Taylor & Francis
Taylor & Francis Group

Critical Reviews in Plant Sciences

ISSN: 0735-2689 (Print) 1549-7836 (Online) Journal homepage: http://www.tandfonline.com/loi/bpts20

Genomic and Chemical Diversity in Cannabis

Ryan C. Lynch, Daniela Vergara, Silas Tittes, Kristin White, C. J. Schwartz, Matthew J. Gibbs, Travis C. Ruthenburg, Kymron deCesare, Donald P. Land & Nolan C. Kane

To cite this article: Ryan C. Lynch, Daniela Vergara, Silas Tittes, Kristin White, C. J. Schwartz, Matthew J. Gibbs, Travis C. Ruthenburg, Kymron deCesare, Donald P. Land & Nolan C. Kane (2016) Genomic and Chemical Diversity in Cannabis, Critical Reviews in Plant Sciences, 35:5-6, 349-363, DOI: 10.1080/07352689.2016.1265363

To link to this article: https://doi.org/10.1080/07352689.2016.1265363

CRITICAL REVIEWS IN PLANT SCIENCES
2016, VOL. 35, NOS. 5–6, 349–363
http://dx.doi.org/10.1080/07352689.2016.1265363

Taylor & Francis
Taylor & Francis Group

Genomic and Chemical Diversity in *Cannabis*

Ryan C. Lynch[a,e], Daniela Vergara[a], Silas Tittes[a], Kristin White[a], C. J. Schwartz[b], Matthew J. Gibbs[b], Travis C. Ruthenburg[c,d], Kymron deCesare[c], Donald P. Land[c], and Nolan C. Kane[a]

[a]Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, Colorado, USA; [b]Marigene Inc, Fort Collins, Colorado, USA; [c]Steep Hill Labs Inc., Berkeley, California, USA; [d]SC Laboratories Inc., Seattle, Washington, USA; [e]Medicinal Genomics Corporation, Woburn, Massachusetts, USA

ABSTRACT

Plants of the *Cannabis* genus are the only prolific producers of phytocannabinoids, compounds that strongly interact with the evolutionarily ancient endocannabinoid receptors shared by most bilaterian taxa. For millennia, the plant has been cultivated not only 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 340 *Cannabis* varieties, including fiber and seed oil hemp, high cannabinoid drug-types, and feral populations. These analyses demonstrate the existence of at least three major groups of diversity with European hemp varieties more closely related to narrow leaflet drug-types (NLDTs) than to broad leaflet drug-types (BLDTs). The BLDT group appears to encompass less diversity than the NLDT, which reflects the larger geographic range of NLDTs, and suggests a more recent origin of domestication of the BLDTs. As well as being genetically distinct, hemp, NLDT, and BLDT genetic groups produce 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 industry.

KEYWORDS

Cannabinoids; *Cannabis*; hemp; marijuana; phylogeny; population structure; terpenoids

I. Introduction

Plants of the genus *Cannabis* (Cannabaceae; hemp, drug-type) have been used for thousands of years for fiber, nutritional seed oil, and medicinal or psychoactive effects. Archeological evidence for hemp fiber textile production in China dates to at least as early as 6000 years ago (Li, 1973), but possibly as early as 12,000 years ago (Russo, 2011), suggesting that *Cannabis* was one of the first domesticated fiber plants. Archeological evidence for medicinal or shamanistic use of *Cannabis* has been found at Indian, central Asian, and Middle-Eastern sites (Russo, 2007), further illustrating the widespread extent of *Cannabis* utilization throughout the human history. A central Asian site of domestication is often cited (Schultes *et al.*, 1974), although, genetic analyses suggest that two independent domestication events may have occurred separately (Hillig, 2005).

Cannabis plants are usually annual wind-pollinated dioecious herbs, though individuals may live more than a year in sub-tropical climates (Cherniak, 1982) and monoecious populations exist (de Meijer *et al.*, 2003). The taxonomic composition of the genus remains unresolved with two species (*Cannabis indica* and *Cannabis sativa*) commonly cited (Hillig, 2005); although *Cannabis ruderalis* is sometimes proposed as a third species that contains northern short-day or auto-flowering plants (Small and Cronquist, 1976). Monospecific treatment of the genus as *C. sativa* L. is also common (van Bakel *et al.*, 2011) and various alternative nomenclature schemes (e.g. *C. sativa* subsp. *indica* var. *kafiristanica*) are sometimes cited (Schultes *et al.*, 1974). Even though an extensive monograph on the genus has recently been published (Small, 2015a), limited genetic and

CONTACT Ryan C. Lynch rlynch@colorado.edu Department of Ecology and Evolutionary Biology, University of Colorado Boulder, 1900 Pleasant Street, 334 UCB, Boulder, CO 80309-0334, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/bpts.

Supplemental materials for this article can be accessed on the publisher's website.

experimental data leave the questions of taxonomy unresolved (Clarke and Merlin, 2015; Small, 2015b).

The geographical and ecological range of *Cannabis* is unusually broad, with cultivated populations growing outdoors on every continent, except Antarctica, in a wide range of environments from sub-arctic to temperate to tropical, and from sea level to over 3000 m elevation (Clarke and Merlin, 2013; Glanzman, 2015). Feral or wild populations are also found as far north as the edge of the Arctic Circle in Eurasia, but they are most common in well-drained soils of temperate continental ecosystems in Eurasia and North America, while tropical populations are absent or rare (Clarke and Merlin, 2013). The species contains extensive phytochemical diversity, particularly in cannabinoid and terpenoid profiles (Hillig and Mahlberg, 2004; Hillig, 2005), and it also shows extensive diversity of morphological and life-history characteristics, further fueling debate regarding the taxonomic status and origins of *Cannabis* domestication.

