Population genomics of a natural *Cannabis sativa* L. collection from Iran identifies novel genetic loci for flowering time, morphology, sex and chemotyping

Mahboubeh Mostafaei Dehnavi [1, 2], Annabelle Damerum [1, $], Sadegh Taheri [3], Ali Ebadi [2], Shadab Panahi [2], George Hodgin [4], Brian Brandley [4], Seyed Alireza Salami [2, 5*], Gail Taylor [1*]

1. Department of Plant Sciences, University of California, Davis, CA, USA
2. Department of Horticultural Science, Faculty of Agriculture, University of Tehran, Karaj, Iran
3. Department of Animal Science, Faculty of Agriculture, Ferdowsi University of Mashhad, Mashhad, Iran
4. Biopharmaceutical Research Company, Castroville, CA
5. Industrial and Medical Cannabis Research Institute (IMCRI), Tehran, Iran

[$] Current address: Zymo Research Corp, Irvine, CA, USA

[=*] Corresponding authors: asalami@ut.ac.ir; gtaylor@ucdavis.edu

Abstract

Future breeding and selection of *Cannabis sativa* L. for drug production and industrial purposes require a source of germplasm with wide genetic variation, such as that found in wild relatives and progenitors of highly cultivated plants. Limited directional selection and breeding have occurred in this crop, especially informed by molecular markers. Here, we investigated the population genomics of a natural cannabis collection of male and female individuals from differing climatic zones in Iran. Using Genotyping-By-Sequencing (GBS), we sequenced 228 genotypes from 35 populations. The results obtained from GBS were used to perform association analysis identifying links between genotype and important phenotypes, including inflorescence characteristics, flowering time, plant morphology, tetrahydrocannabinol (THC) content, cannabidiol (CBD) content and sex. Approximately 23,266 significant SNPs of high quality were detected to establish associations between markers and traits, and population structure showed that Iranian cannabis plants fall into five groups. A comparison of Iranian samples from this study to global data suggests that the Iranian population is distinctive and, in general, is closer to marijuana than to hemp, although some populations in this collection are closer to hemp. The GWAS results showed that novel genetic loci, not previously identified, contribute to sex, yield and chemotype traits in cannabis and are worthy of further study.

Keywords: Genotyping-By-Sequencing, GWAS, Cannabis Breeding, Genetic Diversity, Population Structure, Candidate Gene, Phenotype, Sex.

Introduction

Cannabis sativa L. (cannabis) from the Cannabinaceae family can be used as a source of both pharmacologic drugs for the treatment of tumors, schizophrenia and other medical conditions, but also as fiber and oil, depending on the quantities of tetrahydrocannabinol (THC) and cannabidiol (CBD) within any particular plant, landrace or cultivar (Fischedick J, 2015; Andre et al., 2016). *C. sativa* has a long history of cultivation, and it has been suggested that the global cannabis market may be valued annually at over $300 billion in coming years, as many US states and global nations de-regulate the use of this plant-based chemical for pharmaceutical and recreational use (Weiblen et al., 2015; Grassa et al., 2021; Mostafaei Dehnavi et al., 2022). Much remains to be discovered concerning the diversity of genetic and chemical signatures across the species of *C. sativa*, and it seems likely that wild populations of previously uncharacterized *C. sativa* can provide a valuable source of natural genetic variation as foundational resources for future directed breeding programs (Zhang et al. 2020; Mostafaei Dehnavi et al. 2022). The plant is an annual species, primarily dioecious and exhibits high levels of heterozygosity. It has a diploid genome (2n=20) estimated to be 843 Mb for male plants and 818 Mb for female plants. Additionally, the species is believed to

possess almost 30,000 genes (Jenkins and Orsburn, 2019b, 2019a; Henry et al., 2020; Hurgobin et al., 2021).

Depending on type and cannabinoid yield, in particular, THC: CBD ratio, this species can be defined as industrial hemp (a major source of textiles, food, and oilseed) or marijuana (medical cannabis or a recreational drug) (Piluzza et al., 2013; Ren et al., 2021; Mostafaei Dehnavi et al., 2022). There is significant potential in the development of cannabis and its derivatives, particularly CBD, in the treatment of melanoma, a type of skin cancer and epilepsy-related syndromes such as Lennox-Gastaut syndrome and Dravet syndrome (Bachari et al., 2020; Brunetti et al., 2020).

Cannabis breeding to date has been mostly outside of the public domain; therefore, the true genetic diversity of commercial varieties is unknown (Hurgobin et al., 2021). Unraveling the genetic information of natural cannabis populations facilitates breeding programs for different industrial and medical purposes (Mostafaei Dehnavi et al., 2022). Genotyping by sequencing (GBS) is a highly multiplexed and high-throughput method for determining the genetic structure of an individual or a population (Elshire et al., 2011; Sonah et al., 2013). This technique has played a significant role in the advancement of our understanding of the genetic diversity, evolution and breeding of cannabis (Sawler et al., 2015). Cannabis, as a complex plant species, has a high degree of genetic diversity, which has made it an ideal candidate for GBS studies. One of the main goals of GBS studies in cannabis is to understand links between phenotype and genotype and identify the genetic markers associated with important and complex traits (Pootakham et al., 2015).

By phenotyping a large number of plants, researchers can identify the genetic variations that are associated with specific traits and ultimately help to improve the breeding process. The phenotyping of traits such as sex, flowering time, cannabinoids production, flower structure, agronomic-related and disease and pest resistance is crucial for the cannabis industry (McKernan et al., 2020; Petit et al., 2020a; Pépin et al., 2021). Accurately identifying and characterizing these traits can inform breeding programs for the development of high-quality cannabis varieties with desirable traits. Early sex determination and the identification of molecular markers associated with sex are critical tools for cannabis growers and breeders looking to produce high-quality and high-yielding crops (Faux et al., 2016; Prentout et al., 2019). Furthermore, understanding the flowering time of a plant is a significant characteristic for not only optimizing yield and determining harvest time, but also

the fiber quality and cannabinoids produced (Toth et al., 2022). Flower structure and plant height can also affect the cultivation process and crop yield and cannabinoid production, such as THC and CBD, is a key factor in the medicinal and recreational use of cannabis (Gonçalves et al., 2019; Deguchi et al., 2022; Mostafaei Dehnavi et al., 2022). Therefore, the precise phenotyping of these traits can improve crop management, increase yield and enhance the overall quality of cannabis (Woods et al., 2021).

Another important application of GBS in cannabis is in the identification and characterization of landrace cultivars. These cultivars have unique genetic profiles and are important sources of diversity for breeding programs (Mostafaei Dehnavi et al., 2022). GBS studies have helped identify the genetic relationships between landraces and characterize their genetic diversity (Soorni et al., 2017).

Genetic investigation of different natural resources assists in the development of pre-breeding and the identification of new varieties for research purposes and enables the initiation of a pipeline of novel discovery toward commercialization (Gali et al., 2019; Kovalchuk et al., 2020). In this study, we conducted genotyping of both male and female cannabis genomes to enhance our comprehension of cannabis sex evolution, as well as cannabinoid expression. We used GBS to characterize natural cannabis plant material obtained from various regions of Iran. We employed a set of 23,266 significant SNP markers, which were linked to various essential features, such as THC and CBD content (which distinguishes the drug from the hemp chemotype), sex expression, flowering time, female inflorescence features and some other morpho-physiological traits such as plant height, number of nodes, number of leaves, internode length and footstalk diameter. The investigation offers insight into population structure, genetic relationship, and genetic diversity of the cannabis species.

MATERIALS AND METHODS

Plant material and field experiment

A total of 35 natural cannabis populations sourced from various locations in Iran, obtained from CGRC (www.medcannabase.org), were grown in the research field of the University of Tehran. The cultivation followed a randomized complete block design with three replicates per population. The

separation between each block was set at 2.5 m. Within each block, three rows were arranged for each plot (population), each extending 10 m in length, and with a row spacing of 60 cm and plant spacing of 90 cm, with a total of 10 plants planted across each row. A drip irrigation system was implemented, and plants were grown in soil amended with compost and fertilized with a balanced, water-soluble fertilizer (N, P, K) (Bhattarai and Midmore, 2014). All plants were grown under natural light conditions from April to September 2019. Throughout the growth period, daytime temperatures ranged from 31-36°C and nighttime temperatures ranged from 20-24°C, along with average daytime relative humidity fluctuated between 29% and 43%, and nighttime relative humidity ranged from 47% and 65%. To counter the impact of extreme heat and high evaporation during the leaf formation phase, the site received regular irrigation of 3-4 hours. After leaf growth, irrigation occurred three times a week, each session lasting 4-5 hours. The plants were grown to maturity, at which point they were harvested. Specific details for each population are given in **Table 1**. Collection sites and various climatic zones of the studied populations, are shown in **Figure 1**.

TABLE 1. Panel of 228 *Cannabis sativa* L. and geographical and ecological parameters of the cannabis populations studied.

Origin	Province	Region	Code	Population Size	Elevation (m)	Longitude (E)	Latitude (N)	Annual Rainfall (mm)	Annual Avg. temp. (C)
Iran	Zanjan	Abhar	Abh-01	8	1543	49.22	36.28	301.31	12.39
Iran	Qazvin	Qazvin	Qzv-01	4	1315	49.86	36 47	311.85	13.91
Iran	Qazvin	Qazvin	Qzv-02	12	1315	50.12	36.74	310.2	14.2
Iran	Hamadan	Samen	Sam-01	7	1858	48.71	34.20	324.79	13.53
Iran	Hamadan	Samen	Sam-02	3	1858	48.19	34.50	309.2	13.10
Iran	Khuzestan	Dezful	Dez-01	4	144	48.42	32.38	389.40	24.56
Iran	Sistan & Baluchistan	Zahedan	Zah-01	3	1352	60.86	29.49	73.58	19.30
Iran	Hamadan	Malayer	Mal-01	6	1729	48.82	34.29	456.98	12.43
Iran	Kurdistan	Saghez	Sqz-01	7	1480	46.26	36.24	439.37	11.20
Iran	Kurdistan	Sannandaj	San-02	6	1464	46.97	35.33	437.9	14.3
Iran	Kerman	Sirjan	Sir-01	10	1754	55.68	29.43	138.52	17.81
Iran	Hamadan	Nahavand	Nhv-01	7	1666	48.25	34.15	385.29	14.72
Iran	Fars	Fars	Frs-01	8	1390	53.71	27.38	312.3	16.86
Iran	Ardabil	Mugan plain	Dshm-01	4	1339	47.87	39.66	398.43	9.12
Iran	Kerman	Kerman	Krmn-01	4	1761	56.58	30.15	123.43	17.01
Iran	Kermanshah	Kermanshah	Krsh-01	6	1389	47.03	34.19	402.63	15.51
Iran	Kurdistan	Baneh	Ban-01	2	1503	45.53	35.59	660.88	14.26
Iran	West Azerbaijan	Urmiyeh	Urm-01	7	1362	45.07	37.54	327.1	11.2
Iran	Kermanshah	Gahvareh	Gahv-01	5	1476	46.41	34.34	402.63	15.51
Iran	Markazi	Arak	Ark-02	2	1722	49.68	34.09	326.6	13.7
Iran	Markazi	Mahalat	Mahl-01	5	1746	50.44	33.90	280.9	14.3
Iran	South Khorasan	Boshrouyeh	Bsh-01	6	881	57.43	34.03	79.98	21.07
Iran	Qom	Qom	Qom-02	2	933	50.87	34.65	131.9	18.3
Iran	Esfahan	Radan	Rad-01	6	1571	52.52	33.20	127.87	16.4
Iran	South Khorasan	Tabas	Tab-01	9	981	56.92	33.59	79.98	21.07
Iran	Semnan	Shahroud	Shah-01	8	1381	54.96	36.42	137.8	18.2
Iran	West Azerbaijan	Uromiyeh	Urm-02	5	1328	45.30	37.39	327.1	11.2
Iran	East Azerbaijan	Tabriz	Tabr-01	6	1345	46.14	38.70	272.2	12.1
Iran	Sistan & Baluchistan	Saravan	Sarav-01	12	1352	62.33	27.36	73.58	19.30
Iran	Markazi	Khomein	Khom-01	10	1798	50.07	33.63	296.32	13.6
Iran	Kurdistan	Sannandaj	San-03	12	1464	47.52	35.78	437.9	14.3
Iran	Esfahan	Kashan	Kash-02	6	949	51.44	33.98	136.5	19.7
Iran	Fars	Shiraz	Shrz-03	7	1488	52.36	29.33	329.3	18.0
Iran	Fars	Shiraz	Shrz-04	13	1488	52.36	29.33	329.3	18.0
Iran	North Khorasan	Bojnoord	Boj-01	5	1112	57.18	37.29	255.1	13.2

Phenotyping of the GWAS Panel

For each population, randomly selected plants of the middle rows for each genotype were labeled, and the phenotypic variation of 13 traits was characterized. The number of plants employed per population varied based on availability, as highlighted in **Table 1**. Phenotypes assessed included inflorescence characteristics, flowering time, plant morphology, sex and biochemical trait analysis. Traits were (i) number of days that elapsed from germination to the initiation of flowering (DTF), (ii) the number of days from germination to appearance of approximately 50% flowering within a population (DT50), (iii) plant height (H; m) from the soil surface to the topmost terminal inflorescence before harvest, (iv) sex expression as a binary variable, i.e., male vs. female (M, F), (v) crown length, (CL; measurement of the length of the main stem from the soil surface to the lowest branch, cm), inflorescence features: (vi) inflorescence length (iL; measurement of the length of the main inflorescence in both male and female inflorescences, cm) as well as the (vii) number of lateral pistillate inflorescences (Li), (viii) internode length (intL, measurement of the length between two nodes, cm), (ix) number of nodes (nN), (x) footstalk diameter (FD) at the widest part of the base with digital calipers (cm), (xi) number of leaves (nL) and (xii) analysis of the Δ9- tetrahydrocannabinol (THC Q, %) and (xiii) cannabidiol (CBD Q, %) content of the plant material using HPLC. Air-dried pistillate inflorescences were collected prior to the seed development were analyzed for content (% dry weight (DW)) of the cannabinoid compounds Δ9-THC and CBD. Refer to our publication (Mostafaei Dehnavi et al., 2022) for information on the preparation of samples and the HPLC analysis of cannabinoids. Some characteristics, including Li, nL, DT50 and CL, were only measured in female individuals. Furthermore, it is noteworthy that all plants included in this study exhibited dioecious characteristics, with either male or female flowers. As a result, we recorded the sex expression (male or female) of each individual in every population.

