

1 **Genetic Diversity, Population Structure, and Cannabinoid Variation in Feral *Cannabis*** 2 ***sativa* Germplasm from the United States**

3 Ademola Aina¹, Jonathan P Wenger², Eliot Stanton³, Chandrani Gon Majumdar⁴, Mahmoud
4 ElSohly⁴, George D Weiblen², Shelby Ellison^{1*}

5 ¹Department of Plant and Agroecosystem Sciences, University of Wisconsin, Madison

6 ²Department of Plant and Microbial Biology, University of Minnesota

7 ³Biotechnology Center, University of Wisconsin, Madison

8 ⁴National Center for Natural Products Research, University of Mississippi

9

10 *Corresponding Author: slrepinski@wisc.edu

11

12 **Keywords:** *Cannabis*, Cannabinoids, Genetic diversity, Feral, Conservation, Germplasm

13

14 **Abstract**

15 *Cannabis sativa* is one of the earliest plants to be domesticated for fiber, food and medicine.
16 Seed from *Cannabis* grown for industrial purposes during the 18th through 20th centuries
17 have escaped production and established feralized populations across the United States. To
18 maximize the potential of feral *Cannabis* germplasm, determining the genetic structure and
19 cannabinoid profile is crucial for selection and breeding of new compliant regionally adapted
20 hemp cultivars. To resolve this, a collection of feral *Cannabis*, comprising 760 plants across
21 twelve US states were sequenced using Genotyping-by-Sequencings (GBS), genotyped at the
22 *cannabinoid synthase* (*CBDAS*) gene, and subject to gas chromatography-mass spectrometry
23 (GC-MS) to assess cannabinoid profiles. Clustering analyses by ADMIXTURE and Principal
24 Component Analysis (PCA) stratified the germplasm into five clusters (Mississippi-River,
25 West North Central-b, West North Central-a, New York, and Indiana). The cannabinoid

genotyping assay resolved the feral collections into Type I - C^XC^X (6%), Type II - C^FC^X (15%), and Type III - C^FC^F (78%). Total cannabinoid content ranged from 0.21% to 4.73%. The assessment of genetic diversity, population structure, and cannabinoid profile of the US feral *Cannabis* collection provides critical information and germplasm resources to develop new and improve existing hemp cultivars.

INTRODUCTION

Cannabis sativa L. is a versatile crop that has been cultivated for thousands of years for its bast fiber (stem) and hurd fiber (stem pith), edible grain (seed) and cannabinoids (flowers). *Cannabis* plants produce a substantial number of chemical compounds known as cannabinoids. These cannabinoids comprise C21 terpenophenolic compounds responsible for the medicinal and psychotropic effects obtained from *Cannabis* use¹⁻³. There are over a hundred cannabinoids⁴ but the two main molecules of interest have always been delta-9-tetrahydrocannabinol (THC) and cannabidiol (CBD), both derived from the same precursor, cannabigerolic acid (CBGA)⁵. Consumption of high THC (drug-type *Cannabis*) results in intoxicating and/or analgesic effects⁶, while hemp-type *Cannabis* contains predominantly non-intoxicating cannabinoids (CBD). A single locus with codominant alleles purportedly described the pattern of inheritance for THC to CBD ratios in *Cannabis* plants^{7,8}. Recent findings have resolved mutually exclusive chromosome 7 haplotypes exhibiting orthogonal genetic and physical linkage of functional vs. non-functional *THCAS* and *CBDAS* loci⁹⁻¹¹. Classification of *Cannabis* as hemp or drug-type, is based on the concentration of THC, with 0.3% THC as the allowable upper threshold of industrial hemp (hemp-type *Cannabis*) as defined by the 2014 and 2018 Farm Bills^{12,13}.

Hemp was introduced to North America in the early 17th century¹⁴, and was used to make ropes, grain bags, Conestoga wagon covers, and clothing. In the early 20th century, a USDA agriculturalist Lyster Dewey, began researching industrial hemp production to help meet the US demand for a domestic fiber source with initial breeding efforts aimed at improving fiber quality and yield^{15,16}. During this period, industrial hemp production was primarily concentrated in the Midwest region of the US due to favorable growing conditions and accessible hemp processing mills. Production peaked in 1943 when approximately 176,000 acres were cultivated to supply fiber for the World War II war effort¹⁷. After the war, all *Cannabis* effectively became impractical to grow because of the resumed enforcement of the 1937 Marihuana Tax Act and production rapidly halted. Despite the cessation of cultivation, feralized industrial hemp populations were established after escaping production and continued reproducing in the wild across the major prior production regions in the US¹⁸. Geographically, these remnants of feral *Cannabis* populations are widely dispersed and grow in a wide range of climatic conditions. Feral *Cannabis* thrives under full exposure to sunlight and in nitrogen-rich, well-drained soils where its extensive and deep root structure absorbs enough water for the shoots to withstand hot and dry weather¹⁹. As of 2025, iNaturalist and Global Biodiversity Information Facility (GBIF) online databases currently recorded, 1,795 and 1,207 human observations of feral *Cannabis* in the US, respectively. Many of these observations occurred in the Midwest and Northeast areas of the US where they were previously cultivated in large acreages. Feral hemp is commonly found growing in disturbed, nutrient-rich lands including pastures, feed lots, farmyards, cultivated field borders, open woods, flood plains, dumpsites, roadsides, and railroad rights-of-way.

The recent legalization of hemp, with its many potential industrial and medicinal applications, has reinvigorated the scientific investigation of *Cannabis*. Significant scientific

75 advancements have been made in *Cannabis* genetics and genomics over the last decade. This
 76 includes the completion of a several high-quality genome assemblies^{9,20,21}, understanding
 77 genetic relatedness of major market classes, and studies on the genetic underpinnings of
 78 cannabinoid inheritance, sex determination, and many agronomic and morphological traits^{22–}
 79 ²⁴. *Cannabis* research and breeding efforts lag behind that of many modern crops. The
 80 urgency of creating and disseminating industrial hemp genetic resources cannot be
 81 overemphasized, with these tasks falling on genebanks to ensure adequate germplasm supply.
 82 Despite many germplasm repositories across the globe, only a handful of them show records
 83 of deliberate conservation of *Cannabis* accessions. Governments and universities have been
 84 the custodians of large-scale *ex-situ* germplasm collections for centuries. One such
 85 government-owned collection is the USDA-ARS National Plant Germplasm System (NPGS),
 86 the world's largest supplier of germplasm, holding 617,467 accessions representing 17,482
 87 species²⁵. As a result of federal prohibition starting in 1937, hemp germplasm has been
 88 excluded from such conservation until the 2021 establishment of the USDA-ARS hemp
 89 germplasm repository in Geneva, New York. The current collection holds 581 *Cannabis*
 90 *sativa* accessions of which 57 are available for distribution.

91

92 Escaped, naturalized, and regionally adapted feral cannabis populations that exist across the
 93 US represent a potential source of untapped genetic diversity. These populations potentially
 94 harbor a wide range of genetic variation lost in cultivated crops, providing opportunities to
 95 introduce new traits for improved crop performance. As crops become feral, traits suited to
 96 local conditions like drought, extreme temperatures, poor soil, could be selected for. These
 97 adaptations are valuable for developing climate-resilient crops by incorporating genetic
 98 material from feral populations via breeding. Moreover, studying feral populations offers
 99 insights into the evolutionary processes of crop domestication and artificial selection. Several

studies have described the population structure and chemical profiles in diverse *C. sativa* collections^{26–32} by using next-generation sequencing methods on mostly outsourced drug type *Cannabis* from Canada, Europe or Asian origin. Previous studies from four US states (Minnesota, Nebraska, Kansas and Colorado) have explored feral *Cannabis* diversity and found feral plants to be a rich source of genetic diversity^{33–35}. The goal of this study is to populate the recently established USDA-ARS hemp germplasm repository in Geneva, NY, with genetically characterized, diverse, and compliant plant material to provide a valuable resource for the scientific community. To complete this goal we executed a large-scale collection effort to collect, genotype, and chemotype unique feral *Cannabis* germplasm from across the US to support research efforts, conserve biodiversity and develop regionally adapted, compliant hemp cultivars.

RESULTS

Sample collection

Seeds and floral tissues from 1,821 feral plants were collected across 12 states and 91 unique populations, in the late summer or early fall of 2022 and 2023 (Figure 1a and 1b; Supplemental Table 1). The largest collection of samples came from Wisconsin (n = 589) and the smallest from Colorado (n = 11).

SNP distribution

Among the 91 sampled populations, 5-20 individual plants were selected for Genotyping-By-Sequencing (GBS). A total of 2,739,241,114 barcoded reads were aligned with the *Cannabis* reference genome⁹ resulting in identification of 159,382 polymorphic SNPs. After filtering, the final dataset included 21,037 high quality SNPs. SNP counts ranged from 1,386 on

chromosome 7 to 2,829 on chromosome 1 (Supplemental Table 2), with well distributed coverage across all chromosomes (Supplementary Figure 1).