One distinctive feature of the *Cannabis* genus is the production of a tremendous diversity of compounds called *cannabinoids*, they are so named because they are not produced in high levels in any other plant species (Bauer *et al.*, 2008). Cannabinoids are a group of at least 74 known C21 terpenophenolic compounds (ElSohly and Slade, 2005; Radwan *et al.*, 2008) responsible for many reported medicinal and psychoactive effects of *Cannabis* consumption (Poklis *et al.*, 2010). Some estimates for the total number of phytocannabinoids range to well over a hundred (Mehmedic *et al.*, 2010), though this number includes breakdown products as well as compounds found at extremely low levels. The plants produce a non-psychoactive carboxylic acid form of these compounds, which requires heating to convert cannabinoids into the psychoactive decarboxylated forms. Interestingly, these compounds have pronounced neurological effects on a wide range of vertebrate and invertebrate taxa suggesting an ancient origin of the endocannabinoid receptors, perhaps as old as the last common ancestor of all extant bilaterians over 500 MYA (McPartland *et al.*, 2006). The plant compounds thus produced have the potential to affect a broad range of metazoans, though their ecological functions in nature are not well understood. Indeed, the suggested roles for these compounds include many biotic and abiotic defenses, such as suppression of pathogens and herbivores, protection from UV radiation damage, and attraction of seed dispersers. These hypotheses about the selective benefits of cannabinoid production remain speculative, as none has been conclusively verified to date. We know more, however, about the evolutionary forces during cultivation and domestication.

In particular, high delta-9-tetrahydrocannabinolic acid (THCA) content has been selected for (Mechoulam and Gaoni, 1967). When heated, THCA is converted to delta-9-tetrahydrocannabinol (THC), which has potent psychoactive (Volkow *et al.*, 2014), appetite-stimulating (Berry and Mechoulam, 2002), analgesic (Zogopoulos *et al.*, 2013) and antiemetic (Tramèr *et al.*, 2001) effects. These effects are mediated through interactions with human endocannabinoid CB1 receptors found in the brain (Di Marzo *et al.*, 2004), and CB2 receptors, which are concentrated in peripheral tissues (Pacher and Mechoulam, 2011). Other THC receptor binding locations are hypothesized as well (De Petrocellis *et al.*, 2011). After several decades of accelerated clandestine cultivation technique and breeding improvements, some modern lines can now yield dried unpollinated pistillate inflorescence material that contains over 30% THCA by dry weight (Swift *et al.*, 2013). However, other cannabinoids may also be present in high concentrations. In particular, high cannabidiolic acid (CBDA) plants are used in some hashish preparations (Rustichelli *et al.*, 1996; Hanuš *et al.*, 2016) and are presently in high demand as an antiseizure therapy (Devinsky *et al.*, 2014). In contrast with THC, which acts as a partial agonist of the CB1 and CB2 receptors, CBD does not have strong psychoactive properties as THC, but instead it has antagonist activity on agonists of the CB1 and CB2 receptors (Pertwee, 2008). Thus, the two most abundant cannabinoids produced in *Cannabis* have, to some degree, opposing neurological effects.

THCA and CBDA are alternative products of a shared precursor, cannabigerolic acid (CBGA) (Fellermeier *et al.*, 2001). A single locus with co-dominant alleles was proposed to explain patterns of inheritance for THCA to CBDA ratios (de Meijer *et al.*, 2003; Staginnus *et al.*, 2014). However more recent quantitative trait loci (QTL) mapping experiments (Weiblen *et al.*, 2015), expression studies (Onofri *et al.*, 2015), and genomic analyses (van Bakel *et al.*, 2011) paint a more complex scenario with several linked paralogs responsible for the various THCA and CBDA phenotypes. Other cannabinoids such as cannabigerol (CBG) (Borrelli *et al.*, 2014), cannabichromene (CBC) (Izzo *et al.*, 2012), and delta-9-tetrahydrocannabivarin (THCV) (Mcpartland *et al.*, 2015) demonstrate pharmacological promise, and can also be produced at high levels by the plant (de Meijer and Hammond, 2005; de Meijer and Hammond, 2016; de Meijer *et al.*, 2008). Additionally, *Cannabis* secondary metabolites such as terpenoids and flavonoids likely contribute to therapeutic or psychoactive effects (Russo, 2011). For example, β-myrcene, humulene, and linalool are proposed to produce sedative effects associated with specific varieties (Hazekamp and Fischedick, 2012).

In this study, plants that produce low levels of total cannabinoids are herein referred to as hemp, while high cannabinoid producing varietals are described as drug-

type plants. Legal definitions often use a maximum THCA threshold to delineate hemp from drug-types, thus some high CBDA-producing varieties are categorized as hemp. However, this definition ignores the broader traditional usage of hemp for fibers or seed oils and the longstanding presence of CBDA-producing alleles in some drug-type populations (Rustichelli *et al.*, 1996; Hanuš *et al.*, 2016). Additionally, hemp varieties have a distinct set of growth characteristics (Anderson, 1980), with fiber varieties reaching up to 6 m in height during a growing season, exhibiting reduced flower set, increased internodal spacing, and lower total cannabinoid concentration per unit mass compared to drug-type relative. Despite the widespread prohibition of drug-type *Cannabis* cultivation from the 1930s to present (Bonnie and Whitebread, 1970), hemp cultivation and breeding continued in parts of Europe and China through this period, and experienced a brief comeback during the World War II in the USA through the Hemp for Victory campaign. Studies to date have found that hemp varieties are genetically distinct from drug-type varieties (van Bakel *et al.*, 2011), though, interestingly, Hillig (2005) found broad leaflet Southeastern (SE) Asian hemp landraces to be more closely related to Asian drug-type varieties than to European hemp varieties.

Cannabis has a diploid genome ($2n = 20$), and an XY/XX chromosomal sex-determining system (Divashuk *et al.*, 2014). The genome size is estimated to be 818 Mb for female plants and 843 Mb for male plants (Sakamoto *et al.*, 1998). Currently, a draft genome consisting of 60,029 scaffolds is available for the Purple Kush (PK) drug-type variety from the National Center for Biotechnology Information (NCBI). Additional whole-genome data are available from the NCBI for the Finola and USO31 hemp varieties. Various reduced representation genome, gene, and RNA sequence data are also available from the NCBI. Presently *Cannabis* is the only multi-billion-dollar crop without a sequence-based genetic linkage or physical genome map. Indeed, the first genetic map for the species was only recently published, providing for the first time, quantitative trait mapping of cannabinoid content and other traits (Weiblen *et al.*, 2015).

Initial studies of *Cannabis* genetic diversity examined either many samples with few molecular markers (Hillig, 2005) or whole genome wide data for relatively few samples types (van Bakel *et al.*, 2011). Sawler *et al.* (2015) recently published a survey of *Cannabis* genomic diversity, using a reduced genomic representation strategy to evaluate 81 marijuana (drug-type) and 43 hemp varieties. The aim of this present study is to assess the genomic diversity, and phylogenetic relationships among 340 total *Cannabis* plants that have distinct phenotypes, and that were described *a priori* by plant breeders as various landraces, indica, sativa, hemp and drug-types, as well as commercially available hemp and drug-types with unclear pedigrees. We have combined data from existing sources and generated new data to create the largest sample set of *Cannabis* genomic sequence data published to date. These data and analyses will continue to facilitate the development of modernized breeding and quality assurance tools, which are lacking in the nascent legal *Cannabis* industry.