DNA Extraction, Library Preparation and Genotyping of the GWAS Panel

The GBS method was utilized to genotype the GWAS panel, following the protocol delineated by Elshire et al. (2011). At the juvenile stage before sexual differentiation, leaf tissue was collected from a labeled single plant of each population, and small segments of the tissue were placed into 2 ml vials and freeze-dried. High molecular weight DNA was isolated from approximately 25 mg of freeze-dried tissue following a modified cetyl trimethyl ammonium

bromide (CTAB) protocol (Doyle & Doyle, 1987; Cullings, 1992), which included a step for RNase treatment, to remove any potential RNA contamination, as RNA can inhibit the DNA sequencing library preparation (https://dnatech.genomecenter.ucdavis.edu/faqs/which-dna-isolation-protocols-do-you-recommend-for-illumina-sequencing/). DNA extracts were quantified using a Qubit™ Fluorometer (ThermoFisher Scientific). Individual DNA samples were diluted to 10 ng/µl using 0.5 M Tris-EDTA (TE) buffer, pH 8.0. As the sex of the plants was unknown, to cover all allelic variation within populations and sexes, the genomic DNA was extracted from all available plants per population. 100 ng of each genomic DNA template (in a 10 µl volume) was used for library construction using a single digestion with restriction enzyme *ApekI* and ligated to unique 4-8 sequence barcode adapters. Five µl aliquots of adapter-ligated DNA samples were pooled in a single tube to produce 96-plex libraries. The pooled DNA was PCR-amplified using *Phusion® High-Fidelity PCR Kit* (NEB®), followed by purification with a *Monarch®* PCR & DNA Cleanup Kit (NEB®). Standard experimental conditions, as described by Elshire et al. (2011), were followed for restriction, ligation, and PCR amplification. The purified DNA library was quantified and validated using a Bioanalyzer (Agilent Technologies), and the 96-plex libraries were sequenced on a single lane of Illumina HiSeq™ 4000 platform as single-end 100 (SR100) base pair reads.

Demultiplexing, Data quality control, and Read filtering

The Stacks pipeline was used for GBS data analysis (Catchen et al., 2013). Demultiplexing and trimming the sequence reads were performed using *process_radtags* script, which trims adapter sequences and filters low-quality reads <50 bases. Samples with <100,000 reads were removed before analysis. To elucidate the genetic relationship among Iranian cannabis and marijuana and fibre type accessions, we integrated our data with two public datasets. This included marijuana data consisting of 81 samples and hemp data consisting of 43 samples originally prepared by Sawler et al. (2015) and obtained from the NCBI SRA BioProject: PRJNA285813. Additionally, we incorporated 95 cannabis samples including 70 from Iran, 2 from Afghanistan and 26 accessions provided by CGN and IPK, as previously reported by Soorni et al. (2017), and accessed from the BioProject: PRJNA419020.

Mapping, SNP Variant Calling and SNP Filtering

We used a reference-based pipeline for sample alignment to generate consensus sequences. Trimmed sequence reads were aligned to the reference *C. sativa* 'CBDRx' assembly (cs10 v.1.0) as the most complete and contiguous chromosome-level assembly available at the time of analysis (Hurgobin et al., 2021; Ren et al., 2021), using the sequence alignment tool bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml#the-bowtie2-aligner) with the very-sensitive-local settings. This resulted in an average mapping rate of approximately 80%. The *ref_map.pl* script within the Stacks environment was utilized to call genetic variants (SNPs), and the "*populations"* program of the Stacks pipeline was used to filter the identified SNPs and estimate population genetics statistics like the fixation index (F_{ST}) for genetic relationship analysis and Hardy-Weinberg equilibrium (hwe). The filtration criteria applied were as follows: requiring a locus to be present in a minimum of 10 populations for processing; setting a minimum of 50% individuals per population to process a locus for that population; necessitating a minimum of 50% individuals across populations for locus processing; and specifying a minimum minor allele frequency (MAF) of 0.05 for processing nucleotide sites at a locus. PLINK V 1.9 (Purcell et al., 2007) was used for further filtering for both datasets (derived from this study and the publicly available data). Individuals with a genotyping call rate < 99%, SNPs with a genotyping call rate < 99%, and those exhibiting significant deviations from Hardy-Weinberg equilibrium (P-value $<10^{-6}$) were excluded from the analysis (Taheri et al., 2023). Following this filtering process, the genetic data from both groups were combined, and only the SNPs that were common to both groups were selected. The set of obtained SNPs was used for subsequent analysis, including population structure analysis, heterozygosity and F_{ST} analysis and association analysis.

Population Structure Analysis

Admixture 1.3.0 (Alexander et al., 2009) was utilized to estimate the most likely number of clusters (K) into which the accessions could be grouped and their degree of admixture. The value of K that best fits the data was determined based on the lowest cross-validation (CV) error. Accessions were assigned to clusters based on the probabilities of belonging to one of the clusters derived from the matrix of contributions, Q. *Admixture* was run for each possible group number (K = 1 to 10). In addition, to visualize the genetic relationship and similarity among samples, a principal component analysis was carried out on a combined dataset of 431 samples.

The analysis utilized ggplot2 (V3.4.4) for plotting (Villanueva and Chen, 2019), plotrix (V3.8.4) for zooming the plot (Lemon et al., 2015), and tidyverse (V2.0.0) for eliminating duplicate samples (Wickham et al., 2019) in R (V4.3.1). Following the quality control filtration process, this dataset consisted of 196 Iranian samples from the current study, 93 cannabis samples previously studied by Soorni et al. (2017), as well as 47 individuals from a hemp population and 95 individuals from a marijuana population studied by Sawler et al. (2015).

Heterozygosity and F_{ST} analysis

Heterozygosity was estimated for each individual using PLINK V1.90 and then averaged within each group (Purcell et al., 2007). Additionally, we used R (V4.3.1; R Core Team, 2018) to generate plots.

Due to the limited number of individuals within some populations, we classified the studied Iranian populations into four larger groups based on their geographical distribution and associated climatic patterns. These groups include east and southeast, northeast, south and west and northwest populations. The F_{ST} value was calculated by fsthet package (Flanagan and Jones, 2017) in R (V4.3.1) for each pair of populations to measure genetic differentiation among populations. F_{ST} analysis was conducted three times: firstly, among the four geographic populations of Iranian samples; secondly, using combined data from this study and two public datasets containing previously studied Iranian samples, as well as other marijuana and hemp populations (NCBI SRA BioProject: PRJNA419020 and PRJNA285813), and in the third analysis, all Iranian populations from this study were treated as a single population and combined with two above-named public collections. The F_{ST} plots were created in R (V4.3.1) using the qqman package (Turner, 2014) to visualize the relationships between populations based on the F_{ST} values. We then conducted a thorough analysis of significant SNPs for each pair to identify the specific SNP markers contributing to the observed differences.

IBD Test

We employed PLINK V1.90 software to conduct pairwise IBD analysis, investigating first-degree and second-degree relationships among individuals by assessing the proportion of SNPs where zero, one, or two shared IBD alleles were present, represented by Z0, Z1, and Z2,

respectively. Subsequently, relatedness was quantified using the PI_HAT parameter, indicating the proportion of SNPs in IBD between individual pairs (Arab et al., 2019).

Association Analysis

To identify the associations between genetic varaints and trait performance, GWAS was carried out to estimate SNP effects. Studies indicates that employing a linear mixed model that incorporates population and family structures is currently the most effective approach for mitigating the impact of population stratification (Price et al., 2010). The statistical model used for GWAS analysis is based on the mixed linear model as follows:

$$y = Xb + Z\gamma + Mu + e$$

Where y is a vector representing the phenotype, b is a vector representing fixed effects (group), γ is a vector representing fixed effects of markers, and u is a vector representing random effects (e.g., PC1, PC2, and PI_HAT). X, Z, and M are matrices relating observations to the effects of fixed factors, fixed SNP effects, and random genetic effects, respectively, and e is a vector representing random residuals with e ~ N(0, I σe2).

The analysis initially utilized a mixed linear model through GCTA V1.94.1 software (Yang et al., 2011). Additionally, for undertaking GWAS for sex as a qualititative trait, we employed PLINK V1.90 with case and control anaylsis. Subsequently, we applied Bonferroni testing using PLINK's --assoc, --perm, and --adjust functions. Finally, Manhattan plots and the Q-Q plots were constructed in R (V4.3.1) using the qqman package (Turner, 2014) to visualize the genome-wide association signals.

Furthermore, GCTA V1.94.1 software was utilized to measure the SNP-based heritibility (h^2) for each trait (Zhu and Zhou, 2020). The variance of total additive genetic effects is defined as $\sigma_g^2 =$ p . σ_β^2. The GCTA software was used to estimate the variance components σ_g^2 and σ_e^2. The SNP heritability is estimated as follows:

$$h_g^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$$

Functional annotation of candidate SNPs

In addition, functional annotation of the candidate SNPs identified in the GWAS analysis was performed. For each trait, the significant markers were compared and annotated using the annotated reference genome (cs10, GCF_900626175.1, https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Cannabis_sativa/100/) and the annotated genes were identified using the NCBI Genome Data Viewer (https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_900626175.2) (Jenkins and Orsburn, 2019b).

RESULTS

Phenotyping, genotyping and data quality control

A normal distribution was observed for all traits except for DT50, THC Q, CBD Q, nL, and DTF (**Figure 2**). Sex was not included as it is not feasible due to its non-quantitative nature. We observed great phenotypic variation within our cannabis populations. The footstalk diameter exhibited substantial diversity, ranging from 1.62 to 3.66, with an average of 2.54, showing the highest coefficient of variation at 31.32%. In contrast, THC and CBD concentrations displayed narrower variations, ranging from 1.17 to 3.13% for THC and 0.96 to 6.69% for CBD, with averages of 1.96% and 1.65%, respectively. Consequently, they exhibited the lowest coefficients of variation at 0.51% and 0.52%, respectively (**Table 2**). The phenotypic data collected for the GWAS panel, along with the SNP-based heritability (h^2) measurments are summarized in **Table 2**. A total of 1.456 billion reads of 100 bp length were obtained from the four sequencing lanes on HiSeq™ 4000. After quality control and trimming for barcode adapter sequences, an average of approximately 4 million reads per sample were retained for the first (3,883,851 reads/sample; 80.6% of original reads), second (3,854,236 reads/sample; 82.8%), and third (3,935,551 reads/sample; 79.5%) libraries, while approximately 10 million reads (10,774,545 reads/sample; 64.3%) remained for the fourth library, which included only 35 pooled samples. The number of reads retained per sample ranged between 108K and 104M reads, with samples with <100K reads removed. The average quality scores, Q30 ratio and guanine–cytosine (GC) content of the reads were ~39, ~96% and 44.1%, respectively. On average, 80.11% of the reads aligned with the *C. sativa* cs10 reference genome (Hurgobin et al., 2021). Differences in sequencing depth

298 across regions, excessive PCR amplification, short read length, or issues with the sequencing
299 platform may have contributed to the variations observed in the percentage of reads mapped to
300 the reference. Specifically, the minimum mapping percentage recorded was 50.09%, which was
301 observed in the case of sample ID 218, representing one of the individuals from population Bsh-
302 01. Following filtration, a total of 23,266 high-quality SNPs were identified across Iranian
303 samples, which were subsequently selected for the analysis of population structure and marker-
304 trait association. Additionally, a set of 25,112 informative SNPs were retained after applying
305 filtration to the combined datasets, which included the previously sequenced Iranian samples and
306 collections of hemp and marijuana.

