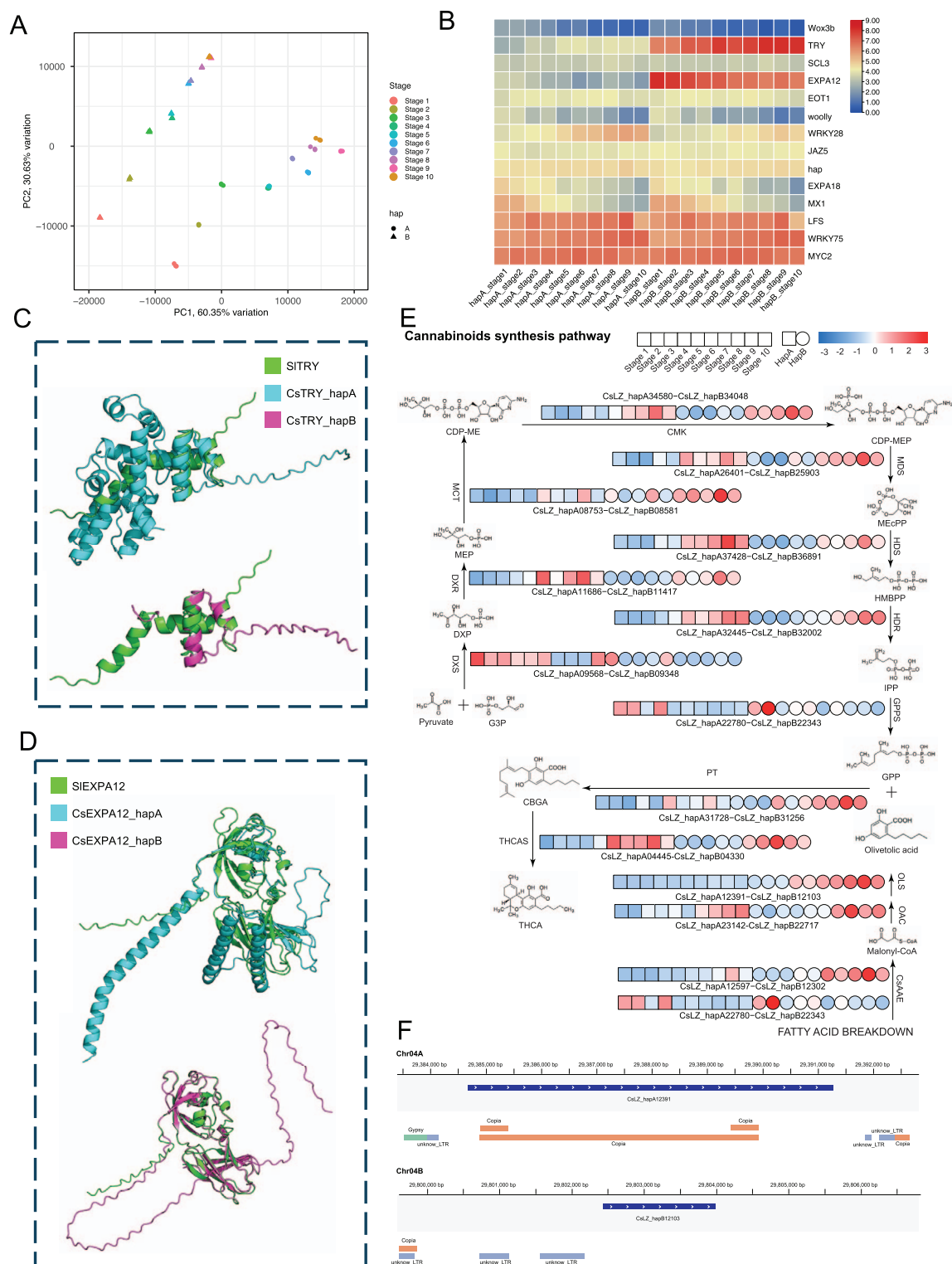


Figure 8 Allele-specific regulation and structural divergence shape trichome development and cannabinoid biosynthesis. (A) PCA of the glandular trichomes RNA-seq data. (B) Heatmap of expression levels of cannabis homologs of tomato trichome development regulators. (C) Protein structure of SITRY, CsTRY_hapA, and CsTRY_hapB. (D) Protein structure of SIEXPA12, CsEXPA12_hapA, and CsEXPA12_hapB. (E) Heatmap of cannabinoid biosynthesis pathway gene expression during glandular trichome development. (F) TE insertion in OLS. OLS_hapA with TE insertion (up) and OLS_hapB without TE insertion (down).

cis-regulatory elements, such as transcription factor binding sites [29]. For example, TEs modulate gene expression by influencing transcription factor binding in the cotton genus [30]. In *Brassica*,

TEs reshape local expression landscapes through methylation and structural variation [31]. Similar observations were reported in *P. densiflora* [5], where LTR retrotransposons affect resin

biosynthetic genes, and modulate DNA methylation and expression of vanillin biosynthetic genes in *Vanilla planifolia* [32]. In CSLZ, TE-associated genes were enriched in lipid, terpene, and wax biosynthetic pathways that contribute to cuticle formation and environmental adaptation. Notably, an LTR insertion in the intron of LTP4-like gene of haplotype B coincided with transcriptional repression, while three consecutive TE insertions in the intron of *OLS-HapA* likely resulted in near-complete transcriptional silencing of the ‘allele’. These findings are consistent with evidence from *Arabidopsis* and tomato showing that intronic regions, being longer and more permissive, frequently accommodate TE insertions that modulate gene regulation [33]. Collectively, our results highlight TEs as dynamic agents of allelic regulatory divergence and metabolic innovation in ‘cannabis’. Through both disruptive and *cis*-regulatory effects, recent TE activity has rewired expression landscapes in a haplotype-specific manner, fostering transcriptional plasticity and chemotypic diversity across tissues and developmental stages.

