Phylogenetic resolution of the *Cannabis* genus reveals extensive admixture

Anna Halpin-McCormick[1], Karolina Heyduk[3], Michael B. Kantar[1], Nick Batora[2], Rishi R. Masalia[2], Kerin Law[2], Eleanor J. Kuntz[2]
[1]Department of Tropical Plant and Soil Sciences, University of Hawaii at Manoa, Honolulu, HI 96822
[2]LeafWorks Inc, 130 S. High St, Sebastopol, CA 95472
[3]School of Life Sciences, University of Hawai‘i at Mānoa, Honolulu, HI 96822

Abstract

Cannabis sativa L. (Cannabaceae) is an annual flowering herb of Eurasian origin that has been associated with humans for over 26,000 years. Multiple independent domestications occurred with different events leading to use as food, fiber, and medicine, with human intervention likely accelerating a division in the genus with varietals broadly known today as either hemp-type or drug-type. Using publicly available sequence data, we assessed genome wide diversity and population relationships across seven independently developed datasets. Phylogenetic analysis was conducted, with data sources providing a unique sampling of *Cannabis* varieties, with landrace and modern cultivars represented. Comparing nucleotide diversity over chromosomal length in landrace and domesticated modern drug-type varieties, genomic regions with decreased diversity where humans may have selected for specific traits were identified. Population structure was evident based on use type. Evidence of hybridization was extensive across the datasets. In a subset of landrace individuals where geographic origin was known, population separation was observed between varieties collected from Northern India in the Hindu Kush Mountains and Myanmar. The use of publicly available data provides an initial impression of the complexity within the *Cannabis* genus and adds to our understanding of the genetics underlying evolutionary history and population stratification, which will be critical for future crop improvement for any potential human use.

Key Words: *Cannabis sativa,* Genome Scan, Public Data, Medicinal Plants, Fiber Plants

Introduction

Cannabis sativa L. (Cannabaceae) is an annual flowering herb that has been domesticated multiple times independently in Central Eurasia [1] over its 26,000 year association with humans, providing agronomic value as a source of food, fiber and medicine [2] [3]. *Cannabis* is typically a dioecious plant [4] with female inflorescences being the site of concentrated production of secondary metabolites in particular cannabinoids and terpenes. There are several outstanding questions in the field of *Cannabis* taxonomy, with publications supporting different regions of origin, with ethnographic data favoring a Central Eurasian origin [3] and fossil pollen dating favoring the Northeastern Tibetan Plateau origin [5]. Additionally, a variable number of species have been described (*C. indica, C. sativa, C. ruderalis*) [6] [1] [3], with no widely accepted consensus in the relationships of ecotypes of *Cannabis.* Further, unhybridized landraces, defined as locally adapted distinctive populations which have undergone long periods of selection [7] have become difficult to obtain and identify [8]. Wild and feral populations of *Cannabis* which have not been influenced by hybridization have become increasingly rare [8] and it has been suggested that unaltered wild populations may no longer exist [6]. Human cultivation coupled with climate fluctuations and habitat fragmentation have likely impacted current phenotypes and may have increased isolation among wild/feral *Cannabis* populations and domesticated cultivars [3] [9]. Today, *Cannabis* is broadly divided into fiber-type, grain-type and drug-type cultivars [3].

While *Cannabis* is popularly known for its recreational uses, its medicinal capacity is increasingly being explored. Secondary metabolites such as tetrahydrocannabinol (THC) and cannabidiol (CBD) have received much attention for their potential in pain management [10], as a multiple sclerosis treatment [11], for epilepsy (Charlotte's Web (CW2A) US Plant Patent No. PP30,639 P2) [12], for reduction in nausea and vomiting in chemotherapy patients [13] as well as an appetite stimulant for patients with HIV/AIDS [14]. While medicinal uses are well documented, due to the prohibition and classification of *Cannabis* as a Schedule I narcotic, research on the genus has been restricted [15] by limiting licenses for who can grow and conduct research. The reduction of restrictions in the 2018 United States Farm Bill coupled with a loosening of state level regulatory restrictions and the expansion of scientific and medical research licenses by the Drug Enforcement Administration in 2020 have led to a steady increase in the number of research groups focusing their attention on *Cannabis*. This is reflected in the Web of Science - Plant Science database, which saw an increase of publications from 50 articles in 2018 to 119 in 2021.

The taxonomy of *Cannabis* is a currently evolving field, with recent debate moving towards a monotypic description of the genus [16] with secondary metabolite profiles being used as a basis for population stratification [17] [18]. While the secondary chemistry of the plant may give a perspective on the diversity within the genus, *Cannabis* has been bred for many different purposes and uses for humans. Secondary chemistry therefore reflects only a small aspect of the complexity of the *Cannabis* genome. Through the use of comparative genomics, we can gain insight into the breadth of diversity of *Cannabis* that could be of use for breeding for industrial and medical purposes. Unfortunately, due to prohibition, lack of permitted cultivation, restrictions on trade across borders and lack of plant passport infrastructure, breeding of *Cannabis* has not benefited from the tools and advances of the last century. For example, to date there have been limited efforts to create an inbred hybrid system which is typical of species with this type of mating system. However, there have been some successes (e.g. Oregon CBD Inc., Aurora Cannabis Inc.).

Advances in sequencing technology have resulted in the release of several genomic datasets in the public domain (Table 1). Broad phylogenetic comparisons of the ecotypes and heterotic groups within the *Cannabis* genus is still lacking with previous approaches limited to gene family phylogenies [19] [20] [21] [22] [23] or limited, regional specific sampling [24] [25]. Drug control laws and prohibition have constrained formal documentation often resulting in unverifiable, anecdotal origins of a given cultivar or plant [7], further constraining genetic diversity analysis. Phylogenies provide insight into evolutionary relationships and facilitate the identification and selection of genetically distinct cultivars for breeding practices. Additionally, understanding genetic relationships can provide greater transparency surrounding cultivar name fidelity in the industry, which represents an ongoing issue (i.e.. the naming problem). Prohibition and the lack of cultivar name fidelity, specifically the inconsistency in genotype or plant line paired with inconsistent naming on the label, places *Cannabis* in the conundrum of having to identify evolutionary relationships among modern lines based on ethnohistorical information and anecdotal oral histories. Molecular analyses would provide a more robust characterization of population structure and reveal relationships within the genus that could facilitate crop improvement. Describing *Cannabis* cultivars through genetic fingerprinting will complement phenotypic observations and label claims offering a path to establishing varietal consistency and therapeutic reliability, as well as aid in the transition towards decriminalization and legalization.

Past publications over the last decade have utilized the Purple Kush (drug-type, female) [26] and Finola (hemp-type, male) [26] reference genome assemblies in their analyses (e.g., [24]). Recently, the International Cannabis Genomics Research Consortium (ICGRC) has proposed that the most complete and contagious chromosome-level assembly, cs10 (CBD-dominant cultivar) be used as the reference for all cannabis genomics [27] [15] [28]. In this study we used seven independent datasets – six available in the public domain and one new dataset. We conducted a comparative genomic analysis of 2,496 samples to understand the population structure present in the different collections and identify potential samples that can be explored as the basis for heterotic groups for breeding potential for the cannabis industry as well as to explore the overall genetic diversity and the genetic landscape surrounding genes of agronomic importance.

Results

Homogenizing Multiple Public Datasets

We approach the subject of *Cannabis* taxonomy by examining genetic diversity and population structure, bringing together genomic data from seven different sources (Table 1). Currently there are ten publicly available datasets (Table 1) with seven of these utilized in this study. Due to large sample size (e.g., Phylos Biosciences (n=845) and LeafWorks (n=498)) and high SNP count post filtering (e.g., Soorni *et al.* dataset), some datasets were able to be analyzed in more depth. Data from the University of Colorado Boulder [18] (PRJNA 310948) and Dalhousie University [29] (PRJNA285813) were not used in this study. PRJNA 310948 appears to be a duplicate of samples from PRJNA317659 both released in 2016. Therefore, only one of these (PRJNA317659) was used. The genomic data made available by Medicinal Genomics' Kannapedia site contains samples that have been sequenced across a variety of platforms (CannSNP90, StrainSEEK v1, v2, v3 and Whole-Genome Sequencing). When aligned to the cs10 reference genome, no common SNPs were found across these 753 samples and they were therefore excluded from nuclear SNP analyses. These 753 samples were however included in the haplogroup assessments. Data from

Dalhousie University was unable to be used as the program Trimmomatic [30] could not remove adapter sequences from the reads despite all adapter references being tested (TruSeq-2 and TruSeq-3) and the Illumina Universal Adapter sequence being identified by FastQC.

Homogenization of the sequence data was possible by aligning all datasets to one reference genome, the CBDRx cultivar, recently renamed cs10 [15] [28]. Previous analyses performed in published datasets have utilized different reference genomes (e.g. [31]) making cross comparison of analyses difficult. Reference genomes for varieties such as Jamaican Lion (female) (GCA_012923425.1) and Jamaican Lion male (GCA_013030025.1) [32], Purple Kush (drug-type, female) [26], Finola (hemp-type, male) [26], Pineapple Banana Bubba Kush (drug-type, male) (GCA_002090435.1), LA Confidential (drug-type, female) (GCA_001510005.1), Chemdog91 (drug-type, female) (GCA_001509995.1), Cannatonic (female) (GCA_001865755.1) and JL (feral variety, female) (GCA_013030365.1) [33] have been released, however with variable assembly quality and completeness (contig versus chromosomal level assembly). Additionally, some releases have yet to be peer reviewed and remain BioRxiv releases (McKernan et al., 2020).

Prior to cs10 no chromosomal level high-quality reference sequence was available. In this study, SNP sets derived from cs10 alignment were used to assess genetic diversity between datasets as well as across chromosomes and domesticated and landrace samples. Analysis of nuclear SNPs post cs10 alignment was possible for the Soorni *et al.,* Phylos Biosciences, LeafWorks, Sunrise Genetics, Lynch *et al*., Medicinal Genomics (n=61) and Courtagen Life Sciences datasets. Due to the variety of different sequencing strategies taken across these datasets, it was not possible to identify common SNPs across all datasets. Therefore, each dataset was analyzed independently, with a range of SNPs from 279, (Phylos Biosciences) to 33, 346 (Soorni et al*.,* (2017)) (see Table 2 and Table 3).

Genetic Diversity Across Datasets

Exploring genetic diversity patterns between datasets (Supplemental Figure 1A) and per chromosome (Supplemental Figure 1A & B), there is no chromosomal specific depression in genetic diversity observed that may indicate selection patterns. However, when examining across individual datasets there are clear differences in diversity between the different samplings (Supplementary Figure 1A). As sample sizes are robust, this suggests that the type of sequencing approach taken (length of reads and platform) and, or the sequencing depth and coverage across chromosomes bias the genetic diversity observed. Library preparation and the regions that were targeted by individual groups for sequencing may have additionally contributed to this bias.

We also investigated genetic diversity of the synthase gene regions of Chromosome 7 (Supplementary Figure 1C). This region is of agronomic interest as prior reports had revealed 10 cannabinoid synthase genes located here (25 – 61.3 Mb) [15]. Further genome annotations revealed 14 cannabinoid synthase gene sequences found within this region in two distinct clusters (https://gdb.supercann.net/index.php/genomebrowser/cbdrx-18). Synthase gene Region 1 (~25-27 Mb) contains seven synthase gene sequences with Region 2 (~30-31 Mb) also containing seven synthase gene sequences (Table 4). Region 1 is composed of fragmented tetrahydrocannabinolic acid synthase (THCAS) gene sequences, all of which are non-functional, while Region 2 contains cannabidiolic acid synthase (CBDAS) and CBDAS-like fragments predominantly with one functional CBDAS gene sequence. See Table 4 for precise start and end positions for these

synthase genes with the genes for THCAS (fragmented non-functional pseudogenes), CBDAS-like and functional CBDAS indicated [28]. The presence of non-functional THCAS fragments in the cs10 cultivar is responsible for its high CBD content [28]. We see high values for pi in the areas flanking the synthase gene regions (Supplementary Figure 1C) while in contrast, within synthase gene Regions 1 and 2 there are low values for pi within certain datasets (e.g., Soorni *et al.* and University of Colorado Boulder). This suggests a trend towards selection in these regions, however due to mixed sampling and variable coverage within each dataset, any selection signals may be partially masked.

Phylogenetic Analyses and Hybridization

Branching pattern in a phylogenetic tree is a hypothesis and reflection of how species or groups may have evolved from a common ancestor. By examining such patterns we can begin to understand the evolutionary steps that have shaped the *Cannabis* genus. Maximum Likelihood (ML) phylogenies were assembled for the LeafWorks (n = 498) (Figure 1A), Phylos Biosciences (n = 845) (Figure 1C) and Soorni *et al.* (n = 98) (Figure 2A) datasets. The LeafWorks and Phylos Biosciences datasets contain global sampling whereas the Soorni *et al.* dataset provides a regional perspective, predominantly consisting of germplasm collected from Iran [24]. We performed phylogenetic analysis on these three datasets specifically due to their large sample number which may give broader insights into the phylogenetic structure of the genus. Hierarchical Clustering of Principal Components (HCPC) was performed on all datasets, with a median of five and a range of three to six clusters observed in each dataset (Supplemental Figures 2A, 3A, 4A, 5B, 6B, 7B, 8B).

In the LeafWorks (n = 498) dataset the ML phylogeny (Figure 1A) partitioned samples into ten clades. Clades 1 - 7 and Clade 9 have bootstrap support of over 90, with Clades 8 and 10 having low support (63 and 52 respectively). The LeafWorks phylogeny reveals four clades partitioning the majority of individuals (Clades 4, 6, 9 and 10) (454 of 498). In the LeafWorks phylogeny, most landrace samples are seen in Clades 5 and 9 and 10, however landrace varieties are also observed in all clades to varying degrees (Figure 1A, blue highlighted tips). The LeafWorks fastSTRUCTURE [34] analysis reveals high levels of admixture in all populations and is a feature that is present across all clades (Figure 1C). At k=2 fastSTRUCTURE analysis reveals clustering by use-type, with samples from Clade 10 (THC-Dominant) showing clear separation (Figure 1C). Additionally, at k=3 and k=4 clustering of Landrace and Hemp-type samples is observed, with k=5 resolving separation between these Landrace and Hemp type samples (Figure 1C). Hierarchical clustering also reveals clustering by use-type in Hemp and THC-Dominant samples (Supplemental Figure 3A) with PCA analysis additionally revealing separation of the four main clades, Clades 4, 6, 9 and 10 (Supplemental Figure 3B).

The phylogeny for the Phylos Biosciences dataset (n = 844), identified three main clades (Clades 4, 8 and 9) (744/844) with support for six less populated subclades (Figure 1C), with bootstrap supports for some clades low (73). The Phylos Biosciences HCPC analysis supports clustering by use-type observed for THC-Dominant, CBD-Dominant, Landrace and Hemp-type samples (Supplemental Figure 4A). PCA analysis also reveals separation between the three main clade partitions (Clades 4, 8 and 9) (Supplemental Figure 4B). Of note, Clades 6 and 8 house a majority of landrace specific accessions (68/107), highlighted in blue at the branch tips (Figure 1C). The Phylos Biosciences fastSTRUCTURE analysis at k=5 reveals high levels of admixture between

the populations (Figure 1D), however no clear pattern is observed with use type in this analysis (k=2-5).

