

Genomic variation in 271 diverse accessions of *Cannabis sativa* L. in China

Baishi Wang · Shan Gao · Xue Wang · Quanzheng Yun · Hao Nie · Ning Li · Yonggang Li · Deri Feng · Linsen Li · Tianye Li · Xu Lu · Jingjing Liu · Xuge Wei · Yanhong Mu · Qiqi Wu · Jinyuan Hu · Fangru Liu · Rui Feng · Jianyu Yin · Qikun Zhang · Ying Zhang · Fu Ren

Received: 27 March 2026 / Accepted: 2 September 2026

© The Author(s) 2026

Abstract

Cannabis sativa L., a multipurpose plant with historical significance in food, textile, and medicinal applications, exhibits undercharacterized genomic diversity within Chinese germplasm despite its widespread cultivation. To address this knowledge gap, we performed whole-genome resequencing of 271 geographically diverse accessions spanning 29 populations across 12 Chinese provinces, achieving an average depth of 7.08×. Comparative genomic analysis against the reference assembly revealed 2,699,120 single nucleotide polymorphisms (SNPs) and 42,201 insertion-deletion mutations (InDels), providing the most comprehensive variant catalog for Chinese cannabis to date. Phylogenetic reconstruction delineated five distinct genetic clusters, with principal component analysis (PCA) and population structure results demonstrating strong concordance between genetic relatedness and geographical origins. Genome-wide selective sweep scans revealed distinct signatures of selection across populations, identifying 3,299 candidate genes underlying north-south population differentiation in Xinjiang, 3,215 genes at the Xinjiang-Xizang divergence interface, and 3,404 genes distinguishing southwestern versus northeastern domestication trajectories. Genome-wide association studies (GWAS) identified CBDAS as a robust candidate gene associated with CBD content, with significant signals localized on chromosome 1. For THC, while several suggestive loci were identified, no definitive candidate gene could be confidently assigned based on functional annotations, highlighting the complex genetic architecture of THC biosynthesis. This study establishes a foundational genomic resource for molecular breeding programs while offering critical insights into the genetic basis of chemotypic variation, with dual implications for narcotics control and agricultural improvement.

Keywords *Cannabis sativa* L. · Whole-genome resequencing · Population genomics · Genome-wide association study (GWAS) · Cannabinoid biosynthesis

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Baishi Wang, Shan Gao, Xue Wang and Quanzheng Yun contributed equally to this work.



Introduction

Cannabis sativa L. (Cannabaceae), a diploid ($2n=20$) annual dioecious herb, represents one of humanity's earliest domesticated crops with multipurpose applications spanning nutrition, industry, and therapeutics [1]. While phylogeographic evidence supports an East Asian origin, anthropogenic dispersal has established its pan-continental distribution across Eurasia, with China's primary cultivation centers strategically located in Yunnan, Heilongjiang, Xinjiang, Inner Mongolia, and Henan provinces [2, 3]. This ecological plasticity enables the species to thrive across diverse latitudes (18°N – 55°N), underpinning its tripartite utility: seed-derived oils rich in essential fatty acids, bast fibers with exceptional tensile strength, and specialized metabolites including over 150 cannabinoids [4–6].

China is one of the origin and genetic variation centers of cannabis (*C. sativa* L.), with a history of utilizing the plant for fiber, oil, and medicinal purposes spanning several millennia [7–9]. Archaeological evidence indicates that cannabis was employed for medicinal and shamanistic practices in northwestern China as early as approximately 700 BC (ca. 2,700 years ago) [10–12]. Cultivated cannabis was domesticated from wild ancestors and, through long-term selection, cultivation, and adaptation to diverse regional ecological environments, has evolved into various landraces and geographically distinct varieties [9]. Currently, cannabis is cultivated in all provinces of China except Taiwan, Hainan, Guangdong, and Fujian. Farmers in regions such as Sichuan, Guangxi, and Shanxi commonly grow seed-type cannabis for oil production and edible seeds [13]. Fiber-type cannabis is extensively cultivated in Yunnan province, Qinghai province, and northeastern China [7]. Furthermore, abundant wild cannabis resources are distributed across Xinjiang, Tibet, Inner Mongolia, Yunnan, as well as North and Northeast China [2, 3]. Notably, China's narcotics regulations prohibit cultivation of drug-type cannabis ($\text{THC} \geq 3 \text{ g/kg}$), aligning with international conventions controlling THC-rich variants [14]. Given the widespread cultivation of fiber- and seed-type cannabis across multiple provinces and the existence of abundant wild cannabis resources, accurate chemotype differentiation through genomic approaches is essential for effective regulatory enforcement within China's traditional cannabis cultivation practices.

The origin and domestication of cannabis has long been a subject of debate and a focus of scientific research. Archaeobotanical evidence traces cannabis domestication to the northeastern Qinghai-Tibet Plateau ($\sim 12,000$ BP) [15], though conflicting hypotheses propose Central Asian [16] or low-latitude origins [17]. In a landmark pan-genome analysis, Lynch et al. characterized 193 genomes from 144 *Cannabis* accessions, establishing a foundational framework for comprehensive cannabis genomic research [18]. Notably, the study acknowledges persistent underrepresentation of Asian germplasm in current genomic repositories and conservation collections, revealing a geographically skewed genomic sampling strategy. This data gap is compounded by the probable existence of undomesticated *Cannabis* progenitor species in Asia, whose ecological distribution patterns and conservation priorities remain poorly documented - critical lacunae that continue confounding taxonomic delineation and population stratification within *C. sativa* L. Recent advances in whole-genome resequencing of 110 globally distributed accessions have revealed a prehistoric domestication event in East Asia dating to the Early Neolithic epoch [1]. Nevertheless, crucial research gaps persist regarding: (1) hybridization mechanisms shaping Chinese landrace diversity; (2) multi-phase domestication trajectories across altitudinal gradients; and (3) genomic signatures of artificial selection pressures. These knowledge deficits are primarily attributable to the inadequate inclusion of representative Chinese cultivars in existing germplasm repositories—a major obstacle to unraveling the regionally distinct domestication processes that may have occurred in the putative center of cannabis origin.

Early genetic analyses employed a variety of markers, including random amplified polymorphic DNAs (RAPD) [19], simple sequence repeats (SSRs) [20–22], inter simple sequence repeat (ISSR) [23], and amplified fragment length polymorphisms (AFLP) [24, 25], to assess diversity of *C. sativa*, while chloroplast and rDNA sequencing provided some phylogenetic insights [26–29]. However, the impact of these studies has been constrained by the limited size of the population and a restricted number of genetic markers. Recently, the progress in genome resequencing has enabled scientists to investigate genetic variation within populations of both wild and cultivated

plants with greater precision, such as citrus [30], castor [31], rapeseed [32], sweet cherry [33], wheat [34], grapevine [35], and cauliflower [36]. This facilitates the examination of evolutionary history and enables informed molecular improvements in crop development going forward.

Recent advances in genome resequencing have revolutionized our understanding of *C. sativa*'s genetic architecture, domestication history, and secondary metabolite biosynthesis. Early genomic studies were limited by fragmented assemblies and sparse sampling, but chromosome-scale genomes have enabled precise identification of structural variations and cannabinoid-related loci [37–39]. Large-scale resequencing of 110 globally distributed accessions revealed a single East Asian domestication origin during the Early Neolithic, challenging earlier hypotheses of multiple domestication centers [1]. Resequencing has further elucidated the genetic basis of chemotypic differentiation, particularly in THC and CBD biosynthesis. The *THCAS* and *CBDAS* genes, located in the cannabinoid synthase cluster (reported to be on chromosome 7 in the CS10 assembly), show strong selective sweeps in drug-type cultivars [40]. Despite these advances, critical gaps persist regarding the genomic signatures of ecological adaptation in wild populations and the role of introgressive hybridization in shaping chemotypic diversity.

Despite progress, two critical limitations hinder cannabis genetic research: (1) Insufficient representation of Chinese landraces in global diversity panels, and (2) Incomplete characterization of THC/CBD regulatory networks. To address these gaps, we analyzed 271 re-sequenced accessions (29 populations, 12 Chinese provinces) encompassing wild, landrace, and cultivated genotypes. Through integrated GWAS and phylogenetic approaches, we aim to resolve the genetic architecture of cannabinoid biosynthesis, elucidate ecological adaptation patterns, and establish a molecular framework for germplasm utilization. This study provides the most comprehensive genomic resource for Chinese cannabis to date, offering actionable insights for molecular breeding and narcotics control.

Materials and methods

Sampling, DNA extraction and sequencing

A total of 271 *C. sativa* accessions representing diverse geographical origins were collected in the field from 29 sampling sites across 12 provinces in China, with the number of accessions per site ranging from 6 to 30 (detailed sampling information is provided in Additional File 1: Table S1). In China, wild cannabis is not classified as an endangered plant species, but rather as a controlled plant species. Under the authorization of the Narcotics Control Bureau of the Ministry of Public Security of China, the Institute of Forensic Science of the Ministry of Public Security (with authors Ying Zhang, Shan Gao, and Hao Nie all affiliated with this institute) completed the collection, species identification, physicochemical determination, and DNA extraction. The Institute of Forensic Science possesses legal and comprehensive qualifications and capabilities for the qualitative identification of cannabis. The subsequent genome sequencing and bioinformatics analysis were carried out by Shenyang Medical College, and this part of the work did not involve the plants themselves. The remaining *C. sativa* samples are all stored in the Cannabis Sample Bank of the Institute of Forensic Science of the Ministry of Public Security.

Genomic DNA was isolated from fresh leaf tissues using a modified cetyltrimethylammonium bromide (CTAB) protocol [41]. DNA purity and concentration were quantified via NanoDrop2000 Spectrophotometer (Thermo Fisher Scientific), followed by library construction with Illumina TruSeq Nano DNA kits according to manufacturer specifications. Paired-end sequencing (2×150 bp) was performed on an Illumina NovaSeq 6000 platform (Kaitai Biotechnology, Hangzhou, China). Raw reads underwent stringent quality control: (1) removal of adaptor-contaminated reads; (2) exclusion of reads containing $\geq 10\%$ ambiguous bases (N); (3) trimming of terminal bases with Phred quality scores < 20 ; (4) elimination of reads with $> 40\%$ low-quality bases ($Q20 < 70\%$) using Fastp v0.23.2 [42].

Read mapping, variation calling and annotation

High-quality paired-end reads were aligned to the *C. sativa* reference genome (NCBI accession: GCA_013030365.1) using BWA-MEM v0.7.17 [43] with default parameters. Resultant Sequence Alignment/Map (SAM) files were converted to sorted BAM formats using SAMtools v1.12 [44, 45], retaining only reads with Phred-scaled mapping quality scores ≥ 20 (q20). PCR duplicates were systematically removed via Picard Tools v2.27.4 (Broad Institute) MarkDuplicates module, followed by BAM index generation.

Genetic variants were identified through the GATK v4.2.6.1 HaplotypeCaller pipeline [46] using joint calling mode with parameters `--emit-ref-confidence GVCF` and `--min-base-quality-score 20`. Further functional annotation of single-nucleotide polymorphisms (SNPs) and insertion-deletion variants (InDels) was performed with snpEff v5.1 [47] using a self-built database. A stringent filtering pipeline was applied to ensure variant reliability: Minor allele frequency (MAF): SNPs with $MAF < 0.05$ were excluded; Mapping quality: Sites with $MapQ < 20$ were discarded; Missing data: Variants showing $> 50\%$ missing genotypes across samples were removed.