The cannabis genome exhibits widespread structural variation, with numerous inversions, translocations, and duplications differentiating the two haplotypes. Many of these directly impact genes encoding key metabolic enzymes, such as GERD and OSC. At the *GERD* locus, the HapB allele exhibits both a truncation and the disruption of an MYC-binding motif in its promoter, which abolish its expression and leave the HapA allele as the sole active copy. This finding is consistent with evidence from *Taraxacum kok-saghyz* (Russian dandelion), where the transcription factor TksMYC2 binds to the promoter of the germacrene A oxidase gene via a homologous motif, and overexpression of TksMYC2 enhances germacrene A oxidase expression in tobacco [34]. Given that germacrene A and germacrene D are structural isomers, it is plausible that the promoter mutation in *GERD*-HapB disrupts MYC transcription factor binding, thereby leading to the loss of its expression. Furthermore, haplotype-specific expansion of *OSC* in *HapA* contributes to differential triterpenoid biosynthesis, suggesting that tandem duplication followed by asymmetric retention and regulatory remodeling drives chemical diversification. The similar patterns were observed in *P. densiflora*, where large-scale SVs underlie allelic diversity in TPS and phenylpropanoid gene families, facilitating differential accumulation of resin acids and volatiles across tissues [5]. Likewise, expansion of *ADS* correlates with enhanced artemisinin content in *A. annua* [13]. Collectively, these patterns highlight structural rearrangements as an important mechanism linking haplotype divergence to metabolic specialization.

At the regulatory level, *cis*-element variation shapes allelic expression landscapes and contributes to tissue-specific metabolic specialization. In cannabis, promoter polymorphisms appear to account for allele-specific transcriptional differences through the loss of key regulatory motifs. For example, transcriptional inactivation of one 3GT allele coincides with the loss of ERE/GCC-box motifs, which are well-known positive *cis*-acting elements enhancing transcription in *Arabidopsis* [35, 36] and other dicots [37]. Similarly, an HCT allele is inactivated due to the absence of a light-responsive Box4 element, which is otherwise conserved in the phenylpropanoid gene promoters of *S. tuberosum* [38] and *Agrostis stolonifera* [39]. This is consistent with the high-frequency variation observed in the promoter regions of ASE genes associated with hybrid vigor in maize parents [40], providing

new perspectives for investigating how promoter region variation between alleles influences gene expression.

Integration of haplotype-resolved and tissue-specific transcriptomes revealed that ASE is tightly aligned with spatial metabolic partitioning. For example, glandular trichomes preferentially express HapB alleles of trichome regulators, including CsTRY and CsEXPA12, which are predicted to produce functional proteins. In contrast, the corresponding HapA alleles encode structurally divergent proteins, CsTRY-HapA encodes a protein with a distinct predicted tertiary structure, and CsEXPA12-HapA lacks the C-terminal expansin domain required for full activity. Similarly, metabolic polarization is observed for pathway genes, with OLS-HapB driving cannabinoid biosynthesis in trichomes, whereas HapA alleles dominate phenylpropanoid and lignin pathways in structural tissues. This developmental and haplotype-specific partitioning likely represents an adaptive strategy to balance growth, defense, and reproduction, paralleling observations in *V. planifolia*, where phased alleles and gene dosage shape vanillin and phenolic compound biosynthesis [32]. In potato, haplotype-level analyses of starch and steroidal glycoalkaloid (SGA) pathways identified 2169 haplotypes across 119 genes, ~70% of which are transcriptionally active in a tissue-specific manner. Structural variation among these haplotypes contributes to differential expression and metabolic partitioning between the biosynthesis of starch and SGA. Together, these findings illustrate how haplotype diversity, transcriptional specialization, and metabolic partitioning collectively shape phenotypic diversification and functional adaptation in plants [41].

The pervasive structural and regulatory asymmetry observed between the two CSLZ haplotypes provides valuable insight into the evolutionary forces shaping the cannabis genome. The maintenance of such extensive allelic divergence, ranging from presence/absence variation to ASE, suggests that balancing selection may play a key role in preserving genetic diversity in wild populations [42]. By maintaining two functionally distinct haplotypes, wild cannabis potentially benefits from a broader repertoire of biosynthetic and defense capabilities, a genomic architecture that underpins heterosis (hybrid vigor) [43]. For example, divergent alleles of resistance genes or metabolic enzymes allow the plant to respond to a wider array of environmental stresses. This landscape contrasts with domesticated varieties, where intense artificial selection for specific traits (e.g. high THC or fiber yield) has likely driven the fixation of favorable alleles and reduced heterozygosity in target regions [44, 45]. Whereas domestication favors genetic uniformity, our data suggest that wild populations rely on ‘allelic plasticity’, the dynamic regulation of divergent alleles—to cope with fluctuating environments. Thus, the haplotype-resolved asymmetry uncovered here represents an ancestral adaptive strategy that may have been partially diminished or reshaped during the breeding of modern cultivars [45].

Together, these findings establish a coherent framework linking haplotype-specific genome architecture, structural and regulatory divergence, and specialized metabolism in wild cannabis. The allelic perspective uncovered here reveals how genome phasing can illuminate hidden layers of regulation underlying metabolic diversity. Beyond evolutionary insights, this framework provides a foundation for precision breeding and metabolic engineering, which enables targeted manipulation of favorable haplotypes to optimize cannabinoid and terpene production for medicinal and industrial applications.