The Soorni *et al.* dataset (n = 94) focusing on Iranian germplasm yielded seven clades in the ML phylogeny, with all nodes with the exception of two showing high bootstrap supports above 90 (Figure 2A). Comparing the ML phylogeny (Figure 2A) to Principal Component Analysis (PCA) (Supplemental Figure 2B) and HCPC (Supplemental Figure 2A), we see complete congruence of clade partitions. The phylogeny from the Soorni *et al.* dataset shows landraces are found entirely in Clades 5, 6 and 7 (highlighted in blue) with high bootstrap support (100) (Figure 2A). For the Soorni *et al.* dataset, fastSTRUCTURE analysis at k=3 separates Clades 5, 6 and 7 from the basal clades and revealed low levels of admixture between these clade partitions (Figure 2C). Additionally at k=3-5 distinct clustering by use-type is observed with CBD-dominant samples clustering distinctly from THC-Dominant samples (Figure 2C).

Within the Soorni *et al.* dataset, nucleotide diversity and Tajima's D between the three largest clades, namely Clade 4 (n = 9), Clade 6 (n = 18) and Clade 7 (n = 56) revealed no difference in overall nucleotide diversity (Supplemental Figure 2C). However, there was a trend towards higher relative values of Tajima's D moving from Clade 4 to Clades 6 and Clade 7 (Supplemental Figure 2D), with high values for Tajima's D indicative of balancing selection. More specifically nucleotide diversity at the 25–31 Mb region of Chromosome 7 was also examined between these three clade partitions (Figure 2B). Again, there were higher values for pi in the areas flanking the synthase gene regions on chromosome 7. While in contrast, within synthase gene regions 1 and 2 there are low values for pi (Figure 2B).

Due to low sample size and low representation of use-types in the remaining datasets, Maximum Likelihood phylogenies were not constructed for four of the seven datasets namely, Sunrise Genetics (n=25) (Supplementary Figure 5, Supplementary Table 11), University of Colorado Boulder (n=162) (Supplementary Figure 6, Supplementary Table 12), Courtagen Life Sciences (n=58) (Supplementary Figure 7, Supplementary Table 13) and Medicinal Genomics (n=61) (Supplementary Figure 8, Supplementary Table 14). Analysis of admixture using fastSTRUCTURE analysis for the University of Colorado Boulder dataset at k=5 (Supplemental Figure 6C) showed partial clustering by use-type with the majority of CBD-dominant samples grouped together. For the remaining datasets where nuclear SNP sets were subjected to fastSTRUCTURE analysis, no clear trends in clustering were observed across use-types (Supplemental Figures 5C, 7C and 8C).

SplitsTree analysis showed high within clade hybridization in the LeafWorks dataset, as seen by the dense line network present particularly in samples from Clade 10 (Supplemental Figure 9). SplitsTree analysis provides unrooted trees, therefore clade associations for each sample were incorporated in the input (nexus) file to facilitate interpretations. Differential hybridization across clades was similarly visible in the Phylos Biosciences dataset though less pronounced, with the more recently emerged clades (Clades 8 and 9) also showing higher levels of hybridization (Supplemental Figure 10). The Soorni *et al.* dataset also revealed high levels of within clade hybridization, more specifically in the more recently expanded clade (Clade 7) and showed higher levels of hybridization then more basal clades (Supplemental Figure 11). Hybridization analysis together with fastSTRUCTURE analysis can aid in the identification of individuals with high and

low levels of admixture and provide insights into the degree of introgression that may have occurred between populations. Such an approach could aid in the identification of landrace accessions (yellow branch tips, Supplemental Figures 9, 10 & 11) which are less admixed and facilitate the conservation of regionally adapted populations.

Comparison by Use-Type

We evaluated sample clustering associated with use-type by partitioning cultivars into CBD-dominant, THC-dominant, balanced CBD:THC, Hemp-type and Landrace. This was possible in the LeafWorks, Phylos Biosciences and Soorni *et al.* datasets as these datasets contained representatives in these categories (see Supplemental Tables 6, 7 and 8). In the LeafWorks (Figure 3A) and Soorni *et al.* datasets (Figure 3C), there is a clear separation of individuals by use-type for THC-dominant, CBD-dominant and Landrace samples by PCA. This is also seen in the Phylos Biosciences dataset (Figure 3B) though the separation is less clear. HCPC in the LeafWorks dataset (Supplemental Figure 3A) also supports the clustering of CBD-dominant and Hemp plants as indicated in green (n =23). Different datasets of varying SNP size were additionally capable of distinguishing use-type (namely CBD-dominant and THC-dominant plants) (Supplemental Figure 6A, Supplemental Figure 7A and Supplemental Figure 8A) and may correspond with the domestication history of these types. The Sunrise Genetics dataset contained only THC-Dominant and unknown samples and therefore could not be used to distinguish use-type association (Supplemental Figure 5A).

Comparison by Domestication Status

The LeafWorks and Phylos Biosciences datasets both contained known landrace and modern cultivars. The LeafWorks dataset contained 101 landrace samples and 397 known modern accessions (Supplemental Table 6), whereas the Phylos Biosciences dataset contained 107 landrace samples and 679 modern accessions (Supplemental Table 7). The Phylos Biosciences dataset revealed a correlation with landrace samples clustering together in the HCPC analysis (Supplemental Figure 4A), predominantly in Clade 8. Landraces consistently show a tighter clustering in the PCA compared to the other use-types in both the LeafWorks (Figure 3A) and Phylos Biosciences datasets (Figure 3B). Examining differential nucleotide diversity between landrace and domesticated samples in the LeafWorks (Figure 4A/B) and Phylos Biosciences datasets (Figure 4C/D) revealed many genomic regions which may revel insights into genes or traits differentially selected for between these partitions. SNP count per dataset pre and post filtering for landrace and domesticated samples for the LeafWorks and Phylos Biosciences datasets is detailed in Supplemental Table 10.

Geographic information was included for 26 landraces in the LeafWorks dataset, providing insights into the present-day geographic structure of samples in Asia (Figure 5). The 26 samples from nine different locations included sampling from Pakistan (n = 3), (including North West Balochistan (n = 1)), the Hindu Kush Mountains (n= 6), Lolab Valley Kashmir (n = 4), Wailing Valley Malana (n= 1), Hunza (n= 2), India (n = 2), Chintal (n =2), Myanmar (n= 4) and Cambodia (n = 2) (Figure 5A).

The PCA revealed separate clustering for the Lolab Valley and Hindu Kush samples from the Myanmar samples (Figure 5C) with the Hindu Kush samples nested within the Lolab Valley ellipse. There are low levels of admixture based on fastSTRUCTURE across the three regions at

k=3 with samples clustering by their geographic locations (Figure 5E). Despite being close in geography to each other it would appear the Hindu Kush, Lolab Valley and Myanmar samples have remained isolated. This challenges the notion that wild populations no longer exist in *Cannabis (*Small & Cronquist, 1976) and indicates that there may be areas of untapped genetic diversity for exploration. The SplitsTree hybridization network analysis also shows clustering by geography, with high levels of hybridization in Lolab Valley accessions (Figure 5F) but less hybridization in both the Myanmar and Hindu Kush samples. The individuals from these three geographical regions were further partitioned and examined specifically for their nucleotide diversity (Figure 5D). Nucleotide diversity analysis shows numerous regions with differential values for pi (e.g., proximal part 12.5 - 50 Mbp of Chromosome 5, center of Chromosome 6 18.75 - 37.5 Mbp and proximal part 12.5 - 37.5 Mbp of Chromosome 7). HCPC analysis divided the samples into three clusters namely group 1 (n =10), group 2 (n= 6) and group 3 (n= 10). Examining the geographic sets that have more than two samples, the Myanmar samples exclusively cluster together in group 1, while two-thirds of Hindu Kush samples (#1, #3, #4, #5) cluster together in group 3 (Figure 5B). The remaining Hindu Kush samples (#2 and #6) were found in groups 1 and 2, respectively. The Lolab Valley samples were observed in groups 1 and 3, with two samples split between these groups. The samples from Pakistan (n = 2), Hunza (n = 2), Chintal (n = 2) and Cambodia (n = 2) did not show any clustering associated with their geographical origins, and these samples were spread across these three groups.

Haplogroup Assessment

A maximum likelihood phylogeny was constructed on 126 whole chloroplast sequences which were provided by LeafWorks Inc. Alignment of these sequences revealed 97.6% pairwise identity and a maximum likelihood phylogeny revealed two distinct groups with bootstrap supports of 87 for Clade1 (n= 2) and 67 for Clade 2 (n=124) (Figure 6). In the datasets other then LeafWorks there were no whole chloroplast sequences available. Instead we created organellar SNP sets by aligning to reference sequences for the 18S ribosome (JF317360.1) chloroplast (KY084475.1) and mitochondria (MT361980.1) [35] [36] [28]. Supplemental Table 5 shows the Phylogenetic Model fit for each dataset from the ModelTest software detailing total SNP count pre and post filtering for all datasets used in haplotype analysis.

There was a variable number of haplogroups across the ribosomal alignments (Supplemental Figures 12 (A/B), 14 (A/B), 15 (A/B), 16 (A/B), 17 (A/B), 18 (A/B)). The Courtagen Life Sciences dataset (3 SNPs) (Supplemental Figure 17A) and Sunrise Genetics dataset (2 SNPs) (Supplemental Figure 15A) showed two groups and the Phylos Biosciences (3 SNPs) (Supplemental Figure 14A), Medicinal Genomics 61 (10 SNPs) (Supplemental Figure 18A) and Soorni *et al.* (Supplemental Figure 12A) supported three haplogroups. Alignments to the chloroplast reference sequence (Supplemental Figures 12 (C/D), 13 (A/B), 14 (C/D), 15 (C/D), 16 (C/D), 17 (C/D), 18 (C/D), 19 (A/B) also revealed a variable number of haplogroups, with support for two groups in the University of Colorado dataset (61 SNPs) (Supplemental Figure 16E), Soorni *et al.* (8 SNPs) (Supplemental Figure 12C) and LeafWorks datasets (15 SNPs) (Supplemental Figure 13A). For the mitochondrial genome alignments (Supplemental Figures 12 (E/F), 13 (C/D), 14 (E/F), 16 (E/F), 17 (E/F), 18 (E/F), 19 (C/D)), two groups are similarly observed in the Medicinal Genomics 753 (42 SNPs) (Supplemental Figure 19C) and Soorni *et al.* dataset (25 SNPs) (Supplemental Figure 12E) datasets, whereas three groups are seen in the LeafWorks dataset (9 SNPs) (Supplemental Figure 13C).

Discussion

Despite the expansion of the *Cannabis* industry and academic inquiry, there is still little known about the genetic history of *Cannabis*. While it is commonly understood that *Cannabis* can be divided into the putative species *indica, sativa* and *ruderalis* [3], to date no genetic studies have been able to distinguish these types [37] [29] [18]. These designations are often based on morphological characteristics (i.e., *C. indica,* shorter with a woody stalk and *C. sativa,* taller with a fibrous stalk) [9] or chemical compositions of the plants (Type I, II, III) based on two chemical compounds (THC & CBD) [38] as well as rare types which have high levels of CBGA (Type IV) or no detectable cannabinoids (Type V) [39]. Despite this, retail purchase is marketed based on these purported species delineations (*indica* or *sativa*) while researchers still debate the number of species, if indeed more than one exists [3]. In order to better understand extant germplasm, we evaluated genetic diversity and population structure in seven different collections of *Cannabis* where genomic data were available. These collections consisted of privately bred THC-dominant samples, public hemp samples, landrace and wild collections. While the datasets were sequenced at different times, we were able to make comparisons by aligning to a common reference, cs10 [28]). Understanding genetic diversity within each collection can provide clues as to what variation is present and aides in the evaluation of the genetic basis of trait variation and speciation. Allelic variation conserved in landraces also can help us understand how domestication may have affected selection of traits associated with adaptation to local environments. We explored clade partitions within each dataset to distinguish clusters of varieties sharing common ancestry. Genome wide polymorphisms additionally provide a genetic basis for the characterization of populations and the preliminary partitioning of samples into heterotic groups which can inform strategic management of breeding programs to avoid the loss of genetic diversity and enhance the breeding gene pool and plant productivity. Establishing the number of clades present in different germplasm groups can aid in the selection of individuals for breeding, where individuals which share a more distant relationship are selected and crossed to maximize heterotic effects.

Addressing the Naming Problem

The naming problem in *Cannabis* refers to the unreliable naming of cultivars which often does not reflect any pedigree. This leads to quality issues and causes problems for both the producer and consumer. Phylogenetic analysis and population stratification facilitates explorations of cultivar name fidelity which is an ongoing issue within the industry. A recent examination of 22 different hemp accessions expressing high cannabidiol (CBD) destined for the commercial market revealed little consistency either genetically or chemically across several varieties of the same name [40]. Investigating name fidelity in the Phylos Biosciences dataset was possible for the Sour Diesel cultivar with 14 samples (Supplemental Figure 20) associated with this name, making it the most represented cultivar in the dataset. Twelve Sour Diesel samples were observed in Clade 4 with the remaining 2 observed in clade 9 (Figure 2C, Table 7). The Phylos Biosciences dataset also contained 23 samples associated with the cookies lineage (Supplemental Figure 21), namely; Animal Cookies (n =5), Cookies (n = 2), Girl Scout Cookies (n = 5) (full name or GSC/G.S.C), GSC Mint (n = 1), Fortune Cookies (n = 1), Forum Cut Cookies (n = 1), Panda Cookies (n = 1), Phantom Cookies (n = 1), Platinum Cookies (n = 2), Platinum GSC (n = 1), Platinum OG Cookies (n = 1), Xmas Cookies (n = 1)). Overall, we see 14/23 cookies lineage samples in Clade 9, 3 in Clade 8, 1 in Clade 7, 4 in Clade 6 and 1 in Clade 4. More specifically all 5 Animal Cookies are observed in Clade 9 indicating good consistency in name fidelity for this variety.

Name fidelity was also examined in the LeafWorks dataset where 12 Blue Dream samples were represented from ten different sources (Supplemental Figure 22). Of these, 7/12 placed in Clade 1 (blue_dream samples #1, 3, 4, 6, 7, 9 & 10), 1 sample in Clade 6 (blue_dream_5), Clade 9 (blue_dream_11) and Clade 10 (blue_dream_2). The remaining two samples (blue_dream # 8 & 12) were unplaced in the phylogenetic tree. Examining differences in cultivar name location using phylogenetic approaches, where the degree of genetic similarity dictates sample placement, offers a means by which cultivar names can be interrogated and would allow inconsistencies in variety naming to be identified.

