1   Genomic and Chemical Diversity in *Cannabis*

2

3   Ryan C. Lynch*[1,5], Daniela Vergara[1], Silas Tittes[1], Kristin White[1], C. J. Schwartz[2], Matthew J.

4   Gibbs[2], Travis C. Ruthenburg[3,4], Kymron deCesare[3], Donald P. Land[3] and Nolan C. Kane*[1]

5

6   [1] University of Colorado Boulder, Department of Ecology and Evolutionary Biology

7   [2] Marigene Inc.

8   [3] Steep Hill Labs Inc.

9   [4] current address: SC Laboratories Inc. 4000 Airport Way South, Seattle, WA 98108

10   [5] current address: Medicinal Genomics Corporation, 12 Gill St. Suite 3700, Woburn, MA 01801

11

12   * Correspondence: nolan.kane@colorado.edu or rlynch@colorado.edu

13

14

15

16

17

18

19

20

21

22

23

24

Abstract

Plants of the *Cannabis* genus are the only producers of phytocannabinoids, terpenoid compounds that strongly interact with evolutionarily ancient endocannabinoid receptors shared by most bilaterian taxa. For millennia, the plant has been cultivated for these compounds, but also for food, rope, paper, and clothing. Today, specialized varieties yielding high-quality textile fibers, nutritional seed oil or high cannabinoid content are cultivated across the globe. However, the genetic identities and histories of these diverse populations remain largely obscured. We analyzed the nuclear genomic diversity among 339 *Cannabis* varieties, and demonstrate the existence of at least three major groups of diversity. As well as being genetically distinct, each group produces unique cannabinoid and terpenoid content profiles. This combined analysis of population genomic and trait variation informs our understanding of the potential uses of different genetic variants for medicine and agriculture, providing valuable insights and tools for a rapidly emerging, valuable legal industry.

47   Plants of the genus *Cannabis* (Cannabaceae; hemp, drug-type) have been used for thousands of

48   years for fiber, nutritional seed oil and medicinal or psychoactive effects. Archaeological evidence

49   for hemp fiber textile production in China dates to at least as early as 6,000 years ago [1], but possibly

50   as early as 12,000 years ago[2], suggesting *Cannabis* was one of the first domesticated fiber plants.

51   Archeological evidence for medicinal or shamanistic use of *Cannabis* has been found at Indian,

52   central-Asian and middle-eastern sites[3], further illustrating the widespread extent of *Cannabis*

53   utilization throughout human history. A central Asian site of domestication is often cited[4], although

54   genetic analyses suggest two independent domestication events may have occurred separately[5].

55   *Cannabis* plants are usually annual wind-pollinated dioecious herbs, though individuals

56   may live more than a year in subtropical climates[6] and monoecious populations exist[7]. The

57   taxonomic composition of the genus remains unresolved, with two species (*C. indica* and *C. sativa*)

58   commonly cited[8], although *C. ruderalis* is sometimes proposed as a third species that contains

59   northern short-day or auto-flowering plants[9]. Monospecific treatment of the genus as *Cannabis*

60   *sativa* L. is also common[10] and various alternative nomenclature schemes (e.g. *Cannabis sativa*

61   subsp. *indica* var. *kafiristanica*) are sometimes referenced (reviewed in[4]). Even though an

62   extensive monograph on the genus has recently been published[11], limited genetic and experimental

63   data leaves the questions of taxonomy unresolved[12,13].

64   The geographical and ecological range of *Cannabis* is unusually broad, with cultivated

65   populations growing outdoors on every continent except Antarctica in a wide range of

66   environments from sub-arctic to temperate to tropical, and from sea level to over 3,000 meters

67   elevation[14,15]. Feral or wild populations are also found as far north as the edge of the Arctic Circle

68   in Eurasia, but are most common in well drained soils of temperate continental ecosystems in

69   Eurasia and North America, while tropical populations are absent or rare[14]. Perhaps unsurprising,

70   given this diversity of habitats, the species contains extensive phytochemical diversity, particularly

71   in cannabinoid and terpenoid profiles[5,16], and also shows extensive diversity of morphological and

72   life-history characteristics, further fueling debate regarding the taxonomic status and origins of

73   *Cannabis* domestication.

74   One distinctive feature of the *Cannabis* genus is the production of a tremendous diversity

75   of compounds called *cannabinoids,* so named because they are not produced at high levels in any

76   other plant species[17]. Cannabinoids are a group of at least 74 known C21 terpenophenolic

77   compounds[18,19] responsible for many reported medicinal and psychoactive effects of *Cannabis*

78   consumption[20]. Some estimates for the total number of phytocannabinoids range to well over a

79   hundred[21], though this number includes breakdown products as well as compounds found at

80   extremely low levels. The plants produce a non-psychoactive carboxylic acid form of these

81   compounds, with heating required to convert cannabinoids into the psychoactive decarboxylated

82   forms. Interestingly, these compounds have pronounced neurological effects on a wide range of

83   vertebrate and invertebrate taxa, suggesting an ancient origin of the endocannabinoid receptors,

84   perhaps as old as the last common ancestor of all extant bilaterians, over 500 MYA[22]. The plant

85   compounds thus produced have the potential to affect a broad range of metazoans, though their

86   ecological functions in nature are not well understood. Indeed, suggested roles for these

87   compounds include many biotic and abiotic defenses, such as suppression of pathogens and

88   herbivores, protection from UV radiation damage, and attraction of seed dispersers. These

89   hypotheses about the selective benefits of cannabinoid production remain speculative, as none

90   have been conclusively verified to date. We do know more, however, about the more recent

91   evolution of the plants under human cultivation.

92   High delta-9-tetrahydrocannabinolic acid (THCA)[23] content has been selected for in many

93   strains due to its potential to be converted to delta-9-tetrahydrocannabinol (THC), which has potent

94   psychoactive[24], appetite-stimulating[25], analgesic[26] and antiemetic[27] effects. These effects are

95   mediated through interactions with human endocannabinoid CB1 receptors found in the brain [28],

96   and CB2 receptors, which are concentrated in peripheral tissues[29]. Other THC receptor binding

97   locations are hypothesized as well[30]. After several decades of accelerated clandestine cultivation

98   technique and breeding improvements, modern strains can now yield dried un-pollenated pistillate

99   inflorescence material that contains over 30% THCA by dry-weight[31]. However, other

100 cannabinoids may also be present in high concentrations. In particular, high cannabidiolic acid

101 (CBDA) plants were historically used in some hashish preparations[32] and are presently in high

102 demand as an anti-seizure therapy[33]. In contrast with THC, which acts as a partial agonist of the

103 CB1 and CB2 receptors, CBD does not have as strong psychoactive properties, and instead has

104 antagonist activity on agonists of the CB1- and CB2-receptors[34]. Thus, the two most abundant

105 cannabinoids produced in *Cannabis* have, to some degree, opposing neurological effects.

106   THCA and CBDA are alternative products of a shared precursor, CBGA[35]. A single locus

107 with co-dominant alleles was proposed to explain patterns of inheritance for THCA to CBDA

108 ratios[7,36]. However more recent quantitative trait loci (QTL) mapping experiments[37], expression

109 studies[38] and genomic analyses[10] paint a more complex scenario with several linked paralogs

110 responsible for the various THCA and CBDA phenotypes. Other cannabinoids such as

111 cannabigerol (CBG)[39], cannabichromene (CBC)[40] and delta-9-tetrahydocannabivarin (THCV)[41]

112 demonstrate pharmacological promise, and can also be produced at high levels by the plant[42–44].

113 Additionally, *Cannabis* secondary metabolites such as terpenoids and flavonoids likely contribute

114   to therapeutic or psychoactive effects[2], such as β-myrcene, humulene and linalool proposed to

115   produce sedative effects associated with specific strains[45].

116   In this study, plants that produce low levels of total cannabinoids are herein referred to as

117   hemp, while high cannabinoid producing varietals are described as drug-type strains. Legal

118   definitions often use a maximum THCA threshold to delineate hemp from drug-types, thus some

119   high CBDA producing strains are categorized as hemp. However this definition ignores the

120   broader traditional usage of hemp for fibers or seed oils and historical presence of CBDA-

121   producing alleles in some drug-type populations[32]. Additionally, hemp strains have a distinct set

122   of growth characteristics, with fiber varieties reaching up to 6 meters in height during a growing

123   season, exhibiting reduced flower set, increased internodal spacing and lower total cannabinoid

124   concentration per unit mass compared to drug-type relative. Despite the widespread prohibition of

125   drug-type *Cannabis* cultivation from the 1930s to present, hemp cultivation and breeding

126   continued in parts of Europe and China though this period, and experienced a brief comeback

127   during World War II in the USA through the Hemp for Victory campaign. Studies to date have

128   found hemp varieties are genetically distinct from drug-type strains[10], though interestingly Hillig[5]

129   found broad leaflet southeastern Asian hemp landraces to be more closely related to Asian drug-

130   type strains than to European hemp strains.