Population Structure

The cross-validation method by ADMIXTURE was unable to resolve the optimum number of clusters (Supplemental Figure 2) for the samples. However, the Bayesian Information Criterion (BIC) value indicated a rapid decline until $K = 5$ clusters, after which the value began to rise, suggesting that five groups should be retained (Figure 2a). Following ADMIXTURE analysis, individual feral *Cannabis* accessions were assigned into a group based on their maximum proportion of membership (Q) value (Figure 2b). To avoid groups with recent admixture between populations or cultivated *Cannabis*, a > 70 percent membership threshold was used to assign subpopulations identity, retaining 346 samples (Supplementary Table 3). Group 1 (Mississippi-River) had the largest number of samples with a total of 177 accessions largely composed of feral collections from eight states (MN, WI, IL, MO, SD, IA, NE and KS) and dominated by accessions from WI (57), MN (52), and IL (36) (Supplementary Table 4). Group 2 (West North Central-a) comprised 96 accessions from SD, NE, IA, ND, CO, KS, MO, IL and MN, however SD (40 accessions), NE (17 accessions), and ND (12 accessions) were the most represented in this group. Group 3 (New York) comprised 14 samples from New York. Group 4 (West North Central-b) comprised 37 samples, with accessions from SD (11), IA (8) and NE (7), similar to Group 2. Finally, group 5 (Indiana) comprised 22 accessions from Indiana. Principal component analysis (PCA) of the ~21K SNPs sorted the feral *Cannabis* accessions into the same five major groups (Figure 3a). The majority of the variation explained came from New York and Indiana, respectively, while the other groups clustered tightly. Samples mostly grouped by states with a few minor exceptions (Figure 3b and 3c).

149

150 **Genetic Differentiation**

151 The F_{ST} values between the inferred groups ranged from 0.07 to 0.39 (Table 1). Pairwise
152 genetic differentiation was highest between the New York and Indiana groups ($F_{ST} = 0.39$),
153 with very low values ($F_{ST} = 0.07$ - 0.09) between Mississippi-River, West North Central-a and
154 West North Central-b. Sliding window F_{ST} analysis revealed extremely high F_{ST} values on
155 chromosome 4 and the X chromosome between groups (Figure 4a and 4b, Supplementary
156 Figure 3). The SNPs associated with elevated F_{ST} (65 – 90 Mb) were analyzed separately by
157 PCA showing three discrete clusters (Supplementary Figure 4a and 4b) explaining 63.9% of
158 the observed variation. Inspection of this region indicated a large genome rearrangement with
159 respect to the Cs10 genome that has been observed previously in feral samples^{21,34}. The
160 inversion genotype (+/+) was characterized for each of the five groups and was found at 93%
161 and 91% in the West North Central-a (Group 2) and Indiana (Group 5) populations,
162 respectively (Table 2).

163

164 **Genetic Diversity Analysis**

165 Minor allele frequency (MAF) observed and expected heterozygosity (H_o and H_e),
166 polymorphism information content (PIC), and the inbreeding coefficient (F_{IS}) are used as
167 indicators of the genetic diversity within a population³⁶. The average H_o , H_e , MAF, PIC, F_{IS}
168 of the 346 accessions were 0.20, 0.20, 0.19, 0.26 and -0.03, respectively (Table 3). Analysis
169 of the genetic diversity of the five *Cannabis* groups showed MAF ranged between 0.12 -
170 0.22, while polymorphic information content (PIC) was between 0.15 - 0.30. The H_o , MAF
171 and PIC of the New York population were lower than those of the Midwestern populations.
172 Hierarchical Analysis of Molecular Variance (AMOVA) based on ADMIXTURE groupings
173 found approximately 9% variation between groups, 20% between accessions that were

clustered within groups and extensive variation (71%) among all samples (Supplementary Table 5).

Linkage Disequilibrium

LD decay occurred rapidly with increasing physical distance. The average maximum r^2 value was 0.47. As LD decayed to its half maximum ($r^2 = 0.24$), the corresponding physical distance was approximately 6.7 kb (Supplementary Figure 5).

CBDAS Genotyping

Among US feral *Cannabis* collections, CBDAS genotypes were resolved for approximately 1,400 accessions from 90 populations located among 12 states into one of three diploid classes (Supplementary Table 6). Among all accession genotypes, 1120 were homozygous for functional ($C^F C^F$), 91 were homozygous for non-functional ($C^X C^X$), and 208 were heterozygous ($C^F C^X$) comprising study wide allele frequencies of 0.86 and 0.14 for functional (C^F) and non-functional (C^X) CBDAS, respectively (Supplementary Table 7). Among a subset of 53 populations with ≥ 10 resolved genotypes, functional CBDAS (C^F) allele frequency ranged from 0.99 to as low as 0.4 among 36 populations and was fixed at 1.0 among 17 populations (Supplementary Table 7). The CBDAS allele frequency among this population subset varied geographically across a region comprising 10 contiguous upper midwestern states (Figure 5). Populations from which accession genotypes were exclusively homozygous for functional CBDAS were located in all but two states, North Dakota and Indiana, at the northeastern and southeastern margins of the upper Midwest sampling region. A total of five populations with exclusively functional CBDAS alleles were also composed of individuals belonging entirely to the Mississippi River group. In addition, three populations

outside of the upper Midwest, two in Colorado and one in New York, were completely homozygous for functional *CBDAS* (Supplementary Table 8).

Cannabinoid Analysis

Among US feral *Cannabis* collections across the 10 states comprising the upper Midwest sampling region (Figure 1b), female floral tissues (pistillate inflorescences) from 523 accessions were analyzed by gas chromatography for determination of the % dry weight content of eight cannabinoid compounds except for 30 accessions for which determinations included seven compounds (omitting Δ^8 -THC; Supplementary Table 6). When categorized by the ratio of Δ^9 -THC:CBD content, these accessions included 91 THC-type, 233 intermediate-type, and 244 CBD-type feral *Cannabis* plants selected from 61 populations (Supplementary Table 8). Total cannabinoid content ranged 0.23-2.96%, 0.29-4.09%, and 0.21-4.73% among accessions classified as THC-type, intermediate-type, and CBD-type, averaging ($\pm$ SE) 0.98% (0.07), 1.38%, and 1.50%, respectively (Supplementary Table 8). Histograms of total cannabinoid content among 523 feral *Cannabis* accessions are shown in Figure 6, along with the histograms for a subset of accessions meeting criteria for assignment to one of five ADMIXTURE groups.

DISCUSSION

US hemp breeding programs have been legislatively impeded for nearly a century due to federal prohibition under the 1937 Marihuana Tax Act and the Controlled Substances Act of 1970. Federal drug law policies prohibited the cultivation and use of *Cannabis* in any form, crippling potential industrial and medicinal use. Interests in industrial hemp production increased with the passage and expansion of the 2014¹² and 2018 Farm Bill¹³. Still, the

limited availability of complaint and diverse *Cannabis* germplasm represents a major bottleneck for breeding and research efforts.

Although the last legal US hemp crops were planted in 1958^{37,38}, remnants of domesticated industrial hemp have escaped to form naturalized, thriving feral populations across regions of the US where it had been previously cultivated. Additional dispersal of seed by birds, animals, and anthropogenic activities might have resulted in further establishment in regions where hemp had not previously been cultivated. Sites suggestive of such secondary range expansion include roadsides, railroad rights-of-way, livestock feedlots, and agricultural boundaries especially as associated with trees and utility lines. We relied on university collaborators, farmers, and citizen scientists to identify and/or collect feral *Cannabis* plants across the US based on sites recorded in iNaturalist or received by word of mouth. Whereas site identification was essentially straightforward, collection of samples was occasionally challenging because of residual stigma associated with cannabis. Additionally, occurrences within interstate highway corridors, on private property, and accompanied by dense vegetation made collection difficult. Wider identification of feral *Cannabis* occurrence and collection is particularly needed in the West, South, and Northeast geographical regions of the US. Meeting this need will require effective communication about the potential economic benefits of adopting industrial hemp production among existing agricultural crops.

Five subpopulations were identified among US feral *Cannabis* collections. Generally, samples clustered according to state of origin showing that geographical location explains genetic variation. Most notable were two clusters, one containing accessions primarily from Indiana and the other containing accessions from New York. The remaining three clusters

were less variable and contained samples from states along the Mississippi River basin and within West North Central states.

Pairwise comparison of fixation indices (F_{ST}) between subpopulations found genetic distance to be in accordance with geographical location with the most divergent subpopulations from Indiana (group 5) and New York (group 3). The F_{ST} values from this study are higher than those previously recorded in numerous studies focused on drug-type *Cannabis*^{26,27,29,34}, which is likely attributed the broader genetic background of hemp as compared to drug-type *Cannabis*. The least divergent subpopulations ($F_{ST} = 0.07 - 0.09$) may have resulted from a single introgression to the Upper Midwest during the WWII “*hemp for victory*” era followed by dispersal by humans or animals across the sampled regions. Interestingly, many of the collection sites in these regions were either along railroad rights-of-way or livestock facilities. High F_{ST} between the Indiana and New York groups indicate separate introgressions. Indeed, hemp arrived at different geographical regions at distinct periods throughout history with samples arriving in the Northeast as early as the 16th century, Kentucky in the 18th century and Wisconsin during the 20th century^{19,39,40}. The three primary subpopulations identified in the current study may have originated from distinct introgression and feralization events and therefore harbor unique alleles based on origin.