Looking at these results collectively, it is possible to observe areas with consistent placement of a cultivar of interest to address naming fidelity and the likelihood of name accuracy using these evolutionary relationships of genetic relatedness. As such, we observed clustering of similarly named samples in the case of Sour Diesel (Supplemental Figure 20; Clade 4) and Blue Dream (Supplemental Figure 22; Clade 1), while not all samples with these cultivar names clustered together in these clades. This demonstrates the naming problem well where similarly named samples can be quite dissimilar genetically. Further work is needed to determine how pervasive the naming problem is. This work also highlights the importance of genetics to inform label claims, which will be particularly crucial upon legalization when the Federal Drug Administration would require accurate plant label claims as it does for all other natural products sold in the United States.

Domestication and Use-Type

The geographical range of *Cannabis* has been influenced and extended through its long-standing human association [41]. The movement of prehistoric peoples in Eurasia along historic trade routes (e.g. Silk Road) facilitated the distribution of the *Cannabis* plant, with use of the crop by these communities driving selection during domestication [9]. The psychoactivity of THC has played a role in the domestication of this species, with extensive work in the field of ethnobotany revealing long standing histories, cultural associations and knowledge of this trait [2] [42].

Studies exploring genetic history of *Cannabis* have used ethnographic associations to delineate populations. Like the results presented here with nuclear SNPs (Supplementary Figure 2A, 3A, 4A, 5B, 6B, 7B and 8B), Henry *et al*. (2020) reported k = 5 as the number of groups which best delineated samples [43] though this number is variable [44]. In plastid data, previous work representing a mix of wild and domesticated populations in China revealed three distinct groups [45], however this differs from published work from law enforcement agencies which suggested five to eight groups [46] [47]. The chloroplast phylogeny assembled here indicates support for two groups (Figure 6), with the second group representing global population diversity but only a small number of samples (n = 2) in the first group. The hierarchical clustering of principal components also supports this two-group structure that was present in independent datasets (Supplementary Figures 16E and 13B). This seems to indicate that maternal diversity does not appear to correlate with nuclear diversity levels or patterns of population structure, nor exhibit clear geographic patterns reflecting human movement.

In the classic tradition of ethnobotany, we examined if clade partitions were related to human use-type. Examining use-type associations (Figure 3) in tandem with phylogenetic analysis (Figure 1A and Figure 1C) offers a perspective on how *Cannabis* populations may have been influenced by

human mediated selection for traits such as high THC or CBD content as well as potential population delineations within the genus (Supplementary Figure 2A/B, 3A/B, 4A/B). The LeafWorks, Phylos Biosciences and Soorni *et al.* datasets exhibited use-type separation via PCA (Figure 3A, 3B and 3C) and demonstrate how SNP set number and quality can affect the degree of discrimination (see Supplementary Table 6, 7 and 8 for assignment of use-type for these datasets). A broader distribution of genetic variation in THC-dominant cultivars is observed versus a narrower range of genetic variation in landrace samples as seen in the LeafWorks (Figures 3A), and Phylos Biosciences datasets (Figures 3B). This may be indicative of the purported large-scale hybridization that is thought to have occurred in THC-dominant varietals in the US from the 1960s onwards [2]. Functionally, this means that numerous crosses were made for high THC content. Alternatively, this high genetic diversity in THC-dominant cultivars could represent convergent selection, with each lineage being bred in isolation and now released back to the market as regulations relax. However, due to prohibition, there is a lack of available pedigree records, making it difficult to reconcile these two hypotheses.

Unlike Ren *et al.* (2021), our results suggest that landrace samples can be observed in all clade partitions, however the number of landrace samples in a particular clade may be affected by sampling biases. Further, we see differential nucleotide diversity when comparing landrace and domesticated samples across many genomic regions (Figure 4B and D) providing a significant number of loci to be investigated.

Landrace Biogeography and Diversity

The debate around the geography of *Cannabis* domestication is ongoing, with current published evidence suggesting different centers of origin [2] [5]. There is evidence that the species was distributed across Eurasia prior to human use, dispersing from Northeastern Tibetan Plateau west into Europe 6 mya and east into Eastern China by 1.2 mya [5]. This suggests that *Cannabis* could have been domesticated across a number of different sites across a diffuse geography. Here, a subset of georeferenced samples (n = 26) was analyzed separately (Figure 5). Hierarchical clustering revealed three landrace populations which appear to be quite distinct from one another (Figure 5B), with minimal admixture (Figure 5E). The largest amount of diversity was observed in the Lolab Valley samples as opposed to the Hindu Kush or Myanmar samples. Population separation was observed between varieties collected from Northern India in the Hindu Kush Mountains and Myanmar. While this subset of landraces clusters distinctly when analyzed separately (Figure 5B), landraces are broadly represented across most clade partitions (Figure 1A and 1B, blue highlighted tips). This suggests that these ancestral landraces have a shared ancestry with the modern cultivars in which they share clade positional location.

Landrace data also allows the exploration of human mediated selection through the exploration of suppressed or low nucleotide diversity [48] [49] and facilitates the identification of genomic locations containing traits of interest that may have been fixed in modern lines [50]. Examining the diversity and identifying regions which may contribute to gene regulation of the two dominant cannabinoids, CBD and THC, is likely to be significant as the *Cannabis* industry continues to develop, as these are the key quality phenotypes for the industry. Exploring the THCAS and CBDAS regions located on Chromosome 7 we generally see high nucleotide diversity in the regions between the two synthase containing regions when sufficient reads can be aligned, while low nucleotide diversity is observed in the synthase containing regions themselves (Figure 2C &

Supplemental Figure 1C). These differences in diversity suggest positive selection in the two synthase gene regions and this is consistently seen across clades. However, due to variable coverage of these important gene sequences within the datasets examined, a comprehensive survey across individuals for cannabinoid gene content and the assessment of functional variants was limited. This finding complements previous work where the focus has been on specific gene family phylogenies, such as the cannabinoid phylogeny for CBDAS and THCAS variants [51] [23] and terpene synthase phylogenies [21] [52] [53]. For breeders and growers seeking to maximize the concentration of various cannabinoids, working with breeding material of known genetic constitution is key to creating better products. Increasing number of genetic resources and information around varietals should help provide a way to breed cultivars more efficiently.

Strategic Plant Breeding for Use-Type

Cannabis varietals are typically classified as high THC (Type I), containing a ratio of THC to CBD (Type II) or high CBD (Type III) plants [38] as well as the rarer types, with high CBG (Type IV) and complete absence of cannabinoids (Type V) [39]. Hemp varietals are typically classified as Type III, often presenting with a CBD enriched phenotype largely the result of a non-functional copy of the synthase THCAS [54]. However, this is not the case for all hemp varietals, with functional and non-functional THCAS genes known to be present in germplasm presenting challenges for farmers who are required to have less than 0.3% THC at harvest based on current regulations. Breeding with a focus on a particular use-type could help to ensure consistency in secondary chemistry and incorporating an assessment of admixture or hybridization in this selection may expedite the time taken to reach population stability. Examining the levels of admixture (Figure 1B/D, Supplementary Figures 9, 10 and 11) between and within clades or populations can also aid in the identification of less admixed individuals that may offer greater heterotic effects when crossed with other more inbred individuals present in distinct clades [55]. Examining hybridization, it appears that more recently expanded clades present higher levels of hybridization than earlier clades (Supplementary Figures 9, 10 and 11) and may be reflective of the rapidly expanding legal *Cannabis* market with large amounts of related germplasm being crossbred. *Cannabis* focused research and hybrid breeding programs could offer solutions to the production of industrial hemp-type *Cannabis*, offering more sustainable alternatives for the paper and textile industries.

Understanding population stratification and use-type could advance specialty crop development for specific traits in this emerging crop system. Using genomic approaches that improve resolution of the evolutionary processes occurring in the *Cannabis* genus will also provide information about the underlying genetic variation that can be used for improving breeding outcomes. The expression of typically low abundance cannabinoids has potential therapeutic translatability [56] and coupling plant phylogeny with metabolomics could facilitate the identification of such plants with unique genetic and secondary chemistries and would provide unique market classes. Further, understanding historic evolution and the distribution of insect/mammal herbivores may provide insight into where to collect rare cannabinoid profiles as it has been suggested that the expansion of cannabinoid as well as terpenoid gene families has been driven by predation [57]. Breeding for traits such as disease and pest resistance is essential for crop reliability, and it is likely that future breeding practices will focus on the selection and stabilization of cultivars with specific secondary chemistries which will help establish clear market classes.

Leveraging evolutionary history for better breeding programs

Evolutionary plant breeding (EPB) is used to increase the diversity and stability of a crop in a specific environment using natural selection by increasing the frequency of favorable alleles in a heterogenous population [58]. Locally adapted samples such as *Cannabis* landraces have undergone many rounds of selection and are better adapted to their niche and may be a reservoir of genetic diversity containing genes conferring disease resistance among other desirable traits. Such an approach has been taken in barley [59] where bulbous barley Mildew Locus O (*mlo*) genes have been introgressed into other barley cultivars for winter cultivar development and powdery mildew resistance. Our sampling of landraces was limited (Supplementary Tables 6, 7, 8, 12 and 15). If more efforts are made to conserve and characterize landrace *Cannabis* populations, then similar approaches could be taken to identify donor parents and populations which could offer an environmentally friendly source of genetic resistance to pests for introgression into other populations, reducing pesticide and herbicide use for more sustainable cultivation practices. The history of prohibition and local cultivars suggests that there is a large possibility of biopiracy with respect to the developing industry. It will be important to develop equitable distribution and ensure that local communities benefit from the work their communities have done in line the international plant genetic resources treaty [60].

Modernizing the *Cannabis* industry by incorporating a genomic perspective would assist in the selection of promising germplasm for breeding purposes though the stratification of samples into heterotic groups [61]. Identifying clade partitions offers a basis for which heterotic groups can be established to maximize performance for a wide variety of traits and could aid the identification of varietals with rare genetic constitutions. Exploiting the genetic variation between heterotic groups has been found to be successful in maximizing hybrid performance in other crop systems such as wheat [62] and rice [63] and has depended crucially on the clustering of germplasm into such groups [64]. Hybrid breeding is known to boost yields and growth characteristics [65] and implementing such an approach could facilitate crop development, as well as address the characterization of existing *Cannabis* germplasm in the industry today. Analyses of population structure in the *Cannabis* genus could bridge the communication and knowledge gap post prohibition in the recreational market between what dispensaries are growing and selling, what the public are buying and consuming, and bring standardization and reliability in variety consistency to the marketplace. The current classification of *Cannabis* as a Schedule I drug presents barriers to open research which could support the production of safe medical and recreational *Cannabis*.

Public Data Implications

High throughput sequencing technologies have advanced at a rapid pace making it possible to affordably generate large collections of genomic sequences. This had led to the release of several *Cannabis* genomic datasets in the public domain by researchers in both academia and industry (Table 1) over the last decade. However, there is often limited alignment in research objectives and goals between the public and private sectors. From the perspective of academia, there is frequently a need to release and publish data as early as possible. In contrast for the private sector genomic data often has high commercial value and the sharing of data through public papers is therefore not a priority. Public-private partnerships offer a route to harness the diverse resources and expertise present in both sectors and could advance *Cannabis* science. Such collaborations can achieve results beyond the reach of a sector working alone and are increasingly being observed for the acceleration of agricultural productivity [66].

Data handling and analysis presents new logistical challenges and there is often a gap between data generation and data analysis. Making data accessible as well as usable, for example through the release of annotations with assemblies, provides context and furthers interpretations. For genomic sequence data to be analyzed to their fullest it is important that standards are increasingly adopted and required around metadata inclusion and maintenance [67]. In this analysis, datasets varied in their ease of access and the availability of metadata. Data sources were not all archived at a central repository such as NCBI. Those that were not were in NCBI were not straightforward to access both in sequence data and paired metadata. Use-type was assigned from the varietal names supplied from the metadata associated with each dataset, however reliability and consistency in varietal naming is highly variable and thus these names and use-type associations should be treated with caution (the naming problem). While this problem is not unique to *Cannabis*, it is acutely problematic in any species that has high economic value and limited foundational genomic resources.

Caveats

Previous work has used various reference genomes [18] [24] [31] [68] and this reflects the current predicament within the industry where standards are still in development. The cs10 reference genome is the most complete assembly to date, but it contains non-functional THCAS gene sequences. This high CBD expressing variety contains a functional CBDAS gene (CsCBD_09G0011030) and simultaneously eight non-functional fragmented THCAS sequences [28]. It therefore does not allow an examination of functional gene content for THCAS, limiting utility for examination of this locus, which is the most economically important locus in the genome. Using a single reference creates a bias impacting examination of other more minor cannabinoid synthases which may be represented across the various datasets. Reference limitations are being addressed through the utilization of pan genomes and are increasingly becoming available for many crop species [69] [70] [71], however, there is not currently a public pan genome for *Cannabis*, which would greatly facilitate analysis.

An additional challenge when working with public data sources is that analysis of any dataset must often be performed in isolation due to the use of different platforms/approaches by academic and industry sources. Care must be taken in the cross comparison of specific datasets as the amount of shared germplasm as well as dataset quality can influence the breadth and inference potential of the analysis. Additionally, when expanding these observations to conclusions about the genus as whole, it is important to carefully consider germplasm sampling bias which limits the direct comparison of the delineated clades across datasets. This may result in limited or no shared SNPs across datasets and therefore obscures the cross comparison of datasets directly. Higher yielding SNP sets such as the Soorni *et al.* dataset offer greater clarity and demonstrates how SNP set quality and quantity can impact interpretations. This and the variable amount of sequencing data across this study highlight how different datasets can be used for different purposes and how different sequencing approaches (amplicon, reduced representation, whole genome sequencing) can restrict interpretations or bias diversity assessments.

Further, data quality and quantity influence the ability to clearly distinguish use-type. While we were able to explore use-type, across the datasets studied, use-types were not evenly represented (e.g., The Phylos Biosciences dataset contained 107 landrace, 17 hemp, 48 CBD enriched, 624

THC enriched and 49 unknown). This uneven distribution may influence conclusions related to the genus overall and will be subject to change as the amount of data increases. Additionally, these designations of type were assigned by sample name and due to the inconsistencies in naming (the naming problem) in the *Cannabis* industry, may also be subject to change. Currently, classification of use-type (Hemp/Drug-type) or secondary chemistry can be predicted by synthase gene content for cannabidiolic and tetrahydrocannabinolic acid synthases across varieties, but this tells us little about other traits or the rest of the genome. As costs continue to decrease, genomic approaches for examination of *Cannabis* naming will likely become the norm and could overcome the ongoing challenge in the industry of clone and cultivar misidentification.