Genetic diversity and population structure analysis

Phylogenetic reconstruction was performed through a three-stage workflow: (1) Pairwise genetic distance matrices were computed using PLINK v1.9 [48] with identity-by-state (IBS) metrics; (2) Neighbor-joining trees were constructed via PHYLIP v3.697 [49] with 1,000 bootstrap replicates; (3) Final visualization was implemented in Interactive Tree Of Life (iTOL v6.7.4) [50] using circular layout optimization. Population stratification was analyzed through complementary approaches: (1) Bayesian clustering: ADMIXTURE v1.3.0 [51] was executed with $K = 2-10$ ancestral populations, employing 10-fold cross-validation and convergence criteria of $\Delta \ln(L) < 0.001$ over 10,000 iterations; the final K was primarily selected based on the lowest cross-validation error, which also coincided with a peak in the ΔK statistic (Evanno method); (2) Principal component analysis (PCA): Genetic relationship matrices (GRM) were generated in GCTA v1.94.1 [52] using the command: `gcta64 --bfile [prefix] --autosome-num 10 --make-grm --out [grm_prefix]`. Top eigenvectors (PC1-PC2) were extracted following variance standardization. To avoid the confounding effects of sex-linked markers in this dioecious species, all variants located on sex chromosomes were excluded from the population structure and nucleotide diversity (π) analyses. Genome-wide diversity metrics were quantified using: (1) Nucleotide diversity (π): Calculated as the average pairwise differences per site; (2) Population differentiation: Estimated via Weir-Cockerham method Both statistics were computed in 100-kb sliding windows with 10-kb steps using VCFtools v0.1.14 [53], excluding windows containing < 10 SNPs.

Linkage Disequilibrium (LD) decay profiling

To investigate genome-wide linkage disequilibrium (LD) architecture, we implemented a standardized analytical workflow using Haploview v4.2 [54] with the following analytical pipeline: (1) Allelic correlation analysis - systematic calculation of pairwise squared correlation coefficients (r^2) for all biallelic SNPs within contiguous 1-Mb genomic intervals; (2) Multiple testing adjustment - application of Bonferroni-corrected significance threshold ($\alpha_{adj} = 0.05/[n_{\text{SNP}}(n_{\text{SNP}}-1)/2]$) to control family-wise error rate; (3) Haplotype block delineation - identification of LD structures using adaptive window parameters (sliding window approach with maximum span: 1,000 kb, unlimited SNP count, and minimum r^2 threshold: 0). For spatiotemporal pattern characterization, we generated LD decay profiles by modeling r^2 as a function of inter-marker physical distance (1–1,000 kb), visualized through nonlinear regression curves constructed with ggplot2 v3.4.2 (<https://ggplot2.tidyverse.org/>).

Allelic-drift analysis

To investigate population divergence and admixture patterns, we employed TREEMIX v1.13 [55] to reconstruct phylogenetic relationships among populations through covariance matrix analysis of allele frequencies. The modeling process involved initial construction of a maximum-likelihood bifurcating tree followed by iterative addition of migration edges to account for historical gene flow events. Resultant admixture signals were subsequently geospatially visualized through customized mapping using the ggplot2 package in R (<https://ggplot2.tidyverse.org/>), with migration vectors superimposed on a digital elevation model of China. Migration edges (m) were tested from 0 to 5 using TREEMIX. The final optimal number of migration events (m=3) was selected based on the plateau of log-likelihood increase and the saturation of standard errors.

Selective sweep analyses

To characterize genomic diversity and population differentiation, we conducted multilayered evolutionary analyses through the following computational framework: (1) Nucleotide diversity (π) and pairwise fixation index (F_{st}) were computed genome-wide using VCFtools v0.1.14 [53] with 100-kb sliding windows and 10-kb steps across the *C. sativa* reference genome; (2) Selective sweep detection was performed using the cross-population composite likelihood ratio (XP-CLR) method [56] with optimized parameters: window-specific weight (w_1)=0.005 incorporating 200 SNPs per window, maximum physical span=10 Mb, and haplotype frequency threshold (p_0)=0.95. Chromosome-level XP-CLR scores were subsequently converted to 100-kb sliding windows (10-kb step size) through weighted averaging. Significant selection signals were identified using an empirical genome-wide threshold corresponding to the top 5% of windowed XP-CLR scores (95th percentile), effectively controlling false discovery rate at $\alpha=0.05$.

Measurement of THC and CBD

The quantification of THC and CBD was performed using ultra-high-performance liquid chromatography coupled with photodiode array detection (UHPLC-PDA). Dried plant material (100 mg) was homogenized in 5 mL methanol containing 0.1% (v/v) formic acid, sonicated at 40 °C for 30 min, and centrifuged at $12,000 \times g$ for 10 min. The supernatant was filtered through a 0.22- μ m PTFE membrane prior to injection. Chromatographic separation was achieved on a C18 reversed-phase column (2.1 $\times$ 100 mm, 1.7 μ m) with a gradient elution of 0.1% formic acid in water (mobile phase A) and acetonitrile (mobile phase B) at a flow rate of 0.3 mL/min. Detection wavelengths were optimized at 220 nm for THC and 210 nm for CBD based on their maximum absorbance [57].

Quantitation was performed using external calibration curves (5–200 μ g/mL, $R^2 > 0.999$) with certified reference standards (THC: Sigma-Aldrich, #T044; CBD: Cerilliant, #C-045). Method validation included precision (RSD<3.5%), accuracy (recovery 95–105%), and limit of detection (LOD=0.1 μ g/mL) following ICH Q2(R1) guidelines [58]. Matrix effects were evaluated by spiking blank plant extracts with known concentrations of analytes.

GWAS

We performed genome-wide association analyses on *C. sativa* accessions representing 14 geographically distinct populations. Genome-wide association mapping was performed using the high-confidence SNP set ($n=2,699,120$), which had already been pre-filtered across all 271 accessions with MAF>0.05 and genotype missing rate<50%. No additional filtering was applied prior to the association analysis. Genetic associations were detected using GEMMA v0.98.1 [59] under a linear mixed model framework. A genetic relatedness matrix (GRM) was derived from the top three principal components to account for population stratification. To correct for multiple testing, we estimated the number of effective independent SNPs to be approximately 1.56×10^5 following linkage

disequilibrium pruning (PLINK --indep-pairwise 50 10 0.2). This yielded a Bonferroni-corrected genome-wide significance threshold of $0.05/1.56 \times 10^5 \approx 3.2 \times 10^{-7}$. To maintain consistency with the standard Manhattan plot presentation and the empirical visualization threshold applied in our study, we conservatively adopted a unified genome-wide significance threshold of $P < 1.0 \times 10^{-7}$ (i.e., $-\log_{10}(P) > 7$) and a suggestive threshold of $P < 1.0 \times 10^{-6}$ (i.e., $-\log_{10}(P) > 6$) throughout all downstream analyses.

Identification of the candidate genes related to THC and CBD

Genome-wide association analysis was performed using a three-stage filtering approach to identify functionally relevant SNPs. First, SNPs surpassing the unified genome-wide significance ($P < 1.0 \times 10^{-7}$) and suggestive threshold ($P < 1.0 \times 10^{-6}$) as defined in the GWAS section were retained. Candidate genomic regions were subsequently defined through pairwise LD clumping ($r^2 > 0.8$) around these lead SNPs. Subsequently, prioritized variants were subjected to functional characterization, focusing on nonsynonymous SNPs affecting protein coding sequences and regulatory SNPs located within 2-kb upstream promoter regions. Final functional annotation of candidate genes was performed through comparative genomic analysis using BLASTN (E-value $< 1 \times 10^{-10}$) against the *Arabidopsis thaliana* reference genome (TAIR10) and related species genomes from the Ensembl Plants database.

Results

Resequencing 271 accessions of *C. sativa*

The genetic resources of cannabis, especially wild cannabis, are extremely scarce and difficult to collect within China. This research was coordinated by the National Anti-Drug Committee and assisted by all levels of public security anti-drug departments across the country. The project basically covered all areas in China where wild cannabis might exist, and the collection work was highly comprehensive and representative. In this study, a comprehensive panel of 271 *C. sativa* accessions was analyzed, representing diverse ecogeographic origins spanning 29 regions across 12 Chinese provinces: Xizang, Xinjiang, Yunnan, Qinghai, Sichuan, Guangxi, Gansu, Shaanxi, Shandong, Heilongjiang, Anhui, and Inner Mongolia (Table S1). Whole-genome resequencing generated 4.51 Tb of raw sequencing data, with individual samples yielding 8.84–19.68 Gb. Following adapter trimming and quality filtering (Phred score ≥ 20), 1.53 Tb of high-confidence data were retained (4.28–11.21 Gb per accession), achieving an average sequencing depth of $7.08\times$ (range: $5.36\text{--}13.45\times$) with a mean read mapping efficiency of 36.98% (ranging from 23.52% to 95.00% per line).

Variant calling against the *C. sativa* reference genome identified 2,699,120 SNPs and 42,201 InDels (Tables S2–S3). Chromosomal distribution analysis revealed significant heterogeneity in variant density: chromosome 3 harbored the highest SNP count (334,455; density: 3.72 SNPs/kb), while chromosome 10 contained the fewest (162,541; density: 2.98 SNPs/kb) (Fig. 1a). InDel frequency peaked on chromosome 7 (5,173 events; density: 0.06 InDels/kb) (Fig. 1b). Functional annotation classified 74.63% of SNPs (2,014,439) as intergenic, with remaining variants distributed in downstream (7.76%; 209,571), upstream (9.43%; 254,599), exonic (2.83%; 76,292), and intronic regions (5.34%; 144,219) (Table S3). Among coding SNPs, 49.80% (37,995) were synonymous and 50.20% (38,297) nonsynonymous. Similarly, InDels predominated in intergenic regions (63.14%; 26,646), followed by upstream (15.72%; 6,634), downstream (11.51%; 4,857), and exons and introns (9.63%; 4,064) (Table S3).

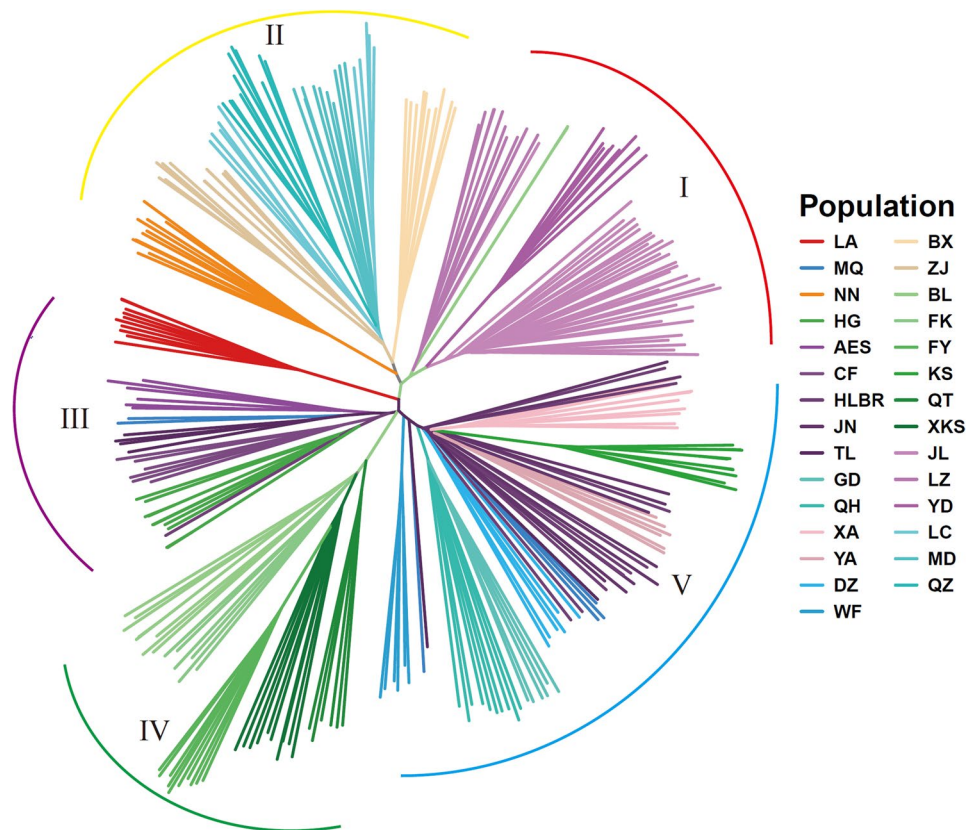
Fig. 1 Genome-wide variant profiling of *Cannabis sativa*. **a** Chromosomal distribution density of single nucleotide polymorphisms (SNPs) across 10 chromosomes. **b** InDels density across chromosomes. Heatmap gradients represent scaled variant density (green: low; red: high)