II. Materials and methods

A. Sample collection

DNA was obtained from numerous sources, including a variety of breeding and production facilities. The variety names, descriptions, and putative origins used in this article were recorded from the providers of the DNA and sequence data (Table S1). For data, which were not previously published, DNA extractions were performed using the Qiagen DNeasy Plant Mini Kit (Valencia, CA) according to the manufacturer's protocol.

B. Whole-genome shotgun (WGS) sequencing

Fifty-seven samples were sequenced using standard Illumina multiplexed library preparation protocols for two 2× 125 HiSeq 2500 lanes and one 2 × 150 NextSeq 500 run. Sequencing efforts were targeted for approximately 4× to 6× coverage of the *Cannabis* genome per sample.

C. Genotype by sequencing (GBS)

One hundred and eighty-two samples were sequenced on two 1 × 100 HiSeq 2500 lanes, following a multiplexed library preparation protocol described previously (Parchman *et al.*, 2012).

D. Publically available data

We obtained three WGS datasets available from the NCBI (van Bakel *et al.*, 2011) and received seven additional WGS datasets from Medicinal Genomics Corporation (www.medicinalgenomics.com), for a total of 67 WGS genomes. GBS data for 143 samples from Sawler *et al.* (2015) were also included in this study.

E. Sequence processing, alignment, and variant calling

Trimmomatic (Bolger *et al.*, 2014) was used to trim any remaining adaptor sequence from raw FASTQ reads and remove sequences with low-quality regions or ambiguous base calls using the following settings:

ILLUMINACLIP:IlluminaAdapters:2:20:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:5:15 MINLEN:100. Trimmed raw reads from a total of 67 WGS samples were then aligned to the only publically available draft genome of PK (JH226140-JH286168) using the Burrows-Wheeler Alignment tool (BWA mem) (Li and Durbin, 2009). Chloroplast and mitochondrial regions were excluded. We collated the individual alignments to produce a single variant call format (vcf) table for all the samples using samtools mpileup-uf | bcftools view – bvcg (Li *et al.*, 2009). We filtered the VCF table to include only high-quality informative SNP sites using VCFtools (Danecek *et al.*, 2011), bash, and awk with the following VCF parameters: Q (>200), GQ (>10), AF1 (.1 to .9), biallelic sites only and no ambiguous bases. The program PLINK (Purcell *et al.*, 2007) was used to filter out low quality data; we required that individuals have a minimum 50% informative sites and that each sites has data for a minimum of 20% of samples. Finally, we used an estimate of expected coverage for the single-copy portion of the genome, based on the estimated genome size and number of reads being aligned. This was adjusted empirically based on total coverage level (across all WGS samples) per SNP site (Figure S1a) and bounded by a 95% Poisson confidence interval (mean 362× coverage). Further removal of repetitive content was achieved by aligning the PK reference to itself with BLASTN and removing all sites that were within the regions of ≥97% identity for ≥500 bp alignments (Figure S1b). These aforementioned processing, alignment, and SNP calling procedures were then preformed separately on the 182 GBS samples generated for this study and the 143 GBS samples previously published (Sawler *et al.*, 2015), which resulted in three VCF tables and filtered SNP sets. GBS SNPs were additionally required to have a minimum of 5× coverage per sample. Due to limited overlap between the SNP sites produced by the two GBS libraries, most downstream analyses were performed separately for each GBS library along with its corresponding set of WGS SNPs. Code used for these analyses is available at https://github.com/KaneLab.

F. SNP analyses

To visualize genetic relationships, divergence, and ancestral hybridization among lineages, a phylogenetic neighbor network was inferred using simple p-distance calculations (Huson and Bryant, 2006). Heterozygosity counts and multidimensional scaling (MDS) analyses were calculated with PLINK (Purcell *et al.*, 2007). Average within- and between-group genetic distances, and a 45 SNP alignment neighbor-joining tree based on p-distances were calculated with MEGA6 (Tamura *et al.*, 2013). Population structure inferences were made through fastSTRUCTURE (Raj *et al.*, 2014) and FLOCK (Duchesne and Turgeon, 2012). Tests for reticulation within the trees and admixture between populations were performed in TreeMix (Pickrell and Pritchard, 2012), using default parameters and 0–10 migration events. F statistic (FST) estimates were calculated with VCFtools (Danecek *et al.*, 2011).

Our complete analysis includes 67 WGS samples, 182 GBS samples generated by us, and 91 GBS samples published by Sawler and collaborators (2015), for a total of 340 accessions. After filtering and conductingquality control, 195 samples remained divided in the following manner: 62 WGS and 133 GBS (Figure 3 and column 3 Table S1). Of the 62 WGS, 15 fell into the BLDT group, 31 into the NLDT, and 16 into the hemp classification. The 133 GBS samples contained 45 BLDT, 82 NLDT, and 6 hemp members (Figure 3 and Table S1). The BLDT, NLDT, and hemp terms are adapted from Clarke and Merlin (2013) and assigned to the FLOCK determined groups based on shared leaf shape characteristics. Therefore, of the total 340 individuals analyzed in this study, our FLOCK analysis includes 195 of them (Figure 3). Of the 143 GBS samples published by Sawler and collaborators (2015) 91 passed our data filters, but these 91 accessions are not included in our FLOCK analysis due to the lack of SNP overlap.