High pairwise F_{ST} scores were found for all comparisons indicating regions of the genome that may be under selective pressures across different regions. The highest F_{ST} values were associated with genomic regions on chromosomes 4 and X (Figure 4; Supplementary Figure 4). Previous work identifies similar peaks and hypothesizes that the regions may be the result of chromosomal inversions maintained at high frequency in the feral US populations³⁴. The *Cannabis* pangenome has confirmed a large-scale inversion on chromosome 4 in a feral

collection from Boone County, Iowa as compared to high cannabinoid cultivars⁴¹. We find this same elevated F_{ST} primarily in samples from the West North Central-a and Indiana groups suggesting it is segregating within US feral populations. Low global F_{ST} observed between West North Central-a and -b groups, indicates these accessions are mostly differentiated by this inversion. Analysis of the 15 Mb region surrounding the chromosome 4 inversion explains 65% of the variation among samples in the PCA (Supplementary Figure 4). Because this inversion is found in some frequency across all subpopulations, it is likely this inversion occurred prior to the introgression of hemp to the United States. Fixation in several populations suggests that it may confer a selective advantage in certain environments.

The expected heterozygosity level (0.20) for the accessions in our study was similar to the observed heterozygosity (0.20) indicating the populations are not experiencing significant selective forces such as inbreeding or genetic draft. Our observed heterozygosity is more similar to what has been previously described in hemp (0.27-0.16) as compared to drug-type cannabis (0.12-0.14)^{26,28}. A lower observed heterozygosity was observed in the New York subpopulation which may be attributed to its small population size and geographic isolation. The negative value of inbreeding coefficients recorded in the MS-River, WNC-a and WNC-b subpopulations, imply increased heterozygosity and higher genetic diversity when compared to the Indiana and New York feral values reflecting random mating and genetic stability among the populations. Polymorphism Information Content (PIC) is a measurement of how well a genetic marker can distinguish different individuals in a population. It is a key indicator of marker quality in genetic studies and is used to predict the likelihood of a marker's heterozygous genotype being passed on to offspring⁴². The PIC marker values per population ranged between 0.15 – 0.30, indicating they are very informative at discriminatory analysis between feral populations.

296

297 High genetic variability from AMOVA as observed within accessions followed by between
298 accessions within groups and lowest between groups suggests potential for trait discovery in
299 breeding programs. The moderate differentiation between groups reiterates the high
300 heterozygosity often recorded in *Cannabis* associated with wind pollination. Within
301 accession AMOVA values among US feral collections was higher than that of the Spanish
302 and Moroccan *Cannabis* germplasm studies^{34,43}. Discrepancies in variation contributed by
303 these studies are consistent with other studies and may be due to differences in sample size,
304 sample composition and types of markers used.

305

306 Linkage disequilibrium decayed rapidly (6.7kb) in the feral germplasm. This was similar to a
307 recent report using a mixture of hemp and drug-type ferals (LD = 6.0 kb)²⁹ and faster than in
308 a recent study of drug-type *Cannabis* averaging between 22.6 to 89.0 kb across
309 chromosomes⁴⁴. Historic breeding efforts during the 18th and 19th centuries that entailed
310 intentional hybridization as well as wind dispersal of hemp pollen over great distances may
311 account for the plunge in LD decay of feral *Cannabis* plants. This finding will be important
312 for finding marker-trait associations and developing genotyping resources to adequately
313 cover haplotype blocks.

314

315 Total cannabinoid content in Midwestern feral *Cannabis* was rather low compared to drug
316 cultivars although somewhat higher than hemp cultivars⁴⁵ with population averages ($\pm$ SE)
317 ranging from 0.62 (0.05) to 4.04 (0.26) (Supplementary Table 8). Most populations contained
318 at least some individual plants with THC content above the US statutory threshold of 0.3%.
319 Such plants were either of intermediate ($C^F C^X$) or THC-type ($C^X C^X$) genotypes as indicated
320 by the presence of non-functional *CBDAS* alleles at low frequency. Approximately one third

of 53 populations with at least 10 resolved accession genotypes were fixed for functional *CBDAS*, with mixed populations varying in non-functional allele frequency from >0.1 to 0.6. These observations are consistent with evidence from Minnesota³⁵ suggesting that midwestern “feral hemp” does not neatly fit the current statutory definition of hemp but is more accurately described as feral *Cannabis*. Genotype frequencies in mixed populations were generally skewed in favor of CBD-type, with a lower frequency of intermediates and the occasional THC-type plant. There is little reason to expect that the original *Cannabis* introductions to the region in the 19th century for fiber production were pure hemp, considering that the chemical nature of intoxicating versus non-intoxicating *Cannabis* (THC versus CBD) was not discovered until decades later^{46,47}. Populations at or near Hardy-Weinberg equilibrium with relatively low yet varying non-functional *CBDAS* allele frequencies might be interpreted as subject to local genetic drift with little or no selection on cannabinoid content and/or little immigration of THC-type genetics from elsewhere. However, several populations deviated from this pattern with a high proportion of intermediate-type plants or a non-equilibrium excess of THC-type plants, which might be interpreted as the result of introgression from more recent drug-type cultivation. Genotyping of historical specimens from herbarium collections might shed light on how and whether cannabinoid genetics has changed over time⁴⁸.

CBDAS genotypes as measured by the PACE assay⁴⁹ accurately predicted each THC:CBD ratio class in most cases (87%). In three cases, cannabinoid content was so close to the limit of detection by gas chromatography (0.01%) that the THC:CBD ratio lacked meaning, and we do not consider them to challenge the accuracy of the genetic assay. However, 33 of 465 accessions (13%) had *CBDAS* genotypes that did not correlate with the THC:CBD ratio even after independent, repeated measures. Such plants merit further investigation of the

possibility that allelic variation at the primer binding sites could skew the assay. This seems more likely than a departure from the simple genetic mechanism of codominant inheritance¹¹ as others have suggested⁵⁰. Another limitation of the PACE assay that we observed when applied to feral *Cannabis* was a relatively high failure rate (20%) in which some accessions failed to amplify (18%) while others (2%) yielded ambiguous fluorescence patterns where it was not possible to distinguish between the intermediate genotype and a homozygous state.

Feral accessions are important genetic reservoirs of useful alleles and superior traits for the improvement of *Cannabis* plants for fiber, grain or cannabinoids^{51,52}. The persistence of these feral *Cannabis* populations in selected habitations and even spread to new localities after decades of naturalization attest to their resilience and importance as new tools for breeding improved cultivars with increased adaptation to biotic and abiotic stress factors. The conservation of these *Cannabis* collections in the USDA-ARS Hemp Germplasm Repository will ensure the long-term security of these genetic resources against any future political or societal pressures and their genetic integrity will be continually maintained through thorough regeneration practices. Conservation of these diverse feral *Cannabis* materials in the genebank promises in contrast to historic absence under decades-long prohibition, to provide access to diverse and compliant *Cannabis* germplasm.

MATERIALS AND METHODS

Germplasm collection

To identify feral *Cannabis* sites, we used the iNaturalist application (a joint initiative by the California Academy of Sciences and the National Geographic Society), an American nonprofit social network of citizen scientists and biologists who identify and share information about biodiversity across the globe. Additional locations were shared by

collaborators via word of mouth. Seeds and floral tissue from 1,815 female hemp plants from 88 unique populations were collected across twelve states in 2022 and 2023 (Figure 1a; Supplementary Table S1; Supplementary Material). An additional feral population from New York was obtained from USDA-ARS (Geneva, NY) and seeds from two Colorado populations were contributed from Colorado State University.

Sample preparation and genotyping

A subset of 760 samples, representing twelve states and ecogeographic US locations, were selected for genotyping. Leaf tissue from individual plants was lyophilized (Labconco, Kansas City, MO, USA), then approximately 20 mg dry weight was homogenized (SPEX SamplePrep, Metuchen, NJ, USA), and genomic DNA was extracted using NucleoSpin 96 Plant II core kit (Macherey-Nagel GmbH & Co., Duren, Germany). DNA quality was checked by 1% agarose electrophoresis and quantified with NanoDrop ND-1000 Spectrophotometer (Thermo Fisher Scientific, USA). The feral hemp DNA quantification, library preparation, and sequencing was constructed using the genotyping-by-sequencing (GBS) protocol⁵³ at the University of Wisconsin-Madison Biotechnology Center. In summary, Illumina adapters and sample-specific barcodes were annealed after restriction enzyme digestion using *ApeKI*. Multiplexed samples were then sequenced on Illumina NovaSeq 6000 (Illumina, Inc., San Diego, CA, USA), generating on average 3.5 million, 150 bp paired end reads per sample.

Variant discovery and data filtering

The Tassel GBS Discovery Pipeline v2⁵⁴ was used for variant detection, using the cs10 reference genome v2⁹ (Genbank assembly accession ID = GCF_900626175.2) with BWA-MEM⁵⁵. Raw VCF files were filtered for SNPs > 10 % missing data, < 5 % minor allele

frequency, > 10 % missing individuals, and finally pruned to reduce redundancy by removing SNPs that are highly correlated with others (--geno 0.1, --maf 0.05, --mind -0.1, --indep-pairwise 50 2 0.2) using PLINK V.1.90b6.21⁵⁶. The final VCF file contained 21,037 SNPs (~21K) used in downstream analyses. Chromosome-wide SNP count and density plots were performed (<https://github.com/YinLiLin/R-CMplot>) in the R environment (version 4.3.2).