131   *Cannabis* has a diploid genome ($2n = 20$), and an XY/XX chromosomal sex-determining

132   system. The genome size is estimated to be 818 Mb for female plants and 843 Mb for male plants[46].

133   Currently a draft genome consisting of 60,029 scaffolds is available for the Purple Kush (PK) drug-

134   type strain from the National Center for Biotechnology Information. Additional whole genome

135   data is available from NCBI for the Finola and USO31 hemp strains. Various reduced

136   representation genome, gene and RNA sequence data are also available from NCBI. Presently

137   *Cannabis* is the only multi-billion dollar legal crop without a sequence-based genetic linkage or

138   physical genome map. Indeed, the first genetic map for the species, using AFLP and microsatellite

139   markers, was only recently published, providing for the first time, quantitative trait mapping of

140   cannabinoid content and other traits[37].

141   Initial studies of *Cannabis* genetic diversity examined either many samples with few

142   molecular markers[5] or whole genome wide data for relatively few samples types[10]. Sawler et al.[47]

143   recently published a survey of *Cannabis* genomic diversity, using a reduced genomic

144   representation strategy to evaluate 81 marijuana (drug-type) and 43 hemp strains. The aim of this

145   present study is to assess the genomic diversity and phylogenetic relationships among 339 total

146   *Cannabis* plants that have distinct phenotypes, and that were described *a priori* by plant breeders

147   as various landraces, *indica*, *sativa*, hemp and drug-types, as well as commercially available

148   hemp and drug-types with unclear pedigrees. We have combined data from existing sources and

149   generated new data to create the largest sample set of *Cannabis* genomic sequence data

150   published to date. These data and analyses will continue to facilitate the development of

151   modernized breeding and quality assurance tools, which are lacking in the nascent legal

152   *Cannabis* industry.

153

154   **Materials and Methods**

155

156   **Sample collection.** DNA was obtained from numerous sources, including a variety of breeding

157   and production facilities. The strain names, descriptions and putative origins used in this paper

158   were recorded from the providers of the DNA and sequence data (Supplementary Table 1). For

159   data not previously published, DNA extractions were performed using the Qiagen DNeasy Plant

160   Mini Kit (Valencia, CA) according to the manufacturer's protocol.

161

162   **Whole genome shotgun (WGS) sequencing.** 60 samples were sequenced using standard

163   Illumina multiplexed library preparation protocols for two 2 x 125 HiSeq 2500 lanes and one 2 x

164   150 NextSeq 500 run. Sequencing efforts were targeted for approximately 4-6x coverage of the

165   *Cannabis* genome per sample.

166

167   **Genotype-by-Sequencing (GBS).** 182 samples were sequenced on two 1 x 100 HiSeq 2500

168   lanes, following a multiplexed library preparation protocol described previously[48].

169

170   **Publically available data.** We obtained three WGS datasets available from NCBI[10] and received

171   seven additional WGS datasets from Medicinal Genomics Corporation

172   (www.medicinalgenomics.com). GBS data for 143 samples from Sawler et al.[47] were also

173   included in this study.

174

175   **Sequence Processing, Alignment and SNP calling.** Trimmomatic[49] was used to trim any

176   remaining adaptor sequence from raw fastq reads and remove sequences with low quality regions

177   or ambiguous base calls using the following settings:

178   ILLUMINACLIP:IlluminaAdapters:2:20:10 LEADING:20 TRAILING:20

179   SLIDINGWINDOW:5:15 MINLEN:100. Trimmed raw reads from the 67 total WGS samples

180   were then aligned to the only publicly available draft genome of PK (JH226140-JH286168)

181   using the Burrows-Wheeler Alignment tool (BWA mem)[50]. Chloroplast and mitochondrial

182   regions were excluded. We collated the individual alignments to produce a single variant call

183   format table (.vcf) for all samples using samtools mpileup -uf | bcftools view –bvcg[51]. We

184   filtered the vcf table to include only high quality informative SNP sites using vcftools[52], bash and

185   awk with the following vcf parameters: Q (>200), GQ (>10), AF1 (.1 - .9), biallelic sites only

186   and no ambiguous bases. Next, data filters were applied through plink[53] to require that

187   individuals have a minimum 50% informative sites and that sites each have data for minimum

188   20% of samples. Finally we used an estimate of expected coverage for the single copy portion of

189   the genome based on the estimated genome size and number of reads being aligned. This was

190   adjusted empirically based on total coverage level (across all WGS samples) per SNP site

191   (Supplementary Figure 1) and bounded by a 95% Poisson confidence interval (mean 362x

192   coverage). Further removal of repetitive content was achieved by aligning the PK reference to

193   itself with BLASTN and removing all sites that were within regions of $\geq$ 97% identity for at $\geq$

194   500 bp alignments. These aforementioned processing, alignment and SNP calling procedures

195   were then preformed separately on the 182 GBS samples generated for this study and the 143

196   GBS samples previously published[47], resulting in three vcf tables and filtered SNP sets. GBS

197   SNPs were additionally required to have a minimum of 5x coverage per sample. Due to limited

198   overlap between the SNP sites produced by the two GBS libraries, most downstream analyses

199   were performed separately for each GBS library along with its corresponding set of WGS SNPs.

200   Code used for these analyses is available at https://github.com/KaneLab.

201

202   **SNP Analyses.** To visualize genetic relationships, divergence, and ancestral hybridization

203   among lineages, a phylogenetic neighbor network was inferred using simple p-distance

204   calculations[54]. Heterozygosity counts and Multidimensional Scaling (MDS) analyses were

205   calculated with Plink[53]. Average within and between group genetic distances, and a 45 SNP

206   alignment neighbor joining tree based on p-distances, were calculated with MEGA6[55].

207   Population structure inferences were made through FastStructure[56] and FLOCK[57]. Tests for

208   reticulation within the trees and admixture between populations were performed in TreeMix[58]

209   F_{ST} estimates were calculated with vcftools[52].

210

211   **Chemical Analyses of Genetic Groups.** The cannabinoid and terpenoid information

212   (chemotype) for a portion of the strains in the genome analysis were generated by Steep Hill

213   Labs (http://steephill.com/). Only strains with matching data in the genomic analysis were

214   analyzed, for a total of 112 individuals from 17 strains from the BLDT group, 278 individuals

215   from 35 unique strains from the NLDT group, and 33 individuals from two strains of hemp, for a

216   total of 423 individuals in this analysis (Supplementary Table 1). This chemotype analysis was

217   performed using high performance liquid chromatography (HPLC) with Agilent (1260 Infinity,

218   Santa Clara, CA) and Shimadzu (Prominence HPLC, Columbia, MD) equipment. Between 400

219   and 600 milligrams of each sample was extracted into methanol, diluted and analyzed by HPLC.

220   A mobile phase consisting of 0.1% formic acid in water and 0.1% formic acid in methanol was

221   used with a gradient starting at 72% methanol and ending at 99% methanol. Terpenoid standards

222   were purchased from Sigma-Aldrich (St. Louis, MO). Cannabinoid standards were purchased

223   from Cerilliant (Round Rock, TX), RESTEK (Bellefonte, PA) and Lipomed (Cambridge, MA).

224   A C18 column from RESTEK (Raptor ARC-18, Bellefonte, PA) or Phenomenex (Kinetex C18,

225   Torrance, CA) was used. Concentrations of cannabinoids without commercially available

226   standards were estimated using published absorptivies[59]. The chemotype data analyzed for this

227   research includes 13 cannabinoids and eight terpenoids. Each compound was quantified using a

228   linear calibration curve. Analytes were measured as percent mass in sample and not corrected for

229   moisture content.

230   We performed a one-way ANOVA for each cannabinoid and terpenoid separately, with

231   the group (NLDT, BLDT, and hemp) as the predictor variable. We used Bonferroni corrections

232   for multiple comparisons. We also implemented a Principal Component Analysis (PCA) with

233   prcomp function in base R, and car was used to visualize 95% confidence ellipses for each group

234   (www.R-project.org). Individuals with missing data values for any cannabinoid or terpenoid

235   were removed. After removing the individuals with missing values, we had a total of 351

236   individuals: 94 BLDT, 229 NLDT, and 28 hemp.