G. Chemical analyses of genetic groups

The cannabinoid and terpenoid information (chemotype) for a portion of the strains in the genome analysis was generated by Steep Hill Labs (http://steephill.com/). Only strains with genetic data as well as this phenotypic data were analyzed. We used a total of 112 individuals from 17 strains from the BLDT group, 278 individuals from 35 unique strains from the NLDT group, and 33 individuals from two strains of hemp, for a total of 423 individuals in this analysis (Table S1). This chemotype analysis was performed using high performance liquid chromatography (HPLC) with Agilent (1260 Infinity, Santa Clara, CA, USA) and Shimadzu (Prominence HPLC, Columbia, MD, USA) equipment. Between 400 and 600 mg of each sample was extracted into methanol, diluted and analyzed by HPLC using a UV/Vis photodiode array detector. All analytes were detected using a local maximum in the absorption spectrum for each compound. Acid cannabinoids were monitored at wavelengths between 260 and 280 nm, neutral cannabinoids were monitored using wavelengths between 215 and 240 nm, and terpenes were monitored using wavelengths between 200 and 215 nm, except for β-myrcene, which was monitored at 225 nm. A mobile phase consisting of 0.1% formic acid in water and 0.1% formic acid in methanol was used with a gradient starting at 72% methanol and ending at 99% methanol. We used a 1.5 min hold to allow the

removal of larger, more nonpolar compounds to elute, then returning to 72% for 1.5 min to recondition the column for the next run. The total run time was 16 min. The column used was a Phenomenex Kinetex 2.6 μm C18 packing in a 3.1 mm × 100 mm package. Terpenoid standards were purchased from Sigma–Aldrich (St. Louis, MO, USA). Cannabinoid standards were purchased from Cerilliant (Round Rock, TX, USA), RESTEK (Bellefonte, PA, USA), and Lipomed (Cambridge, MA, USA). A C18 column from RESTEK (Raptor ARC-18, Bellefonte, PA, USA) or Phenomenex (Kinetex C18, Torrance, CA, USA) was used. Concentrations of cannabinoids without commercially available standards were estimated using published absorptives (Hazekamp *et al.*, 2005). The chemotype data analyzed for this research include 13 cannabinoids and eight terpenoids. Each compound was quantified using a linear calibration curve calibrated from 1 to 500 ppm with dilution factors adjusted to ensure the presence of the most abundant analytes within this range. Analytes below the calibration range were not reported. Analytes were measured as mass by percentage in samples and not corrected for moisture content.

We performed a one-way analysis of variance (ANOVA) for each cannabinoid and terpenoid separately, with the group (NLDT, BLDT, and hemp) as the predictor variable. We used Bonferroni corrections for multiple comparisons. We also implemented a principal component analysis (PCA) with prcomp function in base R, and car function was used to visualize 95% confidence ellipses for each group (www.R-project.org). Individuals with missing data values for any cannabinoid or terpenoid were removed. After removing the individuals with missing values, we had a total of 351 individuals: 94 BLDT, 229 NLDT, and 28 hemp.

III. Results and discussion

A. Sequencing and SNPs

Summary information and raw sequencing libraries are publically available from the NCBI short read archive (Accession: PRJNA310948). Detailed information about all samples can be found in Table S1 and examples of wide and narrow leaflet forms are shown in Figure 1. Of the 466,427,059 nonambiguous base pairs in the PK reference, 77,810,563 bps were removed due to excess self-similarity (≥97% identity and ≥500 bps length). These parameters were chosen to maximize the removal of duplicated regions in the assembly (Figure S1). After this filter, the total single-copy portion of the PK reference within the combined coverage levels for all 67 WGS samples of 326× to 401×, a 95% Poisson confidence interval around a 362× mean, was 71,236,365 bps (Figure S1). After quality (Q), genotype quality (GQ), allele frequency (AF), missing data, biallelic, and ambiguous base filters, the following SNP counts remained: 491,341 WGS, 2,894 GBS (this study), 4,105 GBS from Sawler and collaborators (2015). Forty five single copy SNPs overlapped both GBS datasets and the WGS samples.

B. Phylogenetic relationships

Bifurcating trees are commonly used to model mutation driven divergence and speciation events. Whole

Figure 1. Example of broad leaflet type (A, R4) and narrow leaflet type (B, Super Lemon Haze) varieties.