Population structure and genomic variation

Genetic architecture of the 271 *C. sativa* accessions was investigated through phylogenetic reconstruction, population structure analysis, and principal component analysis (PCA) using high-confidence SNPs. Neighbor-joining (NJ) phylogenetic clustering delineated five major groups (I–V), largely concordant with geographic origins (Fig. 2, Fig. S1): Group I comprised accessions from Southwest China (Xizang: JL, YD, LZ), Group II included Sichuan (BX, ZJ), Yunnan (LC, MD, QZ), and Guangxi (NN) accessions, Group III mainly included Anhui (LA), Heilongjiang (HG) and Inner Mongolia (TL, CF and AES). Group IV contained five populations from Xinjiang (BL, FK, FY, QT, XKS). Group V is the most complex, mainly including the covered population in Xinjiang (KS), Inner Mongolia (HLBR, JN), Gansu (MQ), Shaanxi (XA, YA), Shandong (DZ, WF), Qinghai (QH, GD).

Fig. 2 Neighbor-joining phylogenetic tree constructed using SNPs. The colors of the branches represent different population



Notably, populations from Xinjiang exhibited distinct north-south divergence, with southern (KS) and northern (BL, FK, FY, QT, XKS) accessions segregating into separate groups. Inner Mongolian accessions displayed two clustering across group III (AES, CF and TL) and V (HLBR and JN), suggesting complex migration history.

Our population structure analysis (Fig. 3a) corroborated phylogenetic patterns. At $K=2$, populations partitioned into Southwest/Northeast-Southeast clusters. Optimal clustering ($K=5$, ΔK peak in Fig. S2) resolved: Xizang group: JL, YD, LZ; Northern China group: Inner Mongolia (AES, CF, TL), Heilongjiang (HG), Gansu (MQ); Northern Xinjiang group: BL, FK, FY, QT, XKS; Southwest group: Sichuan (BX, ZJ), Yunnan (LC, MD, QZ), Guangxi (NN); and Transitional group: Shandong (DZ, WF), Qinghai (QH, GD), Inner Mongolia (HLBR, JN), Shaanxi (XA, YA). The Xizang group demonstrated exceptional genetic homogeneity across K -values, supporting its potential status as an evolutionary conservation hotspot and putative origin of Chinese *C. sativa* germplasm. Our PCA results (Fig. 3b) confirmed these divisions found by NJ phylogenetic tree and population structure, with PC1 (6.53% variance) and PC2 (3.80% variance).

Genetic differentiation and population diversity

Genome-wide analyses revealed substantial genetic variation among *C. sativa* populations, with nucleotide diversity ($\theta\pi$) ranging from 2.63×10^{-4} to 4.78×10^{-4} (Fig. 4a, Table S4). Among all the populations, the Xinjiang FK population exhibited the highest diversity ($\theta\pi = 4.78 \times 10^{-4}$), contrasting sharply with the Xinjiang FY population ($\theta\pi = 2.63 \times 10^{-4}$), suggesting heterogeneous selection pressures across Xinjiang's ecogeographic zones. Pairwise F_{st} values (ranging from 0.000 to 0.341) revealed significant genetic differentiation extremes among populations. Notably, the Yunnan QZ and Xinjiang FY populations exhibited the highest divergence ($F_{st} = 0.341$), whereas three population pairs—Inner Mongolia HLBR and Shandong DZ, Inner Mongolia HLBR and Gansu MQ, as well as Gansu MQ and Inner Mongolia TL—showed complete genetic panmixia ($F_{st} = 0.000$) (Table S5).

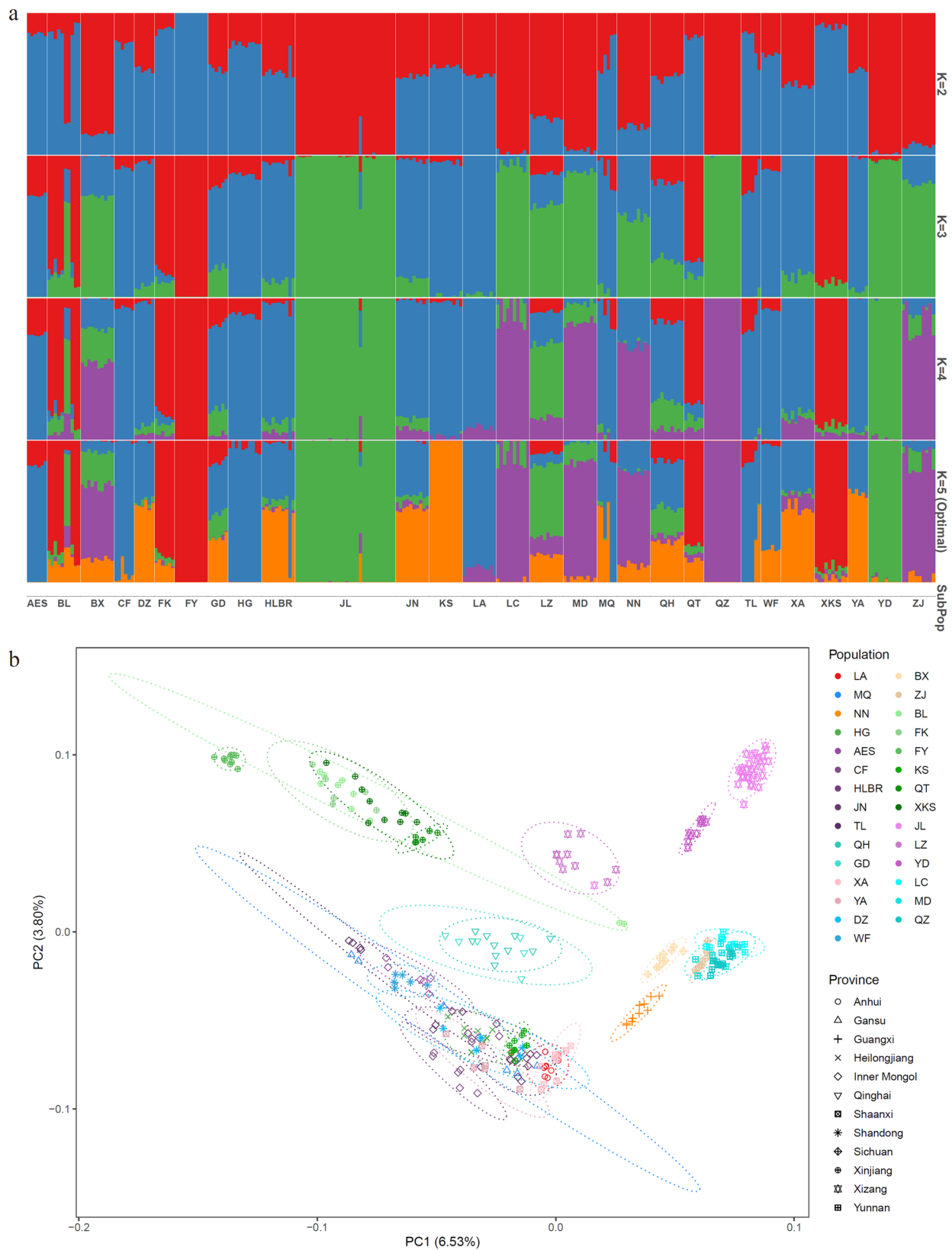


Fig. 3 Population structures and Principal component analysis (PCA) of all *C. sativa* accessions. **a** Population structures estimated through admixture analysis for a presumed number of groups (K) ranging from 2 to 5. **b** PCA of the first two components for *C. sativa* accessions

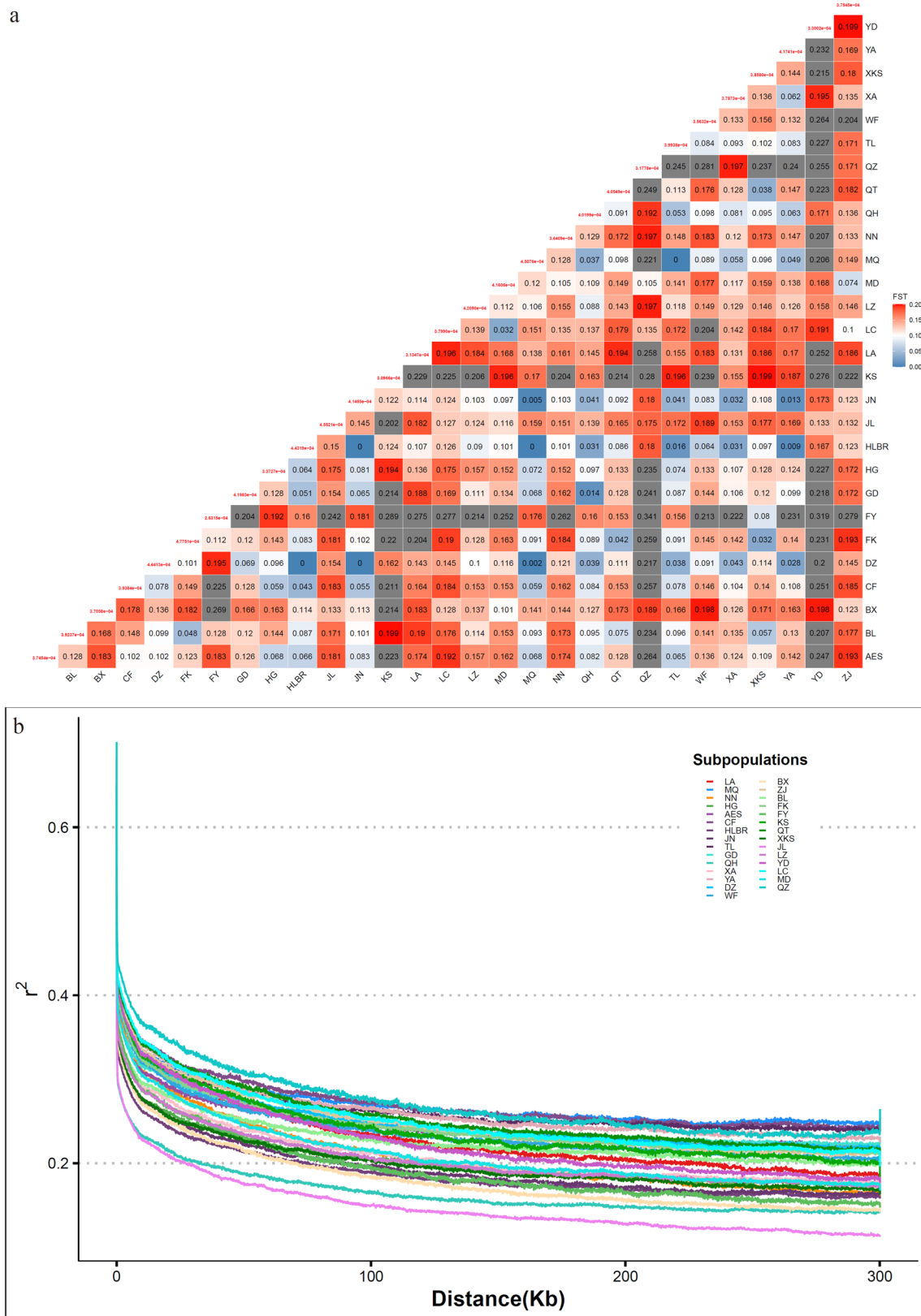


Fig. 4 Nucleotide diversity ($\theta\pi$), genetic divergence (Fst), and LD decay for *C. sativa* populations. **a** Nucleotide diversity ($\theta\pi$) and genetic divergence (Fst) among all analyzed populations. **b** LD decay for each population

Linkage disequilibrium (LD) decay patterns provided evolutionary insights (Fig. 4b). The Xizang Jilong (JL) population exhibited rapid LD decay, characteristic of ancestral populations with prolonged recombination history. This contrasts with slower LD decay in derived populations, supporting the hypothesis that Xizang accessions represent foundational germplasm for Chinese *C. sativa*.