237

238   **Results and Discussion**

239

240   **Sequencing and SNPs.** Summary information and raw sequencing libraries are publically

241   available from the NCBI short read archive (accessions pending). Detailed information about all

242   samples can be found in Supplementary Table 1 and examples of wide and narrow leaflet forms

243   are shown in Figure 1. Of the 466,427,059 non-ambiguous base pairs in the PK reference,

244   77,810,563 bps were removed due to excess self-similarity ($\geq$ 97 % identity and $\geq$ 500 bps

245   length, Supplementary Figure 1). After this filter, the total single copy portion the PK reference

246   within the combined coverage levels for all 67 WGS samples of 326x – 401x, a 95% Poisson

247   confidence interval around a 362x mean, was 71,236,365 bps (Supplementary Figure 1). After

248   quality (Q), genotype quality (GQ), allele frequency (AF), missing data, biallelic and ambiguous

249   base filters, the following SNP counts remained: 491,341 WGS, 2,894 GBS (this study), SNPs

250   4,105 GBS (Sawler[47]). Forty-five SNPs overlapped both GBS datasets, and the WGS samples.
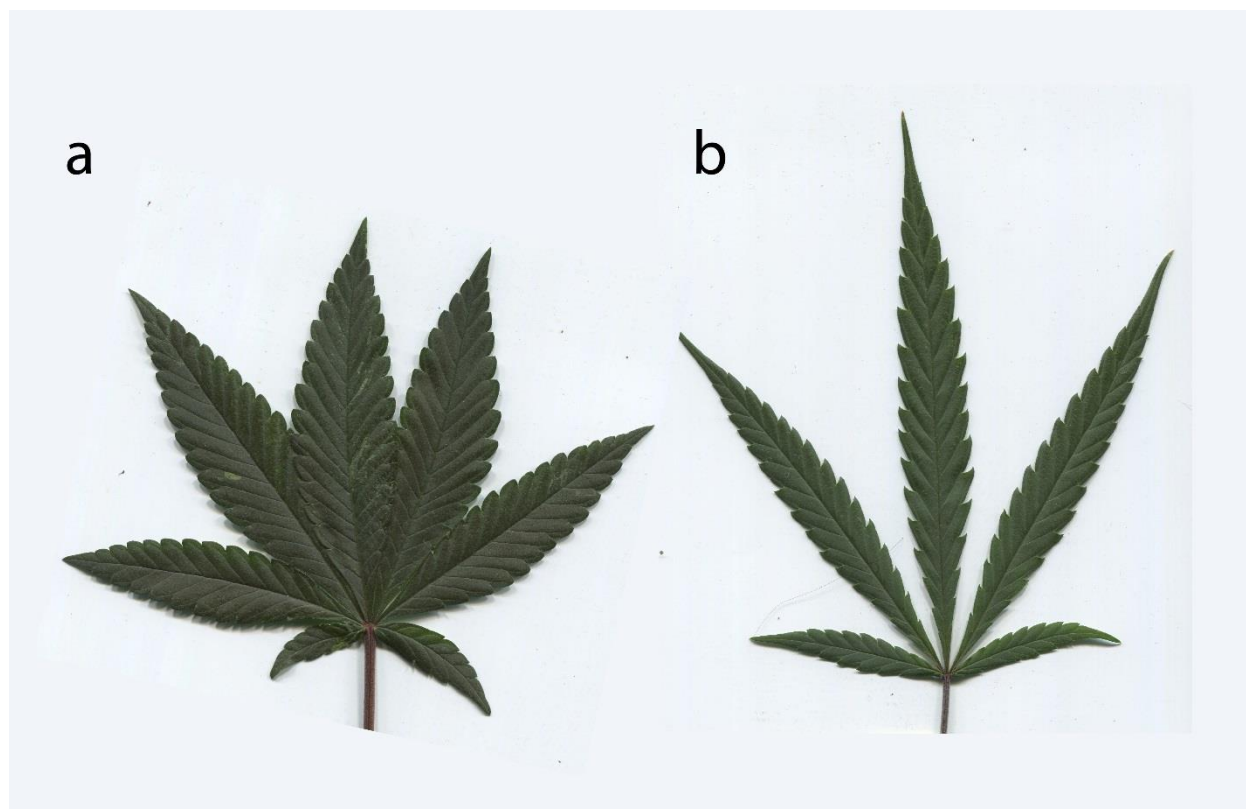
251

Figure 1. Example of broad leaflet type (a, R4) and narrow leaflet type (b, Super Lemon Haze) strains. Photograph credits: D. Vergara.

Phylogenetic Relationships. Bifurcating trees are commonly used to model mutation driven

divergence and speciation events. Whole genome wide sequence datasets include information

about recombination, hybridization, and gene loss or genesis events, some of which may be

incongruent with one and other[54]. Phylogenetic networks can represent incompatible

phylogenetic signals across large character matrices in a visually informative manner. Figure 2

contains 195 *Cannabis* samples including WGS and GBS data, and shows that all European

hemp strains form a distinct clade, separated from drug-type strains by a consistent band of

parallel branches. Broad leaflet drug-type strains clustered with purported Afghan Kush landrace

samples (Supplementary Table 1 and Supplementary Figure 3), while narrow leaflet drug-type

264   strains appear to contain several groups with only faint visible distinctions between them,

265   perhaps influenced by the inclusion of hybrid strains in the analysis.
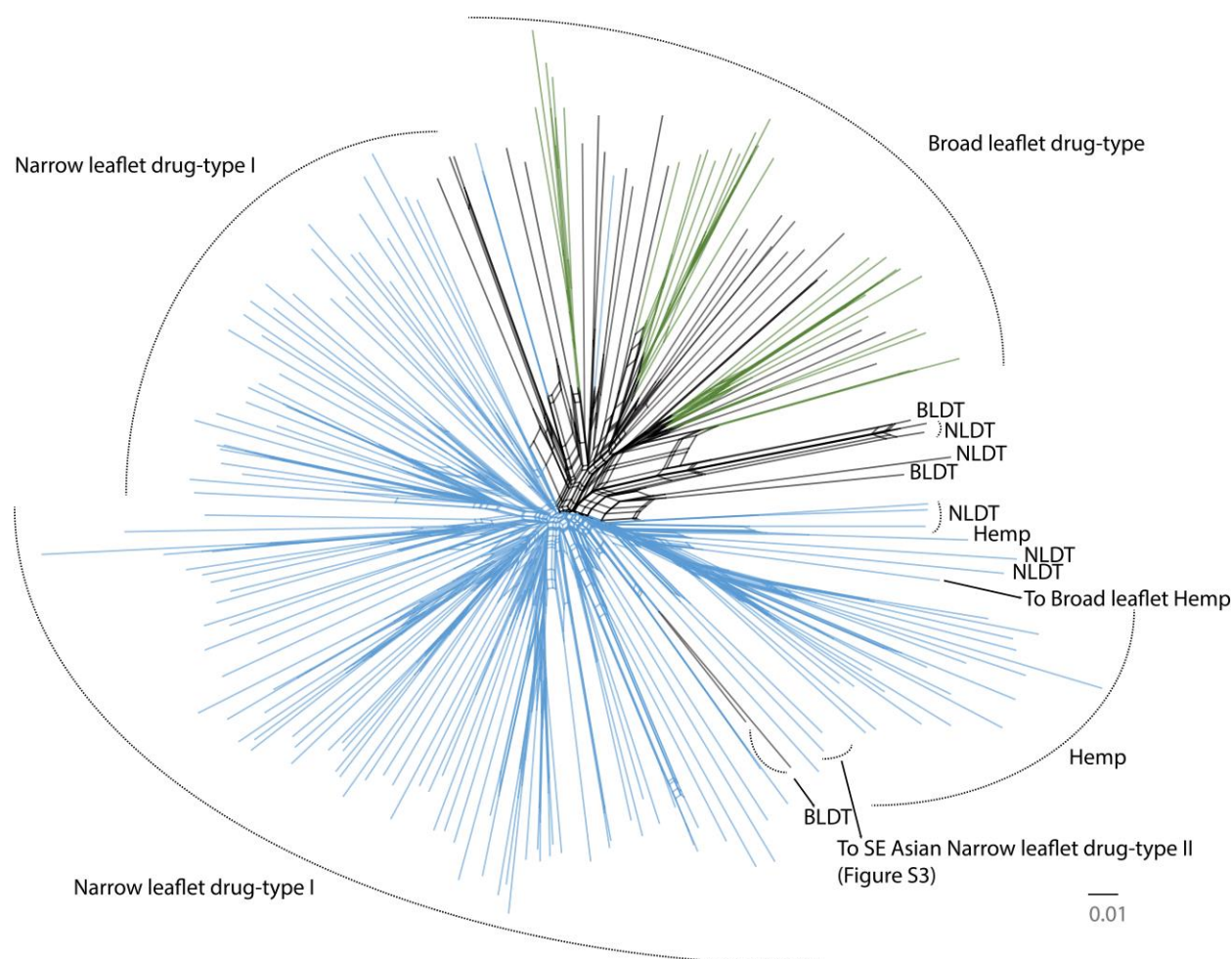