Materials and methods

Plant material and growth conditions

The CSLZ plants used for PacBio sequencing, RNA-seq, and *de novo* genome assembly were propagated and grown from cuttings. Cuttings from mother plants collected from Linzhi County, Tibet, with an average altitude of 3100 m (94.397312° E, 29.617220° N). Cuttings were rooted in the controlled chamber at $24 \pm 2^\circ\text{C}$ with a light–dark cycle of 18/6 h, at 60% relative humidity (RH) under white light at an intensity of $140 \mu\text{mol m}^{-2} \text{s}^{-1}$. After rooting, plantlets were transplanted to a 2.4 l pot growing under the same cultivation conditions. Each vegetative plant was placed in an individual pot and watered daily with 500 ml of a nutrient solution with an N-P-K ratio of 3:1:1; pH 6.8 ± 0.1 . To initiate flowering, the 6-month-old plants were transferred to a peat-perlite-vermiculite medium [5:1:1 (v/v/v)] in 7.6 l pots and grown at $24 \pm 2^\circ\text{C}$ with a light/dark cycle of 12/12 h, at 60% RH under white light at an intensity of $140 \mu\text{mol m}^{-2} \text{s}^{-1}$ in the controlled chamber. Each plant was watered daily with 500 ml of a nutrient solution with an N-P-K ratio of 2:1:5; pH 6.8 ± 0.1 . Genomic DNA was extracted from fresh young leaves for genome sequencing; mixed samples of aboveground tissues (stems, buds, leaves, flowers), and individual samples of female flowers and glandular trichomes were collected for ISO-seq for genome assembly and gene annotation. The roots, stem xylem, stem phloem, leaves, glandular trichomes, and female flowers without trichomes were collected for RNA-seq to compare gene expression differences in different tissues. Female flowers (1–9 weeks after flowering) and glandular trichomes (1–10 weeks after flowering) were collected for RNA-seq to compare gene expression dynamics during their development.

DNA sequencing

Genome DNA (gDNA) was extracted from young leaf tissues of CSLZ samples using the CTAB method. DNA quality and quantity were assessed using three approaches: (i) NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific), (ii) agarose gel electrophoresis, and (iii) Qubit fluorometer (Invitrogen). The extracted DNA was then purified with AMPure PB beads (PacBio, 100-265-900), yielding high-quality gDNA with concentrations $\geq 100 \text{ ng}/\mu\text{l}$ and total amounts $\geq 10 \mu\text{g}$ for subsequent library construction.

Total RNA was extracted from mixed aboveground tissues and individual samples of female flowers and glandular trichomes using the TRIzol method. RNA integrity, concentration, and quantity were assessed using Thermo Fischer Scientific and Agilent 2100 Bioanalyzer (Agilent Technologies), resulting in high-quality RNA with concentrations $\geq 300 \text{ ng}/\mu\text{l}$ and total amounts $\geq 2 \mu\text{g}$ for downstream library preparation.

For whole-genome resequencing, DNA libraries were constructed with an insert size of 300–400 bp and sequenced on the Illumina NovaSeq 6000 platform. Low-quality reads, adapter contamination, and PCR duplicates were removed, yielding $\sim 70.10 \text{ Gb}$ of clean data ($\sim 93\times$ coverage) from a total raw sequencing volume of $\sim 71.92 \text{ Gb}$.

High-molecular-weight genomic DNA for PacBio sequencing was extracted from fresh young leaf tissues of a single plant using the DNeasy Plant Mini Kit. Fragments $> 15 \text{ kb}$ were selected using the BluePippin system, and SMRTbell libraries (30–50 kb) were constructed following PacBio protocols. Sequencing on the PacBio

Sequel II platform generated 71.04 Gb of data ($\sim 94.7\times$ coverage). Additionally, a separate PacBio SMRTbell library was prepared using the SMRTbell® Express Template Prep Kit 2.0 (PacBio, PN 101-853-100) and sequenced on a single Sequel II cell, yielding 29.5 Gb ($\sim 39\times$ coverage).

Hi-C libraries were prepared from young leaf tissues of the same individual using an *in situ* protocol. Briefly, chromatin was cross-linked with formaldehyde, lysed, and digested with the six-base restriction enzyme HindIII. The resulting DNA ends were filled, biotin-labeled, and ligated. Cross-links were reversed, and DNA was sheared by sonication. Biotin-enriched fragments were captured to construct Hi-C libraries, which were subsequently sequenced on the Illumina NovaSeq 6000 platform.

RNA sequencing and experimental design

To ensure statistical robustness, the experiment was conducted using a randomized complete block design. Three independent biological replicates were collected for each tissue type (female flowers, glandular trichomes, and other tissues) at the specified developmental stages. No sample pooling was performed to preserve biological variance.

After passing quality assessment of independent RNA samples from female flowers and glandular hairs, eukaryotic mRNA was enriched using Oligo(dT)-conjugated magnetic beads. The mRNA was then fragmented, and first-strand cDNA was synthesized using hexamer random primers. Second-strand cDNA synthesis was performed by adding buffer, dNTPs, DNA polymerase I, and RNase H. The resulting double-stranded cDNA was purified using AMPure XP beads. Purified cDNA was subsequently end-repaired, A-tailed, and ligated to sequencing adapters, followed by fragment size selection with AMPure XP beads. PCR amplification was performed, and the amplified products were purified with AMPure XP beads to generate the final RNA-seq library. Library quantification was first performed using Qubit 2.0, and fragment size distribution was assessed with the Agilent 2100 Bioanalyzer. After confirming expected insert sizes, library concentration was accurately determined by qPCR to ensure sequencing quality. Paired-end sequencing was then performed on the Illumina HiSeq 2500 platform.