Future perspectives

In this study, nuclear SNP sets for seven datasets were explored, with the datasets having the highest number of taxa chosen to gain insights into the phylogenetic structure of the *Cannabis* genus. Representation within each clade varies across the datasets and is reflective of the different germplasm present in each collection. Use-type associations and separations were observed via PCA in the Soorni *et al.*, LeafWorks and to a lesser extent in the Phylos Biosciences datasets with this in line with prior publications [18]. With the legal status of *Cannabis* now shifting, researchers can begin to examine the effects prohibition on extant *Cannabis* varieties. In the United States, prohibition may have created closed gene pools through the breeding of limited germplasm by limiting germplasm movement. Closed gene pool breeding may have had a role to play in the increased potency of *Cannabis* varieties over time, with increases in THC content from ~4% in 1995 to ~12% in 2014 reported [72]. Analogous to this in the wild, repeated range contractions during the Holocene are thought to have resulted in repeated genetic bottlenecks and likely initiated incomplete allopatric speciation which has led to differences between European (CBD-dominant) and Asian (THC-Dominant) *Cannabis* populations [9].

Commonly known cultivars and cultivar families such as Kush derivatives may perhaps best be thought of and described by a population, rather than a singular plant variety. However, work remains to genetically define and distinguish such cultivars. Currently there is no industry wide system to verify cultivars [37]. Genetic approaches could therefore bring clarity and consistency to the recreational and medicinal markets and would broaden the perspective when ascribing *Cannabis* to factors other than the two typically dominant Cannabinoids, CBD and THC. The approaches taken here provide a basis for the further investigation of the population structure of the *Cannabis* genus. The conservation of wild relatives and naturalized populations of *Cannabis* and the development of germplasm resources may be critical for the preservation of diversity and act as a valuable source of allelic variation for introgression and trait improvement as the industry continues to develop worldwide.

Materials and Methods

Data Acquisition

Data was retrieved for five of the seven datasets utilized here by direct download from the National Center for Biotechnology Information's Sequence Read Archive (NCBI SRA) using the Bioproject IDs in Table 1. A sixth data source with paired fastq for 498 individuals was kindly shared through a public-private partnership in collaboration with LeafWorks Inc. The seventh and last public dataset utilized here was sourced from the Medicinal Genomics' website, where 61 paired fastq

samples were provided for bulk download (Medicinal Genomics 61) and an additional 753 paired fastq (Medicinal Genomics 753) were separately and individually downloaded from each cultivar page. At the time of download (June 2021) the sequencing approach taken for each of these 753 samples (Supplementary Table 15) could not be determined. The metadata associated with each sample has since been updated and assigned the sequencing approach (i.e., CannSNP90, StrainSEEK v1, v2, v3 and Whole-Genome Sequencing) used for each sample. In this publication, with all 753 samples pooled together, no shared nuclear SNPs post filtering were found across all 753 samples for further analysis based on the approach to SNP analysis outlined below. Therefore, these samples were only included in the haplogroup analysis. In regards to type of sequence data available, single reads were provided in the Soorni *et al.*, University of Colorado Boulder and Sawler *et al.* datasets. The remaining datasets (Sunrise Genetics, Courtagen Life Sciences, Medicinal Genomics, LeafWorks, Phylos Biosciences) contained paired reads. Only forward reads were aligned to cs10 for the Phylos Biosciences dataset in this study.

Sample names were assigned to individual samples as supplied by authors in supplemental materials of publication or through the metadata supplied through NCBI. Use-type associations were assigned to the LeafWorks samples through searching sample names on leafly.com or wikileaf.com. For the Phylos Biosciences dataset, samples were assigned to use-type from the Genotype report page for each sample on the Phylos Biosciences website. Use-type association was possible for some samples in the Soorni *et al.* dataset due to a recent publication which focused on the chemistry of these wild population samplings in Iran [73]. The Soorni *et al.* dataset therefore represents regional sampling, while the remaining datasets represent global samplings (Supplementary Tables 6, 7, 8, 11, 12, 13, 14 &15).

Data Processing

Where demultiplexing was required (PRJNA285813) [29], barcodes were acquired from the supplemental materials and removed using the sabre (version 1.0) software. All dataset fastq files were checked for adapter sequence content using the fastqc (version 0.11.8) software [74]. Where adapters were present, Trimmomatic (version 0.39) [30] was implemented to remove these sequence elements. Reads were then aligned to the high CBD expressing varietal CBDRx (cs10) genome reference [28] using bwa-mem (version 0.7.17) [75]. Samtools (version 1.9) (H. Li et al., 2009) was used to convert sam files to bam files and mapped reads were sorted for a mapping quality of 30 or above. BCFtools (version 1.9) [76] was used with the mpileup and call commands to generate VCF files. Using VCFtools (version 0.1.16) [77] samples were filtered for a minor allele frequency of 0.05, Hardy-Weinberg Equilibrium (0.05), and a maximum missingness of 10% (0.9). After filtering, data were analysed using the SNPRelate [78] FactoMineR (Lê et al., 2008) and factoextra [79] packages in R studio (version 1.4.1106) (R Core Team, 2013).

Population Diversity

VCF files for known domesticates and landraces were separately merged into their respective VCF files and strictly filtered in VCFtools with a minor allele frequency of 0.05, Hardy-Weinberg Equilibrium (0.05), and a maximum missingness of 10% (0.9). For nucleotide diversity and Tajima's D, VCFtools was used with a 10,000 bp sliding window across the strictly filtered files for each dataset. Specific chromosomal regions of interest were focused upon by restricting the location range plotted in R studio using the ggplot package [80].

Population Structure and Phylogenetic Analysis

VCFtools was used to generate MAP and PED files. These were then used to generate BED, BIM, and FAM files in the software PLINK (version 1.9) [81]. For each dataset fastSTRUCTURE (version 1.0) was then used to assign individuals into populations [34]. In addition, each dataset was examined for clustering by use-type using principal component analysis (PCA) in SNPRelate [78]. Only bi-allelic SNPs further filtered for linkage disequilibrium (0.2) were used for the Principal Component Analysis (PCA) and Hierarchical Clustering on Principal Components (HCPC) as represented in figures by hierarchical clustering dendograms (Supplementary Figures 2A,3A, 4A, 5B, 6B, 7B and 8B). All SNPs were used to make the Maximum Likelihood phylogenetic trees (Figure 1A & B, Figure 2A). Maximum Likelihood (ML) phylogenetic trees were constructed for the Soorni *et al*. (Figure 2A), LeafWorks (Figure 1A) and Phylos Biosciences (Figure 1B) datasets due to large sample size and high SNP count post filtering. For the remaining datasets, ML phylogenetic trees were not constructed due to low sample size (between 25 – 162 individuals) and low representation of Landrace, Hemp and CBD-Dominant samples, which may confound genus wide interpretations (Sunrise Genetics (Supplemental Figure 5), University of Colorado Boulder (Supplemental Figure 6), Courtagen Life Sciences (Supplemental Figure 7), and Medicinal Genomics (Supplemental Figure 8)).

To look for potential hybridization a nexus file was used as input in the software SplitsTree4 [82] and a Splits Hybridization Network was generated using the RECOMB2007 method, with relationships displayed using the equal angle method.

For phylogenies, VCF files were converted to nexus and fasta format using the software package vcf2phylip (version 2.6 (https://github.com/edgardomortiz/vcf2phylip). Multiple sequence alignment was performed using MAFFT (version 7.475) [83] and this was submitted to the software Modeltest-ng (version 0.1.6) [84] to best evaluate the substitution model to be used. The results for which can be seen in Table 5. The appropriate model selection was then submitted to IQ-TREE (version 2.0.7) to estimate the phylogeny with the -B 1000 flag for bootstrap support. Trees were visualized in FigTree (Version1.4.4) (Figure 1A & B)

To explore haplo-group assignment, samples were aligned to reference sequences for the 18S ribosomal subunit (JF317360.1, 1,701 bp), the chloroplast (KY084475.1, 153,945 bp a Chinese Hemp variety) and the mitochondria (MT361980.1, 415,806 bp Cultivar: Kompolti). Additionally, a maximum likelihood phylogeny was constructed on 126 whole chloroplast sequences which were provided by LeafWorks Inc. Using the ModelTest software the GTR+G4 model was selected as the best substitution model. A phylogenetic tree was generated using IQ-TREE (version 2.0.7) to estimate the phylogeny with the -B 1000 flag for bootstrap support. Trees were visualized in FigTree (Version1.4.4) (Figure 6).

Main Figure Legends

Figure 1. Examining Maximum Likelihood phylogenetic structure and admixture in the LeafWorks and Phylos Biosciences datasets (A) Maximum Likelihood tree for the LeafWorks dataset constructed from 1,405 nuclear SNPs from 498 samples. Landrace samples are highlighted in blue at the branch tips **(B)** Maximum Likelihood phylogenetic tree for the Phylos Biosciences dataset constructed from 279 nuclear SNPs from 844 samples (**C)** Visualization of population structure and admixture for the LeafWorks dataset using the fastSTRUCTURE software (k=2-5) **(D)** Visualization of population structure and admixture for the Phylos Biosciences dataset using the fastSTRUCTURE software (k=2-5).

Figure 2. Results for the Soorni *et al*. dataset (**A**) Maximum Likelihood (ML) phylogenetic tree constructed from 33,629 nuclear single nucleotide polymorphisms (SNPs) from 94 samples consisting of 16 from CGN Genebank, 10 from IPK Genebank and 68 landraces (highlighted in blue) sampled across Iran. Colour codes correspond to the main supported clades **(B)** Comparison of clade partitions for nucleotide diversity at the 25-31 Mb regions of Chromosome 7 (NC_044378.1), with Clade 4 (n=9) from 66,260 SNPs, Clade 6 (n=18) from 60,618 SNPs and Clade 7 (n=56) from 48,795 SNPs **(C)** Visualization of population structure and admixture with fastSTRUCTURE software (k=2-5).

Figure 3. Examination of use-type association across three datasets (A) Principal component analysis (PCA) from 520 nuclear SNPs for the LeafWorks dataset **(B)** PCA from 213 SNPs Phylos Biosciences dataset **(C)** PCA from 6,865 nuclear SNPs for the Soorni *et al.* dataset where cannabinoid content could be determined due to recent publication for 31/94 samples.

Figure 4. Nucleotide diversity for landrace and domesticated partitions for the LeafWorks and Phylos Biosciences datasets **(A)** Nucleotide diversity by chromosome and **(B)** across chromosome length for Domesticated (n=397, 2,096 SNPs) and Landrace (n=101, 2,131 SNPs) samples for the LeafWorks dataset (**C)** Nucleotide diversity by chromosome and **(D)** across chromosome length for Domesticated (n=679, 704 SNPs) and Landrace (n=107, 266 SNPs) samples for the Phylos Biosciences dataset**.**

Figure 5. Landrace accessions from the LeafWorks dataset show separation between Indian and Myanmar populations **(A)** Map detailing the locations of landrace accessions, highlighted are the Hindu Kush Mountains, Lolab Valley and Myanmar **(B)** Hierarchical cluster dendrogram based on 304 SNPs (LD 0.2) across 26 samples of known and trusted origin **(C)** PCA based on 304 SNPs with geographical locations of samples as indicated **(D)** Nucleotide diversity comparison between Hindu Kush Mountains (n=6) 4,304 SNPs, Lobal Valley (n=4) 853 SNPs and Myanmar (n=4) 2,204 SNPs **(E)** Visualization of population structure and admixture using the fastSTRUCTURE software (k=3) **(F)** Network tree visualized using the Splits-Tree software with sample source indicated.

Figure 6. Maximum Likelihood phylogenetic tree for 126 whole chloroplast assemblies. Individuals were aligned using MAFFT. Modeltest-ng revealed the GTR+G4 as the best fit substitution model and IQ-Tree software was used for phylogenetic inference. The resultant tree was visualized using FigTree (Version 1.4.4).

Supplemental Figure Legends

Supplementary Figure 1. Dataset overview **(A)** Nucleotide diversity by chromosome for all 7 genomic datasets for *Cannabis sativa* L. **(B)** Nucleotide diversity across the length of the 10 chromosomes for all 7 genomic datasets **(C)** Nucleotide diversity across datasets for the 25-31Mb region which contains the Cannabinoid synthase genes. Positions of full length and truncated genes are listed in Table 4. Due to low numbers of reads aligning to this partitioned region, not all datasets could be represented in Figure 1C.

Supplemental Figure 2. Soorni *et al.* supplemental nuclear SNP analysis **(A)** Hierarchical cluster dendrogram from 6,865 nuclear SNPs with clade association indicated below **(B)** Principal component analysis (PCA) from 6,865 nuclear SNPs (LD 0.2) indicates complete congruence with clade partitioning **(C)** Nucleotide diversity and **(D)** Tajima's D by clade partitions for the Soorni dataset across the ten chromosomes. Nucleotide diversity and Tajima's D calculated for Clade 4 (n=9) from 66,260 SNPs, Clade 6 (n=18) from 60,618 SNPs and Clade 7 (n=56) from 48,795 SNPs.

Supplemental Figure 3. LeafWorks supplemental nuclear SNP analysis **(A)** Hierarchal cluster dendrogram from 520 nuclear SNPs with use-type and clade association indicated below **(B)** PCA with samples associated with Clade number from Figure 1A as indicated.

Supplemental Figure 4. Phylos Biosciences supplemental nuclear SNP analysis **(A)** Hierarchal cluster dendrogram from 213 nuclear SNPs with use-type and clade association indicated below **(B)** PCA with samples associated with Clade number from Figure 1B as indicated.

Supplemental Figure 5. Nuclear SNP analysis for the Sunrise Genetics dataset for 25 samples **(A)** PCA by use-type based on 1,604 nuclear SNPs. Use-type associations include THC-Dominant (Type I) (n=38) and Unknown (n=12). **(B)** Hierarchal cluster dendrogram with use-type indicated below **(C)** Visualization of population structure and admixture using the fastSTRUCTURE software (k=2-5).

Supplemental Figure 6. Nuclear SNP analysis for the University of Colorado Boulder dataset for 162 samples **(A)** PCA by use-type for 162 samples from 2,223 SNPs. Type associations include Hemp (n=1), Landrace (n=1), THC-Dominant (Type I) (n=162), CBD-Dominant (Type III) (n=11), THC:CBD (Type II) (n=2) and Unknown (n=21). **(B)** Hierarchal cluster dendrogram with use-type indicated below **(C)** Visualization of population structure and admixture using the fastSTRUCTURE software (k=2-5).

Supplemental Figure 7. Nuclear SNP analysis for the Courtagen Life Sciences dataset for 58 samples **(A)** PCA by use-type based on 119 nuclear SNPs. Use-type associations include Hemp (n=1), THC-Dominant (Type I) (n=41), CBD-Dominant (Type III) (n=11) and Unknown (n=5). **(B)** Hierarchal cluster dendrogram with use-type indicated below **(C)** Visualization of population structure and admixture using the fastSTRUCTURE software (k=2-5).

Supplemental Figure 8. Nuclear SNP analysis for the Medicinal Genomics 61 dataset for 61 samples **(A)** PCA by use-type based on 2,267 nuclear SNPs. Use-type associations include Hemp (n=1), THC-Dominant (Type I) (n=47), CBD-Dominant (Type III) (n=5) and Unknown (n=9). (**B)** Hierarchal cluster dendrogram with use-type indicated below **(C)** Visualization of population structure and admixture using the fastSTRUCTURE software (k=2-5).