266

267   Figure 2. Phylogenetic neighbor network of a 2,894 SNP alignment from the single-copy portion
268   of the *Cannabis* genome. Clade names on the periphery were inferred via FLOCK (where K ≥ 3
269   was most likely). Colored branches indicate fastStructure population membership of ≥ 70%
270   assignment (where K=2 was most likely). NLDT = Narrow Leaflet Drug-Type and BLDT =
271   Broad Leaflet Drug-Type. To SE Asian NLDT II points to Dr. Grinspoon and Somali Taxi Cab
272   samples. To Broad Leaflet Hemp points to a Chinese hemp sample. A high-resolution version of
273   this figure that includes each sample name is available from:
274   http://figshare.com/articles/Cannabis_Tree/1585470/1

275

276   We found significantly more heterozygosity in drug-type strains than in hemp varieties

277   (31 % v 22 %, $p < 0.001$, two-tailed Mann-Whitney U-test, Table 1). This likely reflects the

13

278   widespread hybridization of strains in North America during the transition to indoor cultivation

279   of drug-types starting in the 1970s[60], as well as the extensive reliance on clonal propagation for

280   indoor commercial cultivation, which does not require trait stable seed stock. Conversely, fiber

281   and seed oil hemp are grown on multi-acre scales that have necessitated the stabilization of

282   agronomically important traits in seed stocks, likely leading to reduced heterozygosity at some

283   loci.

284

	Mean Within Distances	Heterozygosity %
Hemp	0.195	0.22*
All Drug-types	0.244	0.31*
NLDT	0.237	0.32
BLDT	0.221	0.30
	Mean Between Distances	F_{ST}
Hemp v. All Drug-types	0.273	0.098530
Hemp v. NLDT	0.269	0.091679
Hemp v. BLDT	0.281	0.10131
NLDT v. BLDT	0.258	0.036156

285

286   Table 1. Summary of genetic distance, heterozygosity and F_{st} information for major *Cannabis*
287   groups. * = significantly different (p < 0.001, two-tailed Mann-Whitney U-test).
288

289   **Population Structure.** To determine the statistical likelihood of various population scenarios

290   represented in our samples, we first applied the FLOCK model to our data set of 195 GBS and

291   WGS *Cannabis* samples, which is an iterative reallocation clustering algorithm that does not

292   require non-admixed individuals to make population assignments[57]. Using the K-partitioning

293   method suggested by the authors[57], we determined that K ≥ 3, after testing K values of one to

294   eight (Table 2 and peripheral population names in Figure 2). FLOCK was able to assign all

295   samples to one of the three populations, although it does not calculate admixture proportions.

296   Sample population assignments were largely consistent with the known history of these samples,

14

297 and appear visually consistent with MDS analysis (Supplementary Figure 2). For example all

298 fiber and seed oil hemps were assigned to an exclusive population, with the exception of sample

299 AC/DC, a high CBDA producing variety, with likely hybrid hemp origins (Figure 2, Table 2).

A-train	Original_Sour_Diesel	C36	H11	Schemp
Afghan_Kush	Phantom_Cookies	C37	H5	Skunk_#1
Afghan_Kush	Platinum_OG	Canna_Tsu	Harlequin	Somali_Taxi_Cab
Afghan_Kush	Purple_Kryptonite	Cannatonic	Hash_Plant	Spectrum-11
Afghan_Kush	Purple_Kush	CBD_Diesel	Hawaiian	Spectrum-14
Afghan_Kush	Purple_Urkle	CBD_Shark_F-6	Holy_Grail	Super_Lemon_Haze
Boss_Hogg	R4	CBD-0	Holy_Grail_Kush	Sweet_Afghani_Delicious
Bubba_Kush	San_Fran_Valley_OG_Kush	CBD18	Jack_47	Sweet_Skunk
Char_Tango	Screaming_Haze	Charlottes_Web	Jack_Flash	Tangerine_Haze
Chem_91'	SFV	Cheese_Quake	Jack_Herer	Train_Wreck
Chem91	Skywalker_OG	Cherry	Jack_Herrer	Trainwreck
chemdawg	Snowcap	Cherry_Afghan	Jack_Herer	Violator_Kush
Chocolate_Kush	Snowcap	Chocolope	Jack_Skellington	White_Cookies
Cripped_Out_Cookies	Sour_Diesel	Chocolope	Juicy_Fruit	White_Widdow
Cript_out_Cookies	Sour_Patch_Kush	Colombia_Rio_Negro	Lebanese	White_Widow
Dead_Head_OG	Sour_Willie	Critical_Kush	Lemon_Skunk	Wonder_Woman
Dog_Walker	Sshrek	Critical_Kush	Liberty_Haze	XJ_13
Flo	The_Sauce	Critical_Mass	Lions_Tabernacle	AC/DC
Girl_Scout_Cookie_#6	The_Sauce	Dr._Grinspoon	Low_Ryder	AZ_Star_#1
Girl_Scout_Cookies	Tora_Bora	Durban_Poison	LSD	Carmagnola
Girl_Scout_Cookies	WAF_B	Durban_Poison	Mad_Cow	Carmagnola
Goast_Train_Haze	4-Jack	Easy_Sativa	Mango_Stomper	Carmagnola
Grapefruit	Afghan_Mango	Exodus_Cheese	Maui_Waui	Carmagnola
Guido_OG	Afghan_Mango	G13	Mazar	Carmagnola
Headband	Afghan_Mango	G13_Haze-31	Medical_Mass	Carmagnola
Hindu_Kush	Alaskan_Thunderfuck	Gin	Melon_Gum	Chinese_hemp
Kandy_Kush	Appalachian_Mad_Sun	Girl_Scout_Cookies	Mexican_E	Dagestani_hemp
King_Chem	Auto_AK47	Glass_Slipper	MO	EuroOil_2
King_Louie_Cookies	B-5	Golden_Goat	Nuclear_Fruit	Feral_Kansas
Kool_Aid_Kush	BC_HQ	Grand_Daddy_Purps	Otto	Feral_Nebraska
Kosher_Kush	Black_Cherry	Grape_AK-47	Peaches_and_Cream	Feral_Nebraska
Kosher_Kush	Black_Jack	Grape_Ape	Pineapple	Finola
Kosher_Kush_#1	Blue_Cheese	Grape_Ape	Pineapple_Express	J7
Kunduz	Blue_Dream	Grape_Kush	Pink_Lady	J20
Larry_OG	Blue_Dream	Grape_Kush	Pre-98_Bubba_Kush	J28
Medibud	Blue_OG	Green_Crack	Purps	Kompolti_1
OG_18	Blueberry_DJ	Green_Crack	R4	Kompolti_2
OG_Kush_1	Bubble_Gum	Green_Mandarine	Red_Purps	Sievers_Infinity
Old_Skool_OG	Bubble_Gum_XL	Green_Poison	Rocky_Mountain_Blueberry	US031

300

301 Table 2. Sample names and FLOCK assignment to three groups, represented with different cell
302 colors. Green are BLDT, blue are NLDT and yellow are hemp.

15

303

304   Additionally we applied the admixture model based Bayesian clustering method of

305 fastStructure to the same 195 samples[56]. The most likely population structure analysis of K=2

306 (Figure 2, Supplementary Table 1), shows consistent separation between BLDT and NLDT and

307 hemp strains. Some hemp and NLDT strains were each assigned with near 100% population

308 membership to the same population (Figure 2, light blue samples, Supplementary Table 1),

309 despite the clear separation visualized in the tree and statistically significant mean between-

310 group genetic distance measured (Table 1). The separation of BLDT and NLDT strains into

311 fastStructure populations was stable when hemp samples were excluded from the analysis

312 (Supplementary Table 1). Sawler et al.[47] used fastStructure to delineate hemp from drug-types as

313 the major division of *Cannabis* diversity, and found two drug-type sub-groups within their

314 samples when hemp types were excluded from the analysis. Likewise using a smaller dataset,

315 Lynch[61] found support for K=3, consisting of two separate drug-type populations and hemp

316 types, using the original Structure implementation[62] and the Evanno method to select the best

317 value of K[63]. However, we caution that despite many claims for the availability of "landrace

318 genetics" (strains) from *Cannabis* producers, breeders and seed sellers, these may or may not

319 represent non-admixed individuals[60]—a situation that can be problematic for the Structure and

320 fastStructure approaches[62].