For PacBio Iso-Seq, mixed RNA samples from aboveground tissues (stem, bud, leaf, flower) were used to construct SMRTbell libraries with the SMRTbell® Express Template Prep Kit 2.0 (Pacific Biosciences, PN 101-853-100). The procedure was as follows: (i) reverse transcription using Oligo(dT)-linked primers to synthesize first-strand cDNA; (ii) addition of complementary primers to the template-switching oligo for template conversion; (iii) incorporation of fusion primers with barcodes to amplify cDNA products; (iv) purification of cDNA using ProNex Beads and quantification with Qubit fluorometer (Invitrogen) (v) SMRTbell library construction using the 2.0 kit; and (vi) removal of single-stranded overhangs, DNA damage repair, end repair, A-tailing, and adapter ligation. Libraries were purified with $0.45\times$ AMPure PB beads, size-selected for 1–10 kb fragments, evenly loaded onto SMRT Cells, and sequenced on the PacBio Sequel II/IIe system (Pacific Biosciences, CA, USA).

Genome size estimation

The genome size and heterozygosity of the target plants were estimated using k-mer analysis. Illumina sequencing data quality was

assessed with FastQC (v0.11.9, <https://github.com/s-andrews/FastQC>). The frequency distribution of 17-mer segments from the short reads was calculated using Jellyfish (v2.3.0) [46], and genome size and heterozygosity were subsequently estimated with GenomeScope (v1.0.0) [47].

To experimentally validate the genome size, flow cytometry analysis was performed using fresh young leaves of CSLZ and *O. sativa* ssp. *japonica* (~380 Mb) as the internal reference standard. Nuclei were extracted and stained using the CyStain UV Precise P kit (Sysmex Partec, Germany) following the manufacturer's instructions. Briefly, leaf tissues of the sample and the standard were co-chopped in Nuclei Extraction Buffer to release nuclei, filtered through a nylon mesh, and stained with the Staining Buffer containing 4',6-diamidino-2-phenylindole. The relative fluorescence intensity of the nuclei was measured using a Partec CyFlow Space flow cytometer (Sysmex Partec, Germany) equipped with a UV LED/lamp. The genome size of CSLZ was calculated based on the ratio of the G1 peak mean fluorescence intensity of the sample to that of the internal standard.

De novo assembly and quality assessment of haplotype genomes

A hybrid assembly pipeline was implemented to construct the phased diploid genome (Fig. 1B). To maximize the structural utility of deep long-read coverage, PacBio CLR reads (~94.72×) were first assembled using FALCON (v0.3.0) [48], polished with Arrow and Filon [49], and scaffolded with Mbol Hi-C data using the Juicebox and 3D-DNA pipeline [50] to generate the primary haploid reference (Haploid Genome V1). Simultaneously, PacBio CCS (~39×) and HindIII Hi-C data were assembled using hifiasm (v0.16.1-r375). Specifically, hifiasm integrates Hi-C linkage information to partition the assembly graph into haplotype-resolved string graphs by distinguishing *cis*-chromosomal contacts from *trans*-interactions. These phased contigs were subsequently anchored to the CLR-based primary reference using RagTag (v2.0.1) [51]. Finally, Hi-C interaction matrices were generated using HiC-Pro (v3.1.0) [52], and the assembly underwent manual curation in Juicebox to correct misassemblies and finalize the order and orientation of the phased chromosomes (Phased Genome V2).

Genome assembly quality and completeness were evaluated using Merqury (v1.3) [53], BUSCO (v5.5.0) [54], and Kraken2 (v2.17.1) [55]. For base-level accuracy and phasing consistency, a k-mer database ($k = 21$) was constructed from PacBio HiFi reads using Meryl (v1.3), and consensus QV scores and k-mer completeness were calculated using Merqury. Gene space completeness was assessed using BUSCO with the eudicots_odb10 lineage dataset. To detect potential contamination, taxonomic classification was performed on both the assembled chromosomes and unmapped reads using Kraken2 against the Standard-8 database (encompassing RefSeq archaea, bacteria, plasmid, viral, human, and UniVec_Core sequences). A strict confidence threshold of 0.005 (–confidence 0.005) was applied to assembly classification.

Repetitive sequence and gene annotation

Gene model prediction was performed by integrating three approaches: transcript-based prediction, homologous protein-based prediction, and *de novo* prediction. Both haplotype

genomes were processed separately using GETA (<https://github.com/chenlianfu/geta>). The workflow included the following steps:

(1) The genome was analyzed with itself using RepeatModeler (v1.0.11) [56] to identify high-frequency repeat sequences and construct a species-specific repeat database. For the detailed analysis of TEs and genomic repetitiveness, we employed the EDTA pipeline (v1.9.6), which integrates multiple *de novo* tools (including LTR_FINDER, LTR_harvest, and TIR-Learner) to produce a high-quality, non-redundant TE library. To capture specific repeat classes with higher sensitivity, we additionally employed MITE-Hunter [57] to identify miniature inverted-repeat TEs (MITES) and TRF [58] to annotate tandem repeats (parameters: 2 7 7 80 10 50 500 -f -d -m). The *de novo* libraries from RepeatModeler, MITE-Hunter, and TRF were merged to create a unified database. The genome was then aligned to this comprehensive repeat database using RepeatMasker (v1.331) [56] to identify repetitive elements. The combined results ensured a robust annotation of genomic repeats, which were subsequently masked to facilitate accurate gene prediction.

(2) Raw transcriptome data were filtered with Trimmomatic (v0.33) [59] to remove low-quality reads and adapter sequences. Filtered reads were aligned to the genome using HISAT2 (v2.2.1) [60]. Coverage thresholds were applied based on sequencing depth to obtain reliable transcripts and intron information. High-confidence transcripts were analyzed with TransDecoder (v5.50, <http://github.com/TransDecoder/TransDecoder>) to generate a transcriptome-based gene model.