Population structure and principal component analysis

Population structure was determined using ADMIXTURE⁵⁷ to estimate the genetic ancestries of each feral accession, and cross-validation plots were observed for the optimum number of clusters. ADMIXTURE identifies K genetic clusters, where K is specified by the user ($K2-K40$), from the provided SNP data. For each individual, the ADMIXTURE method estimates the probability of membership to each cluster using Bayesian Information Criterion. Low admixture samples were assigned by setting the membership coefficient to 0.7. Eigenvalues and eigenvectors for genome-wide principal component analysis (PCA) were generated using PLINK open-source analysis toolset and then plotted in the R-software⁵⁸, package “ggplot”⁵⁹.

Genetic differentiation and diversity indices

To measure genome-wide genetic differentiation (F_{ST}) and molecular variance (AMOVA) among identified subgroups, poppr⁶⁰ and Hierfstat⁶¹ R packages were employed. Diversity indices (H_o = Observed heterozygosity, H_e = Expected heterozygosity, MAF = Minor allele frequency, PIC = Polymorphic information content, and F_{IS} = Inbreeding coefficient) were estimated with PLINK. To estimate localized genetic differentiation across the genome, pairwise F_{ST} was calculated using VCFtools with a 100,000 bp, non-overlapping sliding window and plotted in R using package “ggplot”.

Linkage disequilibrium

The LD squared allele frequency correlation (r^2) estimate for each pairwise SNP comparison was generated in TASSEL 5.0 and visualized by plotting r^2 values against physical distance in R package “ggplot”. A nonlinear regression curve was used in the estimation of LD decay. The LD decay rate was determined by the point of intersection with the LD curve and half its maximum value ($r^2 = 0.24$).

CBDAS locus genotyping

To resolve *CBDAS* locus genotypes of accessions, we used the CCP1 primer set⁴⁹ that has been validated to detect and distinguish presence of functional (C^F) and non-functional (C^X) pseudogenous alleles of *CBDAS* that are tightly and opposingly linked to non-functional pseudogenes and a functional allele of the *THCAS* gene^{9,11}. PACE™ 2.0 2X master mix with standard ROX was used according to the manufacturer’s protocol (3CR Bioscience Limited) to produce fluorescently labeled *CBDAS* locus real-time PCR fragments on Bio-Rad CFX96, Roche 980, or Thermo-Fisher QuantStudio qPCR instruments following respective PACE™ protocols for bi-allelic genotyping. Genomic DNA from feral *Cannabis* accessions were prepared as assay templates from dried leaf or floral tissues using either Extract-N-Amp™ Plant PCR (Millipore Sigma) extraction reagents or using a Nucleospin 96 Plant II core kit (Macherey-Nagel GmbH & Co., Duren, Germany) kit according to the respective manufacturer’s protocol.

Chemical cannabinoid diversity

Dried floral tissue samples were analyzed using gas chromatography⁶² to measure the percentage of total inflorescence dry mass of: cannabichromene (CBC), CBD, cannabigerol (CBG), cannabinol (CBN), delta-8-tetrahydrocannabinol (d8-THC), THC, and

tetrahydrocannabivarin (THCV). We report total cannabinoid content as the sum of these individual compounds as well as the percent CBD and THC content. Cannabinoid profiling was conducted on up to five plants of each genotype, as determined by the PACE assay, per population.

REFERENCES

1. ElSohly, M. A. & Slade, D. Chemical constituents of marijuana: The complex mixture of natural cannabinoids. *Life Sci* 78, 539–548 (2005).
2. Radwan, M. *et al.* Isolation and Characterization of New Cannabis Constituents from a High Potency Variety. *Planta Med* 74, 267–272 (2008).
3. Poklis, J. L., Thompson, C. C., Long, K. A., Lichtman, A. H. & Poklis, A. Disposition of Cannabichromene, Cannabidiol, and 9-Tetrahydrocannabinol and its Metabolites in Mouse Brain following Marijuana Inhalation Determined by High-Performance Liquid Chromatography-Tandem Mass Spectrometry. *J Anal Toxicol* 34, 516–520 (2010).
4. Mehmedic, Z. *et al.* Potency Trends of Δ^9 -THC and Other Cannabinoids in Confiscated Cannabis Preparations from 1993 to 2008*. *J Forensic Sci* 55, 1209–1217 (2010).
5. Fellermeier, M., Eisenreich, W., Bacher, A. & Zenk, M. H. Biosynthesis of cannabinoids. *Eur J Biochem* 268, 1596–1604 (2001).
6. Johnson, M. S. & Wallace, J. G. Genomic and Chemical Diversity of Commercially Available High-CBD Industrial Hemp Accessions. *Front Genet* 12, (2021).
7. de Meijer, E. P. M. *et al.* The Inheritance of Chemical Phenotype in *Cannabis sativa* L. *Genetics* 163, 335–346 (2003).
8. Staginuss, C., Zörntlein, S. & de Meijer, E. A PCR marker linked to a THCA synthase Polymorphism is a Reliable Tool to Discriminate Potentially THC-Rich Plants of *Cannabis sativa* L. *J Forensic Sci* 59, 919–926 (2014).
9. Grassa, C. J. *et al.* A new *Cannabis* genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana. *New Phytologist* 230, 1665–1679 (2021).

- 477 10. Grassi, G. & McPartland, J. M. Chemical and Morphological Phenotypes in
478 Breeding of *Cannabis sativa* L. in *Cannabis sativa* L. - *Botany and*
479 *Biotechnology* 137–160 (Springer International Publishing, Cham, 2017).
480 doi:10.1007/978-3-319-54564-6_6.
- 481 11. Weiblen, G. D. *et al.* Gene duplication and divergence affecting drug content in
482 *Cannabis sativa*. *New Phytologist* 208, 1241–1250 (2015).
- 483 12. US Congress. *Agricultural Act of 2014*. 113–79 (Senate and House of
484 Representatives of the United States of America, Washington, 2014).
- 485 13. US Congress. *Agriculture Improvement Act of 2018*. 115–334 (Senate and
486 House of Representatives of the United States of America, 2018).
- 487 14. Fike, J. Industrial Hemp: Renewed Opportunities for an Ancient Crop. *CRC Crit*
488 *Rev Plant Sci* 35, 406–424 (2016).
- 489 15. Schoenrock, R. E. Hemp in Minnesota during the wartime emergency.
490 (Mankato State College, Minnesota, 1966).
- 491 16. Robinson, R. *The Great Book of Hemp: The Complete Guide to the*
492 *Environmental, Commercial and Medicinal Uses of the World's Most*
493 *Extraordinary Plant*. vols 978-089281541–8 (1996).
- 494 17. Wilsie, C. P. & Black, C. A. Hemp production experiments cultural practices
495 and soil requirements. *Iowa State University Digital Repository* 3, (1944).
- 496 18. Haney, A. & Kutscheid, B. B. An Ecological Study of Naturalized Hemp
497 (*Cannabis sativa* L.) in East-Central Illinois. *American Midland Naturalist* 93, 1
498 (1975).
- 499 19. Lyster H. Dewey. *Hemp: USDA Yearbook of Agriculture*. (1914).
- 500 20. Stack, G. M. *et al.* Comparison of Recombination Rate, Reference Bias, and
501 Unique Pangenomic Haplotypes in *Cannabis sativa* Using Seven De Novo
502 Genome Assemblies. *Int J Mol Sci* 26, 1165 (2025).
- 503 21. Lynch, R. C. *et al.* Domesticated cannabinoid synthases amid a wild mosaic
504 cannabis pangenome. Preprint at <https://doi.org/10.1101/2024.05.21.595196>
505 (2024).
- 506 22. Faux, A.-M., Draye, X., Flamand, M.-C., Occre, A. & Bertin, P. Identification
507 of QTLs for sex expression in dioecious and monoecious hemp (*Cannabis*
508 *sativa* L.). *Euphytica* 209, 357–376 (2016).