321   The GBS samples from Sawler et al.[47] appear to contain an additional divergent NLDT

322 clade, with likely SE Asian origins (Supplementary Figures 3 and 4), that did not emerge from

323 our main analyses. Due to very limited overlap between sequence fragments from the two GBS

324 datasets, which results from using different restriction enzymes, we were required to re-analyze

325 the Sawler data in combination with only our 67 WGS samples. A connection was made across

326   the two GBS analyses to this SE Asian NLDT group through two WGS samples (Dr. Grinspoon

327   and Somali Taxi Cab, Figure 2, Supplementary Figures 3) that were included in both sets of GBS

328   analyses. Although only 45 SNPs overlapped between both types of GBS data and the WGS

329   data, a phylogeny of this limited alignment also supports the existence of an additional distinct

330   SE Asian NLDT clade (Supplementary Figure 4). Collectively these analyses lend support to a

331   total lower bound of four *Cannabis* populations, although clearly more extensive sampling with

332   consistent sequencing is required to fully access standing biogeographic diversity.

333

334   **Tests of Tree Models.** To test hypotheses of tree-like evolution for the three genetic groups, we

335   first applied the three-population test for admixture[64], and found no evidence for admixture in

336   any of the pairwise comparisons (positive f statistic values). Next we constructed maximum

337   likelihood trees based on the aggregate SNP frequencies for the three genetic groups and

338   simulated a variety of 'migration' events (0-10), but no simulation produced non-zero migration

339   graph edges (Figure 3). F_{ST} analysis shows little divergence among lineages for most loci, but a

340   substantial number of highly-divergent regions are unique to each clade (Figure 4). This

341   reinforces the importance of using many, high-quality, single-copy regions of the genome, rather

342   than smaller numbers of loci that could lead to less resolution or even misleading results.

343   Although lore[60], Figure 2 and Supplementary Figure 2 strongly suggest at least some individuals

344   have hybrid origins, these tree models for the overall SNP frequencies of the population groups

345   inferred by FLOCK (Table 2) imply each group contains strong genetic signals from ancestral

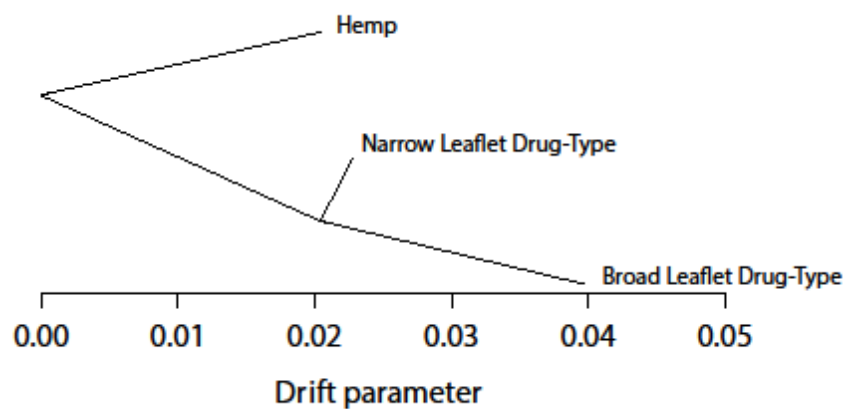346   biogeographic gene pools.

347

348

17

349

350

351

Figure 3. Maximum likelihood tree of three *Cannabis* populations. We found no evidence for extensive admixture or deviations from this tree model.
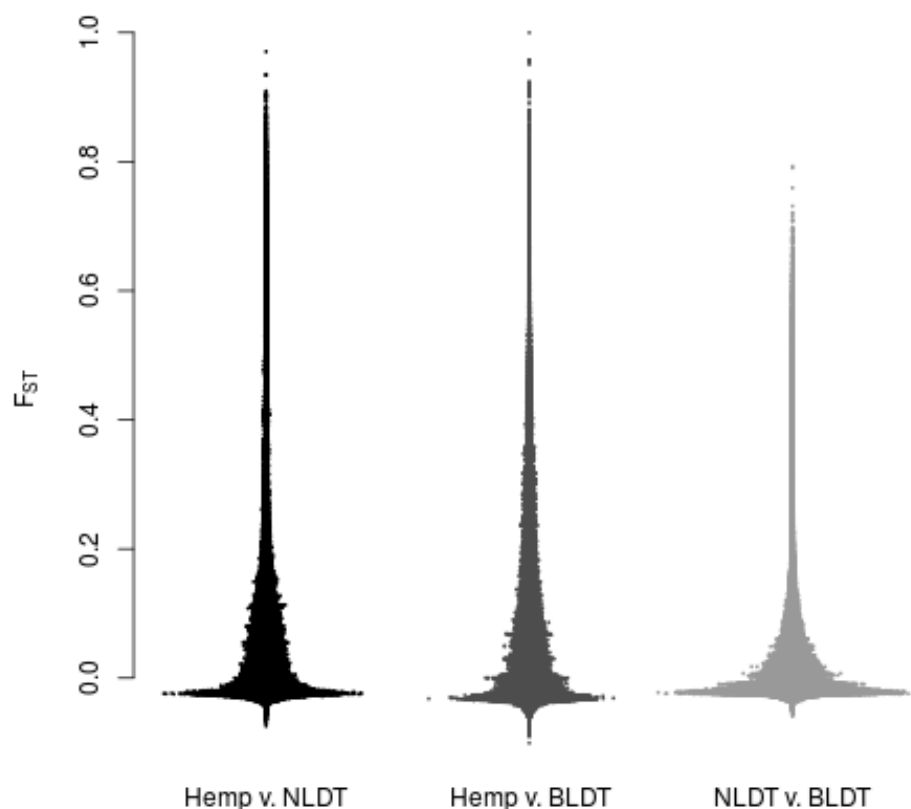
354

18

Figure 4. Distribution of Weir-Cockerham F_{ST} estimates for each population comparison. Each population pair has some portion of segregated sites.

Additional *Cannabis* diversity likely remains to be sampled. Notably absent from all

genome sequence datasets published to date are putative *C. ruderalis*[65] samples. These are short

weedy plants, with free shattering inflorescences found widely from northern Siberia, through

central Asia and into Eastern Europe[66]. Whether these populations represent ancestral, pre-

domesticated wild *Cannabis*, more recent feral escapes or some combination of both remains

unclear. Even though we were not able to sample putative *C. ruderalis* populations, Finola is an

early maturing seed hemp strain from Finland with purported northern Russian landrace

ancestry[60], and Low Ryder and Auto AK-47 are auto-flowering drug-type strains with possible

367 *C. ruderalis* heritage included in our samples (Table 2). Our analyses found Finola fits within the

368 hemp group while Low Ryder and Auto AK-47 are close relatives of each other within the

369 NLDT group (Supplementary Figure 3). Further genomic analyses are required to determine the

370 extent to which *C. ruderalis* populations are genetically distinct from hemp and drug-type

371 groups, and whether they may in fact harbor an ancestral wild-type gene pool from which

372 European hemp varieties were domseticated[5,16].

373 Broad leaflet Asian hemp is also underrepresented, although we included one putative

374 Chinese hemp sample that occupies an area between the core hemp and BLDT populations

375 (Table 2, Figure 2 and Supplementary Figure 2). Hillig's[5] analysis of alloenzymes concluded that

376 Asian hemp strains were more similar to Asian drug-type strains than they were to narrow leaflet

377 European hemp. Likewise, Gao et al.[67] found genetic dissimilarity between European hemp and

378 Chinese hemp, using microsatellites, and showed at least several distinct groups of hemp occur

379 across the vast geography of Asia. Overall, Asian and European hemp strains appear dissimilar

380 genetically, possibly reflecting independent domestication events[66].