(3) A high-quality homologous protein database was constructed using annotated sequences from *P. andersonii*, *T. orientalis*, hops (*H. lupulus*), *A. thaliana*, and CSLZ haploid protein sequences. Gene models from the transcript-based step were aligned to this database using GeneWise (<https://www.ebi.ac.uk/birney/wise2/>) to identify homologous genes.

(4) AUGUSTUS was used to perform *de novo* gene prediction, informed by intron, exon, start codon, and stop codon information obtained from transcript-based and homology-based predictions. Finally, results from all three methods were integrated to produce accurate, high-quality gene models for each haplotype genome.

Genome complexity analysis to visualize the landscape of sequence repetitiveness and distinguish between local and long-range complexity, we employed the alignment-free method AlcoR [61]. Complexity maps were generated for each haplotype using two configurations: a 'local' model (order 5, window 10) to detect short-range repetitiveness, and a 'distant' model (order 12, window 50) to capture broader repetitive structures such as TEs and segmental duplications. The resulting complexity scores were plotted along the chromosomes to identify regions of high and low complexity.

Gene family and genome evolution analysis

Haplotype genome annotations from nine reference species, *O. sativa*, *S. lycopersicum*, *V. vinifera*, *A. thaliana*, *M. notabilis*, *F. macrocarpa*, *P. andersonii*, *T. orientale*, and *H. lupulus* were collected, together with the *Cannabis sativa* (CSLZ) haploid genome annotation generated in this study. In total, 10 species were included for comparative genomic analyses. The annotated datasets for *O. sativa*, *S. lycopersicum*, and *A. thaliana* were

obtained from the Phytozome database (V7.0, ITAG 4.0, and TAIR10, respectively). The annotation of *V. vinifera* was retrieved from EnsemblPlants (PN40024.v4). Genome annotations of *T. orientale*, *P. andersonii*, and *M. notabilis* were downloaded from NCBI (GCA_002914845.1, GCA_002914805.1, and GCF_000414095.1, respectively). The *H. lupulus* annotation was obtained from HopBase [62], and the *F. macrocarpa* annotation was derived from a previously published study (2020) [63].

For each species, the longest transcript per gene was extracted using in-house Python scripts. Orthogroups were identified using OrthoFinder v2.5.4 [64], and 606 single-copy orthologs were extracted to construct a phylogenetic tree using the maximum likelihood method implemented in IQ-TREE v2.2.0.3 [65]. The best-fit substitution model (Q.plant + F + R5) was selected automatically, and branch supports were evaluated with 1500 ultrafast bootstrap replicates (–bb 1500). Four-fold degenerate transversion sites were extracted using ParaAT v2.0 [66], and species divergence times were estimated with the MCMCtree program in PAML v4.9j [67], calibrated using divergence time data from the TimeTree database [68]. Gene family expansion and contraction were analyzed with CAFE5 [69], and phylogenetic trees were visualized using EvolView [70]. KEGG enrichment of expanded gene families was conducted using ClusterProfiler [71].

Speciation and WGD events were inferred based on the distribution of synonymous substitution rates (Ks) among orthologous gene pairs (*C. sativa* vs. *H. lupulus*, *F. macrocarpa*, and *V. vinifera*) and paralogous pairs within the *C. sativa* genome. Syntenic blocks were identified using MCScan (Python version) [72], and gene pairs within these blocks were classified as orthologs or paralogs. Ks values were calculated using the WGD pipeline v1.1 [73].

Structural variant identification

Pairwise genome alignments were performed between the *C. sativa* CSLZ_hapB assemblies and the CSLZ_hapA reference using Minimap2 [74]. SVs were identified from the pairwise alignments using pafTools, a utility included in the Minimap2 package. The alignment results were visualized as genome-wide dot plots using dotPlotly (<https://github.com/tpoorten/dotPlotly/>) and Plotsr (<https://github.com/schneebergerlab/plotsr>). Detected SVs were categorized based on their type and size, and only high-confidence variants were retained for downstream analysis.

Identification of allelic variations and allele-specific expressed genes.

Alleles were defined using a combination of synteny-based and coordinate-based approaches. Syntenic blocks between the two haplotypes were identified using MCScanX [72], and gene pairs with high sequence similarity within each block were designated as allele A and allele B. Gene models sharing identical coding sequences were merged and considered a single allele. For genes not located within syntenic blocks, sequence alignment against the monoploid assembly was performed using GMAP [75], and two genes were considered potential alleles if their genomic coordinates overlapped by more than 50%.

To minimize reference mapping bias, RNA-seq reads were quality-trimmed using Trimmomatic [59], and mapped to the combined diploid reference assembly (containing both Haplotype A and Haplotype B sequences) using Bowtie [76]. This competitive

mapping strategy ensures that reads originating from heterozygous sites are preferentially aligned to their true haplotype origin based on sequence divergence (SNPs and Indels) and optimal alignment scores, thereby minimizing cross-mapping between homoeologs. Transcript abundance was quantified using RSEM [77], which further ensures quantification reliability by handling read mapping uncertainty through an expectation–maximization algorithm. Differential expression analysis was performed using edgeR v4.0.16 [78]. Genes with low expression (read counts <10 in all samples) were filtered out prior to analysis. For ASE analysis, only sites with a minimum coverage of 20 reads and at least 5 reads supporting each allele were retained to ensure statistical power. ASE was identified using a Binomial test followed by FDR correction. Genes were considered ASEs if they met the following criteria: (i) Adjusted *P*-value (FDR) < 0.05 and (ii) $\log_2|\text{Fold Change}|$ (ratio of Allele A/Alelle B expression) > 1.