Phylogeographic patterns and allelic drift dynamics

To analyze population divergences and mixtures among the main source locations, we computed drift parameters based on the genetic variations and constructed trees to illustrate the relationships among populations (Fig. 5a, b). Our results revealed Xizang Jilong (JL) as the basal lineage in phylogenies, establishing this Xizang population as the probable progenitor of domesticated *C. sativa* in China. Integrative analysis of phylogenetic topology, structure coefficients, and diversity metrics supports a southwest-to-northeast expansion pattern with three primary dispersal phases: (1) Primary Radiation: Core origin: Xizang JL → YD → LZ, Southwest expansion: LZ (Xizang) → Sichuan (BX, ZJ), Yunnan (LC, MD, QZ) and Guangxi (NN); (2) Northern Dispersal: Central hub: Xi'an (XA, Shaanxi), Western spread: XA → Southern Xinjiang (KS), Northwest route: XA → Qinghai (GD/QH) → Northern Xinjiang (BL, FK, FY, QT, XKS), Northeast corridor: XA → Yanan (YA) → Jining (JN) → Chifeng (CF) → Tongliao (TL) → AES (Arshan) → HLBR (Hulunbeier) → Heilongjiang (HG); Eastern spread: XA → Lu'an (LA, Anhui); (3) Secondary Radiation: Inner Mongolia nexus: HLBR (Hulunbeier) → Shandong (DZ, WF). Notably, the Kunlun Mountains formed a substantial phylogeographic barrier, restricting gene flow between Xizang and Xinjiang populations. This isolation is reflected in the elevated F_{st} between southern Xinjiang (KS) and Xizang populations (0.187–0.206) compared to intra-regional differentiation (0.127–0.168).

Selective signals of natural and artificial selection

To identify genetic changes shaped by natural and artificial selection, we performed genome-wide selective sweep analysis using XP-CLR across three critical population contrasts. These included: (1) Ecogeographic contrast: Northern versus Southern Xinjiang; (2) Regional isolation: Xinjiang versus Xizang; (3) Domestication axis: Southwestern populations (Xizang-Sichuan-Yunnan) versus Northeastern populations (Inner Mongolia-Heilongjiang). Using 100 kb sliding windows (10 kb step), we identified loci under strong selection containing missense variants. In total, we found 3,299 genes in Xinjiang north-south differentiation (Fig. 6a, Table S6). In addition, a total of 3,215 genes were identified within the sweep regions at Xinjiang-Xizang divergence interface (Fig. 6b, Table S7). Moreover, we found 3,404 genes distinguishing southwestern vs. northeastern domestication pathways (Fig. 6c, Table S8). The few overlapping genes among the three parts indicate that the loci selected for the three regions are different (Fig. S3).

Chemotypic selection in cannabinoid biosynthesis

To dissect genetic determinants of Δ^9 -tetrahydrocannabinol (THC) and cannabidiol (CBD) variation, we quantified cannabinoid content across 14 *C. sativa* populations. Specifically, a subset of 90 individual accessions from these 14 representative populations was selected for phenotyping. These 14 populations were chosen based on their broad geographical distribution across China and their distinct genetic backgrounds revealed by phylogenetic analysis. The GWAS analyses were exclusively performed on these phenotyped individuals to ensure statistical accuracy. The sex of the measured plants was recorded as female to minimize sex-related metabolic interference. The complete phenotype information is provided in Table S9. Based on the measurements, we stratified these accessions into high- (H) and low-content (L) chemotypes (Fig. 7a-b). Using genome-wide F_{st} differentiation analysis, we identified chemotype-associated selective sweeps (Fig. 7c-d).

Based on the findings, the lengths identified in the selected differentiation regions for CBD and THC are 41,199,176 bp and 40,249,195 bp respectively. The corresponding numbers of genes are 1724 and 1238 (Table

Fig. 5 Allelic-drift trajectories illustrating genetic divergence and admixture events across core geographic origins of the sampled accessions. **a** Reconstructed phylogeny depicting primary geographic origins of *C. sativa* accessions and population-level gene-flow dynamics. Gene-flow directions are denoted by arrows, with line coloration corresponding to migration weights scaled by sample size. Horizontal branch lengths reflect cumulative genetic drift magnitude. Scale bars indicate a $10\times$ mean standard error for entries in the sample covariance matrix. **b** Spatial distribution of gene-flow networks across China's biogeographic regions

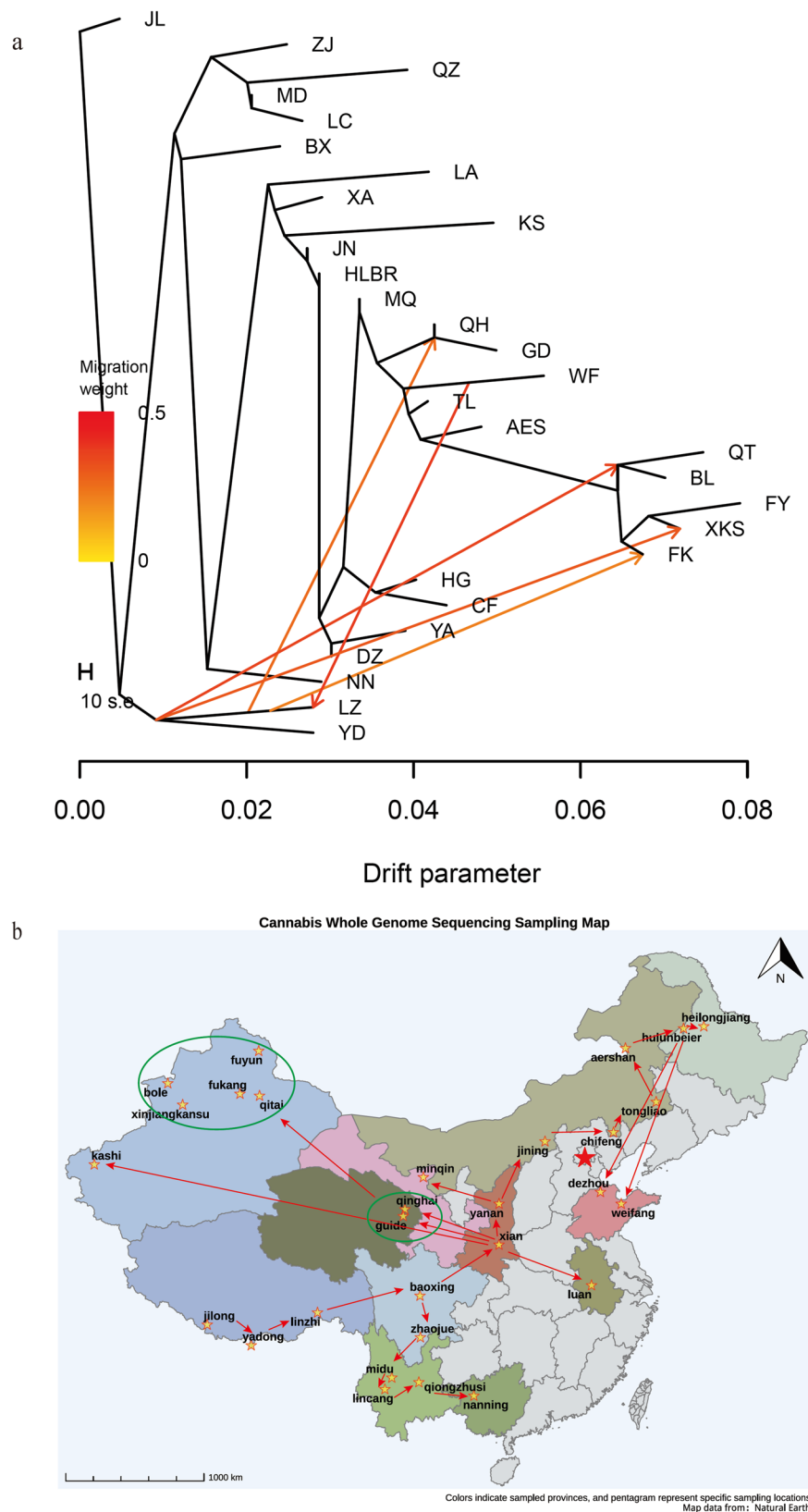


Fig. 6 Genome-wide analysis of selection signatures using XP-CLR driving regional genetic differentiation in *C. sativa* populations. **a** Divergent selection signatures in cannabis populations from northern versus southern Xinjiang (Abbreviated as “Beijiang” and “Nanjiang” respectively). **b** Divergent selection signatures in cannabis populations from Xinjiang versus Xizang. **c** Divergent selection signatures in cannabis populations from Xizang+yunnan+Sichuan versus Neimeng+Heilongjiang