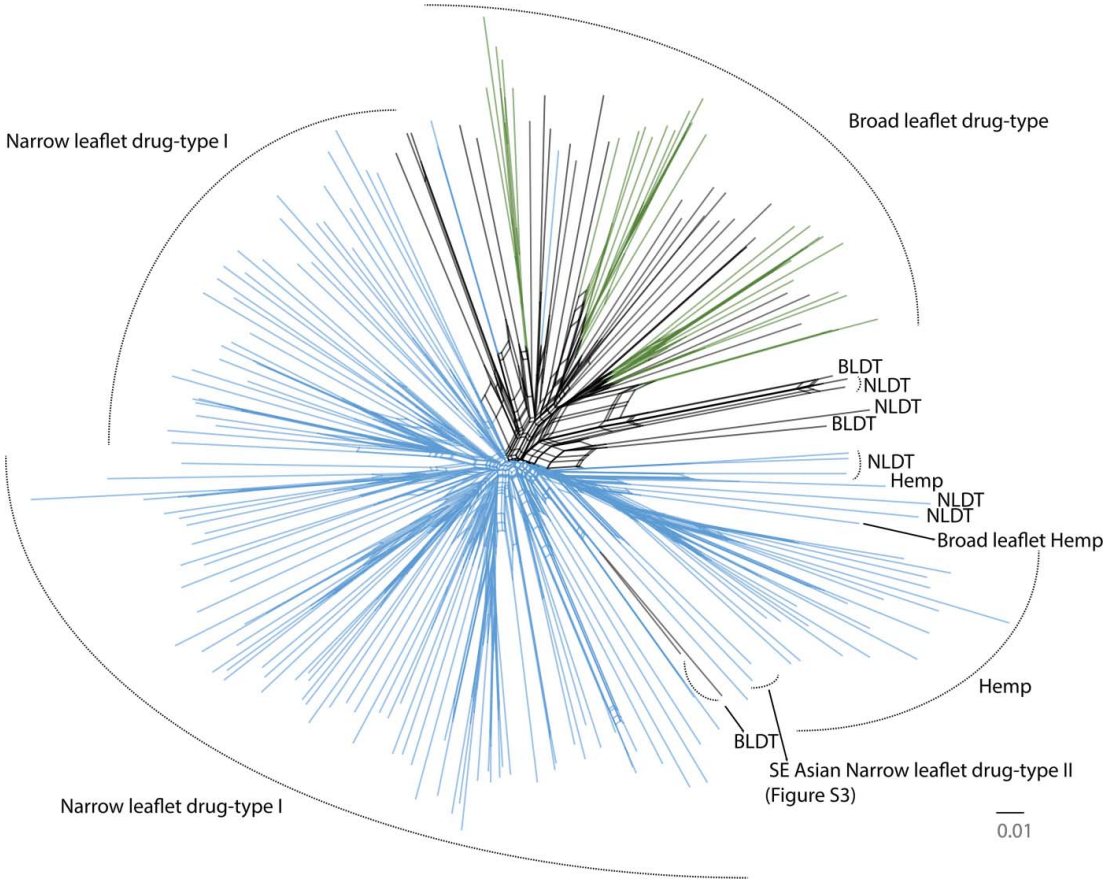

Figure 2. Phylogenetic neighbor network of a 2894 SNP alignment from the single-copy portion of the *Cannabis* genome. Clade names on the periphery were inferred via FLOCK (where $K \geq 3$ was most likely). Colored branches indicate fastSTRUCTURE population membership of ≥70% assignment (where $K = 2$ was most likely). NLDT = Narrow leaflet drug-type (blue clade) and BLDT = Broad leaflet drug-type (green clade). SE Asian NLDT-II refers to Dr. Grinspoon and Somali Taxi Cab samples, which in Figure S3 are part of a distinct SE Asian Narrow leaflet drug-type II group. Broad leaflet hemp points to a Chinese hemp sample. A high-resolution version of this figure that includes each sample name is available in: https://figshare.com/articles/Cannabis_Tree/1585470/4.

genome wide sequence datasets include information about recombination, hybridization, and gene loss, or genesis events, some of which may be incongruent with one another (Huson and Bryant, 2006). 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 varieties form a distinct clade, separated from drug-type varieties by a consistent band of parallel branches. Broad leaflet drug-type varieties clustered with purported Afghan Kush landrace samples (Table S1 and Figure S3), while narrow leaflet drug-type varieties appear to contain several groups with only faint visible distinctions between them, perhaps influenced by the inclusion of hybrid varieties, in the analysis. Our one Chinese hemp sample was more closely related to the BLDT group supporting previous analyses (Hillig, 2005).

We found significantly more heterozygosity in drug-type varieties than in hemp varieties (31% vs 22%, $P < 0.001$, two-tailed Mann–Whitney *U*-test, Table 1). This likely reflects the widespread hybridization of varieties in North America during the transition to indoor cultivation of drug-

Table 1. Summary of genetic distance, heterozygosity, and F_{ST} information for major *Cannabis* groups.

	Mean within distances (%)	Heterozygosity (%)
Hemp	19.5	22*
All drug-types	24.4	31*
NLDT	23.7	32
BLDT	22.1	30
	Mean between distances	F_{ST}
Hemp vs. All drug-types	27.3	0.098530
Hemp vs. NLDT	26.9	0.091679
Hemp vs. BLDT	28.1	0.10131
NLDT vs. BLDT	25.8	0.036156

*Significantly different ($P < 0.001$, two-tailed Mann–Whitney *U*-test). Both genetic distances and heterozygosity levels are reported as percentages of total SNP sites.

type varieties starting in the 1970s (Clarke and Merlin, 2013), as well as the extensive reliance on clonal propagation for indoor commercial cultivation, which does not require trait stable seed stock. Conversely, fiber and seed oil hemp are grown on multi-acre scales that have necessitated the stabilization of agronomically important traits in seed stocks, likely leading to reduced heterozygosity at some loci. Our findings are contrary to those of Sawler *et al.*, (2015), which resulted in hemp samples having significantly more heterozygosity than drug-type varieties. In both analyses, hemp samples numbers were limited, and the two different sequencing library preparation methods make reconciling these conflicting results impossible without further sampling (Vergara *et al.*, 2017).

C. Population structure

To determine the statistical likelihood of various population scenarios represented in our samples, we first applied the FLOCK model to our data set of 195 GBS and WGS *Cannabis* samples, which is an iterative reallocation clustering algorithm that does not require nonadmixed individuals to make population assignments (Duchesne and Turgeon, 2012). Using the *K*-partitioning method suggested by the authors (Duchesne and Turgeon, 2012), we determined that $K \geq 3$, after testing *K* values of one to eight Figure 3 and peripheral population names in Figure 2). FLOCK was able to assign all samples to one of the three identified populations, although it does not calculate admixture proportions. Sample population assignments were largely consistent with the known history of these samples, and appear visually consistent with the MDS analysis (Figure S2). For example all fiber and seed oil hemps were assigned to an exclusive population, with the exception of sample AC/DC, a high CBDA-producing variety, suggesting it has hybrid hemp and drug-type origins (Figures 2 and 3), or that it represents an under sampled and distinct population. Likewise, FLOCK assigned the sole broad leaflet Chinese hemp sample to the hemp group, although it appears genetically distinct from the European hemp cluster, implying this could be a representative of a distinct, divergent Asian hemp lineage (Gao *et al.*, 2014).

Additionally we applied the admixture model-based Bayesian clustering method of fastSTRUCTURE to the same 195 samples (Raj *et al.*, 2014). The most likely population structure analysis of $K = 2$, shows consistent separation between BLDT and NLDT and hemp varieties (Figure 2 branch colors, Table S1). Some hemp and NLDT varieties were each assigned with nearly 100% population membership to the same population (Figure 2 light blue samples, Table S1), despite the clear separation visualized in the tree and statistically significant mean between-group genetic distance measured (Table 1). The separation of BLDT and NLDT varieties into fastSTRUCTURE populations was stable when hemp samples were excluded from the analysis (Table S1). Sawler *et al.* (2015) used fastSTRUCTURE to delineate hemp from drug-type varieties as the major division of *Cannabis* diversity, and found two drug-type sub-groups within their samples when hemp types were excluded from the analysis. Likewise, using a smaller dataset (Lynch, 2015) found support for $K = 3$, consisting of two separate drug-type populations and hemp types, using the original STRUCTURE implementation (Pritchard *et al.*, 2000) and the Evanno method to select the best value of K (Evanno *et al.*, 2005). However, we caution that despite many claims for the availability of "landrace genetics" (varieties) from *Cannabis* producers, breeders, and seed sellers, these may or may not represent nonadmixed individuals (Clarke and Merlin, 2013)—a situation that can be problematic for the STRUCTURE and fastSTRUCTURE approaches (Pritchard *et al.*, 2000).