23. Petit, J. *et al.* Elucidating the Genetic Architecture of Fiber Quality in Hemp (Cannabis sativa L.) Using a Genome-Wide Association Study. *Front Genet* 11, (2020).
24. de Ronne, M., Lapierre, É. & Torkamaneh, D. Genetic insights into agronomic and morphological traits of drug-type cannabis revealed by genome-wide association studies. *Sci Rep* 14, 9162 (2024).
25. GRIN-Global. U.S. National Plant Germplasm System. GRIN-Glob. Web version 11040. GRIN-Global. <https://npgsweb.ars-grin.gov/gringlobal/query/summary> (2023).
26. Sawler, J. *et al.* The Genetic Structure of Marijuana and Hemp. *PLoS One* 10, e0133292 (2015).
27. Lynch, R. C. *et al.* Genomic and Chemical Diversity in Cannabis. *CRC Crit Rev Plant Sci* 35, 349–363 (2016).
28. Soorni, A., Fatahi, R., Haak, D. C., Salami, S. A. & Bombarely, A. Assessment of Genetic Diversity and Population Structure in Iranian Cannabis Germplasm. *Sci Rep* 7, 15668 (2017).
29. Ren, G. *et al.* Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa. *Sci Adv* 7, (2021).
30. Chen, X. *et al.* Whole-genome resequencing of wild and cultivated cannabis reveals the genetic structure and adaptive selection of important traits. *BMC Plant Biol* 22, 371 (2022).
31. Alsaleh, A. & Yilmaz, G. Exploring cannabidiol variations, investigation of genetic diversity, population structure and unveiling male-specific genetic marker in industrial hemp (Cannabis sativa L.). *Genet Resour Crop Evol* (2024) doi:10.1007/s10722-024-02015-1.
32. Halpin-McCormick, A. *et al.* Examining population structure across multiple collections of Cannabis. *Genet Resour Crop Evol* (2024) doi:10.1007/s10722-024-01928-1.
33. Busta, L. *et al.* Chemical and genetic variation in feral Cannabis sativa populations across the Nebraska climate gradient. *Phytochemistry* 200, 113206 (2022).
34. Woods, P., Price, N., Matthews, P. & McKay, J. K. Genome-wide polymorphism and genic selection in feral and domesticated lineages of Cannabis sativa. *G3 Genes|Genomes|Genetics* 13, (2023).

35. Wenger, J. P. *et al.* Validating a predictive model of cannabinoid inheritance with feral, clinical, and industrial *Cannabis sativa*. *Am J Bot* 107, 1423–1432 (2020).
36. Greenbaum, G., Templeton, A. R., Zarmi, Y. & Bar-David, S. Allelic Richness following Population Founding Events – A Stochastic Modeling Framework Incorporating Gene Flow and Genetic Drift. *PLoS One* 9, e115203 (2014).
37. Miller, R. L. *Hemp as a Crop for Missouri Farmers: Markets, Economics, Cultivation, Law. Report to Agriculture Task Force, Missouri House of Representatives.* (1991).
38. Cherney, J. & Small, E. Industrial Hemp in North America: Production, Politics and Potential. *Agronomy* 6, 58 (2016).
39. Humphrey, J. R. Marketing Hemp. Kentucky Agricultural Experimental Station. vol. 221 26–43 (1919).
40. Husbands, J. D. Feasibility of Industrial Hemp Production in the United States Pacific Northwest. *USDA Bulletin* vol. 153 42–42 (1909).
41. Lynch, R. C. *et al.* Domesticated cannabinoid synthases amid a wild mosaic cannabis pangenome. Preprint at <https://doi.org/10.1101/2024.05.21.595196> (2024).
42. Serrote, C. M. L., Reiniger, L. R. S., Silva, K. B., Rabaiolli, S. M. dos S. & Stefanel, C. M. Determining the Polymorphism Information Content of a molecular marker. *Gene* 726, 144175 (2020).
43. Benkirane, C. *et al.* Population structure and genetic diversity of Moroccan cannabis (*Cannabis sativa* L.) germplasm through simple sequence repeat (SSR) analysis. *Genet Resour Crop Evol* 71, 2037–2051 (2024).
44. de Ronne, M., Lapiere, É. & Torkamaneh, D. Genetic insights into agronomic and morphological traits of drug-type cannabis revealed by genome-wide association studies. *Sci Rep* 14, 9162 (2024).
45. Hennink, S. Optimisation of breeding for agronomic traits in fibre hemp (*Cannabis sativa* L.) by study of parent-offspring relationships. *Euphytica* 78, 69–76 (1994).
46. Roger Adams, Madison Hunt & J.H. Clark. Structure of Cannabidiol, a Product Isolated from the Marihuana Extract of Minnesota Wild Hemp. I. *Journal of American Chemical Society* 62, 196–200 (1940).

- 576 47. Mechoulam, R., A. *et al.* Some aspects of cannabinoid chemistry. *The Botany*
577 *and Chemistry of Cannabis* 95–115 (1970).
- 578 48. Haney, A. & F.A. Bazzaz. Some ecological implications of the distribution of
579 hemp (*Cannabis sativa* L.) in the United States of America. *The Botany and*
580 *Chemistry of Cannabis* 39–48 (1970).
- 581 49. Toth, J. A. *et al.* Development and validation of genetic markers for sex and
582 cannabinoid chemotype in *Cannabis sativa* L. *GCB Bioenergy* 12, 213–222
583 (2020).
- 584 50. Vergara, D. *et al.* Gene copy number is associated with phytochemistry in
585 *Cannabis sativa*. *AoB Plants* 11, (2019).
- 586 51. Mabry, M. E. *et al.* Building a feral future: Open questions in crop ferality.
587 *PLANTS, PEOPLE, PLANET* 5, 635–649 (2023).
- 588 52. Ford, T., Aina, A., Ellison, S., Gordon, T. & Stansell, Z. Utilizing digitized
589 occurrence records of Midwestern feral *Cannabis sativa* to develop ecological
590 niche models. *Ecol Evol* 14, (2024).
- 591 53. Elshire, R. J. *et al.* A Robust, Simple Genotyping-by-Sequencing (GBS)
592 Approach for High Diversity Species. *PLoS One* 6, e19379 (2011).
- 593 54. Glaubitz, J. C. *et al.* TASSEL-GBS: A High-Capacity Genotyping by
594 Sequencing Analysis Pipeline. *PLoS One* 9, e90346 (2014).
- 595 55. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–
596 Wheeler transform. *Bioinformatics* 25, 1754–1760 (2009).
- 597 56. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and
598 Population-Based Linkage Analyses. *The American Journal of Human Genetics*
599 81, 559–575 (2007).
- 600 57. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of
601 ancestry in unrelated individuals. *Genome Res* 19, 1655–1664 (2009).
- 602 58. R Core Team. R Core Team (2023) R: A Language and Environment for
603 Statistical Computing. R Foundation for Statistical Computing, Vienna. (2023).
- 604 59. Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. (Springer-
605 Verlag New York, 2016).
- 606 60. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of molecular variance
607 inferred from metric distances among DNA haplotypes: application to human
608 mitochondrial DNA restriction data. *Genetics* 131, 479–491 (1992).

61. de Meeûs, T. & Goudet, J. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. *Infection, Genetics and Evolution* 7, 731–735 (2007).

62. ElSohly, M. A. *et al.* Potency trends of delta9-THC and other cannabinoids in confiscated marijuana from 1980-1997. *J Forensic Sci* 45, 24–30 (2000).

Acknowledgements

The authors express their sincere appreciation to Joseph Smeenge and Pawan Basnet for assistance with fieldwork and genotyping, citizen scientist contributors to iNaturalist, the S1084 Multistate Research Project and every individual that helped to identify potential collection sites or collected and sent feral *Cannabis* seeds and/or leaf tissue. We also thank the United States Department of Agriculture – National Institute of Food and Agriculture (USDA-NIFA), Supplemental and Alternative Crops (SAC) grants program (2021-38624-35738) and USDA-ARS agreement number 58-5090-2-03 for funding this research. This work was also supported in part by awards 3000032227 and 3000038264 from the Agricultural Growth, Research, and Innovation (AGRI) Program of the Minnesota Department of Agriculture. Lastly, Weiblen is grateful to the offices of the General Counsel and the Vice President for Research at the University of Minnesota, and the U. S. Drug Enforcement Administration for guidance in developing a field protocol for feral *Cannabis* collection.

Author contributions

SE and GW conceived the research plan. AA, JW, ES, CM, ME, GW and SE performed the research and analyzed the data. AA, SE, JW, and GW wrote the manuscript with input from all the authors. All authors edited and approved the final version of the manuscript.

Data availability statement

DNA sequencing data are available through NCBI's BioProject ID: PRJNA1206134. All other data are available upon request or provided in the supplementary material.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FIGURES

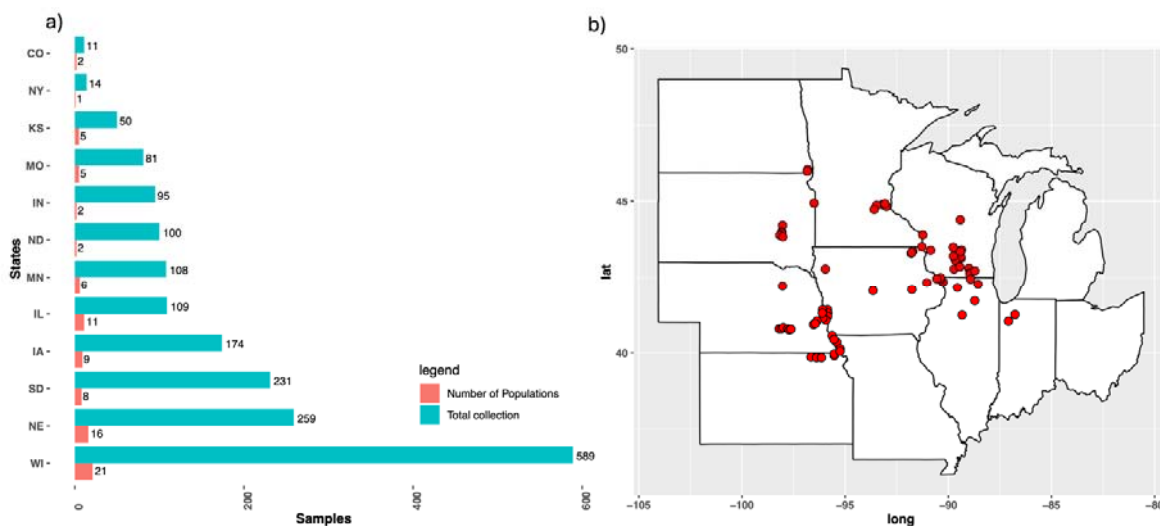


Figure 1. Feral *Cannabis* collection. (a) Bar plot showing total number of individual feral *Cannabis* accessions collected by state alongside the number of populations sampled. (b) Midwestern map showing geographical distribution of sampled locations (Colorado and New York excluded).