381 One major complication obscuring the understanding of *Cannabis* diversity and history is

382 the lack of information about the native range or ranges of *Cannabis*. In addition to divergent

383 breeding efforts and human-vectored transport of seeds, the tendency of *Cannabis* is to escape

384 into feral populations wherever human cultivation occurs in temperate climates[68]. This, coupled

385 with wind pollination biology and no known reproductive barriers, makes the existence of pure

386 wild native *Cannabis* populations unlikely. The weedy tendencies of *Cannabis* are exemplified

387 by the mid-western USA populations of feral hemp that flourish despite the eradication efforts by

388 the Drug Enforcement Agency, which have for decades totaled millions of plants removed per

389 year. A comprehensive evaluation of *Cannabis* diversity, which includes feral and wild Eurasian

390   populations, is required to ascertain if the levels of divergence and gene flow are consistent with

391   one or more origins of domestication[5]. Even if these extant populations are highly admixed with

392   modern varieties, their study promises to offer insight into *Cannabis* ecology and evolution,

393   given how different the selective regime of the feral setting is compared to that of agricultural

394   fields. Considering the similar debates regarding the timing and origins of *Oryza* domestication

395   that remain as of yet unresolved[69], *Cannabis* requires substantially more work to unravel its

396   complicated relationship with humans.

397   'Indica' and 'sativa' are commonly used terms ascribed to plants that have certain

398   characteristics, often related to leaflet morphology and the perceived effects of consuming the

399   plant[8]. However these names are rooted in taxonomic traditions dating to Linnaeus who first

400   classified the genus as monotypic *(Cannabis sativa)* based on hemp specimens from Virginia and

401   Europe[70]. Lamarck subsequently designated *Cannabis indica* to accommodate the shorter stature

402   potent narrow leaflet drug-type plants from the Indian subcontinent[71]. Although currently the

403   term 'indica' is typically used to refer to BLDTs, this biotype from the Hindu-Kush mountains[14]

404   was not documented until a 1929 survey of Afghani agriculture by Vavilov[72]. This absence of

405   historical documentation until the 20th century, a very narrow geographic range, and some

406   evidence for a broader NLDT gene pool (Table 1, Supplementary Figures 3 and 4), suggest a

407   separate and more recent origin of the BLDT clade. This origin could represent a domestication

408   event of a wild or feral BLDT population, or perhaps hybridization events between NLDT and

409   BLDT populations. Final resolution of *Cannabis* taxonomy will require complete assessment of

410   standing global genetic diversity and experimental evaluation of reproductive compatibility

411   across all major genetic groups[73], in conjunction with morphological circumscriptions. Given the

412   current absence of evidence for reproductive barriers, and overall limited genetic distances

413   between hemp and drug-type strains analyzed in this study we suggest continued monotypic

414   treatment of plants in this genus as *Cannabis sativa* L. is warranted.

415

416   **Cannabinoid and Terpenoid Diversity.** THCA and CBDA are the most abundant cannabinoids

417   produced by the majority of strains on the North American market today (Figure 5a), and both

418   compounds show an impressive range of medicinal potential[33,74], although endocannabinoid-

419   based therapy trials have a history of significant rates of study withdraws and adverse effects[75].

420   Historical breeding efforts have resulted in mostly high THCA plants that produce strong

421   intoxicating effects when consumed, and that synthesize only very low levels of alternative

422   cannabinoids (Figure 5b). High CBDA plants have only recently become more available in North

423   America over the last several years in response to demand. Interestingly, these high CBDA-

424   producing plants form several clusters within the both the NLDT and BLDT groups, as well as

425   within the hemp group (Supplementary Table 1), but rarely reach equivalent quantities total

426   cannabinoid production as those found in high THCA plants (Figure 4a). The minor

427   cannabinoids that are commonly assayed, CBGA, CBCA, THCVA and CBDVA are also of

428   interest, despite strains producing high levels of these compounds being largely unavailable for

429   research currently[76]. With at least 74 cannabionoids identified in *Cannabis*, modernized genetic

430   and breeding techniques are required to diversify and optimize *Cannabis* varieties. Efforts should

431   also be made to document and preserve feral, wild and heirloom populations that can serve as

432   reservoirs of cultural and genetic diversity.

433   Aromatic terpenoids impart many of the characteristic fragrances to *Cannabis*, and

434   possibly contribute to the effects of consumption[2]. Terpenoids are synthesized in many plant

435   species, and play a role in relieving various abiotic and biotic stresses through direct and indirect

22

436   mechanisms[2]. Our analysis of strains sharing common genetic groups shows that each group has

437   a distinct terpenoid profile (Figure 5c and Supplementary Figure 5). We found NLDTs to contain

438   significantly more β-myrcene and α-terpinolene than BLDTs, although interestingly the two

439   hemp strains for which we analyzed chemical data for had significantly more β-myrcene than

440   either drug-type group (Figure 5c). Similarly Hillig[77] found NLDTs to yield significantly more

441   β-myrcene than Afghani BLDTs, yet European hemp and un-cultivated accessions labeled as *C.*

442   *ruderalis* contained the highest levels. Hillig also reported that Afghani BLDTs contained the

443   highest levels of guaiol and eudesmol isomers, which we did not measure, although we found

444   BLDTs contained more linalool than NLDTs or hemp. Understanding the ecological functions

445   and evolutionary origins of terpenoids and cannabinoids in *Cannabis* could improve therapeutic

446   potential, and possibly reduce the need for pesticide application during cultivation.
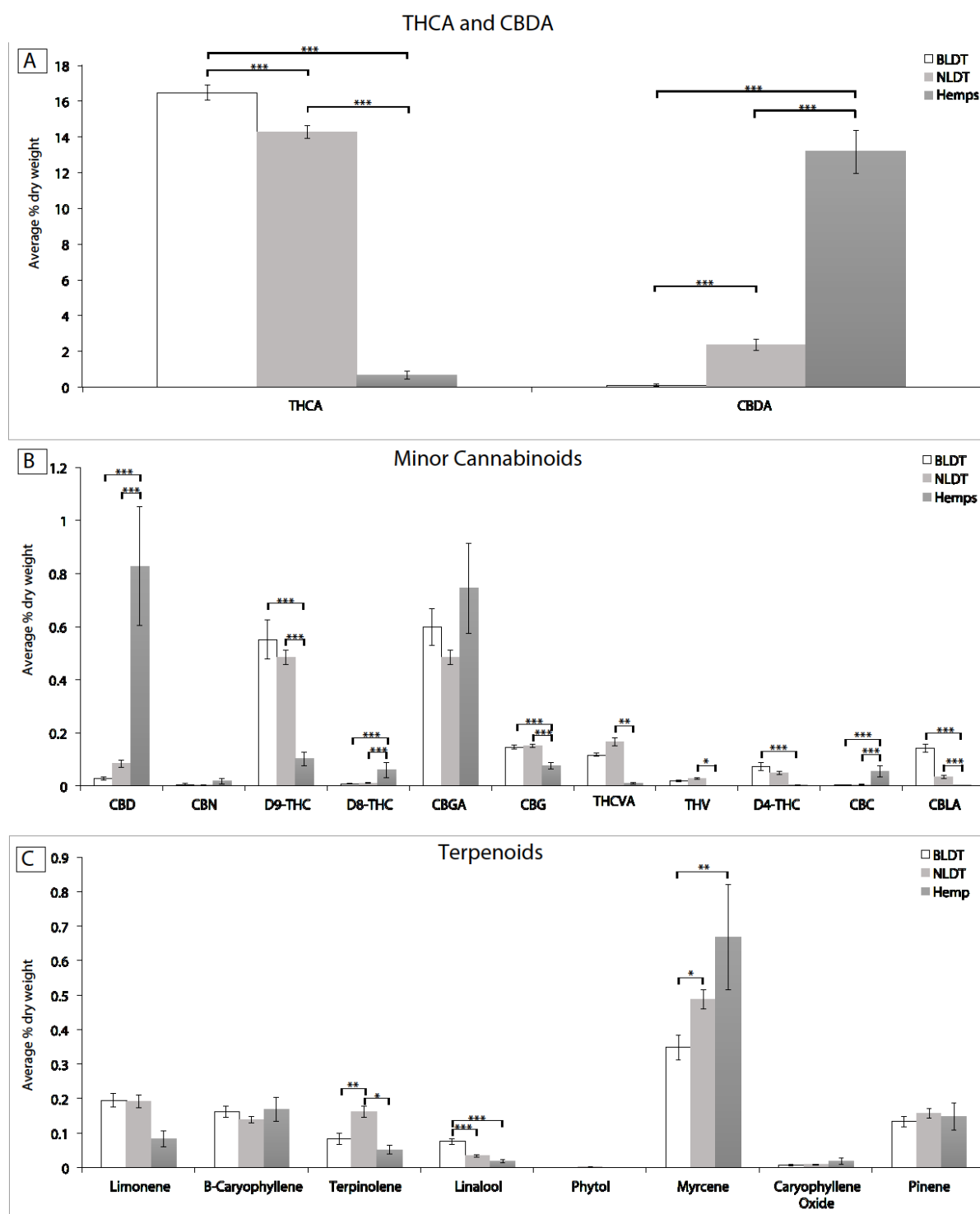
Figure 5. Average percentage of mass for dried and un-pollenated female flowers of *Cannabis* genetic groups. (a) THCA and CBDA cannabinoids (b) Minor cannabinoids (c) Terpenoids. THCA = delta-9-tetrahydrocannabinolic acid. CBDA = cannabidiolic acid. CBD = cannabidiol.