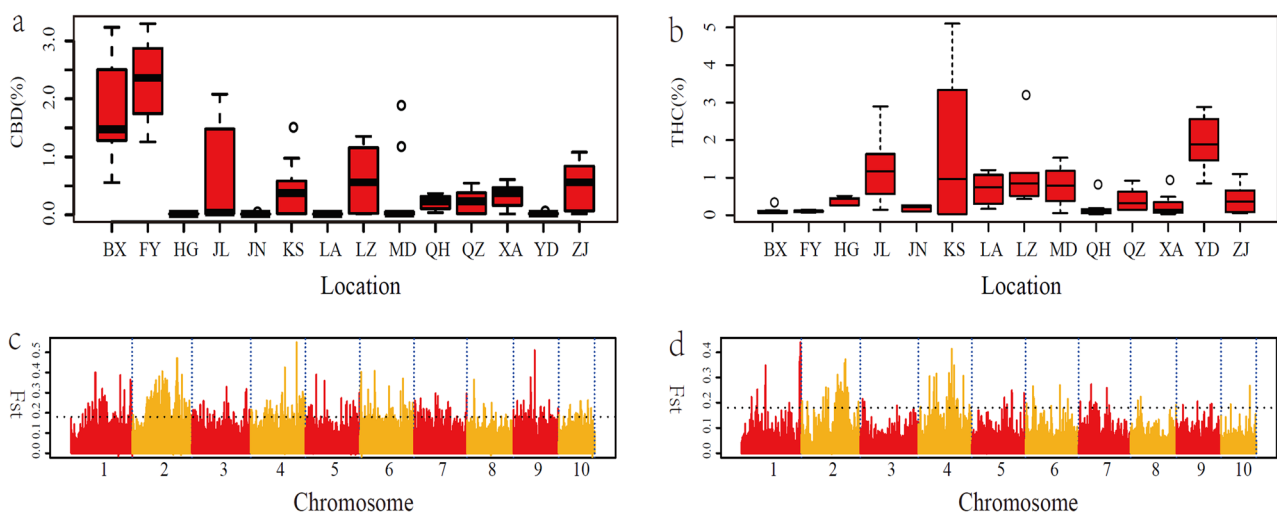
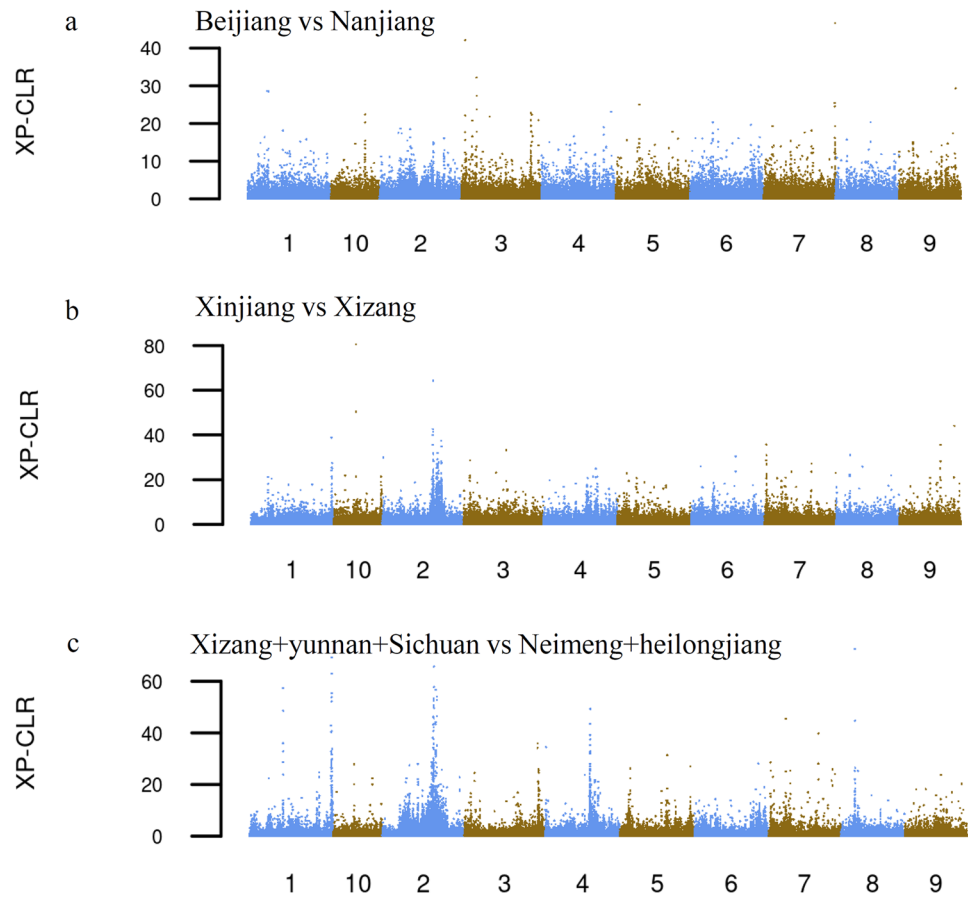


Fig. 7 Comparative genome-wide analysis of genetic differentiation (F_{st}) between high- versus low-CBD and high- versus low-THC chemotypic groups in 14 *C. sativa* populations. **a** Average CBD content between 14 *C. sativa* populations. **b** Average THC content between 14 *C. sativa* populations. **c** Significant association signals identified through F_{st} in regions exhibiting notable CBD-related signatures. **d** Significant association signals identified through F_{st} in regions exhibiting notable THC-related signatures

S10, 11). And the results showed that there are 254 overlapping genes for shared genomic signature of the two parts (Fig. S4). Notably, six key CBD and THC selected-site related genes were annotated as CBDAS and THCAS. Among these genes, there is one THCAS (1fm_FM.07G0003110.PATH1_0_R8) and two CBDAS (1fm_FM.08G0016800.PATH1_0_R4; 5sl_GENE25331_0_R1) genes associated with CBD, while there are two THCAS (8mn_GENE-LOC112093202_0_R9; 1fm_FM.07G0003110.PATH1_0_R8) and one CBDAS (1fm_FM.08G0016800.PATH1_0_R4) genes associated with THC (Table S12). To further explore the potential functions of the candidate genes identified in the chemotype-associated selective sweep regions, we performed Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. The complete results are provided in the supplementary materials (Tables S13–S16; Figs. S5–S6). Among these, for CBD-associated genes, the GO term dolichyl-diphosphooligosaccharide-protein glycotransferase activity (GO:0004579) and the KEGG pathway protein processing in endoplasmic reticulum (map04141) are notable, as they are consistent with the N-linked glycosylation and ER processing of CBDA synthase, which is a glycoprotein requiring proper folding and glycosylation for enzymatic activity. For THC-associated genes, the enrichment results, detailed in the supplementary tables, included a range of GO terms and KEGG pathways (Tables S13–S14; Fig. S5). However, given the high repetitive content of the *C. sativa* genome and the potential for TE-associated annotations in the candidate gene lists, we interpret all enrichment results with caution and emphasize the directly identified THCAS and CBDAS genes as the primary functional evidence for chemotype-associated loci.

GWAS of cannabinoid chemotypes

We performed genome-wide association studies (GWAS) to dissect genetic architecture underlying THC and CBD variation (Fig. 8a–d). Applying the unified threshold of $P < 1.0 \times 10^{-7}$ for genome-wide significance and $P < 1.0 \times 10^{-6}$ for suggestive associations, significant signals were cross-validated against prior selection signals. For CBD association landscape, three genome-wide significant loci were identified, highly overlap with significant *Fst* loci (Fig. 8a–b). Overall, using the suggestive threshold ($P < 1.0 \times 10^{-6}$), we identified 1,176 suggestive SNPs mapping to 294 genes; among these, 140 SNPs within 52 genes surpassed the genome-wide significance threshold ($P < 1.0 \times 10^{-7}$) (Table S17). The significant findings revealed a dispersed distribution of observed signals throughout the genome, with a notable clustering on chromosomes 1, 2, and 3. At the distal end of chromosome 1's right arm, a prominent peak associated with CBD was observed (Fig. 8a, b). Upon local enlargement of this chromosomal region, it was discovered that the gene annotated at the peak corresponds to CBDAS, which plays a functional role in CBD synthesis (Fig. 9a). For THC association architecture, three peaks were also identified. Notably, we detected 323 suggestive SNPs linked to 82 genes, of which 35 SNPs in 17 genes reached genome-wide significance (Fig. 8c, d and Table S18). The significant findings showed a dispersed distribution across the genome, with the exception of chromosomes 6 and 8. The majority of these signals are predominantly localized on chromosome 3 (Table S18). Interestingly, the identification of relevant candidate genes for THC remains elusive based on functional annotations (Fig. 9b).

Discussion

The genomic variation present in *C. sativa* offers a precious opportunity for delving into genetic diversity, gene mapping, selection, and molecular-assisted breeding. Previous research on the genetic diversity of *C. sativa* has mainly concentrated on a small number of restriction site-associated DNA markers and a limited selection of *C. sativa* accessions [19–23]. Previously, Low-density DNA markers like AFLPs and SSRs have been used in cannabis to detect QTLs linked to sex determination and cannabinoid composition [60, 61]. Additionally, some shorter DNA sequences, such as chloroplast genome sequences and ribosomal DNA regions, are also used in genetic analysis of cannabis [29, 62–65]. These studies have provided some initial insights, but it has obvious limitations. The restricted range of markers used may not fully represent the entire genetic complexity of the species, and

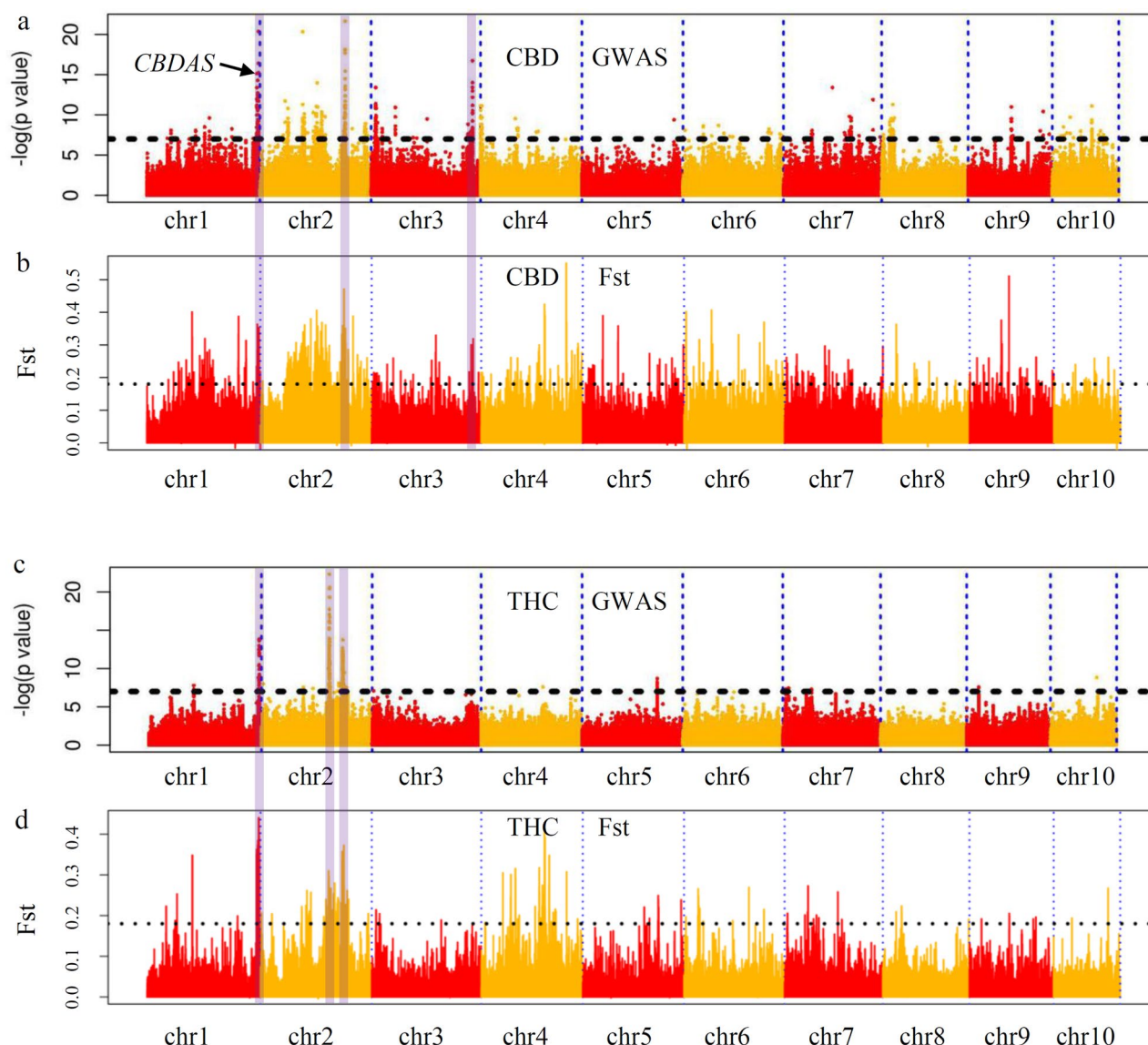


Fig. 8 GWAS mapping of CBD and THC. **a** Manhattan plot of GWAS results for the CBD trait on the 10 chromosomes. **b** Significant association signals identified through Fst in regions exhibiting notable CBD-related signatures. **c** Manhattan plot of GWAS results for the THC trait on the 10 chromosomes. **d** Significant association signals identified through Fst in regions exhibiting notable THC-related signatures

the small number of accessions studied might not accurately reflect the overall genetic diversity present within the diverse populations of *C. sativa*. In the present study, we conducted resequencing and characterization of the genetic variations in 271 *C. sativa* accessions, which were collected from 29 major *C. sativa* production regions of 12 provinces in China.