307

308 **TABLE 2.** Characteristics studied in the cannabis populations (sex trait has not been reported here as it is not a
309 quantitative trait). [a]SD -(standard deviation), [b]CV - (coefficient of variation), estimated as the ratio of the standard
310 deviation to the mean of all populations and [c]h^2- (SNP-based heritability), the proportion of phenotypic variance
311 explained by all measured SNPs.

No.	Trait	Abbr.	Unit	Min.	Max.	Mean	SD[a]	CV[b] (%)	h[2c]
1	Inflorescence length	iL	cm	23.78	31.32	27.86	3.84	14.32	0.426
2	Number of lateral inflorescences	Li	count	12.37	19.44	15.79	2.75	17.77	0.389
3	Number of nodes	nN	count	6.27	9.03	7.52	1.06	14.19	0.462
4	Height	H	m	1.27	2.68	1.92	0.49	27.37	0.341
5	Number of leaves	nL	count	7.57	8.57	8.19	0.54	6.66	0.385
6	Internode length	intL	cm	41.41	59.03	50.43	6.53	13.56	0.314
7	Footstalk diameter	FD	cm	1.62	3.66	2.54	0.73	31.32	0.382
8	Number of days to first blooming	DTF	day	101.55	136.10	121.71	12.87	10.54	0.301
9	Number of days to 50% blooming	DT50	day	117.5	142.35	129.97	11.45	8.96	0.339
10	Crown length	CL	cm	36.28	54.53	44.93	8.55	19.63	0.361
11	Δ9-THC Quantity	THCQ	%	1.17	3.13	1.96	0.92	0.51	0.342
12	CBD Quantity	CBDQ	%	0.96	6.69	1.65	0.70	0.52	0.373

312 **Population Structure Analysis**

313 Principal Component Analysis (PCA) findings indicated a high level of genetic similarity among
314 these populations (Figure 3A). The Sqz-01 population and some individuals of San-02 failed to
315 group with other clusters (Figure 3A). Admixture's cross-validation procedure was used to
316 determine the most likely number of genetic groups (K). The population structure of studied
317 samples was described by testing the probable number of clusters (K) from 1 to 10, with K= 5
318 selected as the optimal representation of ancestral populations based on the lowest cross-entropy

criterion and visualized using a Q estimates bar plot (Figure 3B-C). The PCA analysis of integrated data with two public datasets revealed that while the Iranian samples exhibited distinct genetic differences from the hemp and marijuana populations, they showed generally closer genetic proximity to the marijuana population. However, some individual Iranian samples exhibited a closer genetic resemblance to hemp (Figure 4A). These findings were further supported by the dendrogram plot, which indicated that the genetic distance between the Iranian samples from this study and the previously studied marijuana population was smaller than the distance between the Iranian samples from the current study and the previously studied hemp population (Figure 4B).

Genetic variation and differentiation

To assess the genetic differentiation among Iranian cannabis populations, we calculated the genetic differentiation parameter (F_{ST}) as well as observed and expected heterozygosity for each pairwise comparison. Initially, we generated six F_{ST} plots to compare distinct populations of Iranian samples (**Supplementary Figure 1**). The observed heterozygosity (H_o) for these groups was found as 0.25 (for the east and southeast population), 0.204 (for the northeast population), 0.214 (for the south population), and 0.228 (for the west and northwest). Corresponding, the respective expected heterozygosity (H_e) values were obtained as 0.288, 0.264, 0.286, and 0.296, while the estimated minor allele frequency (MAF) were 0.204, 0.19, 0.203, and 0.208. The details of observed heterozygosity, expected heterozygosity, and minor allele frequency are provided in **Figure 5.**

Our study exhibited higher heterozygosity than that earlier study on Iranian samples observed by Soorni et al. (2017). Our results showed a similar level of heterozygosity to the hemp accessions studied by Sawler et al. (2015), while the earlier Iranian sample study reported the heterozygosity, which is more similar to marijuana accessions studied by Sawler et al. (2015).

Supplementary Figures 1, **2**, **3** present F_{ST} Manhattan plots and correlation plots for all possible pairings of these populations. Significant SNP markers along with their corresponding loci, gene annotations, and the F_{ST} values across all F_{ST} analyses are provided in **Supplementary Tables 2-18**.

The pairwise comparisons among the geography-based Iranian populations in this study revealed that the east and southeast population and northeast populations had the highest number of SNP markers (95 SNPs) (**Supplementary Figure 1C**; **Supplementary Table 2**), followed by the northeast population and south population with 89 SNP markers which related to 32 specific loci (**Supplementary Figure 1F**; **Supplementary Table 3**). On the other hand, the east and southeast population and southern population exhibited the lowest number of SNPs (29 SNPs) that were linked to five specific loci (**Supplementary Figure 1D**; **Supplementary Table 5**).

When comparing Iranian populations with global collections, in contrast to the hemp collection, the marijuana collection exhibited a larger number of SNP markers concerning the geography-based Iranian populations. Among these comparisons, the northeast population of Iran demonstrated the highest number of SNP marker (134 SNPs) spanning 36 loci in comparison to the marijuana collection (**Supplementary Figure 2A**; **Supplementary Table 13**). Furthermore, the northeast population emerged with the highest number of SNP markers (112 SNPs) linked to 49 loci when compared to the hemp collection (**Supplementary Figure 2C**; **Supplementary Table 9**). Overall, the northeast population of Iran appeared to be more distinct not only from other populations within Iran but also from both the hemp and marijuana collections.

In a separate analysis, pairwise comparisons were conducted between the Iranian samples from this study as a single population and previously studied Iranian samples, as well as hemp and marijuana collections. It was found that there were less SNP markers shared between Iranian samples and hemp (107 SNPs) (**Supplementary Figure 3A**; **Supplementary Table 16**) than between Iranian samples and the marijuana collection (113 SNPs) (**Supplementary Figure 3B**; **Supplementary Table 17**). In both of these comparisons, a total of 37 loci were found to exhibit differences (**Supplementary Figure 3A, B**; **Supplementary Tables 16, 17**). Of these, six loci were identified as being common to both comparisons, including LOC115694687, which encodes separase (RefSeq accession: XM_030621771.1), LOC115707169 encoding for histone deacetylase 2 (RefSeq accession: XM_030635038.1), LOC115707184 encoding for spidroin-2 (RefSeq accession: XM_030635066.1), LOC115707237 encoding for serine/threonine protein phosphatase 2A 55 kDa regulatory subunit B (RefSeq accession: XM_030635130.1), LOC115725648 with an uncharacterized description (RefSeq accession: XM_030655228.1), and LOC115725736 encoding for partner of Y14 and mago (RefSeq accession: XM_030655329.1).

It is noteworthy that the comparisons unveiled 134 SNP markers associated with 51 loci that exhibited differences between the Iranian samples of this study and the previously studied Iranian samples. These SNPs were distributed across chromosomes 1 (n=5), 2 (n=11), 4 (n=31), 5 (n=27), 7 (n=6) and 10 (n=54) (**Supplementary Figure 3C**; **Supplementary Table 18**). This suggests the existence of genomic differences between these two sample sets.

In the first F_{ST} analysis conducted among the four geographical-based Iranian populations, the highest and lowest F_{ST} values were observed for the east and southeast: northeast pair (F_{ST}= 0.09) and northeast: west and northwest pair (F_{ST}= 0.024), respectively. A higher F_{ST} value indicates greater genetic differences between populations. Moving on to the second F_{ST} analysis, which incorporated global data, the pairs of marijuana: northeast demonstrated the lowest F_{ST} value (F_{ST}= 0.06), while the pair of hemp: east and southeast exhibited the highest F_{ST} value (F_{ST}= 0.17). These results suggest that Iranian populations displayed higher F_{ST} values compared to hemp rather than marijuana. Furthermore, in the last set of F_{ST} comparisons, it was found that the F_{ST} estimation between Iranian samples as a single population and the hemp collection (F_{ST}= 0.086) was higher compared to the F_{ST} value between Iranian samples and the marijuana population (F_{ST}= 0.062). Additionally, the F_{ST} value between the Iranian samples from this study and the Iranian samples from previous studies was 0.015. Overall, these findings indicate that Iranian populations are genetically closer to marijuana than hemp. Moreover, the SNP markers that revealed differences within Iranian populations were predominantly concentrated on Chromosomes 3, 4, and 1, indicating potential genomic regions that contribute to genetic variation and population differentiation. These specific chromosomes may harbor genes or regulatory elements that play a significant role in shaping the unique genetic landscape of Iranian populations, warranting further investigation to uncover the underlying genetic mechanisms and potential functional implications of these variations.

Genotype and phenotype associations

A total of 191 significant SNP associations were detected for the investigated traits. Among these, 47 SNPs were identified within annotated genes. Overall, the study revealed a total of 47 candidate loci related to the traits under investigation, out of which seven genes remain uncharecterized. Detailed information for all identified significant SNPs, is provided in

406 **Supplementary Table 1** and **Table 3** with associated candidate loci, and annotation information
407 presented specifically for those SNPs located within annotated genes. Specifically, 18 SNPs were
408 found to be associated with H (**Figure 6A**). Among these SNP markers, five markers, two on
409 chromosome 3 (chr3_281101_71, chr3_288389_58), two on chromosome 4 (chr4_327964_19,
410 chr4_382581_26), and one on chromosome 2 (chr2_84187_7) were linked to annotated genes
411 (**Supplementary Table 1**). 19 SNP markers were found to be associated with the nN, while 17
412 SNP markers were associated with the Li. Specifically, for the nN, there were five SNP markers
413 located on chromosomes 2, 4, 5 and 9 that were located within annotated genes (**Figure 6B**).
414 (**Supplementary Table 1**). Similarly, for the Li, there were two markers on chromosomes 5 and
415 9 that were identified within annotated genes (**Figure 6C**; **Supplementary Table 1**). Among 10
416 markers identified as being associated with the iL, a total of four SNP markers positioned on
417 chromosomes 2, 9 and 10 were found to be related to genes (**Figure 6D**; **Supplementary Table**
418 **1**).
419 Significant associations were observed between flowering time-related traits, including DTF and
420 DT50 (**Figures 6E**, **6F**). 16 markers were associated with DTF, while eight markers were found
421 to be associated with DT50 (**Figures 6E and 6F**; **Supplementary Table 1**). The SNPs
422 associated with DTF were found to be linked with candidate loci, including glutamate--
423 glyoxylate aminotransferase 2, YTH domain-containing protein ECT4, ribonuclease H2 subunit
424 B, and one uncharacterized gene. As for the DT50 trait, only one marker located on chromosome
425 5, encoding the putative pentatricopeptide repeat-containing protein At2g02150, was identified
426 (**Supplementary Table 1**).
427 Among 17 SNP markers that were detected to be associated with the FD, six SNPs, two on
428 chromosome 5 and one each on chromosomes 2, 3, 4, and 10 were linked to annotated genes
429 (**Figure 6G**; **Supplementary Table 1**). Also, four SNP markers on chromosomes 3 (n=2), 5
430 (n=1) and 6 (n=1) from a total of seven markers were detected to be associated with annotated
431 genes for the nL trait (**Figure 6H**; **Supplementary Table 1**).
432 For the intL trait, seven markers positioned on chromosomes 3, 4 and 5 were associated with
433 genes (**Figure 7A**; **Supplementary Table 1**). These markers were part of a total of 35 markers
434 identified to be linked with this trait. Regarding the CL trait, among 12 markers that were found
435 to be associated with it, five markers- three on chromosome 3 and two on chromosome 4 were
436 related to annotated genes (**Figure 7B**; **Supplementary Table 1**).

Eight markers, two on chromosome 1, two on chromosome 3, two on chromosome 8 and one each on chromosomes 7 and 10 were detected to be associated with THC Q trait (**Figure 7C**; **Supplementary Table 1**). Furthermore, in the case of CBD Q, a total of 24 SNPs were identified. These markers were distributed across whole genome (**Figure 7D**). The candidate loci identified to be associated with CBD Q trait included the berberine bridge enzyme-like D-2 (chromosome 7), 3-ketoacyl-CoA synthase 19 (chromosome 1), serine/threonine-protein kinase AtPK1/AtPK6 both (chromosome 3), fimbrin-2 (chromosome 4), endonuclease MutS2 (chromosome 5), small nucleolar RNA R71 (chromosome 2), transcription factor MYB14 (chromosome 9), and one uncharacterized locus (chromosome 4) (**Supplementary Table 1**).