Supplemental Figure 9. Network tree visualized using the Splits-Tree software with clade partitions indicated for the LeafWorks dataset (n=498). Landrace samples highlighted in yellow.

Supplemental Figure 10. Network tree visualized using the Splits-Tree software with clade partitions

indicated for the Phylos Biosciences dataset (n=845). Landrace samples highlighted in yellow.

Supplementary Figure 11. Network tree visualized using the SplitsTree software (version 4) reveals high levels of hybridization, particularly within Clade 7 for the Soorni *et al.* dataset (n=94).

Supplemental Figure 12. Haplogroup analysis for the Soorni *et al.* dataset for 94 samples with alignments to reference sequences for the (A-B) 18S Ribosomal subunit (Reference JF317360.1), (C-D) Chloroplast (Reference KY084475.1) and (E-F) Mitochondria (Reference MT361980.1). **(A)** PCA from 3 SNPs (LD 0.2) with samples associated by clade number as indicated **(B)** Hierarchical cluster dendrogram based on 3 SNPs (LD 0.2) with use-type and clade association indicated below **(C)** PCA from 8 SNPs (LD 0.2) with samples associated by clade number as indicated **(D)** Hierarchical cluster dendrogram based on 8 SNPs (LD 0.2) with use-type and clade association indicated below **(E)** PCA from 25 SNPs (LD 0.2) with samples associated by clade number as indicated **(F)** Hierarchical cluster dendrogram based on 25 SNPs (LD 0.2) with use-type and clade association indicated below.

Supplemental Figure 13. Haplogroup analysis for the LeafWorks dataset for 498 samples with alignments to reference sequences for the (A-B) Chloroplast (Reference KY084475.1) and (C-D) Mitochondria (Reference MT361980.1). **(A)** PCA by use-type from 8 SNPs (LD 0.2) with samples associated by clade number as indicated **(B)** Hierarchical cluster dendrogram based on 15 SNPs (LD 0.2) with use-type and clade association indicated below **(C)** PCA by use-type from 9 SNPs (LD 0.2) with samples associated by clade number as indicated **(D)** Hierarchical cluster dendrogram based on 9 SNPs (LD 0.2) with use-type and clade association indicated below.

Supplemental Figure 14. Haplogroup analysis for the Phylos Biosciences dataset for 844 samples with alignments to reference sequences for the 18S Ribosomal subunit (Reference JF317360.1) (A-B), Chloroplast (Reference KY084475.1) (C-D) and Mitochondria (Reference MT361980.1) (E-F). (**A)** PCA from 8 SNPs (LD 0.2) with samples associated by clade number as indicated (n=87) **(B)** Hierarchical cluster dendrogram based on 3 SNPs (LD 0.2) with use-type and clade association indicated below (n=87) **(C)** PCA from 2 SNPs (LD 0.2) with samples associated by clade number as indicated (n=383) **(D)** Hierarchical cluster dendrogram based on 2 SNPs (LD 0.2) with use-type and clade association indicated below (n=383) (**E)** PCA from 24 SNPs (LD 0.2) with samples associated by clade number as indicated (n=845) **(F)** Hierarchical cluster dendrogram based on 24 SNPs (LD 0.2) with use-type and clade association indicated below (n=845).

Supplemental Figure 15. Haplogroup analysis for the Sunrise Genetics dataset for 25 samples with alignments to reference sequences for the 18S Ribosomal subunit (Reference JF317360.1) (A-B) and Chloroplast (Reference KY084475.1) (C-D) **(A)** PCA by use-type from 2 SNPs (LD 0.2) **(B)** Hierarchical cluster dendrogram based on 2 SNPs (LD 0.2) with use type indicated below **(C)** PCA by use-type from 7 SNPs (LD 0.2) **(D)** Hierarchical cluster dendrogram based on 7 SNPs (LD 0.2) with use-type indicated below.

Supplemental Figure 16. Haplogroup analysis for the University of Colorado Boulder dataset for 162 samples with alignments to reference sequences for the 18S Ribosomal subunit (Reference JF317360.1) **(A-B),** Chloroplast (Reference KY084475.1) (**C-D)** and Mitochondria (Reference MT361980.1) (**E-F). (A)** PCA by use-type from 8 SNPs (LD 0.2) **(B)** Hierarchical cluster dendrogram based on 8 SNPs (LD 0.2) **(C)** PCA by use-type from 61 SNPs (LD 0.2) **(D)** Hierarchical cluster dendrogram based on 61 SNPs (LD 0.2) **(E)** PCA by use-type from 28 SNPs (LD 0.2) **(F)** Hierarchical cluster dendrogram based on 28 SNPs (LD 0.2).

Supplemental Figure 17. Haplogroup analysis for the Courtagen Life Sciences dataset for 58 samples with alignments to reference sequences for the 18S Ribosomal subunit (Reference JF317360.1) **(A-B)**,

Chloroplast (Reference KY084475.1) **(C-D)** and Mitochondria (Reference MT361980.1) **(E-F). (A)** PCA by use-type from 3 SNPs (LD 0.2) (n=40) **(B)** Hierarchical cluster dendrogram based on 3 SNPs (LD 0.2) (n=40) **(C)** PCA by use-type from 22 SNPs (LD 0.2) (n=49) **(D)** Hierarchical cluster dendrogram based on 22 SNPs (LD 0.2) (n=49) **(E)** PCA by use-type from 14 SNPs (LD 0.2) (n=52) **(F)** Hierarchical cluster dendrogram based on 14 SNPs (LD 0.2) (n=52).

Supplemental Figure 18. Haplogroup analysis for the Medicinal Genomics 61 dataset for 61 samples with alignments to reference sequences for the 18S Ribosomal subunit (Reference JF317360.1) **(A-B),** Chloroplast (Reference KY084475.1) **(C-D)** and Mitochondria (Reference MT361980.1**) (E-F). (A)** PCA by use-type from 10 SNPs (LD 0.2) (n=47) **(B)** Hierarchical cluster dendrogram based on 10 SNPs (LD 0.2) (n=47) **(C)** PCA by use-type from 17 SNPs (LD 0.2) (n=60) **(D)** Hierarchical cluster dendrogram based on 17 SNPs (LD 0.2) (n=60) **(E)** PCA by use-type from 22 SNPs (LD 0.2) (n=61) **(F)** Hierarchical cluster dendrogram based on 22 SNPs (LD 0.2) (n=61).

Supplemental Figure 19. Haplogroup analysis for the Medicinal Genomics 753 dataset for 753 samples with alignments to reference sequences for the Chloroplast (Reference KY084475.1) **(A-B)** and Mitochondria (Reference MT361980.1) **(C-D). (A)** PCA by use-type from 18 SNPs (LD 0.2) (n=679) **(B)** Hierarchical cluster dendrogram based on 18 SNPs (LD 0.2) (n=679) **(C)** PCA by use-type from 42 SNPs (LD 0.2) (n=744) **(D)** Hierarchical cluster dendrogram based on 42 SNPs (LD 0.2) (n=744).

Supplemental Figure 20. Maximum Likelihood phylogenetic tree for the Phylos Biosciences dataset constructed from 279 nuclear SNPs from 844 samples. Sour Diesel samples (n=12) are highlighted in blue at the branch tips.

Supplemental Figure 21. Maximum Likelihood phylogenetic tree for the Phylos Biosciences dataset constructed from 279 nuclear SNPs from 844 samples. Cookies lineage samples (n =23) are highlighted in blue at the branch tips.

Supplemental Figure 22. Maximum Likelihood tree for the LeafWorks dataset constructed from 1,405 nuclear SNPs from 498 samples. Blue Dream samples (n=12) are highlighted in blue at the branch tips.

Table Legends

Table 1. Data sources used for this project.
Table 2. SNP count per dataset pre and post filtering.
Table 3. SNP counts for each dataset by chromosome following biallelic sorting and Linkage Disequilibrium prune at 0.2 and mapped to CBDRx (cs10) genome.
Table 4. Start/Stop Positions of Cannabinoid Synthase Genes (THCAS, CBDAS)
Table 5. Phylogenetic Model fit for each dataset from ModelTest.
Supplemental Table 6. Cultivar name, use-type, clade association and domestication classifications for the LeafWorks data set.
Supplemental Table 7. SSR ID, Cultivar name, use-type, clade association and domestication classifications for the Phylos Biosciences data set.
Supplemental Table 8. SSR ID, Cultivar name, use-type, clade association and domestication classifications for the Soorni *et al.* data set.
Supplemental Table 9. Sample number, total SNP count pre and post filtering for all datasets used for Haplotype analysis.
Supplemental Table 10. Partition specific (Landrace and Domesticates) SNP count per dataset pre and post filtering
Supplemental Table 11. SSR ID, Cultivar name and use-type association for the Sunrise Genetics data set.

Supplemental Table 12. SSR ID, Cultivar name and use-type association for the University of Colorado Boulder data set.
Supplemental Table 13. SSR ID, Cultivar name and use-type association for the Courtagen Life Sciences data set.
Supplemental Table 14. Sample ID, Cultivar name and use-type association for the Medicinal Genomics (n=61) data set.
Supplemental Table 15. Sample ID, RSP ID, Cultivar name and use-type association for the Medicinal Genomics (n=753) data set.

References

[1] K. W. Hillig, "Genetic evidence for speciation in Cannabis (Cannabaceae)," *Genet. Resour. Crop Evol.*, vol. 52, no. 2, pp. 161–180, Mar. 2005, doi: 10.1007/s10722-003-4452-y.

[2] R. C. Clarke and M. D. Merlin, "Cannabis Domestication, Breeding History, Present-day Genetic Diversity, and Future Prospects," *CRC. Crit. Rev. Plant Sci.*, vol. 35, no. 5–6, pp. 293–327, Nov. 2016, doi: 10.1080/07352689.2016.1267498.

[3] M. Clarke & Merlin, "Cannabis: Evolution and Ethnobotany," *Univ. Calif. Press*, 2013.

[4] Z. K. Punja and J. E. Holmes, "Hermaphroditism in Marijuana (Cannabis sativa L.) Inflorescences – Impact on Floral Morphology, Seed Formation, Progeny Sex Ratios, and Genetic Variation," *Front. Plant Sci.*, vol. 11, Jun. 2020, doi: 10.3389/fpls.2020.00718.

[5] J. M. McPartland, W. Hegman, and T. Long, "Cannabis in Asia: its center of origin and early cultivation, based on a synthesis of subfossil pollen and archaeobotanical studies," *Veg. Hist. Archaeobot.*, vol. 28, no. 6, pp. 691–702, Nov. 2019, doi: 10.1007/s00334-019-00731-8.

[6] E. Small and A. Cronquist, "A PRACTICAL AND NATURAL TAXONOMY FOR CANNABIS," *Taxon*, vol. 25, no. 4, pp. 405–435, Aug. 1976, doi: 10.2307/1220524.

[7] C. S. Duvall, "Drug laws, bioprospecting and the agricultural heritage of Cannabis in Africa," *Sp. Polity*, vol. 20, no. 1, pp. 10–25, Jan. 2016, doi: 10.1080/13562576.2016.1138674.

[8] R. C. Clarke, "Cannabis evolution.," *MS thesis, Indiana Univ. Bloom. IN.*, 1987.

[9] J. M. McPartland, "Cannabis Systematics at the Levels of Family, Genus, and Species," *Cannabis Cannabinoid Res.*, vol. 3, no. 1, pp. 203–212, Oct. 2018, doi: 10.1089/can.2018.0039.

[10] J. M. Walker and S. M. Huang, "Cannabinoid analgesia," *Pharmacol. Ther.*, vol. 95, no. 2, pp. 127–135, Aug. 2002, doi: 10.1016/S0163-7258(02)00252-8.

[11] K. B. Svendsen, T. S. Jensen, and F. W. Bach, "Does the cannabinoid dronabinol reduce central pain in multiple sclerosis? Randomised double blind placebo controlled crossover trial," *BMJ*, vol. 329, no. 7460, p. 253, Jul. 2004, doi: 10.1136/bmj.38149.566979.AE.

[12] E. Perucca, "Cannabinoids in the Treatment of Epilepsy: Hard Evidence at Last?," *J. Epilepsy Res.*, vol. 7, no. 2, pp. 61–76, Dec. 2017, doi: 10.14581/jer.17012.

[13] L. A. Parker, E. M. Rock, and C. L. Limebeer, "Regulation of nausea and vomiting by cannabinoids," *Br. J. Pharmacol.*, vol. 163, no. 7, pp. 1411–1422, Aug. 2011, doi: 10.1111/j.1476-5381.2010.01176.x.

[14] M. Badowski and S. Perez, "Clinical utility of dronabinol in the treatment of weight loss associated with HIV and AIDS," *HIV/AIDS - Res. Palliat. Care*, p. 37, Feb. 2016, doi: 10.2147/HIV.S81420.

[15] B. Hurgobin *et al.*, "Recent advances in Cannabis sativa genomics research," *New Phytol.*, vol. 230, no. 1, pp. 73–89, Apr. 2021, doi: 10.1111/nph.17140.

[16] J. M. McPartland and E. Small, "A classification of endangered high-THC cannabis (Cannabis sativa subsp. indica) domesticates and their wild relatives," *PhytoKeys*, vol. 144, pp. 81–112, Apr. 2020, doi: 10.3897/phytokeys.144.46700.

[17] E. P. M. de Meijer *et al.*, "The inheritance of chemical phenotype in Cannabis sativa L.," *Genetics*, vol. 163, no. 1, pp. 335–46, Jan. 2003, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12586720.