In our study, a comparison between the genomes of *C. sativa* and the reference genome identified a combined total of 2,699,120 SNPs and 42,201 InDels. The results showed that most SNP markers fell in intergenic regions. Stop gained mutations were less frequent. There were more SNPs upstream and downstream, indicating a higher frequency of SNPs near gene promoters and terminators. Similar results have been observed in mango [66] and rapeseed [32]. Our phylogenetic analyses classified the 271 *C. sativa* accessions into five major groups (I-V). Accessions from Xizang, located in southwestern China, formed a distinct cluster (Group I), exhibiting close genetic relationships within the group and significant genetic distance from other populations. Accessions

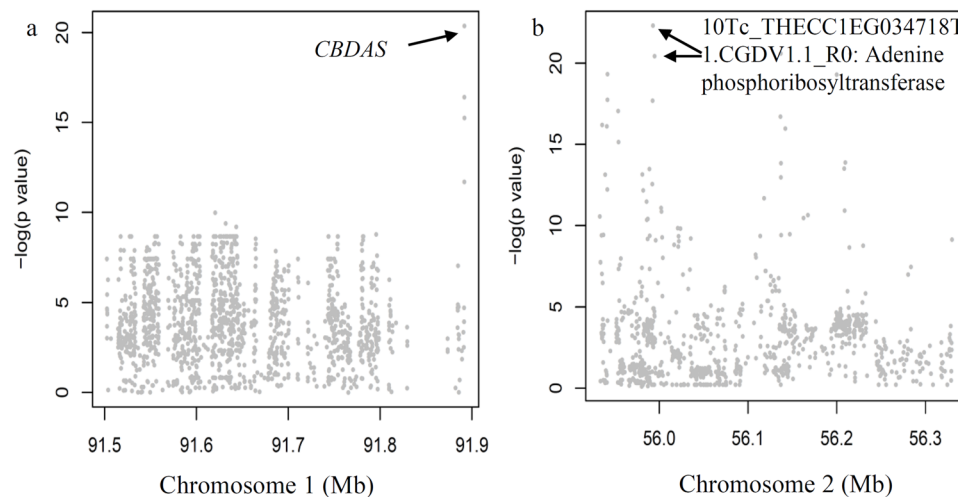


Fig. 9 High-resolution local magnification of the GWAS peak regions for CBD and THC. Panel (a) focuses on the telomeric region of Chromosome 1 (91.5–91.9 Mb) where the most significant CBD-associated signal is located; the peak corresponds to the CBDAS gene, encoding cannabidiolic acid synthase, the key enzyme for CBD biosynthesis. Panel (b) focuses on the telomeric region of Chromosome 2 (56.0–56.8 Mb) where the most significant THC-associated signal is located; the peak region contains gene annotations, but no definitive candidate gene for THC biosynthesis was identified based on functional annotation. X-axis coordinates indicate physical positions in megabases (Mb). This is a local magnification view; genome-wide Manhattan plots are presented in Fig. 8

originating from Sichuan, Yunnan, and Guangxi clustered together as Group II. Group III comprised accessions from other Chinese regions, including Anhui, Heilongjiang, and Inner Mongolia (TL, CF, AES). The five populations sampled from northern Xinjiang formed Group IV. Group V was the most complex, primarily including cultivated populations from southern Xinjiang, Inner Mongolia (HLBR, JN), Gansu, Shaanxi, Shandong, and Qinghai. Notably, the genetic relationship between the Xizang accessions (Group I) and those from Sichuan and Yunnan (Group II) is relatively close, corresponding with their geographical proximity. This phylogenetic pattern suggests that the initial dissemination point of cultivated cannabis within China was likely in Xizang, followed by gradual spread to adjacent areas such as Sichuan, Yunnan, and Guangxi, and subsequently further inland. Although geographically proximate, Xizang and Xinjiang accessions exhibit substantial genetic divergence based on the phylogenetic tree. This observed genetic distance may be explained by a scenario where wild cannabis originating in Xizang hybridized with domesticated fiber-type varieties. These hybridized forms could then have spread to Gansu and Qinghai via the Silk Road before reaching Xinjiang. Our population structure and PCA results are consistent with the phylogenetic tree. The samples from the same province tend to cluster together. However, interestingly, cannabis samples from Kashgar (KS) in Xinjiang show a distant genetic relationship with other Xinjiang population samples (BL, FK, XKS and QT). Considering that the Tianshan Mountains divide Xinjiang into two parts [67], where KS is located in southern Xinjiang and BL, FK, XKS and QT are located in northern Xinjiang. There may be geographical isolation between southern and northern Xinjiang, which could explain why the genetic relationship of KS is distant from other Xinjiang populations.

The P_i (π) value of population diversity is used to analyze the level of nucleotide diversity within subgroups. A larger P_i value indicates a higher level of nucleotide diversity [68]. Our results showed the nucleotide diversity ($\theta\pi$) for the tested populations varied from 2.63×10^{-4} to 4.78×10^{-4} , indicating a significant difference in genetic diversity among *C. sativa* populations. We observed that the Xinjiang FK population exhibited the highest nucleotide diversity ($\theta\pi = 4.78 \times 10^{-4}$), while Xizang populations (JL, YD, LZ) showed moderate diversity levels (range: 3.30 – 4.55×10^{-4}). After analyzing the LD values of *C. sativa* in different provinces, we found that the LD values of *C. sativa* in Xizang are relatively low, which is characteristic of ancestral populations with longer recombination history. Therefore, the basal phylogenetic position of Xizang JL population, its rapid LD decay,

and its geographic proximity to the putative South/Central Asian origin collectively support the status of Xizang as the origin and dissemination center of domestic *C. sativa* in China, rather than nucleotide diversity alone. Previous studies have also reached similar conclusions, thus strengthening our findings [1, 3].

The origin of *C. sativa* has always been a focal point for scientists and is also the subject of controversy. Previously, the common belief was that *C. sativa* had its origins in Central Asia [10, 69]. It has been suggested that Central and Southeast Asia may have been potential regions for its natural origin and initial domestication, playing a significant role in its evolutionary history [16]. However, recent microfossil data indicate that the northeastern Tibetan Plateau may actually be the center of origin for *C. sativa* [15]. Furthermore, in recent years, a large-scale whole-genome resequencing study has been conducted on 110 *C. sativa* resources, and the results show that *C. sativa* underwent single-origin domestication in East Asia during the Early Neolithic [1]. Due to factors such as difficulties in sample collection, there has been very little research on the origin and spread of Chinese cannabis in the past, which also poses a challenge to the protection and rational development of genetic resources of Chinese cannabis. Based on the allelic variation across the 271 genomes to calculate the drift parameters, we have created a China map of the major gene-flow paths. Previous studies have suggested that wild Cannabis progenitors may have originated in Central Asia and/or the Tibetan Plateau [10, 15, 69]. Geographically, JL of Xizang is located adjacent to South Asia (India), which supports a potential early dispersal route from this region. Our research results also confirm that the early cannabis in China originated from JL. Subsequently, the JL population further differentiated into the YD and LZ populations. Then it spread from YD and LZ to Sichuan, Yunnan and Guangxi. Subsequently, it spread from southwest China to Shaanxi and then expanded to the northwest regions such as Gansu, Qinghai, and Xinjiang with Shaanxi as the center; at the same time, it spread to Inner Mongolia and Heilongjiang in the northern regions and Anhui in the eastern regions. Additionally, our study indicated that the cannabis population from Shandong was likely transmitted from HLBR in Inner Mongolia. Xi'an (formerly known as Chang'an) served as the historical starting point of the Silk Road during the Han and Tang dynasties, and held a strategic hub position in the history of cultural exchanges between Asia and Europe. The results of this study confirm that this region also played a crucial node role in the early dissemination network of cannabis germplasm resources - its geographical location connected the agricultural areas of the Central Plains with the trade routes in the Western Regions, facilitating the bidirectional exchange and regional circulation of cannabis germplasm.

Selective sweep is a genetic phenomenon driven by natural selection [70]. Selective sweep analysis can reveal the key driving forces (natural selection vs. artificial selection) of cannabis evolution and provide targets for resource conservation (such as preserving wild relatives) and breeding (such as targeted improvement) [32]. Our analysis results of selective sweep indicate that there are selected sites between northern Xinjiang and southern Xinjiang, Xinjiang and Xizang, and Xizang+Sichuan+Yunnan and Inner Mongolia+Heilongjiang. Moreover, the intersection genes of the three parts are few, which indicates that there are differences in the selected sites among the three regions. There is geographical isolation between northern and southern Xinjiang [67]. The populations in Xinjiang and Xizang have the farthest distance on the evolutionary tree, and perhaps they are respectively the end and the beginning of the spread of cannabis in China. While Xizang+Sichuan+Yunnan and Inner Mongolia+Heilongjiang respectively represent the north-south differentiation regions of cannabis. The above situations may be the reasons for the selective sweep of cannabis. Similar to the relevant studies conducted on other species [32, 35], our study detected a large number of genetic footprints through selective-sweep analyses. These imprints were formed by the combined effect of natural selection and artificial selection. These include changes in genes involved in hormone synthesis and signaling, as well as stress response pathways, which contribute to the differences observed between northern Xinjiang and southern Xinjiang, Xinjiang and Xizang, and Xizang+Sichuan+Yunnan and Inner Mongolia+Heilongjiang.

The genetic divergence of cannabinoid chemotypes in Cannabis populations is primarily driven by selective pressures acting on functional polymorphisms within the THCA synthase (THCAS) and CBDA synthase (CBDAS) gene clusters [37, 71]. By analyzing the selected sites of high/low THC and high/low CBD strains of hemp, this study revealed the genetic selection signals that were significantly related to the content of secondary metabolites. Our research findings indicate that the lengths of the selected differentiation regions for CBD and

THC are 41,199,176 bp and 40,249,195 bp respectively, corresponding to 1,724 and 1,238 genes respectively. Among them, six genes are annotated as CBDAS and THCAS. In the samples with differences in CBD content, strong selection signals were observed in the regions near the CBDAS gene located on chromosomes 1, 2, and 3. CBDAS has high homology with THCAS but catalyzes different products. The differences in allele frequencies of CBDAS (such as frame-shift mutations caused by InDel mutations) may be the key to the phenotypic differentiation of CBD [60]. Interestingly, this study also found that cytochrome P450 gene clusters (such as the CYP71 family and CYP73 family) are associated with the regulation of CBD content, suggesting that, in addition to CBDAS, genes related to oxidative modification may also affect the accumulation of final products through metabolic pathways in a coordinated manner [72]. In contrast, in high-THC strains, strong selection signals were observed in the regions near the THCAS gene located on chromosome 2. This is consistent with previous studies: as a key enzyme in the biosynthesis of THCA, SNPs in the promoter and coding regions of THCAS have been confirmed to be significantly associated with THC content [71]. Notably, in high-THC strains, there is a significant selective sweep in the upstream regulatory region of THCAS (such as MYB transcription factor binding sites), suggesting that artificial selection may enhance THC production by increasing gene expression. It is worth noting that the selection sites for THC and CBD are spatially separated in the genome, which may reflect the differentiation in breeding goals: high-THC strains are mostly used for recreational or specific medical purposes (such as analgesia), while high-CBD strains are more favored in the medicinal market due to their non-psychoactive properties [73]. This directional pressure of artificial selection may lead to the gradual divergence of genetic structures in the two types of samples.