It appears that THC and CBD concentrations have complex genetic architectures that extend beyond the already identified cannabinoid synthase genes and are distributed across various chromosomes of the whole genome.

The SNPs associated with sex determination were primarily distributed on the sex chromosome (cs10 v.1.0 chromosome 1 and cs10 v.2.0 chromosome 10), previously identified by Prentout et al. 2019 and McKernan et al. 2020, but additional SNPs were also identified at different positions. Eight SNP markers discovered for this trait, which are located on chromosomes 1 (n=3), 2 (n=2), 9 (n=1) and 10 (n=2). Among these, two SNPs located across chromosomes 2 and 9 were found to be specifically linked to candidate genes associated with sex (**Table 3**; **Figure 7E**). The Manhattan plots, along with the corresponding quantile-quantile (Q-Q) plots for each trait, were presented in Figures **6A-6H** and **7A-7E**.

457 **TABLE 3**. Functional annotations of the significantly associated SNPs for sex trait in Iranian cannabis collection
458

CHR	SNP ID	Position	*P*-value	MAF	Gene ID	RefSeq Accession	Gene Annotation
1	786516_36	4175218	0.000969	0.33			
1	786516_49	4175231	0.000969	0.33			
1	786516_53	4175235	0.000969	0.33			
2	78363_79	87241293	0.000352	0.42	LOC115721262	XM_030650525.2	uncharacterized LOC115721262
2	109511_15	101011485	0.000414	0.12			
9	585232_75	31767032	0.00059	0.24	LOC115722331	XM_030651522.2	CRM-domain containing factor CFM3A, chloroplastic/mitochondrial
10	666147_74	42986208	0.000558	0.27			
10	689517_88	55869380	0.000621	0.33			

459 CHR, chromosome; MAF, minor allele frequency.

460

DISCUSSION

This study adds significantly to our limited understanding of the population genomics of *Cannabis sativa* and provides novel insights into gene-trait associations for a natural population originating from contrasting climatic zones across Iran. The scarcity of such studies likely results from the historical constraints in accessing wide cannabis populations that capture natural genetic diversity (Campbell et al., 2019; Petit et al., 2020b; Hurgobin et al., 2021; Mostafaei Dehnavi et al., 2022). These insights are new, with few previously published GWAS and population genomics studies available for this largely undomesticated crop, and they add important knowledge to our developing understanding of the genomics of medicinal and industrial characterization of the crop. They will help underpin directed breeding programs to enhance traits of interest for commercial production (Hurgobin et al., 2021) and also add new insight to the population genomics and domestication history recently reported for *C. sativa* (Ren et al., 2021) that did not include data from the wild populations of Iran, such as those reported and characterized here.

The investigation of genetic variation between domesticated and natural populations is crucial for comprehending the patterns of local adaptation and identifying the genetic sources of desirable characteristics (Mostafaei Dehnavi et al., 2022). Studies on wild landraces are emerging, for example, for the wild population of Cannabis in China (Chen et al., 2022), which revealed five distinct groups and the population genomics for this important site of Cannabis domestication is important, since China is believed may possibly be one of the main centers of origin for this crop (Zhang et al., 2018; Gao et al., 2020).

Genotyping of this Iranian wild collection using GBS allowed us to conduct genetic diversity (genetic distance), population structure and genome-wide association analysis among these native populations, such as has been performed in wild (feral) cannabis collections as well as in other species, where wild progenitor populations have been used to inform breeding (Piluzza et al., 2013; Labate et al., 2014; Arbelaez et al., 2015; Ivanizs et al., 2019; Schwabe et al., 2021; Busta et al., 2022; Muli et al., 2022; Blois et al., 2023; Pandey et al., 2023). The results of population structure revealed the presence of five clusters, in contrast to the two genetic clusters reported in the previous investigation of Iranian cannabis populations conducted by Soorni *et al.*

(2017). The estimation of K is influenced by two crucial factors: the number of populations and the genetic dispersion among them (Zhang et al., 2022). The observed variation may be attributed to the inclusion of a broader range of populations situated in diverse climatic zones and a larger sample size per population. Additionally, the Sqz-01 population did not form a cluster with other population groups. As previously stated, this population stands out due to distinctive morphophysiological characteristics, such as dwarf stature, early flowering, and compact inflorescence, which differentiate it from the other populations (Mostafaei Dehnavi et al., 2022).

Due to its predominantly wind-pollinated dioecious nature, cannabis is a highly heterozygous and outcrossing species (Lynch et al. 2016; Welling et al. 2020). Sawler *et al.* (2015) noted greater heterozygosity in hemp than in marijuana. Heterozygosity across distinct geographically-based Iranian populations showed a similarity in the heterozygosity to that of hemp accessions studied by Sawler et al. (2015), while the earlier study on Iranian samples reported an average heterozygosity which is more similar to that observed in marijuana. This is despite the fact that Lynch *et al.* (2016) observed a significant rise in heterozygosity within drug-type varieties compared to hemp varieties. These distinctions highlight the complex interplay between genetic backgrounds and environmental factors, resulting in diverse heterozygosity patterns within cannabis populations (Zhang et al., 2018). The expected heterozygosity values were higher than the observed heterozygosity in all four geography-based groups of Iranian populations, possibly indicating the impact of inbreeding and reduced genetic variability (Radosavljević et al., 2015). However, the limited sample size of our cannabis collection may also have contributed to these results and increasing the sample size of the population could further improve the results (Nei, 1978; Stevens et al., 2007; Arab et al., 2019; Schmidt et al., 2021).

Lower F_{ST} values between Iranian population pairs ranging from 0.02 to 0.06 indicate no strong genetic differentiation among these populations (Wrights, 1978). This phenomenon could be attributed to the distribution of pollen and seeds and the gene flow between these areas (Cheng *et al.*, 2020). These seed exchanges could also have been facilitated by human activity, particularly for distant locations, as well as by wind pollination and seed dispersal by bird movements (Kitada et al., 2021). In the earlier investigation by Soorni *et al.* (2017), it was similarly noted that the anticipated lack of significant population differentiation is a result of the wind-

pollination characteristic of all known cannabis cultivars, coupled with their notable heterozygosity. To ensure the preservation of their genetic uniformity, numerous marijuana strains are propagated clonally rather than through seed.

GWAS has previously been successful in identifying genotype–phenotype associations in hop (*Humulus lupulus*) and cannabis (Henning et al., 2019). Moreover, one of the earliest studies of the cannabis genome and transcriptome found that the genes responsible for producing THC and CBD were located in different regions of the genome than previously thought and that there have been extensive rearrangements and variations in these regions among different cultivars of the plant (van Bakel et al., 2011). For marker–trait association in our study, multiple proteins across the genome are involved in traits related to cannabinoid yields, such as THC and CBD content. This implies that the production of cannabinoids is not only directly tied to the genes responsible for their synthesis in the cannabinoid pathway (Sawler et al., 2015; Hurgobin et al., 2021), but rather that additional areas of the genome also control these biosynthetic pathways (Marks et al., 2009; Laverty et al., 2019; Welling et al., 2020). This finding aligns with the earlier studies, further supporting the notion that multiple genetic factors contribute to cannabinoid content variation (Stout et al., 2012; Weiblen et al., 2015; Grassa et al., 2018; Braich et al., 2019; Zager et al., 2019; Livingston et al., 2020). Here we contribute further to this literature by confirming some of these important loci, but also identifying other novel loci for targeted future selection and breeding.

Furthermore, an additional SNP marker -816770_13- located on chromosome 1 and associated with the locus LOC115705717 (RefSeq accession: XM_030633131.1), responsible for 3-ketoacyl-CoA synthesis 19, has been found to be linked with CBD concentration. This SNP was previously identified as being connected to Autoflower1, which is involved in regulating the flowering time of hemp (Toth et al., 2022). These findings suggest that certain genes involved in the regulation of flowering time may also be correlated with cannabinoid content. One limitation of this study is that the phenotypic data collected here are only reported for a single growing season and environment, and thus, some caution must be used in generalizing the importance of these findings. However, given that our study confirms earlier identified genetic loci, such as those for flowering linked to cannabinoid production, this gives us confidence that our approach,

549 is worthy and valid but should, in the future, be complemented with additional field trials and
550 analyses to confirm these and other genetic loci.

551 Regarding the sex trait, the significant SNPs were annotated to uncover potential genetic
552 mechanisms associated with sex determination. While a total of eight potential SNPs located on
553 different chromosomes were identified for this trait, the prior study by Soorni *et al.* (2017),
554 which failed to pinpoint distinct alleles for the regions responsible for sex determination. As the
555 results of this study and earlier studies showed, the process of sex determination in cannabis is
556 intricate and is not solely related to sex chromosomes (McKernan et al., 2020). It appears that the
557 sex determination mechanism may be influenced by environmental factors and chemical
558 applications (Lubell and Brand, 2018; Campbell et al., 2021) and involves the participation of
559 other candidate genes like those related to trichome growth, sex determination, hermaphroditism,
560 and photoperiod independence (Hurgobin et al., 2021) or genes involved in regulating
561 phytohormone balance and the development of male flowers in female plants (Petit et al.,
562 2020a). For example, the gene LOC115720754 (RefSeq accession: XM_030649931.1), located
563 on chromosome 2, encodes the Zinc finger protein GAI-ASSOCIATED FACTOR 1, which acts
564 as a transcription factor and a positive regulator of gibberellin (GA) action, homeostasis, and
565 signaling in Arabidopsis (Fukazawa et al., 2014). This candidate gene was found to be associated
566 with flowering time (Colasanti et al., 2006). In plants, GA has a role in flowering control
567 (Goldberg-Moeller et al., 2013), and Petit *et al.* (2020b) linked the control of the flowering
568 pathways in cannabis to that of sex determination, further highlighting the complexity of these
569 inter-linked traits, where further research is warranted.

570 Out of the eight candidate loci identified for plant height, we observed that LOC115709353
571 (RefSeq accession: XM_030637438.1), located on chromosome 3, encodes anaphase-promoting
572 complex subunit 8. In Arabidopsis, this protein is known to play a role in various aspects of
573 development and embryogenesis by regulating the cell cycle, cell division, cell elongation, and
574 endoreduplication control (Eloy et al., 2012; Xu et al., 2019; Saleme et al., 2021). While QTL
575 analysis has been previously conducted for a range of agronomic traits on a population of 375
576 individuals (Woods et al., 2021), it has not been performed for the unique array of morpho-
577 physiological traits presented here. These triats include, the number of nodes, internode length,
578 crown length, and the number of leaves, as well as inflorescence- related features such as the

number of lateral inflorescences and inflorescences length. The identified SNP markers for these traits have not been previously mapped, and consequently, this study stands as the inaugural attempt to assess these specific traits through association analysis, with loci and candidate genes reported, using the recently available genome sequence of *C. sativa*. The significant SNPs identified are novel for a range of traits and are not shared among them. These SNPs have the potential to serve as markers for marker-assisted breeding in cannabis, pending proper validation.

CONCLUSION

Using GBS data from a diverse Iranian cannabis collection of wild germplasm (CGRC), this study has provided significant insights into the genetic variation and differentiation, population structure, and genotype-phenotype associations for this novel germplasm and how it differs from currently available global hemp and marijuana collections. Population structure analysis revealed five distinct groups in the Iranian cannabis collection. Pairwise F_{ST} comparisons identified the northeast population of Iran as the most genetically distinct, making it a priority for future breeding programs. Furthermore, the study confirmed several gene targets for unique traits, including inflorescence features, flowering time, cannabinoid content, sex, and some morphological traits. Together, this study has created a research platform that can link genomic variation and germplasm collection, facilitating selections for molecular breeding. These findings have important implications for improving the quality and productivity of new commercial cannabis varieties through breeding.

FUNDING

This research was funded by BRC (Biopharmaceutical Research Company) as part of a grant to GT at UC Davis and by the Iran National Science Foundation (INSF) under grant number 96014753. Also, the research in the laboratory of GT is supported by the John B Orr endowment.

ACKNOWLEDGMENTS

The authors thank the Iran National Science Foundation (INSF) for their support. The authors also express their profound gratitude to Afshin Peirovi, CIAN Diagnostics, 5330 Spectrum Drive, Suite I, Frederick, MD 21703, USA, for invaluable assistance in the preparation of lyophilized samples.