The GBS samples from Sawler *et al.* (2015) appear to contain an additional divergent NLDT clade, with likely SE Asian origins (Figures S3 and S4) which did not emerge from our main analyses. However, our SplitsTree analysis (Figure 2) does support a possible separation of the NLDT group into two subgroups NLDT-I and NLDT-II, as found by Sawler *et al.*, (2015), with the NLDT-II clade peripheral and separated by the BLDT samples. Due to very limited overlap between sequence fragments from the two GBS datasets, which results from using different restriction enzymes, we could only use the Sawler *et al.*, (2015) data, in combination with only our 67 WGS samples. A connection was made across the two GBS analyses to this SE Asian NLDT group through two WGS samples (Dr. Grinspoon and Somali Taxi Cab, Figure S3) that were included in both sets of GBS analyses. Moreover, although only 45 SNPs overlapped between our GBS data, the Sawler GBS data, and the WGS data, a phylogeny of this limited alignment also supports the existence of an additional distinct SE Asian NLDT clade (Figure S4). Collectively, these analyses lend support to a total lower bound of four *Cannabis* populations, although clearly more extensive sampling with consistent sequencing is required to fully access the standing biogeographic diversity.

D. Tests of tree models

To test hypotheses of tree-like evolution for the three genetic groups, we first applied the three-population

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_Herrer	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_Cookies	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

Figure 3. Sample names and FLOCK assignment of 195 varieties to three groups, represented with different cell colors. Green (left) are BLDT, blue (center) are NLDT, and yellow (lower right) are hemp.

test for admixture (Reich *et al.*, 2009), and found no evidence for admixture in any of the pairwise comparisons (positive f statistic values). Next we constructed maximum likelihood trees based on the aggregate SNP frequencies for the three genetic groups and simulated a variety of "migration" events

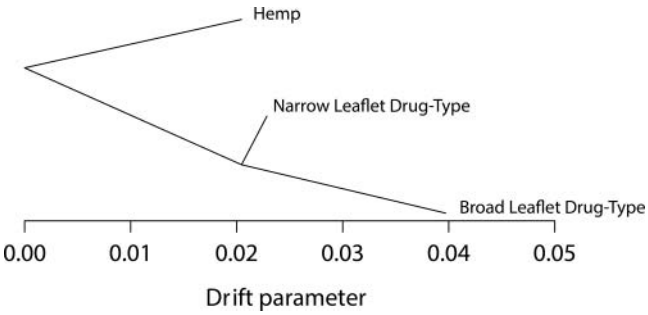

Figure 4. Maximum likelihood tree of three *Cannabis* populations, created in TreeMix. We found no evidence for extensive admixture or deviations from this tree model.

(0–10), but no simulation produced nonzero migration graph edges (Figure 4). F_{ST} analysis shows little divergence among lineages for most loci. However, a substantial number of highly divergent regions are unique to each clade (Figure 5). Although Lore (Clarke and Merlin, 2013), Figures 2 and S2 strongly suggest at least some individuals have hybrid origins due to admixture between the distinct *Cannabis* populations. However, our tree models for the overall SNP frequencies of the population groups inferred by FLOCK (Figure 3) demonstrate that each group still contains strong genetic signals from ancestral biogeographic gene pools. Still, our results and the incongruence between *Cannabis* genomic analyses (Sawler *et al.*, 2015) reinforces the importance of using many high-quality single-copy regions of the genome, rather than smaller numbers of loci that could lead to lower resolution or even misleading results (Vergara *et al.*, 2017).

The divergent regions depicted in Figure 5 establish that hemp and BLDT are the most divergent groups, followed by hemp and NLDT. NLDT and BLDT are the most genetically similar groupings, with many genetically similar loci and lower average F_{ST} (Table 1). The shared SNPs (Figure 4) and F_{ST} values (Table 1, Figure 5) suggest that both drug-type groups, NLDT and BLDT, are the most closely related. This finding is also supported by our FLOCK and fastSTRUCTURE analyses discussed above. Thus, multiple lines of evidence show similar patterns of relatedness between these three groups.

Additional *Cannabis* diversity remains to be sampled. Notably absent from all genome sequence datasets published to date are putative *Cannabis ruderalis* (Janischevsky, 1924) samples. These are short weedy plants, with free shattering inflorescences found widely from Northern Siberia, through Central Asia, and into

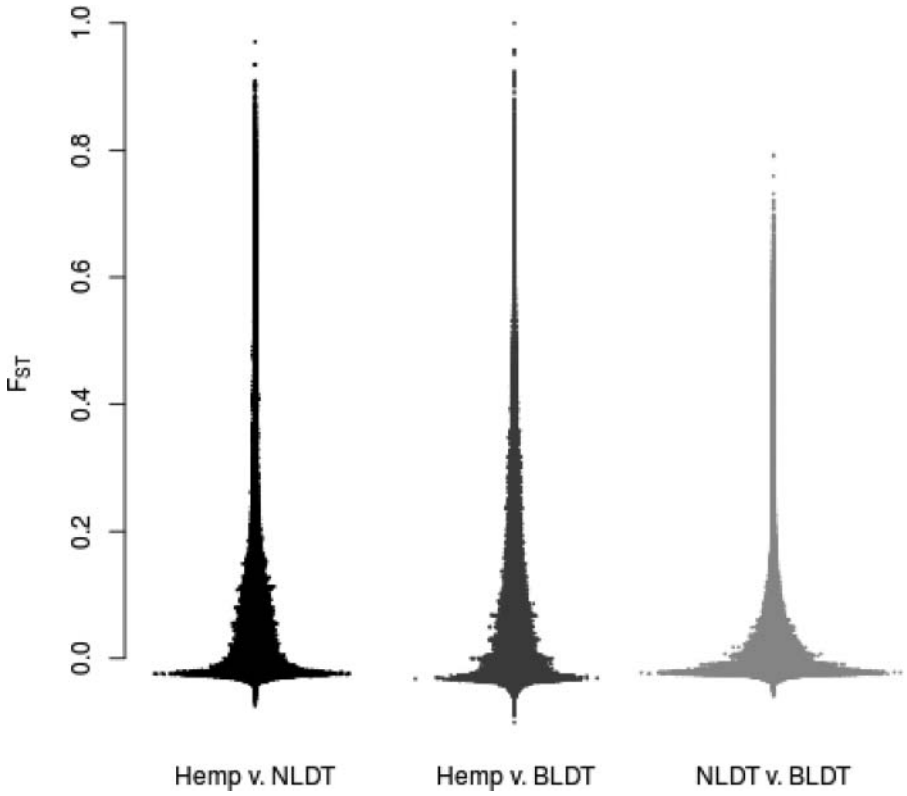

Figure 5. Distribution of Weir-Cockerham F_{ST} estimates for each population comparison. This figure re-enforces that the BLDT and the NLDT groups are the most similar, with the most low-F_{ST} SNPs and the fewest divergent SNPs, while the Hemp and the BLDT are the most different, with the most segregating SNPs.