Functional enrichment analyses of ASE genes were conducted using clusterProfiler [71] for GO and KEGG pathways. Heatmaps of gene expression were generated using pheatmap (<https://github.com/raivokolde/pheatmap>), and PCA was performed using PCAtools (<https://github.com/kevinblighe/PCATools>). Venn diagrams of gene set overlaps were generated with E Venn [79]. Visualization of TE insertions between haplotypes was carried out using IGV [80], and *cis*-regulatory elements in promoter regions were identified based on the PlantCARE database [81].

Transient expression in *N. benthamiana* and metabolite detection

For transient expression, recombinant pEAQ (a virus-based plant expression vector) plasmids were introduced into *Agrobacterium tumefaciens* strain GV3101 via transformation. Positive colonies verified by colony PCR were cultured in Luria-Bertani (LB) medium containing appropriate antibiotics until the OD₆₀₀ reached 0.5–0.7. The bacterial cells were harvested and resuspended in infiltration buffer to an OD₆₀₀ of 0.6–0.7. Strains carrying different gene constructs were mixed in equal proportions and incubated with 200 μ M acetosyringone in the dark for 2–3 h before infiltration into leaves of 3–4-week-old *N. benthamiana* plants. Following infiltration, plants were incubated in the dark for 24 h and then transferred to a greenhouse under normal growth conditions for an additional 3 days. Substrate solutions were infiltrated on the fourth day, and leaf tissues were harvested after another 24 h of dark incubation, lyophilized, and stored for subsequent metabolite analysis.

For metabolite detection, ~40 mg of lyophilized leaf powder was saponified and extracted with 2 ml of 50% ethanol containing 20% KOH at 100°C for 10 min, with 5 α -cholestane used as an internal standard. Metabolites were extracted twice with *n*-hexane, and the combined organic phases were concentrated and derivatized with pyridine and N,O-bis(trimethylsilyl)trifluoroacetamide (BSTFA) containing 1% trimethylchlorosilane (TMCS) at 70°C for 1 h. The derivatized samples were then evaporated to dryness, redissolved in chloroform, and filtered prior to Gas chromatography-mass spectrometry (GC–MS) analysis. GC–MS was using an Agilent 8890 Gas Chromatograph coupled to an Agilent 7000E Triple Quadrupole Mass Spectrometer. Metabolites were identified by comparison with the NIST14 mass spectral library and relatively quantified by normalizing peak areas to the internal standard.

Acknowledgements

The work was supported by the National Natural Science Foundation of China (32561143033, 32470281); the Natural Science Foundation for Distinguished Young Scholars of Fujian Province (2025J010027). This work was also supported by Beijing PARAT-ERA Technology Co., LTD (<http://cloud.paratera.com>) for providing high-performance and AI computing resources.

Author contributions

Y.L. conceived the study; S.C., Z.Z., and Y.L. designed the experiments; S.C., Z.Z., M.S., C.L., Y.Z., X.L., M.H., D.A., and Y.L. performed the experiments. S.C., Z.Z., S.Y., K.T., N.B., and Y.L. analyzed the data. S.C., Z.Z., and Y.L. wrote the manuscript with input from all authors.

Data availability

The data that support the findings of this study have been deposited in both the China National Center for Bioinformation (CNCB) and the National Center for Biotechnology Information (NCBI).

Raw sequencing data: The raw resequencing data generated in this study have been deposited in the CNCB Genome Sequence Archive (GSA) under the accession number CRA033206, and in the NCBI BioProject database under the accession number PRJNA1417719.

Genome assembly data: The haplotype-resolved CSLZ genome assemblies have been deposited in the CNCB Genome Warehouse (GWH) under the accession numbers GWHGZJI000000000.1 (HapA) and GWHGZJJ000000000.1 (HapB). Additionally, the assemblies have been submitted to NCBI GenBank under the accession numbers JBVFIW000000000 (HapA) and JBVFIW000000000 (HapB).

Custom Code: The custom scripts and pipelines used in this study are publicly available on GitHub at https://github.com/caiisen/LZ_phased_genome.

Conflicts of interest statement

The authors declare no competing interests.