Conflict of Interest

Author Annabelle Damerum is employed by Zymo Research Corporation, and authors George Hodgin and Brian Brandley are employed by Biopharmaceutical Research Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19, 1655–1664. doi: 10.1101/gr.094052.109

Andre, C. M., Hausman, J.-F., and Guerriero, G. (2016). Cannabis sativa: The Plant of the Thousand and One Molecules. *Front. Plant Sci.* 7. doi: 10.3389/fpls.2016.00019

Arab, M. M., Marrano, A., Abdollahi-Arpanahi, R., Leslie, C. A., Askari, H., Neale, D. B., et al. (2019). Genome-wide patterns of population structure and association mapping of nut-related traits in Persian walnut populations from Iran using the Axiom J. regia 700K SNP array. *Sci. Rep.* 9, 6376. doi: 10.1038/s41598-019-42940-1

Arbelaez, J. D., Moreno, L. T., Singh, N., Tung, C.-W., Maron, L. G., Ospina, Y., et al. (2015). Development and GBS-genotyping of introgression lines (ILs) using two wild species of rice, O. meridionalis and O. rufipogon, in a common recurrent parent, O. sativa cv. Curinga. *Mol. Breed.* 35, 81. doi: 10.1007/s11032-015-0276-7

Bachari, A., Piva, T. J., Salami, S. A., Jamshidi, N., and Mantri, N. (2020). Roles of Cannabinoids in Melanoma: Evidence from In Vivo Studies. *Int. J. Mol. Sci.* 21, 6040. doi: 10.3390/ijms21176040

Bhattarai, J. H., Surya P., and Midmore, D. J. (2014). Effect of industrial hemp (Cannabis sativa L) planting density on weed suppression, crop growth, physiological responses, and fibre yield in the subtropics. *Renew. Bioresour.* 2, 1. doi: 10.7243/2052-6237-2-1

Blois, L., de Miguel, M., Bert, P., Girollet, N., Ollat, N., Rubio, B., et al. (2023). Genetic structure and first genome□wide insights into the adaptation of a wild relative of grapevine, *Vitis berlandieri*. *Evol. Appl.* 16, 1184–1200. doi: 10.1111/eva.13566

639 Braich, S., Baillie, R. C., Jewell, L. S., Spangenberg, G. C., and Cogan, N. O. I. (2019).
640 Generation of a Comprehensive Transcriptome Atlas and Transcriptome Dynamics in
641 Medicinal Cannabis. *Sci. Rep.* 9, 16583. doi: 10.1038/s41598-019-53023-6

642 Brunetti, P., Pichini, S., Pacifici, R., Busardò, F. P., and del Rio, A. (2020). Herbal Preparations
643 of Medical Cannabis: A Vademecum for Prescribing Doctors. *Medicina (Mex.)* 56, 237.
644 doi: 10.3390/medicina56050237

645 Campbell, B. J., Berrada, A. F., Hudalla, C., Amaducci, S., and McKay, J. K. (2019). Genotype
646 × Environment Interactions of Industrial Hemp Cultivars Highlight Diverse Responses to
647 Environmental Factors. *Agrosystems Geosci. Environ.* 2, 1–11. doi:
648 10.2134/age2018.11.0057

649 Campbell, L. G., Peach, K., and Wizenberg, S. B. (2021). Dioecious hemp (Cannabis sativa L.)
650 plants do not express significant sexually dimorphic morphology in the seedling stage.
651 *Sci. Rep.* 11, 16825. doi: 10.1038/s41598-021-96311-w

652 Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., and Cresko, W. A. (2013). Stacks: an
653 analysis tool set for population genomics. *Mol. Ecol.* 22, 3124–3140. doi:
654 10.1111/mec.12354

655 Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015).
656 Second-generation PLINK: rising to the challenge of larger and richer datasets.
657 *GigaScience* 4, 7. doi: 10.1186/s13742-015-0047-8

658 Chen, X., Guo, H.-Y., Zhang, Q.-Y., Wang, L., Guo, R., Zhan, Y.-X., et al. (2022). Whole-
659 genome resequencing of wild and cultivated cannabis reveals the genetic structure and
660 adaptive selection of important traits. *BMC Plant Biol.* 22, 371. doi: 10.1186/s12870-022-
661 03744-0

662 Cheng, J., Kao, H., and Dong, S. (2020). Population genetic structure and gene flow of rare and
663 endangered Tetraena mongolica Maxim. revealed by reduced representation sequencing.
664 *BMC Plant Biol.* 20, 391. doi: 10.1186/s12870-020-02594-y

665 Colasanti, J., Tremblay, R., Wong, A. Y., Coneva, V., Kozaki, A., and Mable, B. K. (2006). The
666 maize INDETERMINATE1 flowering time regulator defines a highly conserved zinc
667 finger protein family in higher plants. *BMC Genomics* 7, 158. doi: 10.1186/1471-2164-7-
668 158

669 Cullings, K. W. (1992). Design and testing of a plant-specific PCR primer for ecological and
670 evolutionary studies. *Mol. Ecol.* 1, 233–240. doi: 10.1111/j.1365-294X.1992.tb00182.x

671 Deguchi, M., Dhir, S., Potlakayala, S., Dhir, S., Curtis, W. R., and Rudrabhatla, S. (2022). In
672 planta Female Flower Agroinfiltration Alters the Cannabinoid Composition in Industrial
673 Hemp (Cannabis sativa L.). *Front. Plant Sci.* 13, 921970. doi: 10.3389/fpls.2022.921970

674 Dowling, C. A., Melzer, R., and Schilling, S. (2021). Timing is everything: the genetics of
675 flowering time in *Cannabis sativa*. *The Biochemist* 43, 34–38. doi:
676 10.1042/bio_2021_138

677 Doyle, J. J., and Doyle J. L. (1987). A rapid DNA isolation procedure for small quantities of
678 fresh leaf tissue. Phytochemistry Bulletin 19, 11-15
679
680 Eloy, N. B., Gonzalez, N., Van Leene, J., Maleux, K., Vanhaeren, H., De Milde, L., et al. (2012).
681 SAMBA, a plant-specific anaphase-promoting complex/cyclosome regulator is involved
682 in early development and A-type cyclin stabilization. *Proc. Natl. Acad. Sci.* 109, 13853–
683 13858. doi: 10.1073/pnas.1211418109

684 Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., et al. (2011).
685 A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity
686 Species. *PLoS ONE* 6, e19379. doi: 10.1371/journal.pone.0019379

687 Faux, A.-M., Draye, X., Flamand, M.-C., Occre, A., and Bertin, P. (2016). Identification of
688 QTLs for sex expression in dioecious and monoecious hemp (Cannabis sativa L.).
689 *Euphytica* 209, 357–376. doi: 10.1007/s10681-016-1641-2

690 Fischedick J, E. S. (2015). Cannabinoids and Terpenes as Chemotaxonomic Markers in
691 Cannabis. *Nat. Prod. Chem. Res.* 03. doi: 10.4172/2329-6836.1000181

692 Flanagan, S. P., and Jones, A. G. (2017). Constraints on the FST–Heterozygosity Outlier
693 Approach. *J. Hered.* 108, 561–573. doi: 10.1093/jhered/esx048

694 Fukazawa, J., Teramura, H., Murakoshi, S., Nasuno, K., Nishida, N., Ito, T., et al. (2014).
695 DELLAs Function as Coactivators of GAI-ASSOCIATED FACTOR1 in Regulation of
696 Gibberellin Homeostasis and Signaling in *Arabidopsis*. *Plant Cell* 26, 2920–2938. doi:
697 10.1105/tpc.114.125690

698 Gali, K. K., Sackville, A., Tafesse, E. G., Lachagari, V. B. R., McPhee, K., Hybl, M., et al.
699 (2019). Genome-Wide Association Mapping for Agronomic and Seed Quality Traits of
700 Field Pea (Pisum sativum L.). *Front. Plant Sci.* 10, 1538. doi: 10.3389/fpls.2019.01538

701 Gao, S., Wang, B., Xie, S., Xu, X., Zhang, J., Pei, L., et al. (2020). A high-quality reference
702 genome of wild Cannabis sativa. *Hortic. Res.* 7, 73. doi: 10.1038/s41438-020-0295-3

703 Goldberg-Moeller, R., Shalom, L., Shlizerman, L., Samuels, S., Zur, N., Ophir, R., et al. (2013).
704 Effects of gibberellin treatment during flowering induction period on global gene
705 expression and the transcription of flowering-control genes in Citrus buds. *Plant Sci.* 198,
706 46–57. doi: 10.1016/j.plantsci.2012.09.012

707 Gonçalves, J., Rosado, T., Soares, S., Simão, A., Caramelo, D., Luís, Â., et al. (2019). Cannabis
708 and Its Secondary Metabolites: Their Use as Therapeutic Drugs, Toxicological Aspects,
709 and Analytical Determination. *Medicines* 6, 31. doi: 10.3390/medicines6010031

Grassa, C. J., Weiblen, G. D., Wenger, J. P., Dabney, C., Poplawski, S. G., Timothy Motley, S., et al. (2021). A new *Cannabis* genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytol.* 230, 1665–1679. doi: 10.1111/nph.17243

Grassa, C. J., Wenger, J. P., Dabney, C., Poplawski, S. G., Motley, S. T., Michael, T. P., et al. (2018). A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. Genomics. doi: 10.1101/458083

Henning, J., Coggins, J., Hill, S., Hendrix, D., and Townsend, S. (2019). Genome-wide association study on ten traits of economic importance in hop (*Humulus lupulus* L.). *Acta Hortic.*, 93–104. doi: 10.17660/ActaHortic.2019.1236.13

Henry, P., Khatodia, S., Kapoor, K., Gonzales, B., Middleton, A., Hong, K., et al. (2020). 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. Plant Biology. doi: 10.1101/2020.02.16.951459

Hurgobin, B., Tamiru□Oli, M., Welling, M. T., Doblin, M. S., Bacic, A., Whelan, J., et al. (2021). Recent advances in *Cannabis sativa* genomics research. *New Phytol.* 230, 73–89. doi: 10.1111/nph.17140

Ivanizs, L., Monostori, I., Farkas, A., Megyeri, M., Mikó, P., Türkösi, E., et al. (2019). Unlocking the Genetic Diversity and Population Structure of a Wild Gene Source of Wheat, Aegilops biuncialis Vis., and Its Relationship With the Heading Time. *Front. Plant Sci.* 10, 1531. doi: 10.3389/fpls.2019.01531

Jenkins, C., and Orsburn, B. (2019a). Constructing a Draft Map of the Cannabis Proteome. Plant Biology. doi: 10.1101/577635

Jenkins, C., and Orsburn, B. (2019b). The First Publicly Available Annotated Genome for Cannabis plants. Plant Biology. doi: 10.1101/786186

Kitada, S., Nakamichi, R., and Kishino, H. (2021). Understanding population structure in an evolutionary context: population-specific *F* ST and pairwise *F* ST. *G3 GenesGenomesGenetics* 11, jkab316. doi: 10.1093/g3journal/jkab316

Kovalchuk, I., Pellino, M., Rigault, P., van Velzen, R., Ebersbach, J., Ashnest, J. R., et al. (2020). The Genomics of *Cannabis* and Its Close Relatives. *Annu. Rev. Plant Biol.* 71, 713–739. doi: 10.1146/annurev-arplant-081519-040203

Labate, J. A., Robertson, L. D., Strickler, S. R., and Mueller, L. A. (2014). Genetic structure of the four wild tomato species in the *Solanum peruvianum* s.l. species complex. *Genome* 57, 169–180. doi: 10.1139/gen-2014-0003

Laverty, K. U., Stout, J. M., Sullivan, M. J., Shah, H., Gill, N., Holbrook, L., et al. (2019). A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the *THC/CBD acid synthase* loci. *Genome Res.* 29, 146–156. doi: 10.1101/gr.242594.118

747 Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., Tyagi, A., et al.
748 (2015). Package 'plotrix'. Vienna: R Development Core Team
749
750 Livingston, S. J., Quilichini, T. D., Booth, J. K., Wong, D. C. J., Rensing, K. H.,
751 Laflamme□Yonkman, J., et al. (2020). Cannabis glandular trichomes alter morphology
752 and metabolite content during flower maturation. *Plant J.* 101, 37–56. doi:
753 10.1111/tpj.14516

754 Lubell, J. D., and Brand, M. H. (2018). Foliar Sprays of Silver Thiosulfate Produce Male
755 Flowers on Female Hemp Plants. *HortTechnology* 28, 743–747. doi:
756 10.21273/HORTTECH04188-18

757 Lynch, R. C., Vergara, D., Tittes, S., White, K., Schwartz, C. J., Gibbs, M. J., et al. (2016).
758 Genomic and Chemical Diversity in *Cannabis*. *Crit. Rev. Plant Sci.* 35, 349–363. doi:
759 10.1080/07352689.2016.1265363

760 Marks, M. D., Tian, L., Wenger, J. P., Omburo, S. N., Soto-Fuentes, W., He, J., et al. (2009).
761 Identification of candidate genes affecting Δ9-tetrahydrocannabinol biosynthesis in
762 Cannabis sativa. *J. Exp. Bot.* 60, 3715–3726. doi: 10.1093/jxb/erp210

763 McKernan, K. J., Helbert, Y., Kane, L. T., Ebling, H., Zhang, L., Liu, B., et al. (2020). Sequence
764 and annotation of 42 cannabis genomes reveals extensive copy number variation in
765 cannabinoid synthesis and pathogen resistance genes. Genomics. doi:
766 10.1101/2020.01.03.894428