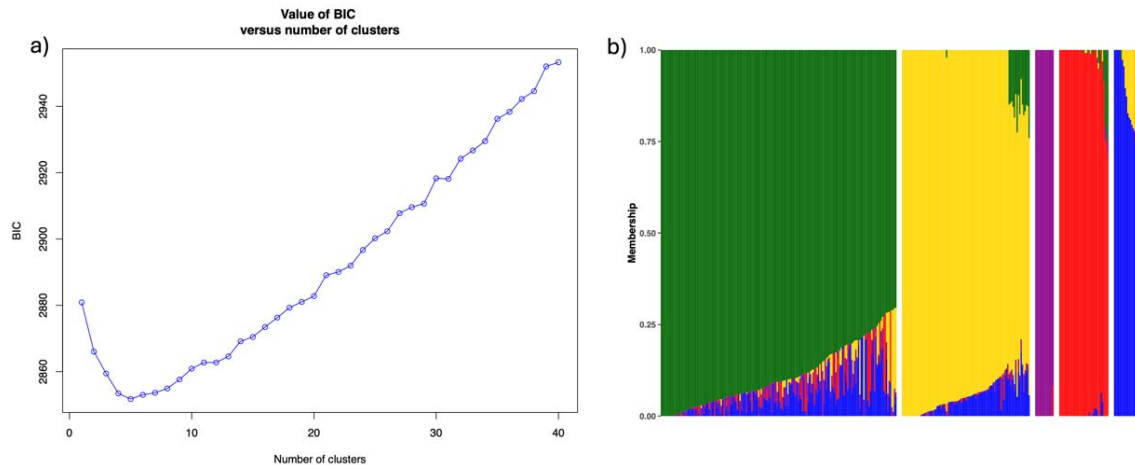


Figure 2. Population ADMIXTURE. (a) Graph of Bayesian Information Criterion versus number of clusters. (b) Admixture ancestry of 346 feral *Cannabis* accessions representing Group 1 (Mississippi-River), Group 2 (West North Central-a), Group 3 (New York), Group 4 (West North Central-b), and Group 5 (Indiana), respectively (from left to right).

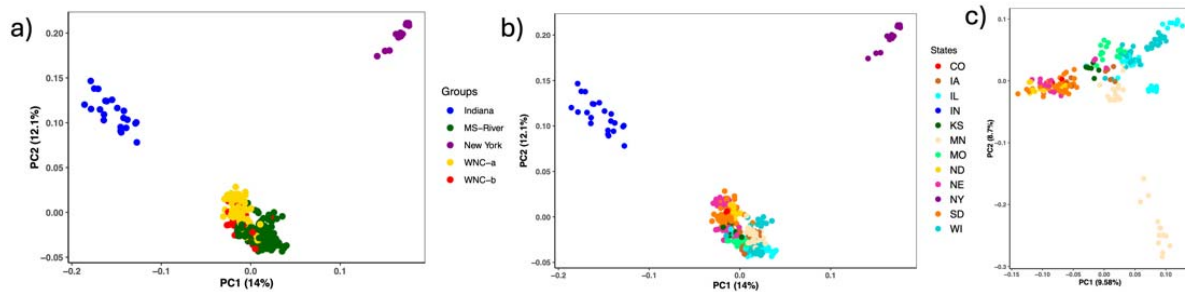


Figure 3. Principle Component Analysis of (a) 346 feral *Cannabis* plants colored by ADMIXTURE clusters, (b) collection states, and (c) collection states after removing Indiana and New York samples.

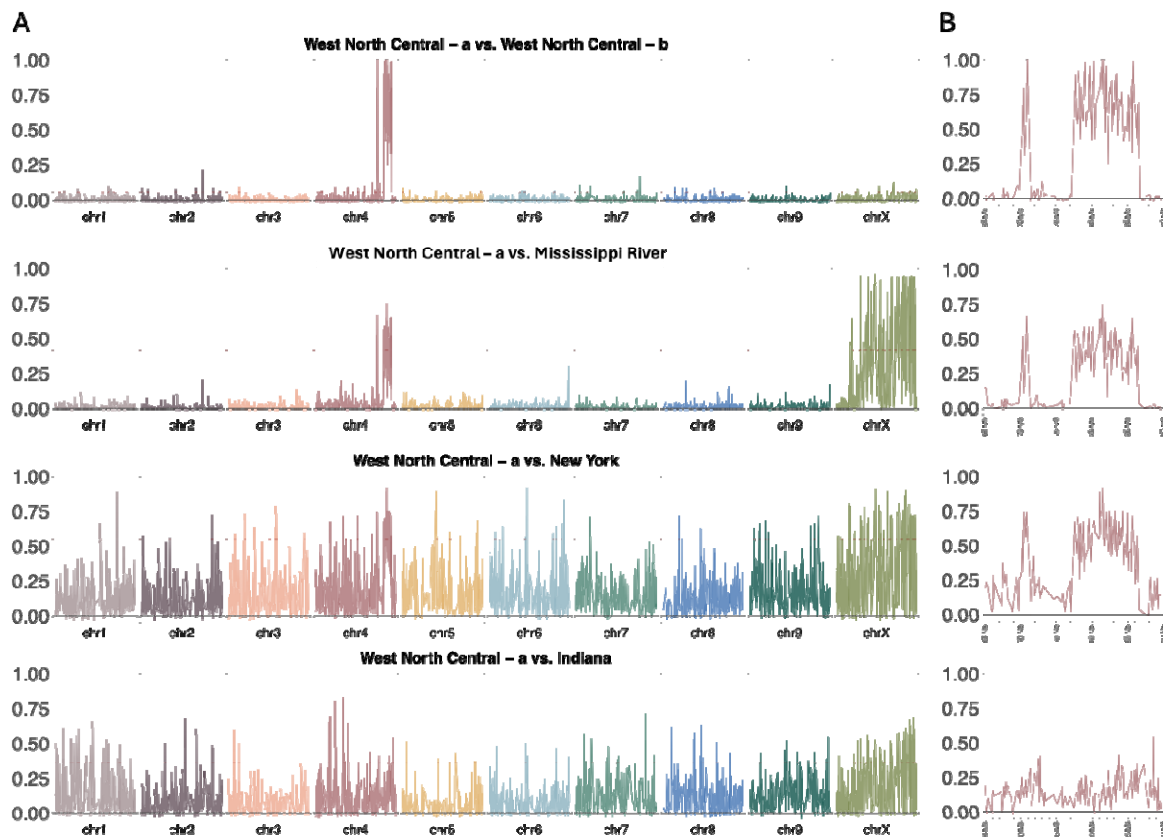


Figure 4. Genetic differentiation between all groups as compared to the “West North Central - a” group as determined by sliding window F_{ST} analysis of 100 kb bins across a) the entire genome and b) a highly differentiated region (65-90 Mb) on Chromosome 4. The red dashed lines indicates the top 5% of F_{ST} values for each comparison.

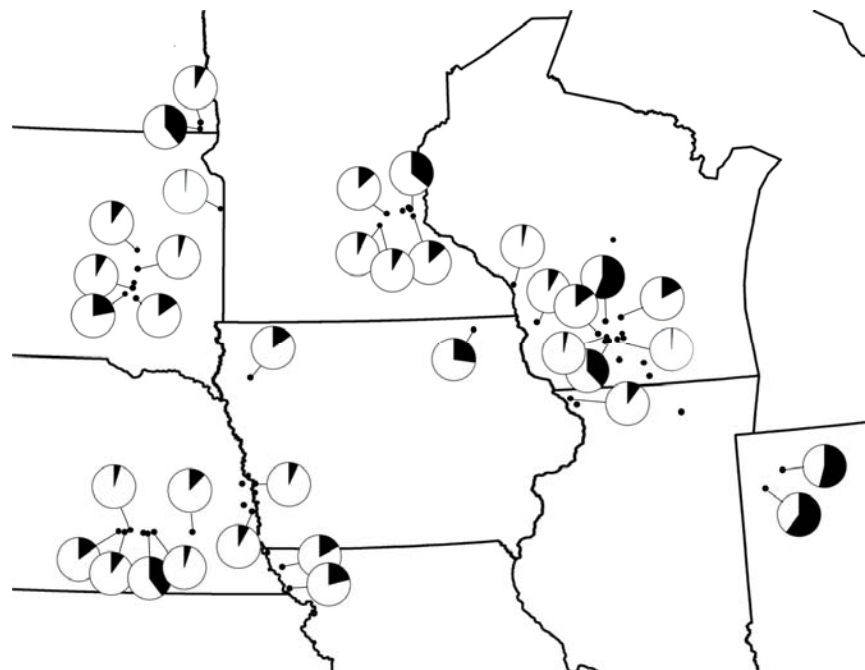


Figure 5. *CBDAS* allele frequency variation among populations with ≥ 10 resolved genotypes. Population localities are indicated by small black circles. Pie-charts display the proportion of functional (C^F ; white) and non-functional (C^X ; black) *CBDAS* alleles observed within populations among individuals for which a *CBDAS* genotype was resolved. Populations not linked to a chart are fixed where $fC^F = 1$.