References

- Khan N. Exploring plant resilience through secondary metabolite profiling: advances in stress response and crop improvement. *Plant Cell Environ.* 2025;48:4823–37
- Liu Y, Li S, Shen M. *et al.* Integration of transcriptome and metabolome provides insights into metabolites and pathways associated with antiproliferative activity of cannabis flower extracts. *Ind Crop Prod.* 2025;223:120239
- Sun X, Jiao C, Schwaninger H. *et al.* Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nat Genet.* 2020;52:1423–32
- Zhou Q, Tang D, Huang W. *et al.* Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat Genet.* 2020;52:1018–23
- Jang M-J, Cho HJ, Park Y-S. *et al.* Haplotype-resolved genome assembly and resequencing analysis provide insights into genome evolution and allelic imbalance in *Pinus densiflora*. *Nat Genet.* 2024;56:2551–61
- Guardo MD, Moretto M, Moser M. *et al.* The haplotype-resolved reference genome of lemon (*Citrus Limon* L. Burm f.). *Tree Genet Genomes.* 2021;17:46
- Nath O, Fletcher SJ, Hayward A. *et al.* A haplotype resolved chromosomal level avocado genome allows analysis of novel avocado genes. *Hortic Res.* 2022;9:uhac157
- Mao J, Wang Y, Wang B. *et al.* High-quality haplotype-resolved genome assembly of cultivated octoploid strawberry. *Hortic Res.* 2023;10:uhad002
- Sun H, Jiao W-B, Krause K. *et al.* Chromosome-scale and haplotype-resolved genome assembly of a tetraploid potato cultivar. *Nat Genet.* 2022;54:342–8
- Zhang X, Chen S, Shi L. *et al.* Haplotype-resolved genome assembly provides insights into evolutionary history of the tea plant *Camellia sinensis*. *Nat Genet.* 2021b;53:1250–9
- Li C, Yuan Y, Nie Z. *et al.* The haplotype-resolved telomere-to-telomere genome and OMICS analyses reveal genetic responses to tapping in rubber tree. *Nat Commun.* 2025;16:6255
- Xia Z, Fan W, Liu D. *et al.* Haplotype-resolved chromosomal-level genome assembly reveals regulatory variations in mulberry fruit anthocyanin content. *Hortic Res.* 2024;11:uhae120
- Liao B, Shen X, Xiang L. *et al.* Allele-aware chromosome-level genome assembly of *Artemisia annua* reveals the correlation between ADS expansion and artemisinin yield. *Mol Plant.* 2022;15:1310–28
- Zhang H, He Q, Xing L. *et al.* The haplotype-resolved genome assembly of autotetraploid rhubarb *Rheum officinale* provides insights into its genome evolution and massive accumulation of anthraquinones. *Plant Commun.* 2024;5:100677
- Li H-L, Wu L, Dong Z. *et al.* Haplotype-resolved genome of diploid ginger (*Zingiber officinale*) and its unique gingerol biosynthetic pathway. *Hortic Res.* 2021;8:189
- Lynch RC, Padgitt-Cobb LK, Garfinkel AR. *et al.* Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome. *Nature.* 2025;643:1001–10
- Sánchez-Pérez R, Neilson EH. The case for sporadic cyanogenic glycoside evolution in plants. *Curr Opin Plant Biol.* 2024;81:102608
- Boncan DAT, Tsang SSK, Li C. *et al.* Terpenes and terpenoids in plants: interactions with environment and insects. *Int J Mol Sci.* 2020;21:7382
- Hanuš LO, Hod Y. Terpenes/terpenoids in *cannabis*: are they important? *Med Cannabis Cannabinoids.* 2020;3:25–60
- Sharma S, Chaurasia S, Dinday S. *et al.* High-level biosynthesis of enantiopure germacrene D in yeast. *Appl Microbiol Biotechnol.* 2024;108:50
- Kushiro T, Shibuya M, Ebizuka Y. β -Amyrin synthase: cloning of oxidosqualene cyclase that catalyzes the formation of the most popular triterpene among higher plants. *Eur J Biochem.* 1998;256:238–44
- Ferreira MLF, Serra P, Casati P. Recent advances on the roles of flavonoids as plant protective molecules after UV and high light exposure. *Physiol Plant.* 2021;173:736–49
- Mierziak J, Kostyn K, Kulma A. Flavonoids as important molecules of plant interactions with the environment. *Molecules.* 2014;19:16240–65