[18] R. C. Lynch *et al.*, "Genomic and Chemical Diversity in Cannabis," *CRC. Crit. Rev. Plant*

Sci., vol. 35, no. 5–6, pp. 349–363, Nov. 2016, doi: 10.1080/07352689.2016.1265363.
[19] G. Guerriero *et al.*, "Transcriptomic profiling of hemp bast fibres at different developmental stages.," *Sci. Rep.*, vol. 7, no. 1, p. 4961, 2017, doi: 10.1038/s41598-017-05200-8.
[20] G. Guerriero *et al.*, "Identification of the aquaporin gene family in Cannabis sativa and evidence for the accumulation of silicon in its tissues," *Plant Sci.*, vol. 287, p. 110167, Oct. 2019, doi: 10.1016/j.plantsci.2019.110167.
[21] K. D. Allen, K. McKernan, C. Pauli, J. Roe, A. Torres, and R. Gaudino, "Genomic characterization of the complete terpene synthase gene family from Cannabis sativa," *PLoS One*, vol. 14, no. 9, p. e0222363, Sep. 2019, doi: 10.1371/journal.pone.0222363.
[22] K. J. McKernan *et al.*, "Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes," *bioRxiv*, 2020, doi: 10.1101/2020.01.03.894428.
[23] R. van Velzen and M. E. Schranz, "Origin and Evolution of the Cannabinoid Oxidocyclase Gene Family," *Genome Biol. Evol.*, vol. 13, no. 8, Aug. 2021, doi: 10.1093/gbe/evab130.
[24] A. Soorni, R. Fatahi, D. C. Haak, S. A. Salami, and A. Bombarely, "Assessment of Genetic Diversity and Population Structure in Iranian Cannabis Germplasm," *Sci. Rep.*, vol. 7, no. 1, p. 15668, Dec. 2017, doi: 10.1038/s41598-017-15816-5.
[25] J. Zhang *et al.*, "Genetic Diversity and Population Structure of Cannabis Based on the Genome-Wide Development of Simple Sequence Repeat Markers," *Front. Genet.*, vol. 11, Sep. 2020, doi: 10.3389/fgene.2020.00958.
[26] H. van Bakel *et al.*, "The draft genome and transcriptome of Cannabis sativa," *Genome Biol.*, vol. 12, no. 10, p. R102, 2011, doi: 10.1186/gb-2011-12-10-r102.
[27] T. Maoz, "Making Cannabis History in 2020," *https://www.nrgene.com/blog/making-cannabis-history-in-2020/*, 2020.
[28] C. J. Grassa *et al.*, "A new Cannabis genome assembly associates elevated cannabidiol (CBD) with hemp introgressed into marijuana," *New Phytol.*, p. nph.17243, Feb. 2021, doi: 10.1111/nph.17243.
[29] J. Sawler *et al.*, "The Genetic Structure of Marijuana and Hemp," *PLoS One*, vol. 10, no. 8, p. e0133292, Aug. 2015, doi: 10.1371/journal.pone.0133292.
[30] A. M. Bolger, M. Lohse, and B. Usadel, "Trimmomatic: a flexible trimmer for Illumina sequence data," *Bioinformatics*, vol. 30, no. 15, pp. 2114–2120, Aug. 2014, doi: 10.1093/bioinformatics/btu170.
[31] K. U. Laverty *et al.*, "A physical and genetic map of Cannabis sativa identifies extensive rearrangements at the THC/CBD acid synthase loci," *Genome Res.*, vol. 29, no. 1, pp. 146–156, Jan. 2019, doi: 10.1101/gr.242594.118.
[32] K. J. McKernan *et al.*, "Sequence and annotation of 42 cannabis genomes reveals extensive copy number variation in cannabinoid synthesis and pathogen resistance genes," *bioRxiv*, p. 2020.01.03.894428, Jan. 2020, doi: 10.1101/2020.01.03.894428.
[33] S. Gao *et al.*, "A high-quality reference genome of wild Cannabis sativa," *Hortic. Res.*, vol. 7, no. 1, p. 73, Dec. 2020, doi: 10.1038/s41438-020-0295-3.
[34] A. Raj, M. Stephens, and J. K. Pritchard, "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets," *Genetics*, vol. 197, no. 2, pp. 573–589, Jun. 2014, doi: 10.1534/genetics.114.164350.
[35] G. Siniscalco Gigliano, "Preliminary data on the usefulness of internal transcribed spacer I

(ITS1) sequence in Cannabis sativa L. identification.," *J. Forensic Sci.*, vol. 44, no. 3, pp. 475–7, May 1999, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/10408103.

[36] D. Vergara, K. H. White, K. G. Keepers, and N. C. Kane, "The complete chloroplast genomes of Cannabis sativa and Humulus lupulus.," *Mitochondrial DNA. Part A, DNA mapping, Seq. Anal.*, vol. 27, no. 5, pp. 3793–4, 2016, doi: 10.3109/19401736.2015.1079905.

[37] A. L. Schwabe and M. E. McGlaughlin, "Genetic tools weed out misconceptions of strain reliability in Cannabis sativa: implications for a budding industry," *J. Cannabis Res.*, vol. 1, no. 1, p. 3, Dec. 2019, doi: 10.1186/s42238-019-0001-1.

[38] M. Lewis, E. Russo, and K. Smith, "Pharmacological Foundations of Cannabis Chemovars," *Planta Med.*, vol. 84, no. 04, pp. 225–233, Mar. 2018, doi: 10.1055/s-0043-122240.

[39] A. R. Garfinkel, M. Otten, and S. Crawford, "SNP in Potentially Defunct Tetrahydrocannabinolic Acid Synthase Is a Marker for Cannabigerolic Acid Dominance in Cannabis sativa L.," *Genes (Basel).*, vol. 12, no. 2, p. 228, Feb. 2021, doi: 10.3390/genes12020228.

[40] M. S. Johnson and J. G. Wallace, "Genomic and Chemical Diversity of Commercially Available High-CBD Industrial Hemp Accessions," *Front. Genet.*, vol. 12, Jul. 2021, doi: 10.3389/fgene.2021.682475.

[41] T. Long, M. Wagner, D. Demske, C. Leipe, and P. E. Tarasov, "Cannabis in Eurasia: origin of human use and Bronze Age trans-continental connections," *Veg. Hist. Archaeobot.*, vol. 26, no. 2, pp. 245–258, Mar. 2017, doi: 10.1007/s00334-016-0579-6.

[42] S. Pisanti and M. Bifulco, "Modern History of Medical Cannabis: From Widespread Use to Prohibitionism and Back," *Trends Pharmacol. Sci.*, vol. 38, no. 3, pp. 195–198, Mar. 2017, doi: 10.1016/j.tips.2016.12.002.

[43] P. Henry *et al.*, "A single nucleotide polymorphism assay sheds light on the extent and distribution of genetic diversity, population structure and functional basis of key traits in cultivated north American cannabis," *J. Cannabis Res.*, vol. 2, no. 1, 2020, doi: 10.1186/s42238-020-00036-y.

[44] G. Ren *et al.*, "Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa," *Sci. Adv.*, vol. 7, no. 29, Jul. 2021, doi: 10.1126/sciadv.abg2286.

[45] Q. Zhang *et al.*, "Latitudinal Adaptation and Genetic Insights Into the Origins of Cannabis sativa L.," *Front. Plant Sci.*, vol. 9, Dec. 2018, doi: 10.3389/fpls.2018.01876.

[46] S. Gilmore, R. Peakall, and J. Robertson, "Organelle DNA haplotypes reflect crop-use characteristics and geographic origins of Cannabis sativa," *Forensic Sci. Int.*, vol. 172, no. 2–3, pp. 179–190, 2007, doi: 10.1016/j.forsciint.2006.10.025.

[47] M. G. Roman, D. Gangitano, and R. Houston, "Characterization of new chloroplast markers to determine biogeographical origin and crop type of Cannabis sativa," *Int. J. Legal Med.*, vol. 133, no. 6, pp. 1721–1732, Nov. 2019, doi: 10.1007/s00414-019-02142-w.

[48] A. Liu and J. M. Burke, "Patterns of Nucleotide Diversity in Wild and Cultivated Sunflower," *Genetics*, vol. 173, no. 1, pp. 321–330, May 2006, doi: 10.1534/genetics.105.051110.

[49] S. Zhao, F. Zheng, W. He, H. Wu, S. Pan, and H.-M. Lam, "Impacts of nucleotide fixation during soybean domestication and improvement," *BMC Plant Biol.*, vol. 15, no. 1, p. 81,

2015, doi: 10.1186/s12870-015-0463-z.
[50] L. Hua *et al.*, "LABA1 , a Domestication Gene Associated with Long, Barbed Awns in Wild Rice," *Plant Cell*, vol. 27, no. 7, pp. 1875–1888, Jul. 2015, doi: 10.1105/tpc.15.00260.
[51] C. Onofri, E. P. M. de Meijer, and G. Mandolino, "Sequence heterogeneity of cannabidiolic- and tetrahydrocannabinolic acid-synthase in Cannabis sativa L. and its relationship with chemical phenotype," *Phytochemistry*, vol. 116, pp. 57–68, Aug. 2015, doi: 10.1016/j.phytochem.2015.03.006.
[52] J. J. Zager, I. Lange, N. Srividya, A. Smith, and B. M. Lange, "Gene Networks Underlying Cannabinoid and Terpenoid Accumulation in Cannabis," *Plant Physiol.*, vol. 180, no. 4, pp. 1877–1897, Aug. 2019, doi: 10.1104/pp.18.01506.
[53] J. K. Booth, M. M. S. Yuen, S. Jancsik, L. L. Madilao, J. E. Page, and J. Bohlmann, "Terpene Synthases and Terpene Variation in Cannabis sativa," *Plant Physiol.*, vol. 184, no. 1, pp. 130–147, Sep. 2020, doi: 10.1104/pp.20.00593.
[54] J. Murovec, J. J. Eržen, M. Flajšman, and D. Vodnik, "Analysis of Morphological Traits, Cannabinoid Profiles, THCAS Gene Sequences, and Photosynthesis in Wide and Narrow Leaflet High-Cannabidiol Breeding Populations of Medical Cannabis," *Front. Plant Sci.*, vol. 13, Feb. 2022, doi: 10.3389/fpls.2022.786161.
[55] N. M. Springer and R. M. Stupar, "Allelic variation and heterosis in maize: How do two halves make more than a whole?," *Genome Res.*, vol. 17, no. 3, pp. 264–275, Mar. 2007, doi: 10.1101/gr.5347007.
[56] N. L. Stone, A. J. Murphy, T. J. England, and S. E. O'Sullivan, "A systematic review of minor phytocannabinoids with promising neuroprotective potential," *Br. J. Pharmacol.*, p. bph.15185, Sep. 2020, doi: 10.1111/bph.15185.
[57] S.-H. Park *et al.*, "Contrasting Roles of Cannabidiol as an Insecticide and Rescuing Agent for Ethanol–induced Death in the Tobacco Hornworm Manduca sexta," *Sci. Rep.*, vol. 9, no. 1, p. 10481, Dec. 2019, doi: 10.1038/s41598-019-47017-7.
[58] L. F. Merrick, S. R. Lyon, K. A. Balow, K. M. Murphy, S. S. Jones, and A. H. Carter, "Utilization of Evolutionary Plant Breeding Increases Stability and Adaptation of Winter Wheat Across Diverse Precipitation Zones," *Sustainability*, vol. 12, no. 22, p. 9728, Nov. 2020, doi: 10.3390/su12229728.
[59] A. Dreiseitl, "Specific Resistance of Barley to Powdery Mildew, Its Use and Beyond: A Concise Critical Review," *Genes (Basel).*, vol. 11, no. 9, p. 971, Aug. 2020, doi: 10.3390/genes11090971.
[60] H. D. Cooper, "The International Treaty on Plant Genetic Resources for Food and Agriculture," *Rev. Eur. Community Int. Environ. Law*, vol. 11, no. 1, pp. 1–16, Apr. 2002, doi: 10.1111/1467-9388.00298.
[61] Y. Zhao *et al.*, "Genome-based establishment of a high-yielding heterotic pattern for hybrid wheat breeding," *Proc. Natl. Acad. Sci.*, vol. 112, no. 51, pp. 15624–15629, Dec. 2015, doi: 10.1073/pnas.1514547112.
[62] P. K. Gupta *et al.*, "Hybrid wheat: past, present and future," *Theor. Appl. Genet.*, vol. 132, no. 9, pp. 2463–2483, Sep. 2019, doi: 10.1007/s00122-019-03397-y.
[63] R. H. Wanjari, K. G. Mandal, P. K. Ghosh, T. Adhikari, and N. H. Rao, "Rice in India: Present Status and Strategies to Boost Its Production Through Hybrids," *J. Sustain. Agric.*, vol. 28, no. 1, pp. 19–39, May 2006, doi: 10.1300/J064v28n01_04.
[64] K. Kempe, M. Rubtsova, and M. Gils, "Split-gene system for hybrid wheat seed

production," *Proc. Natl. Acad. Sci.*, vol. 111, no. 25, pp. 9097–9102, Jun. 2014, doi: 10.1073/pnas.1402836111.

[65] S. Peng, G. S. Khush, P. Virk, Q. Tang, and Y. Zou, "Progress in ideotype breeding to increase rice yield potential," *F. Crop. Res.*, vol. 108, no. 1, pp. 32–38, Jul. 2008, doi: 10.1016/j.fcr.2008.04.001.

[66] M. Ferroni and P. Castle, "Public-Private Partnerships and Sustainable Agricultural Development," *Sustainability*, vol. 3, no. 7, pp. 1064–1073, Jul. 2011, doi: 10.3390/su3071064.

[67] T. C. Chao, "Enhancing metadata for research methods in data curation," *Proc. Am. Soc. Inf. Sci. Technol.*, vol. 51, no. 1, pp. 1–4, 2014, doi: 10.1002/meet.2014.14505101103.

[68] D. Jin, P. Henry, J. Shan, and J. Chen, "Classification of cannabis strains in the Canadian market with discriminant analysis of principal components using genome-wide single nucleotide polymorphisms," *PLoS One*, vol. 16, no. 6, p. e0253387, Jun. 2021, doi: 10.1371/journal.pone.0253387.

[69] S. Hübner *et al.*, "Sunflower pan-genome analysis shows that hybridization altered gene content and disease resistance," *Nat. Plants*, vol. 5, no. 1, pp. 54–62, Jan. 2019, doi: 10.1038/s41477-018-0329-0.

[70] J. Li *et al.*, "Cotton pan-genome retrieves the lost sequences and genes during domestication and selection," *Genome Biol.*, vol. 22, no. 1, p. 119, Dec. 2021, doi: 10.1186/s13059-021-02351-w.

[71] R. Della Coletta, Y. Qiu, S. Ou, M. B. Hufford, and C. N. Hirsch, "How the pan-genome is changing crop genomics and improvement," *Genome Biol.*, vol. 22, no. 1, p. 3, Dec. 2021, doi: 10.1186/s13059-020-02224-8.

[72] M. A. ElSohly, Z. Mehmedic, S. Foster, C. Gon, S. Chandra, and J. C. Church, "Changes in Cannabis Potency Over the Last 2 Decades (1995–2014): Analysis of Current Data in the United States," *Biol. Psychiatry*, vol. 79, no. 7, pp. 613–619, Apr. 2016, doi: 10.1016/j.biopsych.2016.01.004.

[73] M. Mostafaei Dehnavi, A. Ebadi, A. Peirovi, G. Taylor, and S. A. Salami, "THC and CBD Fingerprinting of an Elite Cannabis Collection from Iran: Quantifying Diversity to Underpin Future Cannabis Breeding," *Plants*, vol. 11, no. 1, p. 129, Jan. 2022, doi: 10.3390/plants11010129.

[74] S. Andrews, "FastQC: a quality control tool for high throughput sequence data," 2010, [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

[75] H. Li, "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM," *arXiv: Genomics*, 2013.

[76] P. Danecek and S. A. McCarthy, "BCFtools/csq: haplotype-aware variant consequences.," *Bioinformatics*, vol. 33, no. 13, pp. 2037–2039, Jul. 2017, doi: 10.1093/bioinformatics/btx100.

[77] P. Danecek *et al.*, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–2158, Aug. 2011, doi: 10.1093/bioinformatics/btr330.