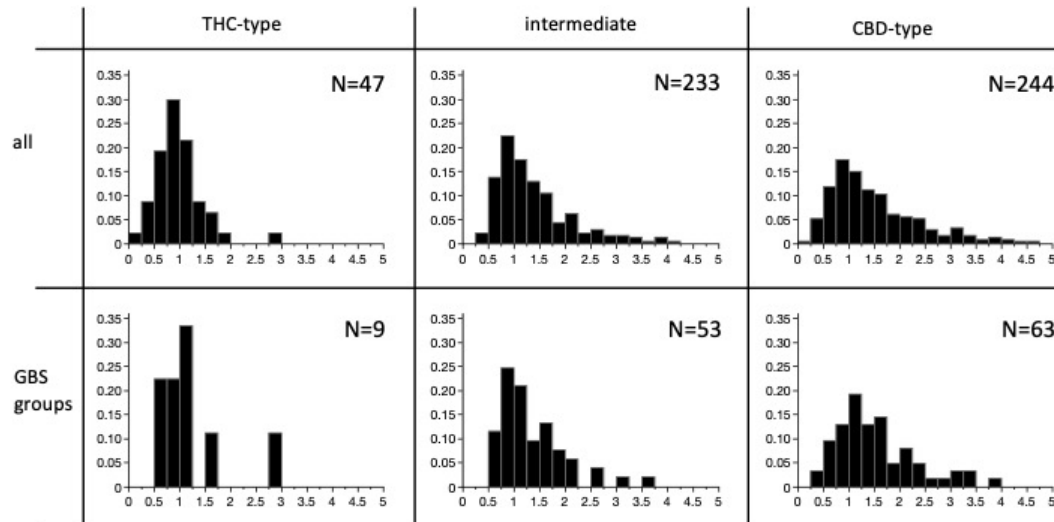


Figure 6. Distribution of total cannabinoid content (% dry weight of “female flower” or pistillate inflorescence) among plants differing in the ratio of THC:CBD. Left, center, and right columns include plants with either an excess of THC (THC-type or type-I), similar amounts of THC and CBD (intermediate or type-II), or an excess of CBD (CBD-type or type-III), respectively. Upper row histograms include all accessions for which cannabinoid content was measured by gas chromatography where N indicates the total number of plants. Lower row histograms include a subset of genotyped plants assigned to each of the three types using the PACE assay. Histograms are proportional to the number of samples sharing $\pm 0.5\%$ cannabinoid content.

TABLES

Table 1. Pairwise comparison of genetic differentiation (F_{ST} values) among the five observed groups.

	MS-River	WNC-a	New York	WNC-b
WNC-a	0.09			
New York	0.22	0.25		
WNC-b	0.07	0.07	0.28	
Indiana	0.17	0.14	0.39	0.20

*Mississippi-River (MS-River), West North Central-a, West North Central-b

Table 2. Frequency of chromosome 4 inversion across subpopulations. -/- indicates same as Cs10 reference (no inversion), +/- is heterozygous for the inversion and ++ is homozygous for the inversion.

ADMIXTURE Group	-/-	+/-	++	Total
-----------------	-----	-----	----	-------

MS-River	0.44	0.46	0.10	177
WNC-a	0.03	0.04	0.93	96
New York	0.79	0.14	0.07	14
WNC-b	0.70	0.03	0.27	37
Indiana	0.00	0.09	0.91	11
Total				335

Table 3. Diversity indices for the entire feral *Cannabis* collection and by subpopulation. *Ho* = Observed heterozygosity, *He* = Expected heterozygosity, MAF = Minor allele frequency, PIC = Polymorphic information content, *F_{IS}* = Inbreeding coefficient.

	<i>Ho</i>	<i>He</i>	MAF	PIC	<i>F_{IS}</i>
MS-River	0.24	0.24	0.22	0.30	-0.08
WNC-a	0.24	0.24	0.22	0.30	-0.05
New York	0.09	0.10	0.12	0.15	0.06
WNC-b	0.24	0.23	0.22	0.30	-0.07
Indiana	0.19	0.20	0.18	0.25	0.00
All samples	0.20	0.20	0.19	0.26	-0.03

Table 4. Cannabinoid content and genotype frequencies among feral *Cannabis* populations. Genotype frequencies are reported for the *CBDAS* locus corresponding to the three major cannabinoid ratio classes, CBD-type ($C^F C^F$), intermediate ($C^F C^X$), and THC-type ($C^X C^X$). Cannabinoid content, including total cannabinoid content (TCC) is reported as the mean percentage of inflorescence dry mass $\pm$ SE.

as the mean percentage of immorescence dry mass = 0.21.							
ADMIXTURE cluster	genotype summary			cannabinoid content			
location	N _g	<i>f</i>	<i>CBDAS</i>	N _c	%THC (SE)	%CBD (SE)	%TCC (SE)
Mississippi R (1)							
WI, Cross Plains	22	0.48	C ^F C ^F	4	0.04 (0.00)	0.58 (0.15)	0.72 (0.16)
T7N R7E	13	0.28	C ^F C ^X	16	0.13 (0.02)	0.36 (0.05)	0.66 (0.08)
	11	0.24	C ^X C ^X	8	0.25 (0.03)	0.02 (0.00)	0.51 (0.06)
WI, West Point	6	0.15	C ^F C ^F	2	0.10 (0.05)	0.59 (0.02)	0.87 (0.15)
T10N R7E	24	0.59	C ^F C ^X	24	0.25 (0.02)	0.74 (0.05)	1.31 (0.08)
	11	0.27	C ^X C ^X	10	0.37 (0.03)	0.04 (0.00)	1.05 (0.09)
WNC-a (2)							
ND, Waldo	44	0.88	C ^F C ^F	7	0.07 (0.01)	1.72 (0.21)	1.94 (0.23)
T130N R49W	4	0.08	C ^F C ^X	6	1.16 (0.07)	1.63 (0.14)	3.19 (0.17)
	2	0.04	C ^X C ^X	-	-	-	-

SD, Cavour	8	0.80	C ^F C ^F	5	0.08 (0.01)	2.18 (0.29)	2.47 (0.32)
T111N R60W	2	0.20	C ^F C ^X	2	1.03 (0.43)	1.40 (0.58)	2.78 (1.13)
	-	-	C ^X C ^X	-	-	-	-
WNC-b (3)							
IA, Remsen	7	0.70	C ^F C ^F	7	0.10 (0.01)	2.17 (0.30)	2.44 (0.33)
T92N R43W	3	0.30	C ^F C ^X	3	0.47 (0.06)	0.78 (0.33)	1.46 (0.43)
	-	-	C ^X C ^X	-	-	-	-
MO, Clark	9	0.75	C ^F C ^F	4	0.09 (0.01)	1.14 (0.25)	1.40 (0.30)
T64N R40W	2	0.17	C ^F C ^X	3	0.59 (0.24)	0.85 (0.22)	1.67 (0.49)
	1	0.08	C ^X C ^X	-	-	-	-
Indiana (4)							
IN, Wayne	12	0.24	C ^F C ^F	11	0.06 (0.01)	0.94 (0.15)	1.12 (0.16)
T32N R3W	21	0.43	C ^F C ^X	28	0.40 (0.02)	0.48 (0.02)	1.02 (0.04)
	16	0.33	C ^X C ^X	10	0.72 (0.07)	0.02 (0.00)	0.90 (0.08)
IN, Barkley	5	0.20	C ^F C ^F	5	0.07 (0.01)	0.99 (0.15)	1.17 (0.17)
T30N R6W	10	0.40	C ^F C ^X	14	0.49 (0.06)	0.57 (0.06)	1.19 (0.12)
	10	0.40	C ^X C ^X	6	0.86 (0.14)	0.03 (0.01)	1.04 (0.16)