Eastern Europe (Clarke and Merlin, 2013). Whether these populations represent wild *Cannabis*, more recent feral escapes, or some combination of both remains unclear. Unfortunately, we did not sample putative *C. ruderalis* populations. However Finola is an early maturing seed hemp variety from Finland with purported northern Russian landrace ancestry (Clarke and Merlin, 2013), and Low Ryder and Auto AK-47 are auto-flowering drug-type varieties with possible *C. ruderalis* heritage included in our samples (Figure 3). Our analyses found Finola fits within the hemp group while Low Ryder and Auto AK-47 are close relatives of each other within the NLDT group (Figure S3). Further genomic analyses are required to determine the extent to which *C. ruderalis* populations are genetically distinct from hemp and drug-type groups, and whether they may, in fact, harbor an ancestral wild-type gene pool from which European hemp varieties were domesticated (Hillig and Mahlberg, 2004; Hillig, 2005).

Broad leaflet Asian hemp is also underrepresented, although we included one putative Chinese hemp sample that occupies an area between the core hemp and BLDT populations (Figures 2, 3 and S2). Hillig's (2005) analysis of alloenzymes concluded that Asian hemp varieties were more similar to Asian drug-type varieties than they were to narrow leaflet European hemp. Likewise, Gao *et al*., (2014) found genetic dissimilarity between European hemp and Chinese hemp using microsatellites, and showed at least several distinct groups of hemp occur across the vast geography of Asia. Overall, Asian and European hemp varieties appear genetically dissimilar, possibly reflecting the independent domestication events (Clarke and Merlin, 2013).

One major complication obscuring the understanding of *Cannabis* diversity and history is the lack of information about the native range or ranges of *Cannabis*. In addition to divergent breeding efforts and human-vectored transport of seeds, the tendency of *Cannabis* is to escape into feral populations wherever human cultivation occurs in temperate climates (Small *et al.*, 2003). This, coupled with wind pollination biology and no known reproductive barriers, makes the existence of pure wild native *Cannabis* populations unlikely. The weedy tendencies of *Cannabis* are exemplified by the Midwestern U.S. populations of feral hemp that flourish despite the eradication efforts by the Drug Enforcement Agency, which have for decades totaled millions of plants removed per year. A comprehensive evaluation of *Cannabis* diversity, which includes feral and wild Eurasian populations is required to ascertain if the levels of divergence and gene flow are consistent with one or more origins of domestication (Hillig, 2005). Even if these extant populations are highly admixed with modern varieties, their study promises to offer insight into *Cannabis* ecology and evolution, given how different the selective regime of the feral setting is compared to that of the agricultural fields. Considering the similar debates regarding the timing and origins of *Oryza* domestication that remains, as of yet, unresolved (Gross and Zhao, 2014), *Cannabis* requires substantially more work to unravel its complicated relationship with humans.

"Indica" and "sativa" are commonly used terms ascribed to plants that have certain characteristics, often related to leaflet morphology and the perceived effects of consuming the plant (Habib *et al.*, 2013). However, these names are rooted in taxonomic traditions dating to Linnaeus who first classified the genus as monotypic (*C. sativa*) based on hemp specimens from Virginia and Europe (Linnaeus, 1753). de Lamarck subsequently designated *C. indica* to accommodate the shorter stature potent narrow leaflet drug-type plants from the Indian subcontinent (de Lamarck, 1783). Although currently the term "indica" is typically used to refer to BLDTs, this biotype from the Hindu Kush mountains (Clarke and Merlin, 2013) was not clearly documented until a 1929 survey of Afghani agriculture by Vavilov (Vavilov and Bukinich, 1929). This absence of historical documentation until the 20th century, a very narrow geographic range, and some evidence for a broader NLDT gene pool (Table 1, Figures S3 and S4) suggest a separate and more recent origin of the BLDT clade. This origin could represent a domestication event of a wild or feral BLDT population, or perhaps hybridization events between NLDT and BLDT populations. The final resolution of *Cannabis* taxonomy will require complete assessment of standing global genetic diversity and experimental evaluation of reproductive compatibility across all major genetic groups (Rieseberg and Willis, 2007), in conjunction with morphological circumscriptions. Given the current absence of evidence for reproductive barriers, and overall limited genetic distances between hemp and drug-type varieties analyzed in this study, we suggest continued monotypic treatment of plants in this genus as *C. sativa* L. is warranted.

E. Cannabinoid and terpenoid diversity

THCA and CBDA are the most abundant cannabinoids produced by the majority of varieties on the North American market today (Figure 6a), and both compounds show an impressive range of medicinal potential (Di Marzo *et al.*, 2004; Devinsky *et al.*, 2014), although endocannabinoid-based therapy trials have a history of significant rates of study withdraws and adverse effects (Wade *et al.*, 2006). Historical breeding efforts have