767 Mostafaei Dehnavi, M., Ebadi, A., Peirovi, A., Taylor, G., and Salami, S. A. (2022). THC and
768 CBD Fingerprinting of an Elite Cannabis Collection from Iran: Quantifying Diversity to
769 Underpin Future Cannabis Breeding. *Plants* 11, 129. doi: 10.3390/plants11010129

770 Muli, J. K., Neondo, J. O., Kamau, P. K., Michuki, G. N., Odari, E., and Budambula, N. L. M.
771 (2022). Genetic diversity and population structure of wild and cultivated Crotalaria
772 species based on genotyping-by-sequencing. *PLOS ONE* 17, e0272955. doi:
773 10.1371/journal.pone.0272955

774 Nei, M. (1978). ESTIMATION OF AVERAGE HETEROZYGOSITY AND GENETIC
775 DISTANCE FROM A SMALL NUMBER OF INDIVIDUALS. *Genetics* 89, 583–590.
776 doi: 10.1093/genetics/89.3.583

777 Pépin, N., Hebert, F. O., and Joly, D. L. (2021). Genome-Wide Characterization of the MLO
778 Gene Family in Cannabis sativa Reveals Two Genes as Strong Candidates for Powdery
779 Mildew Susceptibility. *Front. Plant Sci.* 12, 729261. doi: 10.3389/fpls.2021.729261

780 Petit, J., Salentijn, E. M. J., Paulo, M.-J., Denneboom, C., and Trindade, L. M. (2020a). Genetic
781 Architecture of Flowering Time and Sex Determination in Hemp (Cannabis sativa L.): A
782 Genome-Wide Association Study. *Front. Plant Sci.* 11, 569958. doi:
783 10.3389/fpls.2020.569958

784 Petit, J., Salentijn, E. M. J., Paulo, M.-J., Denneboom, C., van Loo, E. N., and Trindade, L. M.
785 (2020b). Elucidating the Genetic Architecture of Fiber Quality in Hemp (Cannabis sativa
786 L.) Using a Genome-Wide Association Study. *Front. Genet.* 11, 566314. doi:
787 10.3389/fgene.2020.566314

788 Piluzza, G., Delogu, G., Cabras, A., Marceddu, S., and Bullitta, S. (2013). Differentiation
789 between fiber and drug types of hemp (Cannabis sativa L.) from a collection of wild and
790 domesticated accessions. *Genet. Resour. Crop Evol.* 60, 2331–2342. doi:
791 10.1007/s10722-013-0001-5

792 Pootakham, W., Jomchai, N., Ruang-areerate, P., Shearman, J. R., Sonthirod, C., Sangsrakru, D.,
793 et al. (2015). Genome-wide SNP discovery and identification of QTL associated with
794 agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics* 105,
795 288–295. doi: 10.1016/j.ygeno.2015.02.002

796 Prentout, D., Razumova, O., Rhoné, B., Badouin, H., Henri, H., Feng, C., et al. (2019). A high-
797 throughput segregation analysis identifies the sex chromosomes of *Cannabis sativa*.
798 Genomics. doi: 10.1101/721324

799 Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population
800 stratification in genome-wide association studies. *Nat. Rev. Genet.* 11, 459–463. doi:
801 10.1038/nrg2813

802 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007).
803 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage
804 Analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

805 R Core Team (2018). R: A language and environment for statistical computing. R Foundation for
806 Statistical Computing, Vienna, Austria. Available online at https://www.R-project.org/
807
808 Radosavljević, I., Satovic, Z., and Liber, Z. (2015). Causes and consequences of contrasting
809 genetic structure in sympatrically growing and closely related species. *AoB Plants* 7,
810 plv106. doi: 10.1093/aobpla/plv106

811 Ren, G., Zhang, X., Li, Y., Ridout, K., Serrano-Serrano, M. L., Yang, Y., et al. (2021). Large-
812 scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*.
813 *Sci. Adv.* 7, eabg2286. doi: 10.1126/sciadv.abg2286

814 Saleme, M. de L. S., Andrade, I. R., and Eloy, N. B. (2021). The Role of Anaphase-Promoting
815 Complex/Cyclosome (APC/C) in Plant Reproduction. *Front. Plant Sci.* 12, 642934. doi:
816 10.3389/fpls.2021.642934

817 Sawler, J., Stout, J. M., Gardner, K. M., Hudson, D., Vidmar, J., Butler, L., et al. (2015). The
818 Genetic Structure of Marijuana and Hemp. *PLOS ONE* 10, e0133292. doi:
819 10.1371/journal.pone.0133292

820 Schmidt, T. L., Jasper, M., Weeks, A. R., and Hoffmann, A. A. (2021). Unbiased population
821 heterozygosity estimates from genome□wide sequence data. *Methods Ecol. Evol.* 12,
822 1888–1898. doi: 10.1111/2041-210X.13659

823 Sonah, H., Bastien, M., Iquira, E., Tardivel, A., Légaré, G., Boyle, B., et al. (2013). An Improved
824 Genotyping by Sequencing (GBS) Approach Offering Increased Versatility and
825 Efficiency of SNP Discovery and Genotyping. *PLoS ONE* 8, e54603. doi:
826 10.1371/journal.pone.0054603

827 Soorni, A., Fatahi, R., Haak, D. C., Salami, S. A., and Bombarely, A. (2017). Assessment of
828 Genetic Diversity and Population Structure in Iranian Cannabis Germplasm. *Sci. Rep.* 7,
829 15668. doi: 10.1038/s41598-017-15816-5

830 Stevens, M. H. H., Sanchez, M., Lee, J., and Finkel, S. E. (2007). Diversification Rates Increase
831 With Population Size and Resource Concentration in an Unstructured Habitat. *Genetics*
832 177, 2243–2250. doi: 10.1534/genetics.107.076869

833 Stout, J. M., Boubakir, Z., Ambrose, S. J., Purves, R. W., and Page, J. E. (2012). The hexanoyl-
834 CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in
835 Cannabis sativa trichomes: A cytoplasmic acyl-activating enzyme involved in
836 cannabinoid biosynthesis. *Plant J.*, no-no. doi: 10.1111/j.1365-313X.2012.04949.x

837 Taheri, S., Saedi, N., Zerehdaran, S., and Javadmanesh, A. (2023). Identification of selection
838 signatures in *CAPRA HIRCUS* and *CAPRA AEGAGRUS* in Iran. *Anim. Sci. J.* 94, e13864. doi:
839 10.1111/asj.13864

840 Toth, J. A., Stack, G. M., Carlson, C. H., and Smart, L. B. (2022). Identification and mapping of
841 major-effect flowering time loci Autoflower1 and Early1 in Cannabis sativa L. *Front.*
842 *Plant Sci.* 13, 991680. doi: 10.3389/fpls.2022.991680

843 Turner, S. D. (2014). qqman: an R package for visualizing GWAS results using Q-Q and
844 manhattan plots. Bioinformatics. doi: 10.1101/005165

845 van Bakel, H., Stout, J. M., Cote, A. G., Tallon, C. M., Sharpe, A. G., Hughes, T. R., et al.
846 (2011). The draft genome and transcriptome of Cannabis sativa. *Genome Biol.* 12, R102.
847 doi: 10.1186/gb-2011-12-10-r102

848 Villanueva, R. A. M., and Chen, Z. J. (2019). ggplot2: Elegant Graphics for Data Analysis (2nd
849 ed.). *Meas. Interdiscip. Res. Perspect.* 17, 160–167. doi:
850 10.1080/15366367.2019.1565254

851 Weiblen, G. D., Wenger, J. P., Craft, K. J., ElSohly, M. A., Mehmedic, Z., Treiber, E. L., et al.
852 (2015). Gene duplication and divergence affecting drug content in *Cannabis sativa*. *New*
853 *Phytol.* 208, 1241–1250. doi: 10.1111/nph.13562

854 Welling, M. T., Liu, L., Kretzschmar, T., Mauleon, R., Ansari, O., and King, G. J. (2020). An
855 extreme-phenotype genome□wide association study identifies candidate cannabinoid
856 pathway genes in Cannabis. *Sci. Rep.* 10, 18643. doi: 10.1038/s41598-020-75271-7

857 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., et al. (2019).
858 Welcome to the Tidyverse. *J. Open Source Softw.* 4, 1686. doi: 10.21105/joss.01686

859 Woods, P., Campbell, B. J., Nicodemus, T. J., Cahoon, E. B., Mullen, J. L., and McKay, J. K.
860 (2021). Quantitative trait loci controlling agronomic and biochemical traits in *Cannabis*
861 *sativa*. *Genetics* 219, iyab099. doi: 10.1093/genetics/iyab099

862 Xu, R., Xu, J., Wang, L., Niu, B., Copenhaver, G. P., Ma, H., et al. (2019). The Arabidopsis
863 anaphase□promoting complex/cyclosome subunit 8 is required for male meiosis. *New*
864 *Phytol.* 224, 229–241. doi: 10.1111/nph.16014

865 Yang, J., Lee, S. H., Goddard, M. E., and Visscher, P. M. (2011). GCTA: A Tool for Genome-
866 wide Complex Trait Analysis. *Am. J. Hum. Genet.* 88, 76–82. doi:
867 10.1016/j.ajhg.2010.11.011

868 Zager, J. J., Lange, I., Srividya, N., Smith, A., and Lange, B. M. (2019). Gene Networks
869 Underlying Cannabinoid and Terpenoid Accumulation in Cannabis. *Plant Physiol.* 180,
870 1877–1897. doi: 10.1104/pp.18.01506

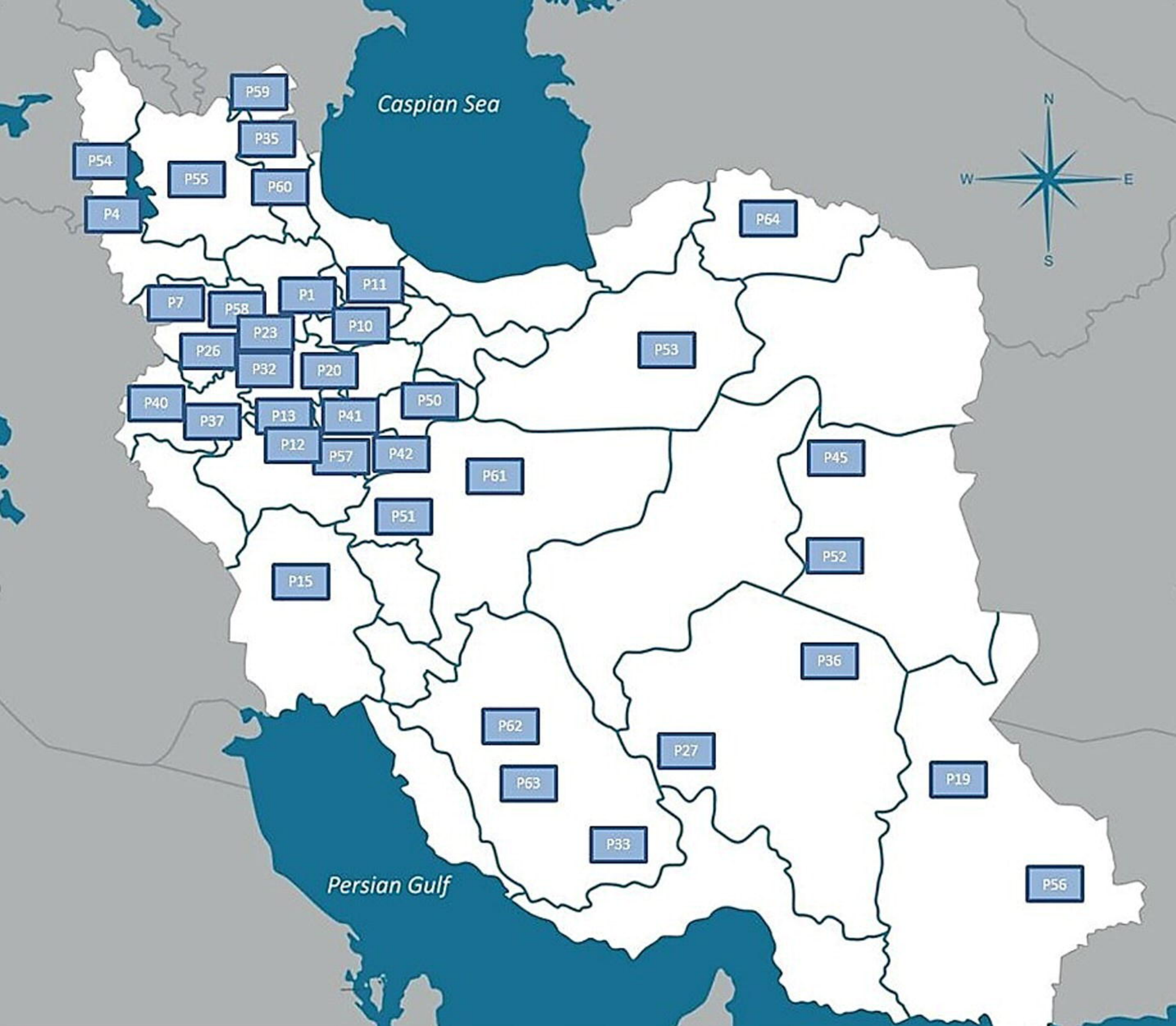
871 Zhang, C., Tang, Y., Tian, D., Huang, Y., Yang, G., Nan, P., et al. (2022). Population genetic
872 structure of Wikstroemia monnula highlights the necessity and feasibility of hierarchical
873 analysis for a highly differentiated species. *Front. Plant Sci.* 13, 962364. doi:
874 10.3389/fpls.2022.962364