24. Chervin C, TTira-Umphon A, Chatelet P. *et al.* Ethylene and other stimuli affect expression of the UDP glucose-flavonoid 3-O-glucosyltransferase in a non-climacteric fruit. *Vitis*. 2015;48:11
25. Zhang Y, Guo S, Zhang Z. *et al.* Functional characterization of anthocyanin biosynthesis-related Dihydroflavonol 4-reductase (DFR) genes in blueberries (*Vaccinium corymbosum*). *Plants*. 2025;14:1449
26. Huang X, Chen W, Zhao Y. *et al.* Deep learning-based quantification and transcriptomic profiling reveal a methyl jasmonate-mediated glandular trichome formation pathway in *Cannabis sativa*. *Plant J*. 2024;118:1155–73
27. Liu Y, Zhu P, Cai S. *et al.* Three novel transcription factors involved in cannabinoid biosynthesis in *Cannabis sativa* L. *Plant Mol Biol*. 2021;106:49–65
28. Sampedro J, Cosgrove DJ. The expansin superfamily. *Genome Biol*. 2005;6:242
29. Benoit M, Drost H-G, Catoni M. *et al.* Environmental and epigenetic regulation of rider retrotransposons in tomato. *PLoS Genet*. 2019;15:e1008370
30. Tian X, Wang R, Liu Z. *et al.* Widespread impact of transposable elements on the evolution of post-transcriptional regulation in the cotton genus *Gossypium*. *Genome Biol*. 2025;26:60
31. Zheng P, Ko C, Ou J. *et al.* Transposable elements drive evolution and perturb gene expression in *Brassica rapa* and *B. oleracea*. *Plant J*. 2025;123:e70452
32. Hasing T, Tang H, Brym M. *et al.* A phased *Vanilla planifolia* genome enables genetic improvement of flavour and production. *Nat Food*. 2020;1:811–9
33. Deneweth J, Van De Peer Y, Vermeirssen V. Nearby transposable elements impact plant stress gene regulatory networks: a meta-analysis in *A. thaliana* and *S. Lycopersicum*. *BMC Genomics*. 2022;23:18
34. Fantini E, Daddiego L, Facella P. *et al.* MYC2 influences rubber and sesquiterpene lactones synthesis in *Taraxacum* species. *Planta*. 2025;262:5
35. Kagale S, Links MG, Rozwadowski K. Genome-wide analysis of ethylene-responsive element binding factor-associated amphiphilic repression motif-containing transcriptional regulators in *Arabidopsis*. *Plant Physiol*. 2010;152:1109–34
36. Licausi F, Ohme-Takagi M, Perata P. APETALA 2/ethylene responsive factor (AP 2/ERF) transcription factors: mediators of stress responses and developmental programs. *New Phytol*. 2013;199:639–49
37. Liu M, Gomes BL, Mila I. *et al.* Comprehensive profiling of ethylene response factor expression identifies ripening-associated ERF genes and their link to key regulators of fruit ripening in tomato. *Plant Physiol*. 2016;170:1732–44
38. Payyavula RS, Navarre DA, Kuhl JC. *et al.* Differential effects of environment on potato phenylpropanoid and carotenoid expression. *BMC Plant Biol*. 2012;12:39
39. Dong D, Yang Z, Ma Y. *et al.* Expression of a hydroxycinnamoyl-CoA shikimate/Quinate hydroxycinnamoyl transferase 4 gene from *Zoysia japonica* (ZjHCT4) causes excessive elongation and lignin composition changes in *Agrostis stolonifera*. *Int J Mol Sci*. 2022;23:9500
40. Zhan W, Cui L, Yang S. *et al.* Natural variations of heterosis-related allele-specific expression genes in promoter regions lead to allele-specific expression in maize. *BMC Genomics*. 2024;25:476
41. Li H, Brouwer M, Pup ED. *et al.* Allelic variation in the autotetraploid potato: genes involved in starch and steroidal glycoalkaloid metabolism as a case study. *BMC Genomics*. 2024;25:274
42. Gaut BS, Seymour DK, Liu Q. *et al.* Demography and its effects on genomic variation in crop domestication. *Nat Plants*. 2018;4:512–20
43. Zhang C, Yang Z, Tang D. *et al.* Genome design of hybrid potato. *Cell*. 2021a;184:3873–3883.e12
44. Alonge M, Wang X, Benoit M. *et al.* Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell*. 2020;182:145–161.e23
45. Ren G, Zhang X, Li Y. *et al.* Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*. *Sci Adv*. 2021;7:eabg2286
46. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics*. 2011;27:764–70
47. Vurture GW, Sedlazeck FJ, Nattestad M. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*. 2017;33:2202–4
48. Chin C-S, Peluso P, Sedlazeck FJ. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods*. 2016;13:1050–4
49. Walker BJ, Abeel T, Shea T. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014;9:e112963
50. Dudchenko O, Batra SS, Omer AD. *et al.* De novo assembly of the *Aedes aegypti* genome using hi-C yields chromosome-length scaffolds. *Science*. 2017;356:92–5
51. Alonge M, Lebeigle L, Kirsche M. *et al.* Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biol*. 2022;23:258
52. Servant N, Varoquaux N, Lajoie BR. *et al.* HiC-pro: an optimized and flexible pipeline for hi-C data processing. *Genome Biol*. 2015;16:259
53. Rhie A, Walenz BP, Koren S. *et al.* Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21:245
54. Tegenfeldt F, Kuznetsov D, Manni M. *et al.* OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic Acids Res*. 2025;53:D516–22
55. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with kraken 2. *Genome Biol*. 2019;20:257
56. Flynn JM, Hubley R, Goubert C. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci*. 2020;117:9451–7
57. Han Y, Wessler SR. MITE-hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res*. 2010;38:e199–9
58. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27:573–80
59. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*. 2014;30:2114–20
60. Kim D, Paggi JM, Park C. *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol*. 2019;37:907–15

61. Silva JM, Qi W, Pinho AJ. *et al.* AlcoR: alignment-free simulation, mapping, and visualization of low-complexity regions in biological data. *GigaScience*. 2022;12:giad101
62. Hill ST, Sudarsanam R, Henning J. *et al.* HopBase: a unified resource for *Humulus* genomics. *Database*. 2017;2017:bax009
63. Zhang X, Wang G, Zhang S. *et al.* Genomes of the banyan tree and pollinator wasp provide insights into fig-wasp coevolution. *Cell*. 2020;183:875–889.e17
64. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol*. 2019;20:238
65. Minh BQ, Schmidt HA, Chernomor O. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol*. 2020;37:1530–4
66. Zhang Z, Xiao J, Wu J. *et al.* ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem Biophys Res Commun*. 2012;419:779–81
67. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics*. 1997;13:555–6
68. Kumar S, Suleski M, Craig JM. *et al.* TimeTree 5: an expanded resource for species divergence times. *Mol Biol Evol*. 2022;39:msac174
69. Mendes FK, Vanderpool D, Fulton B. *et al.* CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics*. 2021;36:5516–8
70. Subramanian B, Gao S, Lercher MJ. *et al.* Evolview v3: a web-server for visualization, annotation, and management of phylogenetic trees. *Nucleic Acids Res*. 2019;47:W270–5
71. Yu G, Wang L-G, Han Y. *et al.* clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16:284–7
72. Tang H, Krishnakumar V, Zeng X. *et al.* JCVI: a versatile toolkit for comparative genomics analysis. *iMeta*. 2024;3:e211
73. Chen H, Zwaenepoel A. Inference of Ancient Polyploidy from Genomic Data. In: Van De Peer Y, ed. *Polyploidy*. Vol. 2545. Springer US: New York, NY, 2023,3–18
74. Li H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics*. 2021;37:4572–4
75. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 2005;21:1859–75
76. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods*. 2012;9:357–9
77. Grabherr MG, Haas BJ, Yassour M. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29:644–52
78. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*. 2010;26:139–40
79. Yang M, Chen T, Liu Y. *et al.* Visualizing set relationships: EVenN's comprehensive approach to Venn diagrams. *iMeta*. 2024;3:e184
80. Robinson JT, Thorvaldsdottir H, Turner D. *et al.* Igv.js: an embeddable JavaScript implementation of the integrative genomics viewer (IGV). *Bioinformatics*. 2023;39:btac830
81. Lescot M. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res*. 2002;30:325–7