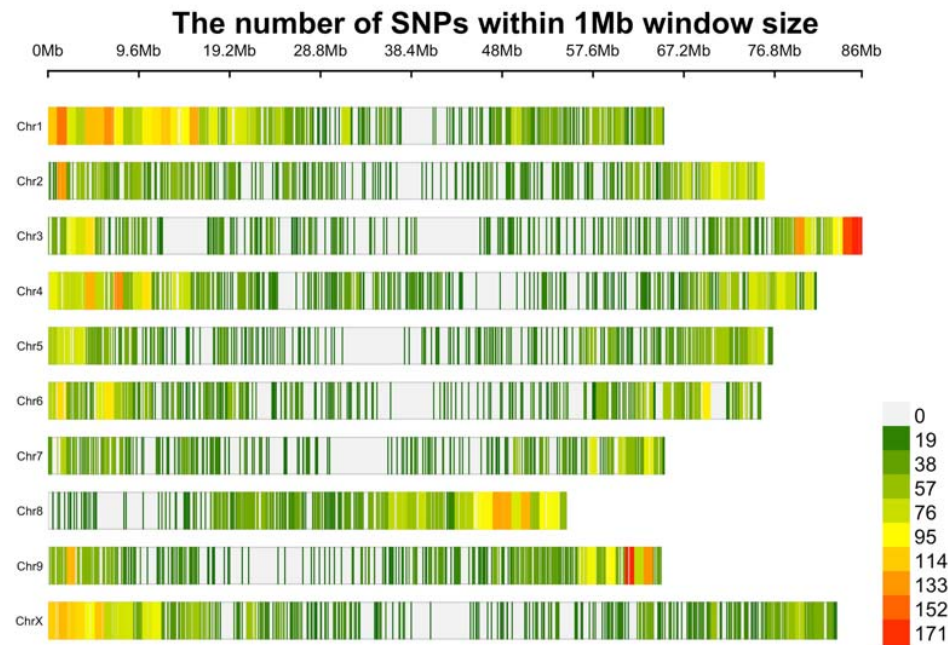
697

698

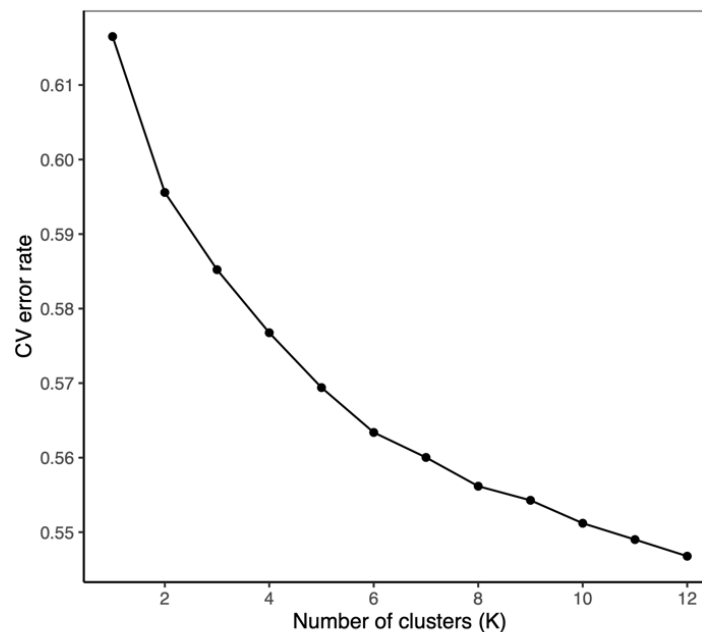
699

700

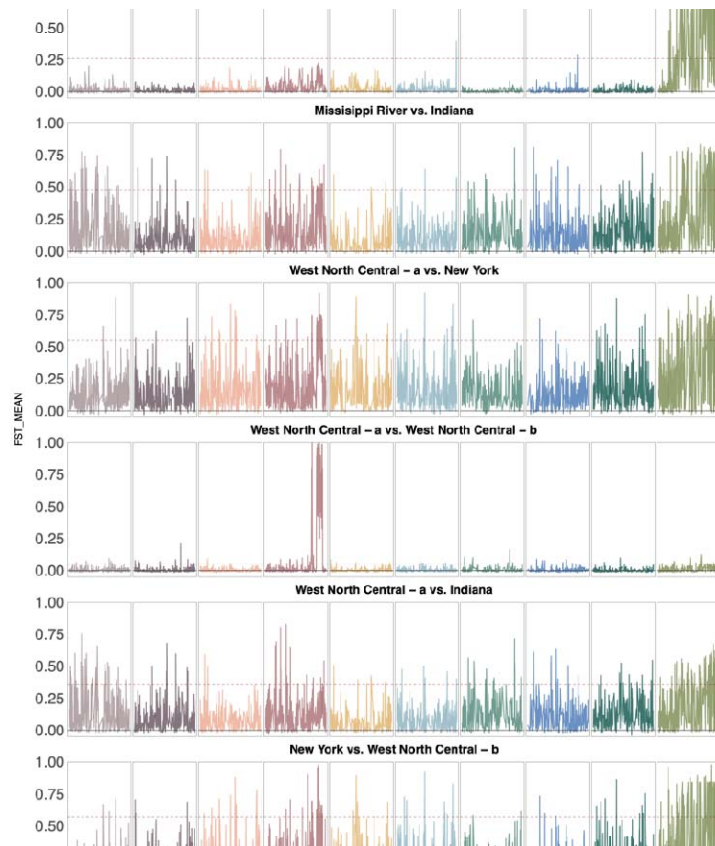
SUPPLEMENTARY FIGURES



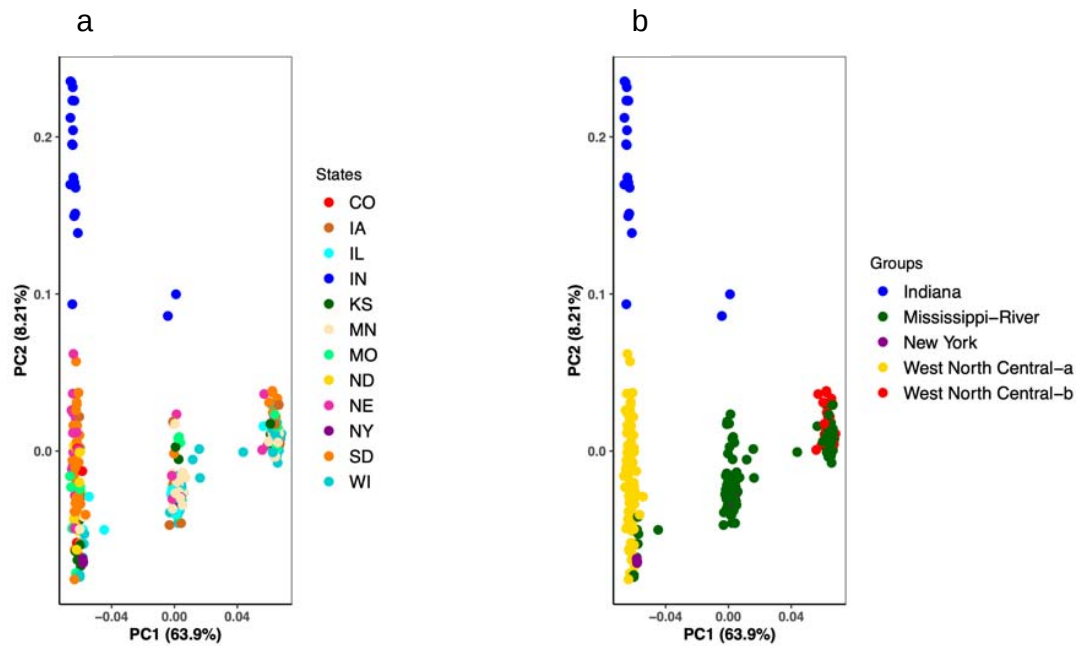
Supplementary Figure 1: SNP density plot of the chromosome-wise number of SNPs within 1 MB window size.



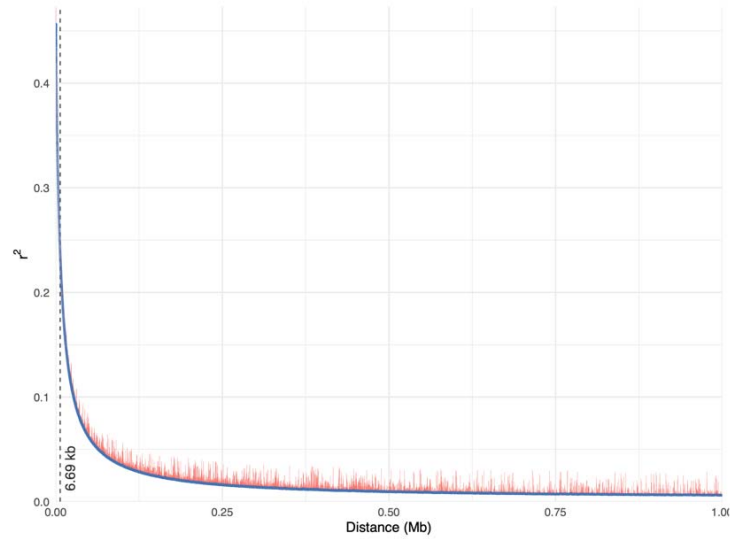
Supplementary Figure 2: Plot of cross-validation error against number of clusters (K).



Supplementary Figure 3: Sliding window FST for all pairwise comparisons of the 5 subgroups.



Supplementary Figure 4: PCA for SNPs within 60-95 Mb region on chromosome 4 colored by a) states and b) ADMIXTURE groups.



Supplementary Figure 5: Genome-wide linkage disequilibrium (LD decay) plot. The X-axis indicates distances, and the Y-axis indicates pairwise correlations (r^2) between all pairs of markers