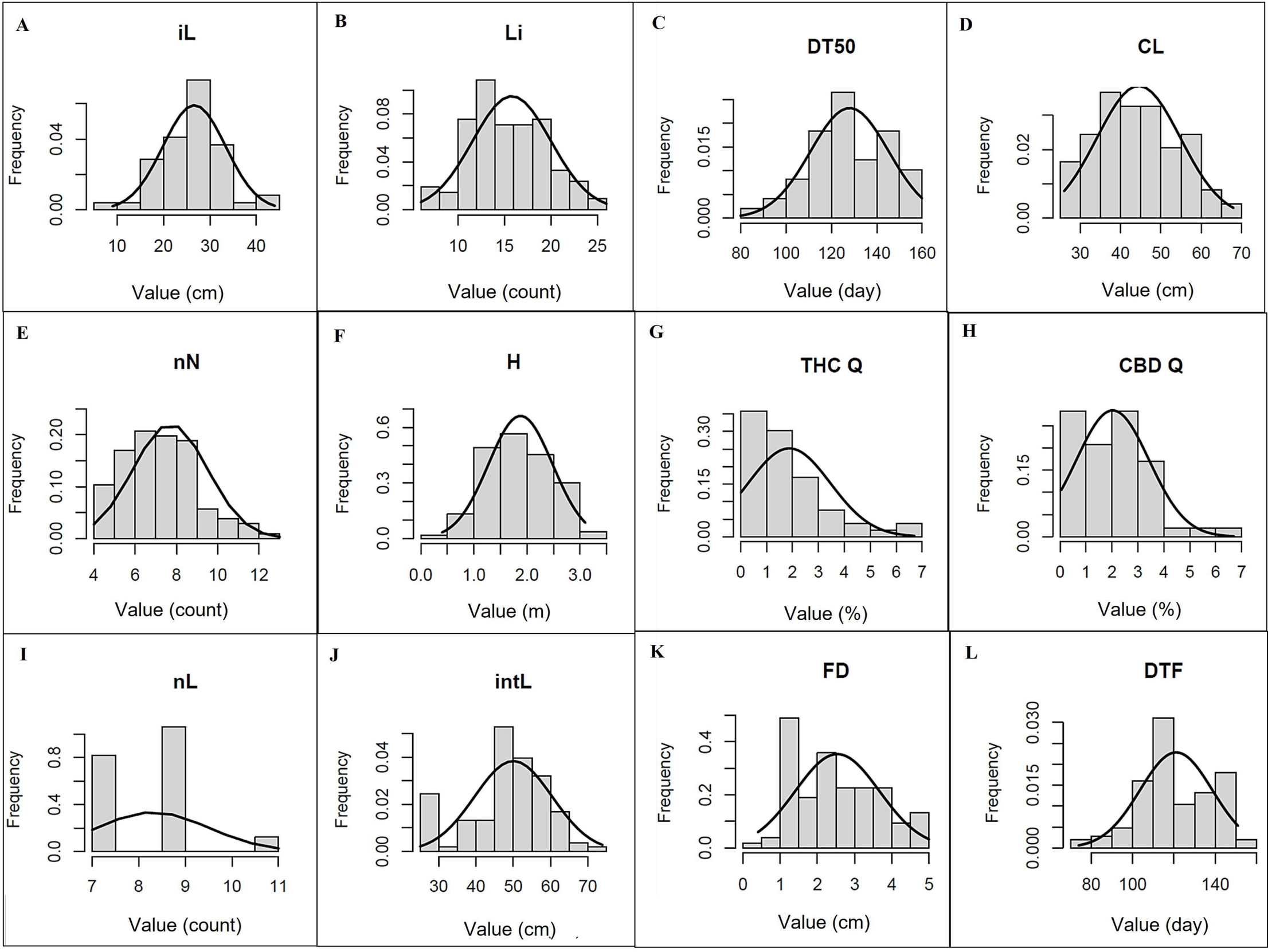
875 Zhang, J., Yan, J., Huang, S., Pan, G., Chang, L., Li, J., et al. (2020). Genetic Diversity and
876 Population Structure of Cannabis Based on the Genome-Wide Development of Simple
877 Sequence Repeat Markers. *Front. Genet.* 11, 958. doi: 10.3389/fgene.2020.00958

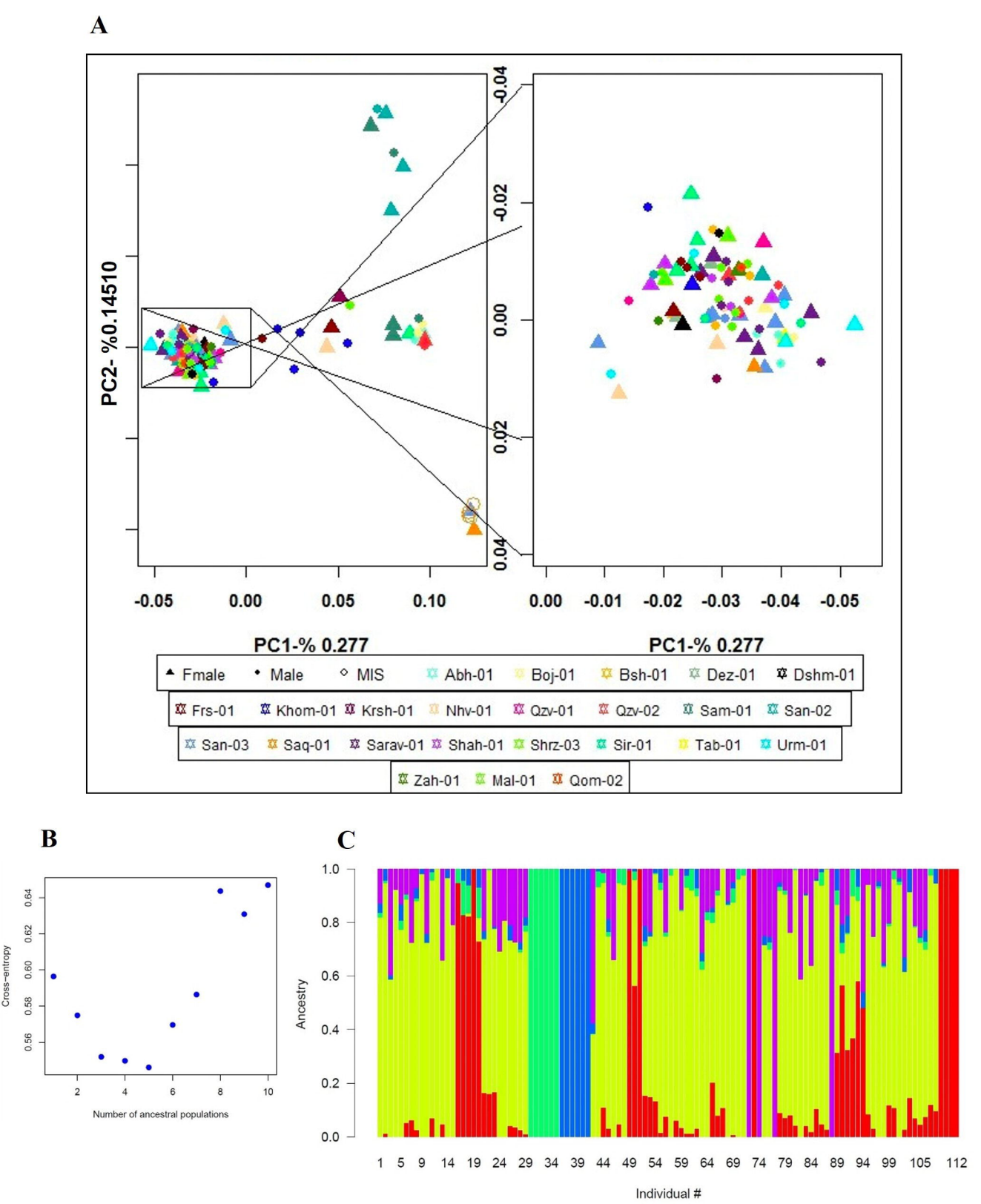
878 Zhang, Q., Chen, X., Guo, H., Trindade, L. M., Salentijn, E. M. J., Guo, R., et al. (2018).
879 Latitudinal Adaptation and Genetic Insights Into the Origins of Cannabis sativa L. *Front.*
880 *Plant Sci.* 9, 1876. doi: 10.3389/fpls.2018.01876

881 Zhu, H., and Zhou, X. (2020). Statistical methods for SNP heritability estimation and partition: A
882 review. *Comput. Struct. Biotechnol. J.* 18, 1557–1568. doi: 10.1016/j.csbj.2020.06.011

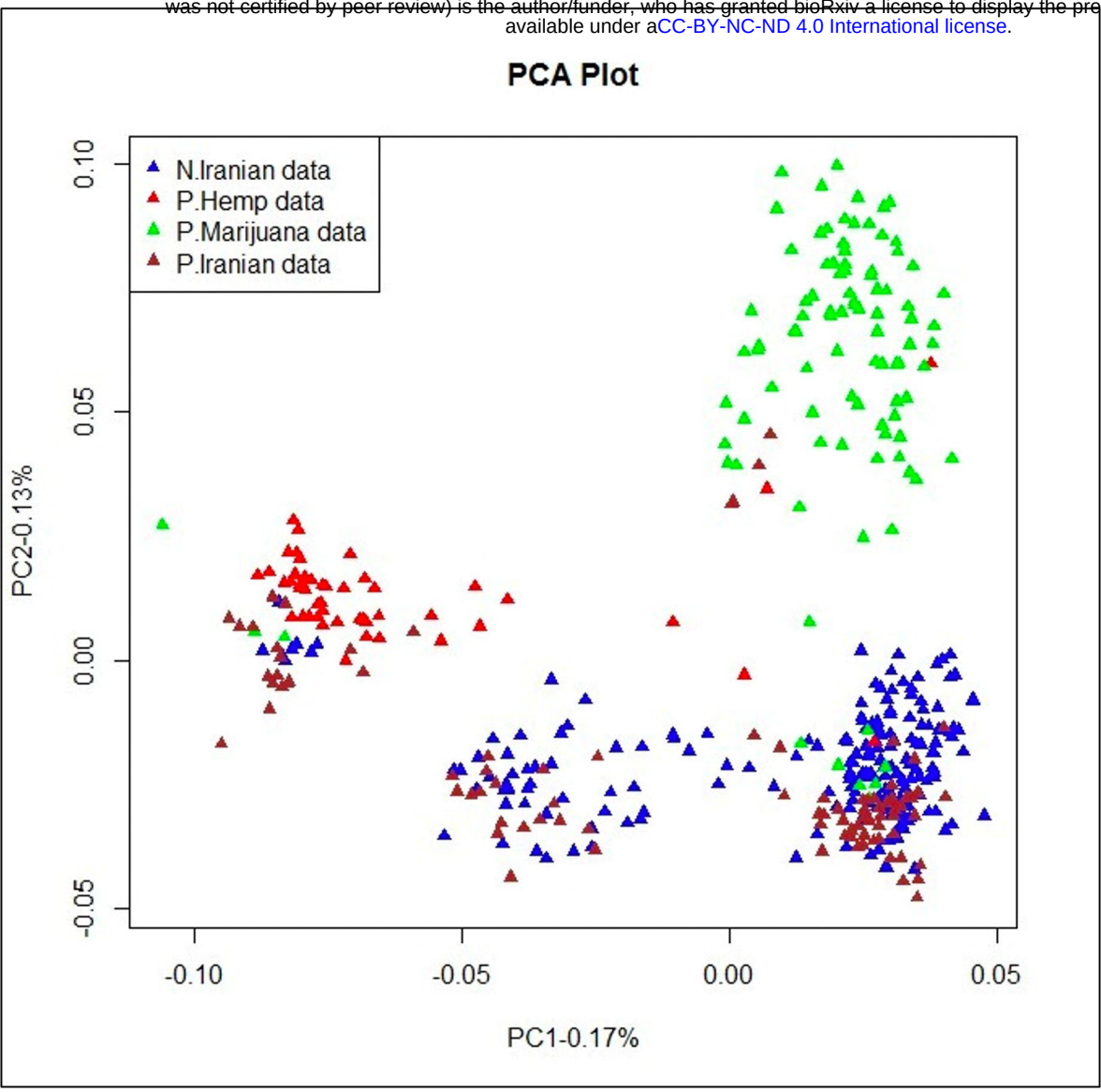
883


Caspian Sea
Persian Gulf
N
W
E
S
P59
P35
P54
P55
P60
P4
P64
P11
P7
P58
P1
P10
P23
P53
P26
P32
P20
P40
P50
P37
P13
P41
P12
P57
P42
P45
P61
P51
P52
P15
P36
P62
P27
P63
P19
P33
P56