GWAS have become an indispensable tool in modern genetics, particularly when it comes to understanding complex traits in crops [36, 74]. Our study identified multiple genetic loci significantly associated with the content of THC and CBD in the cannabis population through GWAS. Among them, our GWAS identified 1,176 significant SNP signals ($-\log_{10} P \geq 7$) associated with 294 genes that are strongly linked with CBD. Among these genes, CBDAS, corresponding to a distinct peak at the end of the right arm of chromosome 1, are the most crucial adaptive genes, whose functions are related to CBD synthesis [37, 75]. Additionally, our GWAS identified 323 suggestive SNP signals ($-\log_{10} P \geq 7$) associated with 82 genes that are strongly linked with THC, which highlights the intricate genetic architecture underlying individual differences in THC sensitivity. Furthermore, the distribution of significant results across the genome, with a notable concentration on chromosome 2, suggests that multiple genetic loci may contribute to variations in response to THC. This underscores the polygenic nature of THC-related traits and emphasizes the need for comprehensive genomic analyses to unravel its genetic underpinnings. Despite these advancements, challenges persist in pinpointing specific candidate genes for THC based on functional annotations. The multifaceted interactions between genes and environmental factors further complicate efforts to elucidate the molecular mechanisms driving individual responses to THC [75, 76].

The genetic polymorphisms of THCAS and CBDAS genes, which serve as key enzymes regulating the biosynthesis of THC and CBD, provide molecular basis for chemotype classification within *Cannabis* species [77, 78]. de Meijer et al. (2003) conducted selfing experiments using homozygous medicinal cannabis strains and observed a 3:1 Mendelian ratio of THC/CBD content in F_1 hybrids. This fundamental discovery established the codominant allelic hypothesis (BT/BD) at the B locus, where drug-type (BT/BT), fiber-type (BD/BD), and medicinal-type (BT/BD) chemotypes correspond to specific genotypic combinations [79]. Kojoma et al. (2006) further identified 62 SNP variations within THCAS across 13 cultivars, establishing a chemotype classification model subsequently validated by physicochemical analyses [80]. Subsequent studies have progressively expanded this field: Genome-wide analysis of 124 accessions by Sawler et al. (2015) revealed 14,031 population-differentiated SNPs, demonstrating the polygenic basis of chemotypic differentiation [81]. Notably, Abdelaziz et al. (2016) identified 47 intra-specific SNPs in *Moroccan* germplasm, with stable sequence variations at THCAS locus reinforcing its critical role in chemotype discrimination [82]. Our study successfully identified SNP polymorphic markers significantly associated with THCAS and CBDAS genes, which provides a high-resolution molecular detection system for achieving DNA genotyping identification within *Cannabis* species in China. Furthermore,

it establishes precise technical discrimination criteria for tracing illicit cannabis cultivation in drug law enforcement and supervising the regulated cultivation of industrial hemp.

Conclusion

In conclusion, this study presents a comprehensive analysis of genetic polymorphism in *C. sativa*, utilizing resequencing data from a diverse collection of 271 germplasm accessions sourced from around China. By examining this extensive dataset, the research reveals the China pattern of genetic variation within *C. sativa*, offering valuable insights into the genomic changes that have occurred during both natural and artificial selection processes. This study provides the largest genetic resource from China for identifying molecular markers linked to key agronomic traits in *C. sativa*. These markers are essential for accelerating genetic improvement and developing varieties with higher yield, better quality, and stronger resilience. The detailed genetic variation map serves as a foundation for future research on the genetic basis of important traits in *C. sativa*.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1186/s12870-026-09909-5>.

Authors' contributions Baishi Wang, Shan Gao, Xue Wang and Quanzheng Yun: Wrote the original draft, collect sample materials, and analyze data. Hao Nie, Ning Li, Yonggang Li, Deri Feng, Linsen Li, Tianye Li, Xu Lu and Jingjing Liu: Perform experiment, data analysis, data verification and visualization. Xuge Wei, Yanhong Mu, Qiqi Wu, Jinyuan Hu, Fangru Liu, Rui Feng and Jianyu Yin: Perform experiment and data analysis, Qikun Zhang, Ying Zhang and Fu Ren Draft writing, review and editing, data analysis, project administration and conceptualization. All authors read and approved the final manuscript.

Funding This study was financially supported by Liaoning Provincial Science and Technology Plan Project (2025JH2/101800069), Team Construction Project of Liaoning Province Education Department (LJ222410164013), Master's Degree Innovation Fund for Science and Technology of Shenyang Medical College (Y20250501).

Data availability All data is available in the main text or the supplementary information. The re-sequencing data of 271 accessions have been uploaded to the National Center for Biotechnology Information (NCBI) with the BioProject accession PRJNA1207598.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

- Ren G, Zhang X, Li Y, Ridout K, Serrano-Serrano ML, Yang Y, Liu A, Ravikanth G, Nawaz MA, Mumtaz AS, et al. Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*. *Sci Adv*. 2021;7(29):eabg2286.
- Xian-you S. Application evaluation of introduced monoecious hemp varieties. *Plant Fiber Sci China*. 2011;33(6):277–80.
- Xie Z, Mi Y, Kong L, Gao M, Chen S, Chen W, Meng X, Sun W, Chen S, Xu Z. *Cannabis sativa*: origin and history, glandular trichome development, and cannabinoid biosynthesis. *Hortic Res*. 2023;10(9):uhad150.
- Veljković VB, Kostić MD, Stamenković OS. Camelina seed harvesting, storing, pretreating, and processing to recover oil: A review. *Ind Crops Prod*. 2022;178:114539.
- Crini G, Lichtfouse E, Chanet G, Morin-Crini N. Applications of hemp in textiles, paper industry, insulation and building materials, horticulture, animal nutrition, food and beverages, nutraceuticals, cosmetics and hygiene, medicine, agrochemistry, energy production and environment: a review. *Environ Chem Lett*. 2020;18(5):1451–76.
- Henson JD, Vitetta L, Hall S. Tetrahydrocannabinol and cannabidiol medicines for chronic pain and mental health conditions. *Inflammopharmacology*. 2022;30(4):1167–78.
- Brand EJ, Zhao Z. Cannabis in Chinese medicine: Are some traditional indications referenced in ancient literature related to Cannabinoids? *Front Pharmacol*. 2017;8:108.
- John MJ, Thomas S. Biofibres and biocomposites. *Carbohydr Polym*. 2008;71(3):343–64.
- Chen X, Guo HY, Zhang QY, Wang L, Guo R, Zhan YX, Lv P, Xu YP, Guo MB, Zhang Y, et al. Whole-genome resequencing of wild and cultivated cannabis reveals the genetic structure and adaptive selection of important traits. *BMC Plant Biol*. 2022;22(1):371.
- Russo EB, Jiang HE, Li X, Sutton A, Carboni A, del Bianco F, Mandolino G, Potter DJ, Zhao YX, Bera S, et al. Phytochemical and genetic analyses of ancient cannabis from Central Asia. *J Exp Bot*. 2008;59(15):4171–82.
- Bai Y, Jiang M, Xie T, Jiang C, Gu M, Zhou X, Yan X, Yuan Y, Huang L. Archaeobotanical evidence of the use of medicinal cannabis in a secular context unearthed from south China. *J Ethnopharmacol*. 2021;275:114114.
- Li H-L. The origin and use of cannabis in eastern asia linguistic-cultural implications. *Econ Bot*. 1974;28(3):293–301.
- Sun X. Research progress on cannabinoids in cannabis (*Cannabis sativa* L.) in China. *Molecules*. 2023;28(9):3806.
- Amaducci S, Scordia D, Liu FH, Zhang Q, Guo H, Testa G, Cosentino SL. Key cultivation techniques for hemp in Europe and China. *Ind Crops Prod*. 2015;68:2–16.
- McPartland JM. Cannabis systematics at the levels of family, genus, and species. *Cannabis Cannabinoid Res*. 2018;3(1):203–12.
- Bonini SA, Premoli M, Tambaro S, Kumar A, Maccarinelli G, Memo M, Mastinu A. *Cannabis sativa*: A comprehensive ethnopharmacological review of a medicinal plant with a long history. *J Ethnopharmacol*. 2018;227:300–15.
- Zhang Q, Chen X, Guo H, Trindade LM, Salentijn EMJ, Guo R, Guo M, Xu Y, Yang M. Latitudinal adaptation and genetic insights into the origins of *Cannabis sativa* L. *Front Plant Sci*. 2018;9:1876.
- Lynch RC, Padgitt-Cobb LK, Garfinkel AR, Knaus BJ, Hartwick NT, Allsing N, Aylward A, Bentz PC, Carey SB, Mamerto A, et al. Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome. *Nature*. 2025;643(8073):1001–10.
- Sakamoto K, Abe T, Matsuyama T, Yoshida S, Ohmido N, Fukui K, Satoh S. RAPD markers encoding retrotransposable elements are linked to the male sex in *Cannabis sativa* L. *Genome*. 2005;48(5):931–6.
- Zhang J, Yan J, Huang S, Pan G, Chang L, Li J, Zhang C, Tang H, Chen A, Peng D, et al. Genetic diversity and population structure of cannabis based on the genome-wide development of simple sequence repeat markers. *Front Genet*. 2020;11:958.
- Gao C, Xin P, Cheng C, Tang Q, Chen P, Wang C, Zang G, Zhao L. Diversity analysis in *Cannabis sativa* based on large-scale development of expressed sequence tag-derived simple sequence repeat markers. *PLoS ONE*. 2014;9(10):e110638.
- Ioannidis K, Tomprou I, Mitsis V, Koropouli P. Genetic Evaluation of in vitro micropropagated and regenerated plants of *Cannabis sativa* L. using SSR molecular markers. *Plants (Basel)*. 2022;11(19):2569.
- Zhang LG, Chang Y, Zhang XF, Guan FZ, Yuan HM, Yu Y, Zhao LJ. Analysis of the genetic diversity of Chinese native *Cannabis sativa* cultivars by using ISSR and chromosome markers. *Genet Mol Res*. 2014;13(4):10490–500.
- McPartland JM: Commentary on: Datwyler SL, Weiblen GD. Weiblen. Genetic variation in hemp and marijuana (*Cannabis sativa* L.) according to amplified fragment length polymorphisms. *J Forensic Sci*. 2006;51:371–5. *J Forensic Sci* 2006, 51(6):1405; author reply 1406.
- Peil A, Flachowsky H, Schumann E, Weber WE. Sex-linked AFLP markers indicate a pseudoautosomal region in hemp (*Cannabis sativa* L). *Theor Appl Genet*. 2003;107(1):102–9.
- Solano J, Anabalon L, Figueroa A, Gangitano D. ITS barcoding using high resolution melting analysis of *Cannabis sativa* drug seizures in Chile: A forensic application. *Forensic Sci Int*. 2020;316:110550.
- Xia RC, Zhang XC, Wang XX, Yang Q, Chen C, Yu H, Qu YL, Wang ZW, Shi Y, Xiang P, et al. Identification of *Cannabis Sativa* L. based on *rbcL* sequence. *Fa yi xue za zhi*. 2021;37(2):187–91.