[78] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir, "A high-performance computing toolset for relatedness and principal component analysis of SNP data," *Bioinformatics*, vol. 28, no. 24, pp. 3326–3328, 2012, doi: 10.1093/bioinformatics/bts606.

[79] F. Kassambara & Mundt, "Factoextra: Extract and Visualize the Results of Multivariate Data Analyses," 2017.

[80] H. Wickham, “ggplot2,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 3, no. 2, pp. 180–185, Mar. 2011, doi: 10.1002/wics.147.
[81] S. Purcell *et al.*, “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.
[82] D. H. Huson, “SplitsTree: analyzing and visualizing evolutionary data,” *Bioinformatics*, vol. 14, no. 1, pp. 68–73, Feb. 1998, doi: 10.1093/bioinformatics/14.1.68.
[83] K. Katoh and D. M. Standley, “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability,” *Mol. Biol. Evol.*, vol. 30, no. 4, pp. 772–780, Apr. 2013, doi: 10.1093/molbev/mst010.
[84] D. Darriba, D. Posada, A. M. Kozlov, A. Stamatakis, B. Morel, and T. Flouri, “ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models,” *Mol. Biol. Evol.*, vol. 37, no. 1, pp. 291–294, Jan. 2020, doi: 10.1093/molbev/msz189.

Table 1. Data sources used for this project. Light grey indicates other public datasets which were not utilized in this study. *†‡ refers to the number of individual samples used in this study out of the total that are publicly available.

Data source	Dataset reference in this text	Bioproject	Number of individuals	Citation
University of Tehran	Soorni *et al.*	PRJNA419020	94	Soorni *et al*., 2017
LeafWorks Inc	**LeafWorks**	**See attached**	**498**	**This manuscript**
Phylos Biosciences	Phylos Biosciences	PRJNA347566	845	https://phylos.bio
Sunrise Genetics	Sunrise Genetics	PRJNA350539	25	NA
University of Colorado Boulder	University of Colorado Boulder	PRJNA317659	162	Lynch *et al*., 2016
University of Colorado Boulder		PRJNA310948	57	Lynch *et al*., 2016
Courtagen Life Sciences	Courtagen Life Sciences	PRJNA297710	58	NA
Medicinal Genomics	Medicinal Genomics (n=61)	NA	61/70	www.medicinalgenomics.com/kannapedia-fastq/
Medicinal Genomics	Medicinal Genomics (n=753)	NA	753	www.kannapedia.net
Dalhouse University	Sawler *et al.*	PRJNA285813	143	Sawler *et al.*, 2015
Total			2,496/2698*†‡	

Table 2. SNP count per dataset pre and post filtering

Dataset	Sample (n)	Total # SNPs	# SNPs post filter	# Bi-allelic SNPs	LD (0.2)
Soorni *et al.*	94	38,195,216	33,629	33,346	6,865
Phylos Biosciences	845	3,735,351	279	279	213
LeafWorks	**498**	**11,149,120**	**1,405**	**1400**	**520**
Sunrise Genetics	25	7,780,326	6,329	6,284	1,604
Lynch *et al.*	162	140,696,175	5,999	5,946	2,223
Medicinal Genomics 61	61	250,960,262	8,716	8,709	2,267
Courtagen Life Sciences	58	479,320,740	311	310	119

Table 3. SNP counts for each dataset by chromosome following biallelic sorting and Linkage Disequilibrium prune at 0.2 and mapped to CBDRx (cs10) genome.

Dataset	**SNPs CHR 1**	**SNPs CHR 2**	**SNPs CHR 3**	**SNPs CHR 4**	**SNPs CHR 5**	**SNPs CHR 6**	**SNPs CHR 7**	**SNPs CHR 8**	**SNPs CHR 9**	**SNPs CHR X**	**SNPs Total**
Soorni *et al.*	917	797	658	780	593	670	552	667	642	589	6,865
LeafWorks	**65**	**63**	**51**	**43**	**51**	**83**	**38**	**32**	**46**	**48**	**520**
Phylos Biosciences (0.99)	30	32	16	32	15	22	15	26	7	18	213
Sunrise Genetics	191	184	176	174	136	178	138	158	117	152	1,604
Lynch *et al.*	336	338	327	304	249	291	264	114	0	0	2,223
Courtagen Life Sciences	13	15	18	5	25	15	7	12	2	7	119
Kannapedia 61	215	209	239	196	329	289	198	166	129	297	2,267

Table 4. Start/Stop Positions of Cannabinoid Synthase Genes (THCAS, CBDAS) as observed from the genome annotation provided by Supercann.net

Chromosome	Start position (bp)	End position (bp)	Size (bp)	Gene description	Supercann.net Gene annotation name
NC_044378.1	25,821,957	25,823,594	1,638	THCA synthase (Fragment)	CsCBD_09G0009540
NC_044378.1	25,848,302	25,871,266	22,965	THC synthase (Fragment)	CsCBD_09G0009550
NC_044378.1	25,982,702	25,983,893	1,192	Tetrahydrocannabinolic acid synthase	CsCBD_09G0009570
NC_044378.1	26,045,537	26,046,424	888	THCA synthase (Fragment)	CsCBD_09G0009580
NC_044378.1	26,085,884	26,086,459	576	Truncated THCA synthase	CsCBD_09G0009590
NC_044378.1	26,086,541	26,087,254	714	THCA synthase (Fragment).	CsCBD_09G0009600
NC_044378.1	26,172,473	26,172,994	522	THCA synthase (Fragment)	CsCBD_09G0009620
NC_044378.1	29,576,301	29,577,125	825	CBDAS-like 2	CsCBD_09G0010520
NC_044378.1	29,577,563	29,577,880	318	CBDAS-like 1.	CsCBD_09G0010530
NC_044378.1	29,635,230	29,636,687	1,458	CDBAS-like 1.	CsCBD_09G0010550
NC_044378.1	29,669,285	29,669,862	578	CBDAS-like 1	CsCBD_09G0010560
NC_044378.1	29,670,027	29,670,575	549	THCAS (Fragment)	CsCBD_09G0010570
NC_044378.1	29,699,660	29,701,184	525	CBDAS-like 2.	CsCBD_09G0010590
NC_044378.1	30,980,797	30,982,759	1,963	Cannabidiolic acid synthase	CsCBD_09G0011030

Table 5. Phylogenetic Model fit for each dataset from ModelTest

Dataset	**Model Fit**
Phylos Biosciences	TPM2uf+G4
University of Tehran	TVM+G4
LeafWorks Inc	TIM2+G4

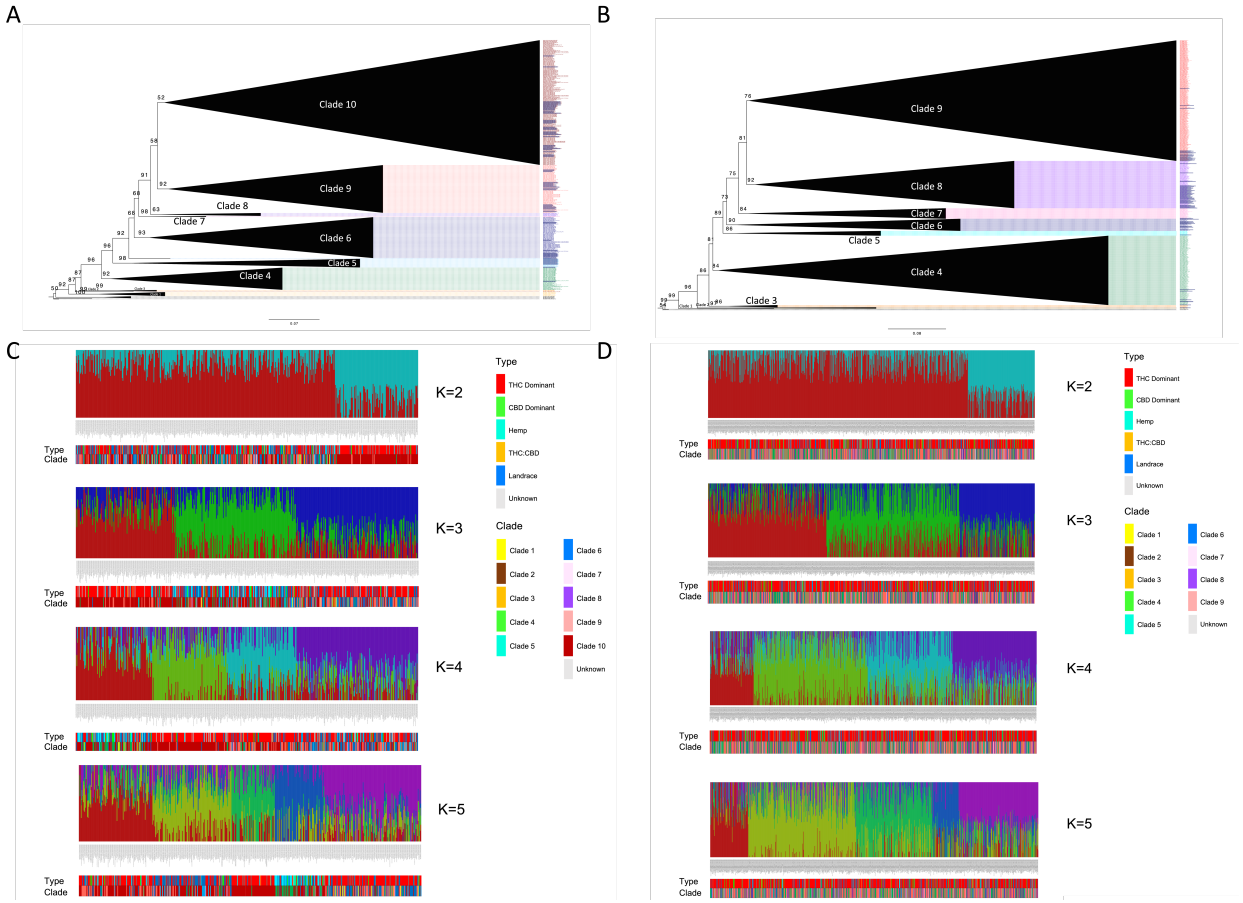

Figure 1. Examining Maximum Likelihood phylogenetic structure and admixture in the LeafWorks and Phylos Biosciences datasets (A) Maximum Likelihood tree for the LeafWorks dataset constructed from 1,405 nuclear SNPs from 498 samples. Landrace samples are highlighted in blue at the branch tips **(B)** Maximum Likelihood phylogenetic tree for the Phylos Biosciences dataset constructed from 279 nuclear SNPs from 844 samples (**C**) Visualization of population structure and admixture for the LeafWorks dataset using the fastSTRUCTURE software (k=2-5) **(D)** Visualization of population structure and admixture for the Phylos Biosciences dataset using the fastSTRUCTURE software (k=2-5).

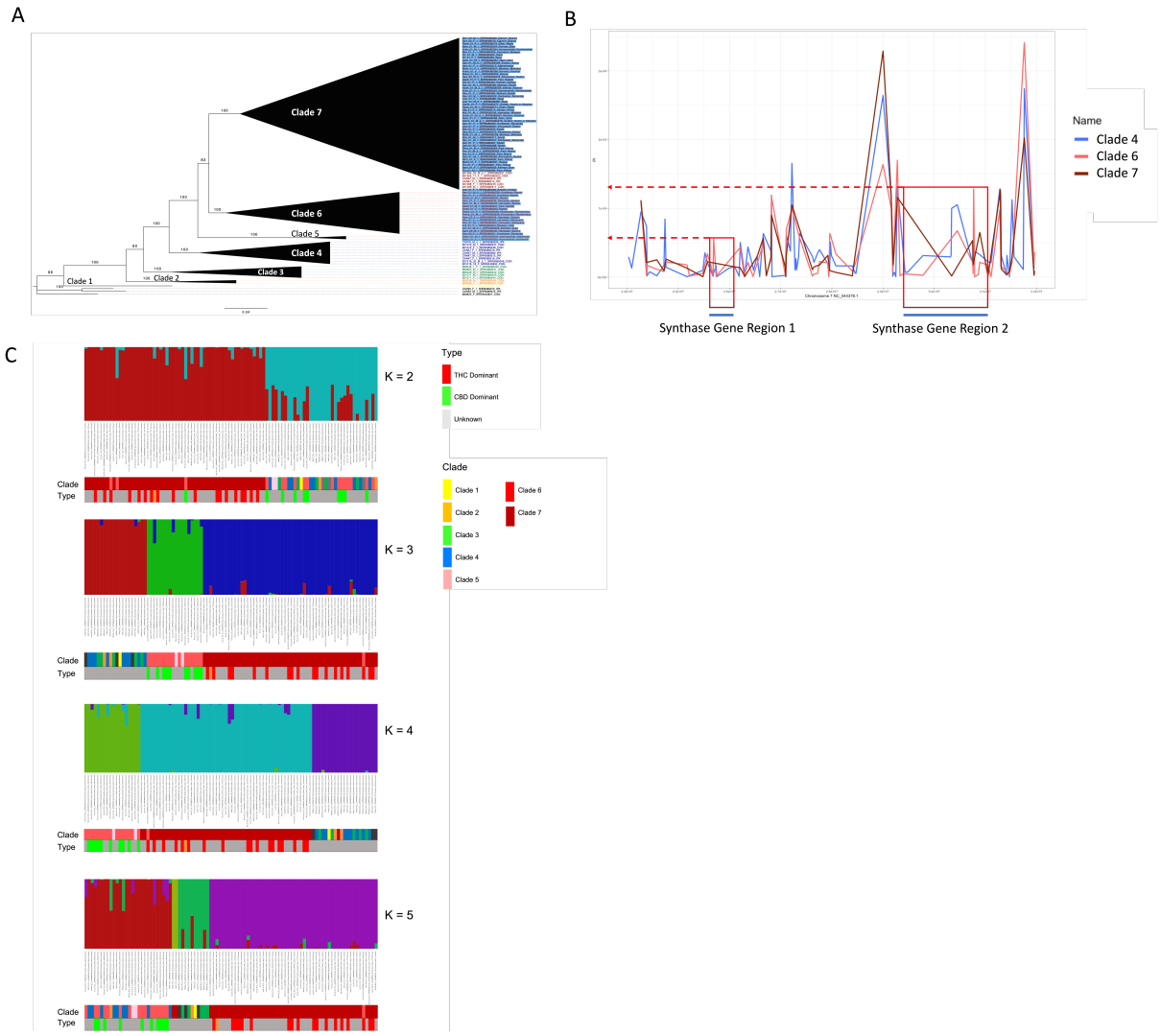

Figure 2. Results for the Soorni *et al.* dataset (**A**) Maximum Likelihood (ML) phylogenetic tree constructed from 33,629 nuclear single nucleotide polymorphisms (SNPs) from 94 samples consisting of 16 from CGN Genebank, 10 from IPK Genebank and 68 landraces (highlighted in blue) sampled across Iran. Colour codes correspond to the main supported clades **(B)** Comparison of clade partitions for nucleotide diversity at the 25-31 Mb regions of Chromosome 7 (NC_044378.1), with Clade 4 (n=9) from 66,260 SNPs, Clade 6 (n=18) from 60,618 SNPs and Clade 7 (n=56) from 48,795 SNPs **(C)** Visualization of population structure and admixture with fastSTRUCTURE software (k=2-5).