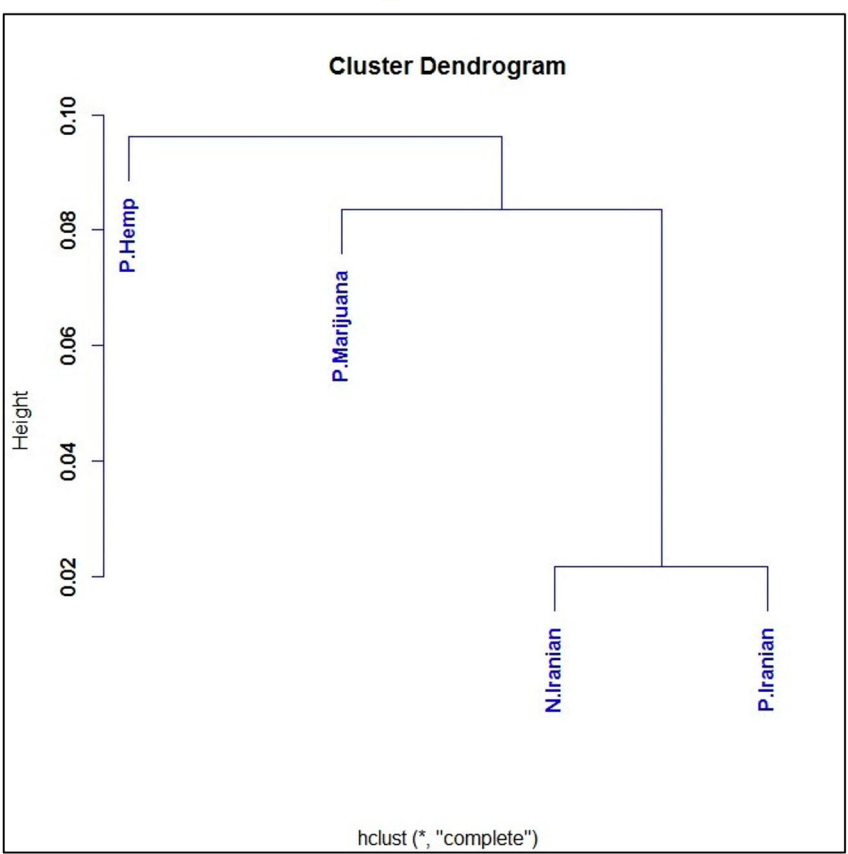

A
iL
Frequency
0.00
0.04
10
20
30
40
Value (cm)
B
Li
Frequency
0.00
0.04
0.08
10
15
20
25
Value (count)
C
DT50
Frequency
0.000
0.015
80
100
120
140
160
Value (day)
D
CL
Frequency
0.00
0.02
30
40
50
60
70
Value (cm)
E
nN
Frequency
0.00
0.10
0.20
4
6
8
10
12
Value (count)
F
H
Frequency
0.0
0.3
0.6
0.0
1.0
2.0
3.0
Value (m)
G
THC Q
Frequency
0.00
0.15
0.30
0
1
2
3
4
5
6
7
Value (%)
H
CBD Q
Frequency
0.00
0.15
0
1
2
3
4
5
6
7
Value (%)
I
nL
Frequency
0.0
0.4
0.8
7
8
9
10
11
Value (count)
J
intL
Frequency
0.00
0.02
0.04
30
40
50
60
70
Value (cm)
K
FD
Frequency
0.0
0.2
0.4
0
1
2
3
4
5
Value (cm)
L
DTF
Frequency
0.000
0.015
0.030
80
100
140
Value (day)

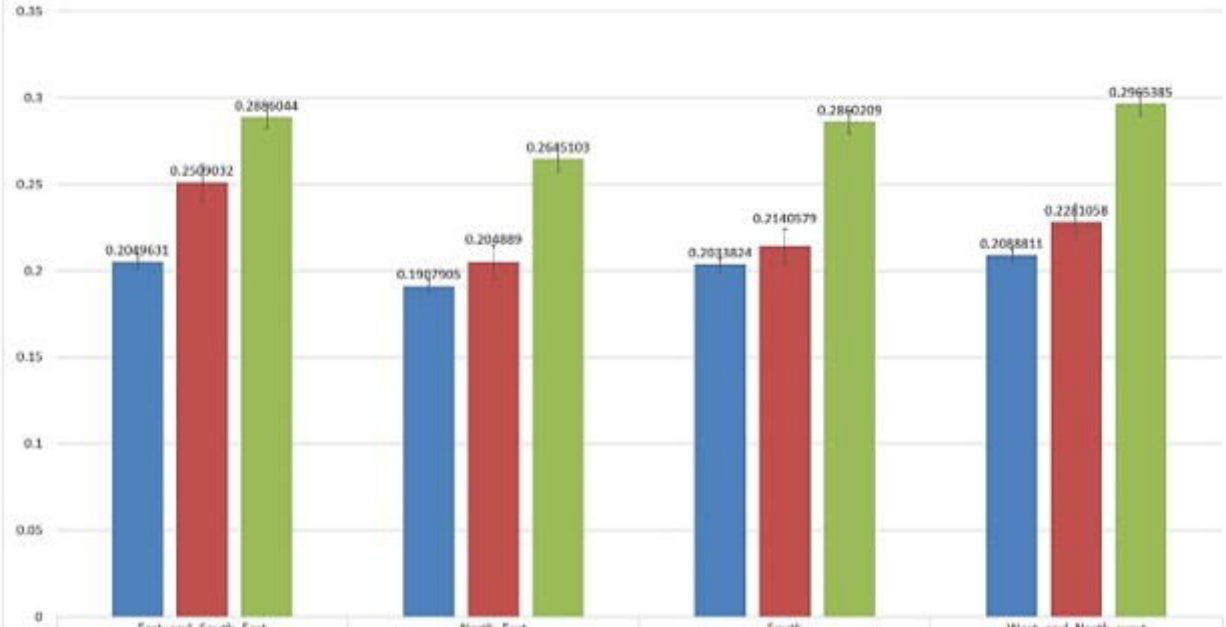

A
PC2- %0.14510
PC1-% 0.277
PC1-% 0.277
Fmale
Male
MIS
Abh-01
Boj-01
Bsh-01
Dez-01
Dshm-01
Frs-01
Khom-01
Krsh-01
Nhv-01
Qzv-01
Qzv-02
Sam-01
San-02
San-03
Saq-01
Sarav-01
Shah-01
Shrz-03
Sir-01
Tab-01
Urm-01
Zah-01
Mal-01
Qom-02
B
Cross-entropy
Number of ancestral populations
C
Ancestry
Individual #

A

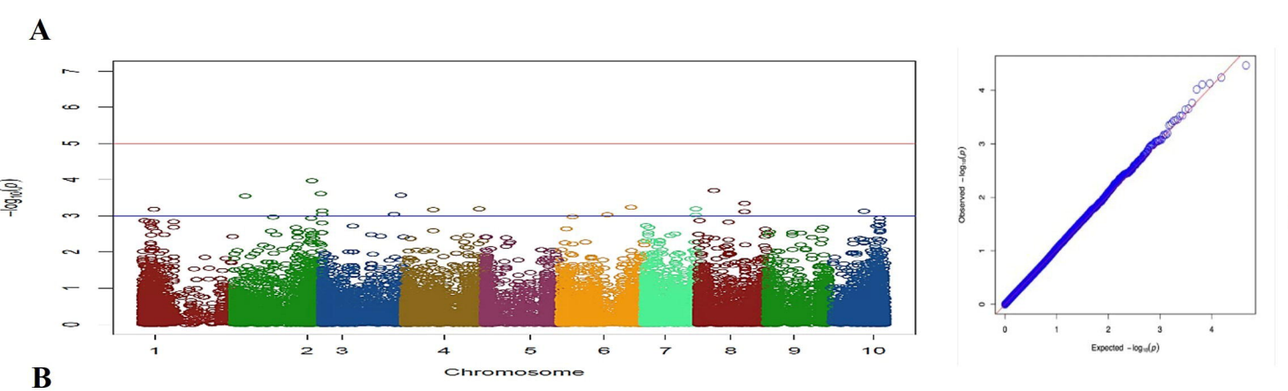


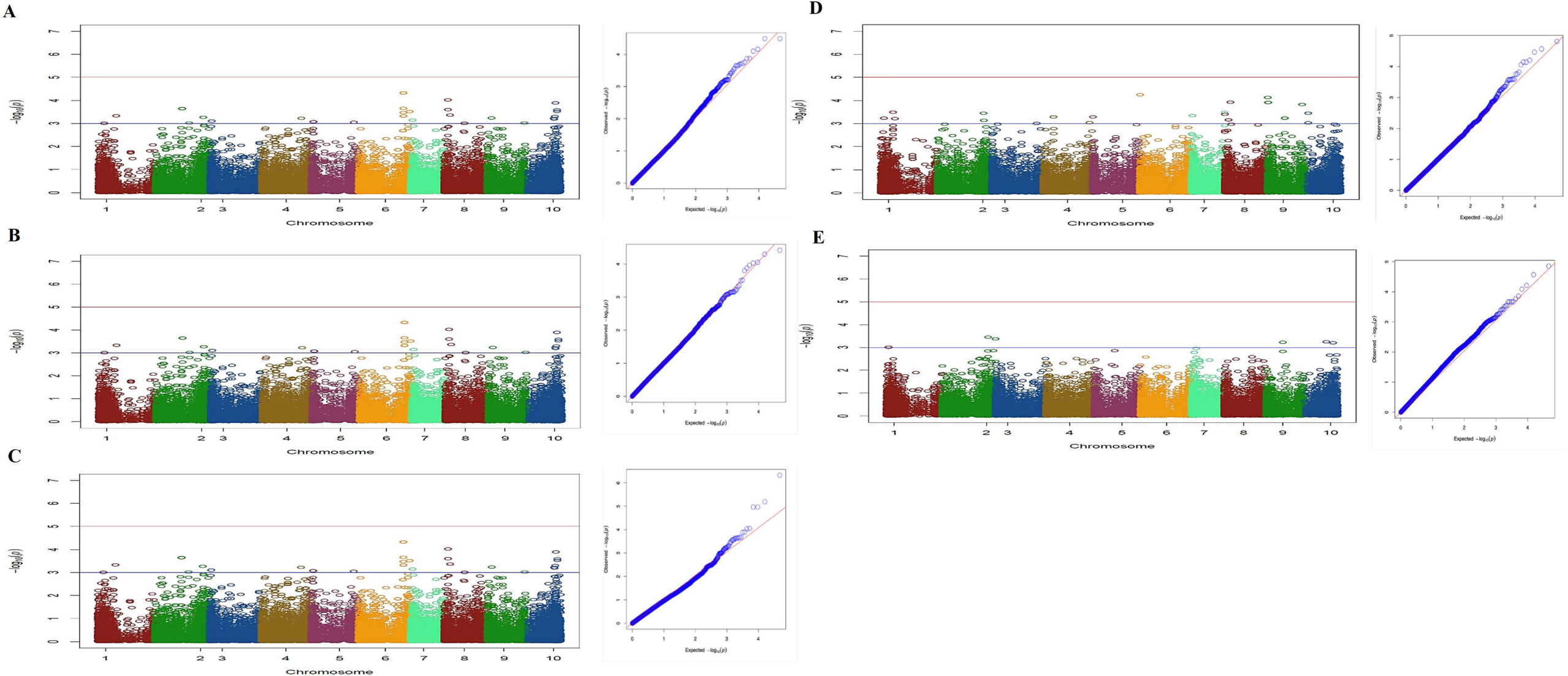
B





	East_and_South_East	North_East	South	West_and_North_west
MAF	0.2049631	0.1907905	0.2033824	0.2088811
O.HET	0.2509032	0.204889	0.2140579	0.2281058
E.HET	0.2886044	0.2645103	0.2860209	0.2965385


A
B
C
D
E
F
G
H
Chromosome
Expected $-\log_{10}(p)$
Observed $-\log_{10}(p)$
$-\log_{10}(p)$


A
B
C
D
E
Chromosome
Expected −log10(p)
Observed −log10(p)