28. Metouekel A, Badrana F, Kachkoul R, Chebaibi M, Akhazzane M, El Moussaoui A, Touil N, El Amri H, El Fahime E, El Kazzouli S, et al. Genetic characterization and chemical identification of moroccan *Cannabis sativa* (L.) Seeds: Extraction, and in vitro and in silico biological evaluation. *Plants (Basel)*. 2024;13(14):1938.
29. Roman MG, Gangitano D, Houston R. Characterization of new chloroplast markers to determine biogeographical origin and crop type of *Cannabis sativa*. *Int J Legal Med*. 2019;133(6):1721–32.
30. Wu GA, Terol J, Ibanez V, Lopez-Garcia A, Perez-Roman E, Borreda C, Domingo C, Tadeo FR, Carbonell-Caballero J, Alonso R, et al. Genomics of the origin and evolution of *Citrus*. *Nature*. 2018;554(7692):311–6.
31. Fan W, Lu J, Pan C, Tan M, Lin Q, Liu W, Li D, Wang L, Hu L, Wang L, et al. Sequencing of Chinese castor lines reveals genetic signatures of selection and yield-associated loci. *Nat Commun*. 2019;10(1):3418.
32. Wu D, Liang Z, Yan T, Xu Y, Xuan L, Tang J, Zhou G, Lohwasser U, Hua S, Wang H, et al. Whole-genome resequencing of a worldwide collection of rapeseed accessions reveals the genetic basis of ecotype divergence. *Mol Plant*. 2019;12(1):30–43.
33. Xanthopoulou A, Manioudaki M, Bazakos C, Kissoudis C, Farsakoglou AM, Karagiannis E, Michailidis M, Polychroniadou C, Zambounis A, Kazantzis K, et al. Whole genome re-sequencing of sweet cherry (*Prunus avium* L.) yields insights into genomic diversity of a fruit species. *Hortic Res-England*. 2020;7(1):60.
34. Wang X, Hu Y, He W, Yu K, Zhang C, Li Y, Yang W, Sun J, Li X, Zheng F, et al. Whole-genome resequencing of the wheat A subgenome progenitor *Triticum urartu* provides insights into its demographic history and geographic adaptation. *Plant Commun*. 2022;3(5):100345.
35. Dong Y, Duan S, Xia Q, Liang Z, Dong X, Margaryan K, Musayev M, Goryslavets S, Zdunić G, Bert PF, et al. Dual domestications and origin of traits in grapevine evolution. *Science*. 2023;379(6635):892–901.
36. Chen R, Chen K, Yao X, Zhang X, Yang Y, Su X, Lyu M, Wang Q, Zhang G, Wang M et al. Genomic analyses reveal the stepwise domestication and genetic mechanism of curd biogenesis in cauliflower. *Nat Genet*. 2024;1235–44.
37. Grassa CJ, Weiblen GD, Wenger JP, Dabney C, Poplawski SG, Timothy Motley S, Michael TP, Schwartz CJ. A new *Cannabis* genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytol*. 2021;230(4):1665–79.
38. Wei H, Yang Z, Niyitanga S, Tao A, Xu J, Fang P, Lin L, Zhang L, Qi J, Ming R, et al. The reference genome of seed hemp (*Cannabis sativa*) provides new insights into fatty acid and vitamin E synthesis. *Plant Commun*. 2024;5(1):100718.
39. Gao S, Wang BS, Xie SS, Xu XY, Zhang J, Pei L, Yu YY, Yang WF, Zhang Y. A high-quality reference genome of wild *Cannabis sativa*. *Hortic Res-England*. 2020;7(1):73.
40. Lavery KU, Stout JM, Sullivan MJ, Shah H, Gill N, Holbrook L, Deikus G, Sebra R, Hughes TR, Page JE, et al. A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci. *Genome Res*. 2019;29(1):146–56.
41. Uzunova M, Ecker W, Weissleder K, Röbbelen G. Mapping the genome of rapeseed (*Brassica napus* L.). I. Construction of an RFLP linkage map and localization of QTLs for seed glucosinolate content. *Theor Appl Genet*. 1995;90(2):194–204.
42. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884–90.
43. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv: Genomics*. 2013.
44. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. Twelve years of SAMtools and BCFtools. *GigaScience*. 2021;10(2):giab008.
45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPPD. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25(16):2078–9.
46. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The genome analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297–303.
47. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80–92.
48. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559–75.
49. Retief JD. Phylogenetic analysis using PHYLIP. *Methods Mol Biol*. 2000;132:243–58.
50. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res*. 2021;49(W1):W293–6.
51. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655–64.
52. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet*. 2011;88(1):76–82.
53. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27(15):2156–8.

54. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21(2):263–5.
55. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet*. 2012;8(11):e1002967.
56. Chen H, Patterson N, Reich D. Population differentiation as a test for selective sweeps. *Genome Res*. 2010;20(3):393–402.
57. Pellati F, Brighenti V, Sperlea J, Marchetti L, Bertelli D, Benvenuti S. New methods for the comprehensive analysis of bioactive compounds in *Cannabis sativa* L. (hemp). *Molecules*. 2018;23(10):393–402.
58. International Council for Harmonisation. Validation of analytical procedures: Text and methodology Q2(R1). International Council for Harmonisation; 2005.
59. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*. 2012;44(7):821–4.
60. Weiblen GD, Wenger JP, Craft KJ, ElSohly MA, Mehmedic Z, Treiber EL, Marks MD. Gene duplication and divergence affecting drug content in *Cannabis sativa*. *New Phytol*. 2015;208(4):1241–50.
61. Faux AM, Draye X, Flamand MC, Occre A, Bertin P. Identification of QTLs for sex expression in dioecious and monoecious hemp (*Cannabis sativa* L). *Euphytica*. 2016;209(2):357–76.
62. Pisupati R, Vergara D, Kane NC. Diversity and evolution of the repetitive genomic content in *Cannabis sativa*. *BMC Genomics*. 2018;19(1):156357–376.
63. Roman MG, Gutierrez R, Houston R. Massively parallel sequencing of *Cannabis sativa* chloroplast hotspots for forensic typing. *J Cannabis Res*. 2022;4(1):13.
64. Cheng YC, Houston R. Evaluation of the trnK-matK-trnK, ycf3, and accD-psal chloroplast regions to differentiate crop type and biogeographical origin of *Cannabis sativa*. *Int J Legal Med*. 2021;135(4):1235–44.
65. Di Nunzio M, Barrot-Feixat C, Gangitano D. Characterization and evaluation of nine *Cannabis sativa* chloroplast SNP markers for crop type determination and biogeographical origin on European samples. *Forensic Sci Int Genet*. 2024;68:102971.
66. Ma X, Wu H, Liu B, Wang S, Zhang Y, Su M, Zheng B, Pan H, Du B, Wang J, et al. Genomic diversity, population structure, and genome-wide association reveal genetic differentiation and trait improvements in mango. *Hortic Res*. 2024;11(7):uhae153.
67. Zhang Q, Singh VP, Li J, Jiang F, Bai Y. Spatio-temporal variations of precipitation extremes in Xinjiang, China. *J Hydrol*. 2012;434–435:7–18.
68. Dutoit L, Burri R, Nater A, Mugal CF, Ellegren H. Genomic distribution and estimation of nucleotide diversity in natural populations: perspectives from the collared flycatcher (*Ficedula albicollis*) genome. *Mol Ecol Resour*. 2017;17(4):586–97.
69. McPartland JM, Hegman W, Long T. Cannabis in Asia: its center of origin and early cultivation, based on a synthesis of subfossil pollen and archaeobotanical studies. *Veg History Archaeobotany*. 2019;28(6):691–702.
70. Rubin CJ, Zody MC, Eriksson J, Meadows JR, Sherwood E, Webster MT, Jiang L, Ingman M, Sharpe T, Ka S, et al. Whole-genome resequencing reveals loci under selection during chicken domestication. *Nature*. 2010;464(7288):587–91.
71. Onofri C, de Meijer EPM, Mandolino G. Sequence heterogeneity of cannabidiolic- and tetrahydrocannabinolic acid-synthase in *Cannabis sativa* L. and its relationship with chemical phenotype. *Phytochemistry*. 2015;116:57–68.
72. Zirpel B, Kayser O, Stehle F. Elucidation of structure-function relationship of THCA and CBDA synthase from *Cannabis sativa* L. *J Biotechnol*. 2018;284:17–26.
73. Russo EB. Taming THC: potential cannabis synergy and phytocannabinoid-terpenoid entourage effects. *Br J Pharmacol*. 2011;163(7):1344–64.
74. Guo Z, Liu G, Röder MS, Reif JC, Ganai MW, Schnurbusch T. Genome-wide association analyses of plant growth traits during the stem elongation phase in wheat. *Plant Biotechnol J*. 2018;16(12):2042–52.
75. Mostafaei Dehnavi M, Damerum A, Taheri S, Ebadi A, Panahi S, Hodgins G, Brandley B, Salami SA, Taylor G. Population genomics of a natural *Cannabis sativa* L. collection from Iran identifies novel genetic loci for flowering time, morphology, sex and chemotyping. *BMC Plant Biol*. 2025;25(1):80.
76. Singh A, Bilichak A, Kovalchuk I. The genetics of Cannabis—genomic variations of key synthases and their effect on cannabinoid content. *Genome*. 2021;64(4):490–501.
77. Hurgobin B, Tamiru-Oli M, Welling MT, Doblin MS, Bacic A, Whelan J, Lewsey MG. Recent advances in *Cannabis sativa* genomics research. *New Phytol*. 2021;230(1):73–89.
78. Henry P, Khatodia S, Kapoor K, Gonzales B, Middleton A, Hong K, Hilyard A, Johnson S, Allen D, Chester Z, et al. A single nucleotide polymorphism assay sheds light on the extent and distribution of genetic diversity, population structure and functional basis of key traits in cultivated north American cannabis. *J Cannabis Res*. 2020;2(1):26.
79. de Meijer EP, Bagatta M, Carboni A, Crucitti P, Moliterni VM, Ranalli P, Mandolino G. The inheritance of chemical phenotype in *Cannabis sativa* L. *Genetics*. 2003;163(1):335–46.

80. Kojoma M, Seki H, Yoshida S, Muranaka T. DNA polymorphisms in the tetrahydrocannabinolic acid (THCA) synthase gene in drug-type and fiber-type *Cannabis sativa* L. Forensic Sci Int. 2006;159(2–3):132–40.
81. Sawler J, Stout JM, Gardner KM, Hudson D, Vidmar J, Butler L, Page JE, Myles S. The genetic structure of marijuana and hemp. PLoS ONE. 2015;10(8):e0133292.
82. Abdelaziz M, El Alaoui MA, Melloul M, Hakki E, Udupa S, Stambouli H, Bouri A, Amine S, Fahime E. Study of Moroccan *Cannabis sativa* DNA polymorphism in the THCA synthase gene from seized Moroccan cannabis resin (Hashish). J Plant Biology Res. 2016;5:1–11.

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

Authors and Affiliations

Baishi Wang^{1,2} · Shan Gao³ · Xue Wang^{1,2} · Quanzheng Yun⁴ · Hao Nie³ · Ning Li^{1,2} · Yonggang Li^{1,2} · Deri Feng^{1,2} · Linsen Li^{1,2} · Tianye Li^{1,2} · Xu Lu^{1,2} · Jingjing Liu^{1,2} · Xuge Wei^{1,2} · Yanhong Mu^{1,2} · Qiqi Wu^{1,2} · Jinyuan Hu^{1,2} · Fangru Liu^{1,2} · Rui Feng^{1,2} · Jianyu Yin^{1,2} · Qikun Zhang⁴ · Ying Zhang³ · Fu Ren^{1,2}

✉ Qikun Zhang
zhangqikun@kaitaibio.com

✉ Ying Zhang
ying_zh@126.com

✉ Fu Ren
rf@symc.edu.cn

¹ Department of Biology, Shenyang Medical College, Shenyang, Liaoning 110034, People's Republic of China

² Liaoning Province Key Laboratory for Phenomics of Human Ethnic Specificity and Critical Illness (LPKL-PHESCI), Shenyang, Liaoning 110034, People's Republic of China

³ Institute of Forensic Science, Ministry of Public Security, No. 17 South Muxidi Lane, Xicheng District, Beijing 100038, People's Republic of China

⁴ Kaitai-bio Company, Hangzhou, Zhejiang 310000, People's Republic of China