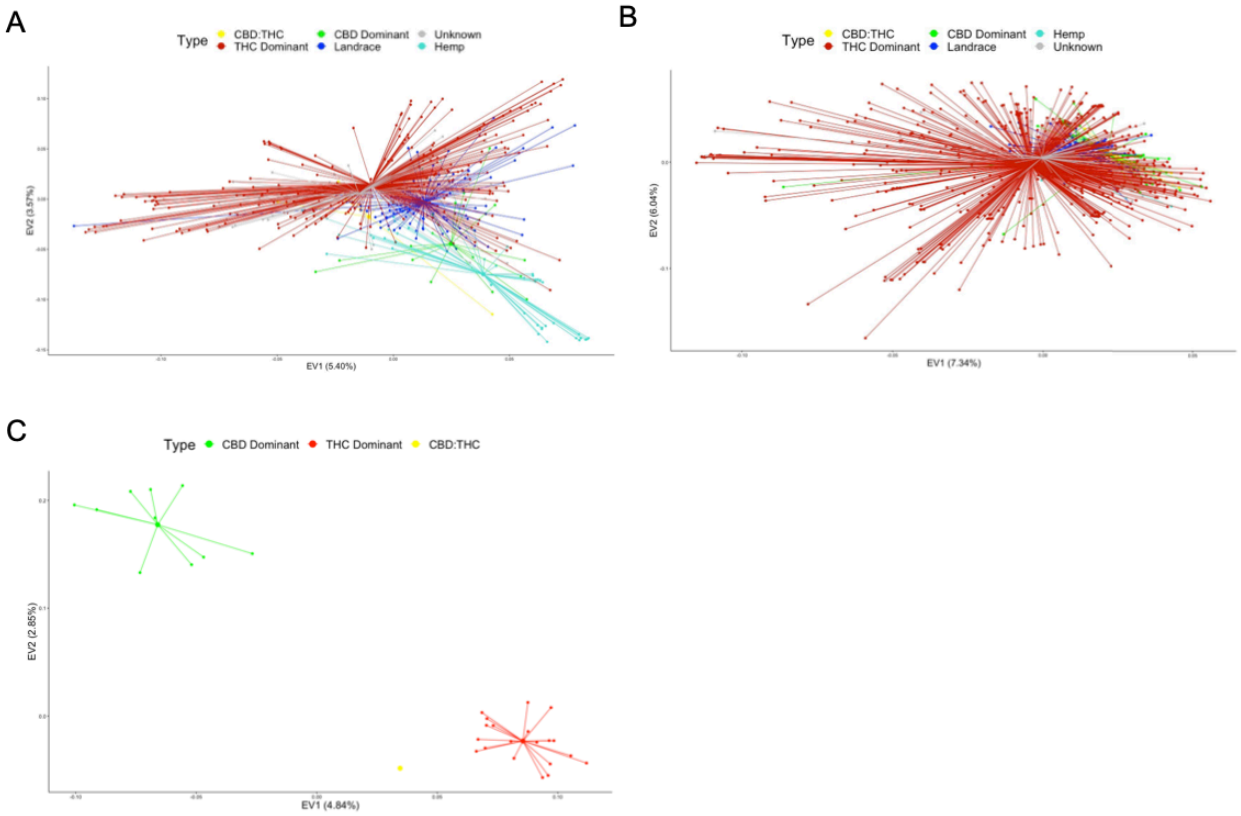

Figure 3. Examination of use-type association across three datasets (A) Principal component analysis (PCA) from 520 nuclear SNPs for the LeafWorks dataset **(B)** PCA from 213 SNPs Phylos Biosciences dataset **(C)** PCA from 6,865 nuclear SNPs for the Soorni *et al.* dataset where cannabinoid content could be determined due to recent publication for 31/94 samples.

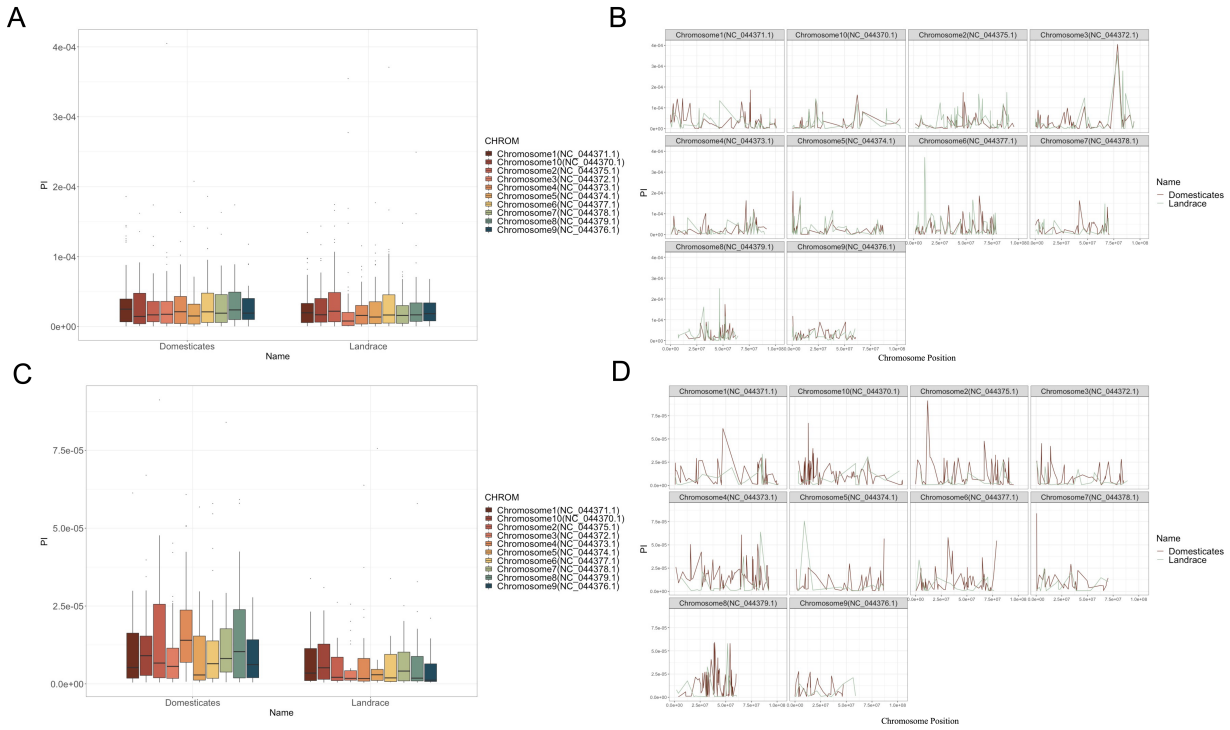

Figure 4. Nucleotide diversity for landrace and domesticated partitions for the LeafWorks and Phylos Biosciences datasets **(A)** Nucleotide diversity by chromosome and **(B)** across chromosome length for Domesticated (n=397, 2,096 SNPs) and Landrace (n=101, 2,131 SNPs) samples for the LeafWorks dataset (**C**) Nucleotide diversity by chromosome and **(D)** across chromosome length for Domesticated (n=679, 704 SNPs) and Landrace (n=107, 266 SNPs) samples for the Phylos Biosciences dataset**.**

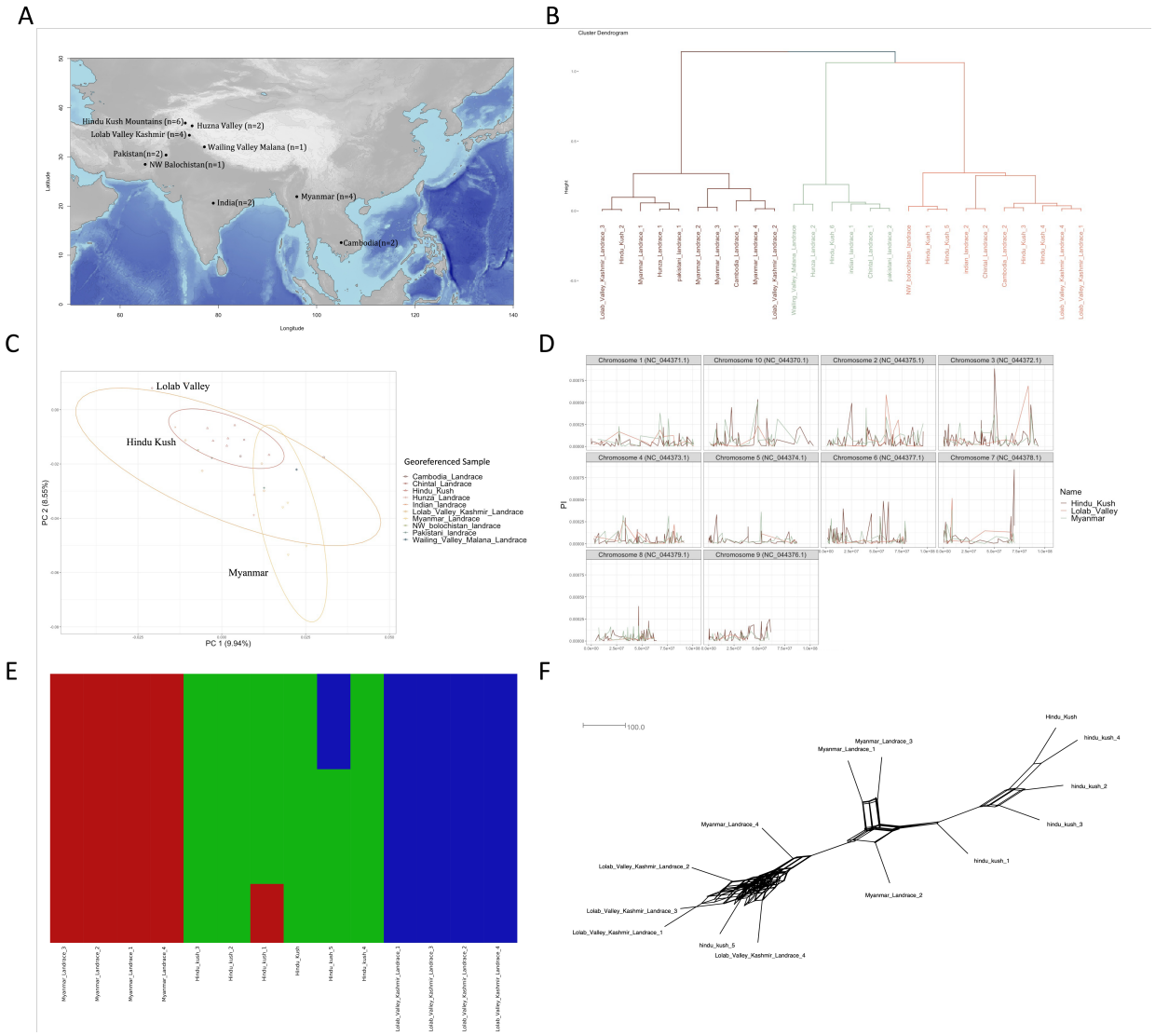

Figure 5. Landrace accessions from the LeafWorks dataset show separation between Indian and Myanmar populations **(A)** Map detailing the locations of landrace accessions, highlighted are the Hindu Kush Mountains, Lolab Valley and Myanmar **(B)** Hierarchical cluster dendrogram based on 304 SNPs (LD 0.2) across 26 samples of known and trusted origin **(C)** PCA based on 304 SNPs with geographical locations of samples as indicated **(D)** Nucleotide diversity comparison between Hindu Kush Mountains (n=6) 4,304 SNPs, Lobal Valley (n=4) 853 SNPs and Myanmar (n=4) 2,204 SNPs **(E)** Visualization of population structure and admixture using the fastSTRUCTURE software (k=3) **(F)** Network tree visualized using the Splits-Tree software with sample source indicated.

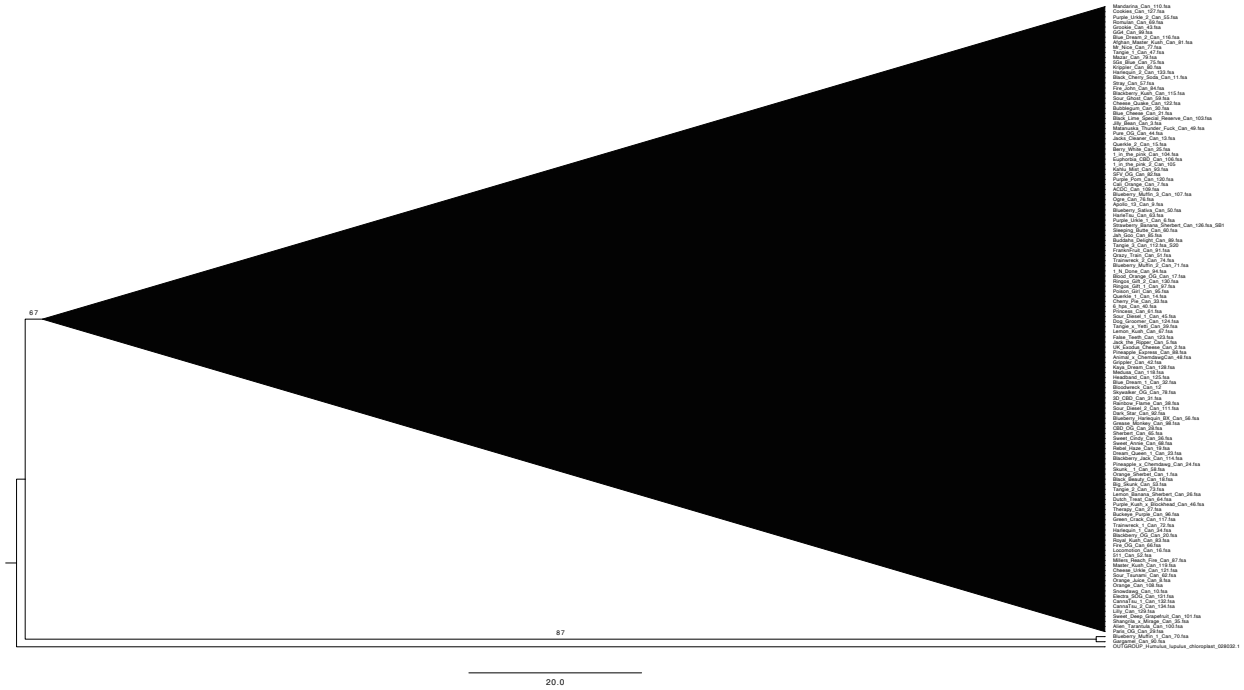

Figure 6. Maximum Likelihood phylogenetic tree for 126 whole chloroplast assemblies. Individuals were aligned using MAFFT. Modeltest-ng revealed the GTR+G4 as the best fit substitution model and IQ-Tree software was used for phylogenetic inference. The resultant tree was visualized using FigTree (Version 1.4.4).