452   CBN = cannabinol. D9-THC = delta-9-tetrahydrocannabinol. D8-THC = delta-8-

453   tetrahydrocannabinol. CBGA = cannabigerolic acid. CBG = cannabigerol. THCVA =

454   Tetrahydrocannabivarin carboxylic acid. THCV = Tetrahydrocannabivarin. D4-THC = delta-4-

455   tetrahydrocannabinol. CBC = cannabichromene. CBLA = cannabicyclolic acid.

456

457   **Conclusions.** *Cannabis* genomics offers a window into the past, but also a road forward.

458   Although historical and clandestine breeding efforts have been clearly successful in many

459   regards[21,31], *Cannabis* lags decades behind other major crop species in many other respects.

460   Developing stable *Cannabis* lines capable of producing the full range of potentially therapeutic

461   cannabinoids is important for the research and medical communities, which currently lacks

462   access to diverse high-quality material in the USA[78].

463   In this paper we extended the initial *Cannabis* genome study[10], by re-mapping WGS and

464   GBS sequence reads to the existing PK draft scaffolds, to understand diversity and evolutionary

465   relationships among the major lineages. Although hybridization of cultivated varieties[60] and

466   human transport of seeds across the globe was hypothesized to have obscured much of the

467   ancestral genetic signal[13], we found significant evidence for apparent ancestral signals in

468   genomic data derived largely from modern cultivated varieties (Table 2, Figures 2 and 3). Re-

469   analysis of previously published GBS data[47] provides additional limited evidence for a fourth

470   group (Supplementary Figures 4 and 5). Interestingly, unique cannabinoid and terpenoid profiles

471   were associated with three of the genetic groups, lending support to their validity, despite the

472   limitations of our sampling scheme. Overall, we hope the publicly available data and analyses

473   from this study will facilitate the continued research on the history of this controversial plant and

474   the development of the agricultural and therapeutic potential of *Cannabis*.

475

495   **References**

496   1.   Li, H. L. An archaeological and historical account of cannabis in China. *Econ. Bot.* **28,** 437–448
497   (1973).

498   2.   Russo, E. B. Taming THC: Potential cannabis synergy and phytocannabinoid-terpenoid entourage
499   effects. *Br. J. Pharmacol.* **163,** 1344–1364 (2011).

500   3.   Russo, E. B. History of cannabis and its preparations in saga, science, and sobriquet. *Chem.*
501   *Biodivers.* **4,** 1614–1648 (2007).

502   4.   Schultes, R. E., Klein, M. W., Plowman, T. & Lockwood, T. Cannabis: an example of taxonomic
503   neglect. *Harvard Univ. Bot. Museum Leafl.* **23,** 337–367 (1974).

504   5.   Hillig, K. W. Genetic evidence for speciation in Cannabis (Cannabaceae). *Genet. Resour. Crop Evol.*
505   **52,** 161–180 (2005).

506   6.   Cherniak, L. *The Great Books of Cannabis vol. I, Book II*. (1982).

507   7.   de Meijer, E. P. M. *et al.* The Inheritance of Chemical Phenotype in Cannabis sativa L. *Genetics*
508   **346,** 335–346 (2003).

509   8.   Habib, R., Finighan, R. & Davenport, S. Testing for Psychoactive Agents. (2013). at
510   <http://liq.wa.gov/publications/Marijuana/BOTEC reports/1c-Testing-for-Psychoactive-Agents-
511   Final.pdf>

512   9.   Small, E. & Cronquist, A. A practical and natural taxonomy for Cannabis. *Taxon* **25,** 405–435
513   (1976).

514   10.   van Bakel, H. *et al.* The draft genome and transcriptome of Cannabis sativa. *Genome Biol.* **12,**
515   R102 (2011).

516   11.   Small, E. Evolution and Classification of Cannabis sativa (Marijuana , Hemp) in Relation to
517   Human Utilization. *Bot. Rev.* **81,** 189–294 (2015).

518   12.   Clarke, R. C. & Merlin, M. D. Letter to the Editor: Small, Ernest. 2015. Evolution and Classification
519   of Cannabis sativa (Marijuana, Hemp) in Relation to Human Utilization. Botanical Review 81(3):
520   189-294. *Bot. Rev.* **81,** 295–305 (2015).

521   13.   Small, E. Response to the Erroneous Critique of my Cannabis Monograph by R. C. Clarke and M.D.
522   Merlin. *Bot. Rev.* **81,** 306–316 (2015).

523   14.   Clarke, R. C. & Merlin, M. D. in *Cannabis Evolution and Ethanobotany* 13–26 (2013).

524   15.   Glanzman, A. Discover Himalaya's Outlawed Marijuana Fields. *Time* (2015). at
525   <http://time.com/3736616/discover-himalayas-illegal-marijuana-fields>

526   16.   Hillig, K. W. & Mahlberg, P. G. A chemotaxonomic analysis of cannabinoid variation in Cannabis
527   (Cannabaceae). *Am. J. Bot.* **91,** 966–975 (2004).

528   17.   Bauer, R., Salo-Ahen, K. & Bauer, O. CB Receptor Ligands from Plants. *Curr. Top. Med. Chem.* **8,**
529   173–186 (2008).

530   18.   Radwan, M. M. *et al.* Isolation and characterization of new cannabis constituents from a high
531   potency variety. *Planta Med.* **74,** 267–272 (2008).

532   19.   ElSohly, M. a. & Slade, D. Chemical constituents of marijuana: The complex mixture of natural
533   cannabinoids. *Life Sci.* **78,** 539–548 (2005).

534   20.   Poklis, J. L., Thompson, C. C., Long, K. a, Lichtman, A. H. & Poklis, A. Disposition of
535   cannabichromene, cannabidiol, and Δ^9-tetrahydrocannabinol and its metabolites in mouse brain
536   following marijuana inhalation determined by high-performance liquid chromatography-tandem
537   mass spectrometry. *J. Anal. Toxicol.* **34,** 516–20 (2010).

538   21.   Mehmedic, Z. *et al.* Potency trends of Δ9-THC and other cannabinoids in confiscated cannabis
539   preparations from 1993 to 2008. *J. Forensic Sci.* **55,** 1209–1217 (2010).

540 22. McPartland, J. M., Matias, I., Di Marzo, V. & Glass, M. Evolutionary origins of the
541   endocannabinoid system. *Gene* **370,** 64–74 (2006).

542 23. Mechoulam, R. & Gaoni, Y. Recent advances in the chemistry of hashish. *Chemie Org. Naturstoffe*
543   **25,** 175–213 (1967).

544 24. Volkow, N. D., Baler, R. D., Compton, W. M. & Weiss, S. R. B. Adverse Health Effects of Marijuana
545   Use. *N. Engl. J. Med.* **370,** 2219–2227 (2014).

546 25. Berry, E. M. & Mechoulam, R. Tetrahydrocannabinol and endocannabinoids in feeding and
547   appetite. *Pharmacol. Ther.* **95,** 185–190 (2002).

548 26. Zogopoulos, P., Vasileiou, I., Patsouris, E. & Theocharis, S. E. The role of endocannabinoids in pain
549   modulation. *Fundam. Clin. Pharmacol.* **27,** 64–80 (2013).

550 27. Tramèr, M. R. *et al.* Cannabinoids for control of chemotherapy induced nausea and vomiting:
551   quantitative systematic review. *BMJ* **323,** 16–21 (2001).

552 28. Di Marzo, V., Bifulco, M. & De Petrocellis, L. The endocannabinoid system and its therapeutic
553   exploitation. *Nat. Rev. Drug Discov.* **3,** 771–784 (2004).

554 29. Pacher, P. & Mechoulam, R. Is lipid signaling through cannabinoid 2 receptors part of a protective
555   system? *Prog. Lipid Res.* **50,** 193–211 (2011).

556 30. De Petrocellis, L. *et al.* Effects of cannabinoids and cannabinoid-enriched *Cannabis* extracts on
557   TRP channels and endocannabinoid metabolic enzymes. *Br. J. Pharmacol.* **163,** 1479–1494 (2011).

558 31. Swift, W., Wong, A., Li, K. M., Arnold, J. C. & McGregor, I. S. Analysis of Cannabis Seizures in NSW,
559   Australia: Cannabis Potency and Cannabinoid Profile. *PLoS One* **8,** 1–9 (2013).

560 32. Rustichelli, C., Ferioli, V., Vezzalini, F., Rossi, M. C. & Gamberini, G. Simultaneous separation and
561   identification of hashish constituents by coupled liquid chromatography-mass spectrometry
562   (HPLC-MS). *Chromatographia* **43,** 129–134 (1996).

563 33. Devinsky, O. *et al.* Cannabidiol: Pharmacology and potential therapeutic role in epilepsy and
564   other neuropsychiatric disorders. *Epilepsia* **55,** 791–802 (2014).

565 34. Pertwee, R. G. The diverse CB1 and CB2 receptor pharmacology of three plant cannabinoids:
566   delta9-tetrahydrocannabinol, cannabidiol and delta9-tetrahydrocannabivarin. *Br. J. Pharmacol.*
567   **153,** 199–215 (2008).

568 35. Fellermeier, M., Eisenreich, W., Bacher, A. & Zenk, M. H. Biosynthesis of cannabinoids. *Eur. J.*
569   *Biochem.* **268,** 1596–1604 (2001).

570 36. Staginnus, C., Zörntlein, S. & de Meijer, E. A PCR marker linked to a THCA synthase polymorphism
571   is a reliable tool to discriminate potentially THC-rich plants of Cannabis sativa L. *J. Forensic Sci.*
572   **59,** 919–26 (2014).

573 37. Weiblen, G. D. *et al.* Gene duplication and divergence affecting drug content in Cannabis sativa.
574   *New Phytol.* **208,** 2141–1250 (2015).

575 38. Onofri, C., de Meijer, E. P. M. & Mandolino, G. Sequence heterogeneity of cannabidiolic- and
576   tetrahydrocannabinolic acid-synthase in Cannabis sativa L. and its relationship with chemical
577   phenotype. *Phytochemistry* **116,** 57–68 (2015).

578 39. Borrelli, F. *et al.* Colon carcinogenesis is inhibited by the TRPM8 antagonist cannabigerol , a

579   Cannabis - derived non-psychotropic cannabinoid. *Carcinogenesis* **35,** 2787–2797 (2014).

580   40.   Izzo, A. a. *et al.* Inhibitory effect of cannabichromene, a major non-psychotropic cannabinoid
581   extracted from Cannabis sativa, on inflammation-induced hypermotility in mice. *Br. J. Pharmacol.*
582   **166,** 1444–1460 (2012).

583   41.   Mcpartland, J. M., Duncan, M., Marzo, V. Di & Pertwee, R. G. Are cannabidiol and Δ 9 -
584   tetrahydrocannabivarin negative modulators of the endocannabinoid system ? A systematic
585   review. 737–753 (2015). doi:10.1111/bph.12944

586   42.   de Meijer, E. P. M., Hammond, K. M. & Sutton, a. The inheritance of chemical phenotype in
587   Cannabis sativa L. (IV): cannabinoid-free plants. *Euphytica* **168,** 95–112 (2009).

588   43.   de Meijer, E. P. M., Hammond, K. M. & Micheler, M. The inheritance of chemical phenotype in
589   Cannabis sativa L. (III): variation in cannabichromene proportion. *Euphytica* **165,** 293–311 (2008).

590   44.   De Meijer, E. P. M., Hammond, K. M. & Sutton, a. The inheritance of chemical phenotype in
591   Cannabis sativa L. (IV): Cannabinoid-free plants. *Euphytica* **168,** 95–112 (2009).

592   45.   Hazekamp, a. & Fischedick, J. T. Cannabis - from cultivar to chemovar. *Drug Test. Anal.* 660–667
593   (2012). doi:10.1002/dta.407

594   46.   Sakamoto, K., Akiyama, Y., Fukui, K., Kamada, H. & Satoh, S. Characterization; Genome Sizes and
595   Morphology of Sex Chromosomes in Hemp (Cannabis sativa L.). *Cytologia (Tokyo).* **63,** 459–464
596   (1998).

597   47.   Sawler, J. *et al.* The Genetic Structure of Marijuana and Hemp. *PLoS One* 1–9 (2015).
598   doi:10.1371/journal.pone.0133292

599   48.   Parchman, T. L., Gompert, Z., Mudge, J. & Schilkey, F. D. Genome-wide association genetics of an
600   adaptive trait in lodgepole pine. *Mol. Ecol.* **21,** 2991–3005 (2012).

601   49.   Bolger, A. M., Lohse, M. & Usadel, B. Genome analysis Trimmomatic : a flexible trimmer for
602   Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

603   50.   Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform.
604   *Bioinformatics* **25,** 1754–1760 (2009).

605   51.   Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079
606   (2009).

607   52.   Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–8 (2011).

608   53.   Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage
609   Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

610   54.   Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol.*
611   *Evol.* **23,** 254–267 (2006).

612   55.   Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary
613   Genetics Analysis Version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).

614   56.   Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: Variational Inference of Population
615   Structure in Large SNP Data Sets. *Genetics* **197,** 573–589 (2014).

616   57.   Duchesne, P. & Turgeon, J. FLOCK Provides Reliable Solutions to the 'Number of Populations'

617   Problem. *J. Hered.* **103,** 734–743 (2012).

618   58.   Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide
619   Allele Frequency Data. *PLoS Genet.* **8,** e1002967 (2012).

620   59.   Hazekamp, A., Peltenburg, A., Verpoorte, R. & Giroud, C. Chromatographic and Spectroscopic
621   Data of Cannabinoids from Cannabis sativa L. *J. Liq. Chromatogr. Relat. Technol.* **28,** 2361–2382
622   (2005).

623   60.   Clarke, R. C. & Merlin, M. D. in *Cannabis Evolution and Ethanobotany* 295–309 (2013).

624   61.   Lynch, R. C. Genomics of Adaptation and Diversification. (2015).

625   62.   Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus
626   genotype data. *Genetics* **155,** 945–959 (2000).

627   63.   Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the
628   software STRUCTURE: A simulation study. *Mol. Ecol.* **14,** 2611–2620 (2005).

629   64.   Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population
630   history. *Nature* **461,** 489–494 (2009).

631   65.   Janischevsky, D. E. Cannabis Ruderalis. *Proc. Saratov* **2,** 14–15 (1924).

632   66.   Clarke, R. C. & Merlin, M. D. in *Cannabis Evolution and Ethanobotany* 311–331 (2013).

633   67.   Gao, C. *et al.* Diversity Analysis in Cannabis sativa Based on Large- Scale Development of
634   Expressed Sequence Tag-Derived Simple Sequence Repeat Markers. **9,** (2014).

635   68.   Small, E., Pocock, T. & Cavers, P. The biology of Canadian weeds. 119. Cannabis sativa L. *Can. J.*
636   *plant Sci.* (2003). at <http://pubs.aic.ca/doi/abs/10.4141/P02-021>

637   69.   Gross, B. L. & Zhao, Z. Archaeological and genetic insights into the origins of domesticated rice.
638   *Proc. Natl. Acad. Sci. U. S. A.* **111,** 6190–7 (2014).

639   70.   Linnaeus, C. *Species Plantarum*. (1753).

640   71.   Lamarck, J. B. Encyclopédie méthodique: botanique. (1783).

641   72.   Vavilov, N. I. & Bukinich, D. D. Agricultural Afganistan. *Bull. Appl. Bot. Genet. Plant Breed.* **33,**
642   (1929).

643   73.   Rieseberg, L. H. & Willis, J. H. Plant speciation. *Science* **317,** 910–914 (2007).

644   74.   Di Marzo, V., Bifulco, M. & De Petrocellis, L. The endocannabinoid system and its therapeutic
645   exploitation. *Nat. Rev. Drug Discov.* **3,** 771–784 (2004).

646   75.   Wade, D. T., Makela, P. M., House, H., Bateman, C. & Robson, P. Long-term use of a cannabis-
647   based medicine in the treatment of spasticity and other symptoms in multiple sclerosis. *Mult.*
648   *Scler.* **12,** 639–645 (2006).

649   76.   Abrams, D. I. *et al.* Vaporization as a smokeless cannabis delivery system: a pilot study. *Clin.*
650   *Pharmacol. Ther.* **82,** 572–578 (2007).

651   77.   Hillig, K. W. A chemotaxonomic analysis of terpenoid variation in Cannabis. *Biochem. Syst. Ecol.*
652   **32,** 875–891 (2004).

653   78.   Nutt, D. J., King, L. a & Nichols, D. E. Treatment Innovation. **14,** 577–585 (2013).

654