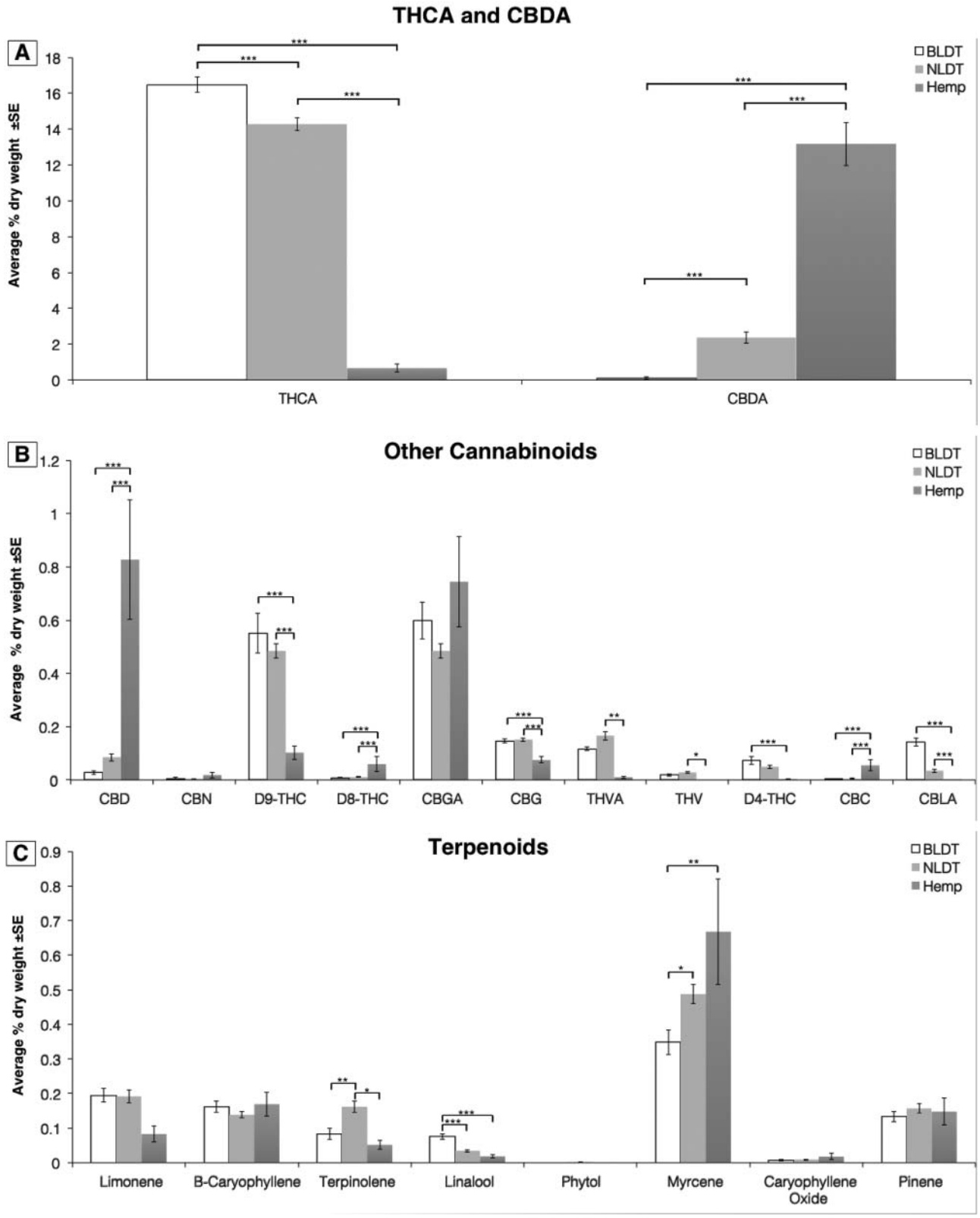

Figure 6. Average percentage and standard error of mass for dried and un-pollinated female inflorescences of *Cannabis* genetic groups. (A) THCA and CBDA cannabinoids (B) Minor cannabinoids (C) Terpenoids. THCA, delta-9-tetrahydrocannabinolic acid; CBDA, cannabidiolic acid; CBD, cannabidiol; CBN, cannabinol; D9-THC, delta-9-tetrahydrocannabinol; D8-THC, delta-8-tetrahydrocannabinol; CBGA, cannabigerolic acid; CBG, cannabigerol; THCVA, Tetrahydrocannabivarin carboxylic acid; THCV, Tetrahydrocannabivarin; D4-THC, delta-4-tetrahydrocannabinol; CBC, cannabichromene; CBLA, cannabicyclolic acid.

resulted in mostly high THCA plants that lead to strong intoxicating effects when consumed, and that synthesize only very low levels of alternative cannabinoids (Figure 6b). High CBDA plants have only recently become more available in North America over the last several years in response to the demand. Interestingly, these high CBDA-producing plants form several clusters within both the NLDT and BLDT groups, as well as within the hemp group (Table S1), but rarely reach equivalent quantities of total cannabinoid production as those found in high THCA plants (Figure 5a). The minor cannabinoids that are commonly assayed, CBGA, CBCA, THCVA, and CBDVA are also of interest, despite varieties producing high levels of these compounds being largely unavailable for research currently (Abrams *et al.*, 2007). With at least 74 cannabinoids identified in *Cannabis*, modernized genetic and breeding techniques are required to diversify and optimize *Cannabis* varieties. Efforts should also be made to document and preserve feral, wild, and heirloom populations that can serve as reservoirs of cultural and genetic diversity.

A caveat of our research is the lack of chemical data for most of our genetic samples. However, the differences in both cannabinoids and terpenoids between the varieties in the FLOCK groups are supported by previous research. Particularly, the higher levels of CBD and CBDA, and lower levels of THC and THCA found in the hemp group, where we have only limited chemotype data, are supported by numerous studies that have found that hemp varieties have high CBD and CBDA relative to THC and THCA (de Meijer *et al.*, 1992; Rustichelli *et al.*, 1998; Mechtler *et al.*, 2004; Datwyler and Weiblen, 2006).

Aromatic terpenoids impart many of the characteristic fragrances to *Cannabis*, and possibly contribute to the effects of consumption (Russo, 2011). Terpenoids are synthesized in many plant species, and play a role in relieving various abiotic and biotic stresses through direct and indirect mechanisms (Holopainen and Gershenzon, 2010). Despite the limitations of our study design where we were only able to connect genotype and chemical phenotype through variety names, our analysis found that each of the three distinct genetic groups shows a distinct terpenoid profile on average (Figures 6c and S5). We found NLDTs to contain significantly more β-myrcene and α-terpinolene than BLDTs, although interestingly the two hemp varieties for which we analyzed chemical data had significantly more β-myrcene than either drug-type groups (Figure 6c). Similarly, Hillig (2004) found NLDTs to yield significantly more β-myrcene than Afghani BLDTs, yet European hemp and un-cultivated accessions labeled as *C. ruderalis* contained the highest levels. Hillig also reported that Afghani BLDTs contained the highest levels of guaiol and eudesmol isomers, which we did not measure, although we found BLDTs contained more linalool than NLDTs or hemp. Understanding the ecological functions and evolutionary origins of terpenoids and cannabinoids in *Cannabis* could improve therapeutic potential, and possibly reduce the need for pesticide application during cultivation.

F. Conclusions

Cannabis genomics not only offers a window into the past, but also a road forward. Although historical and clandestine breeding efforts have been clearly successful in many regards (Mehmedic *et al.*, 2010; Swift *et al.*, 2013), *Cannabis* lags decades behind other major crop species in many other respects. Developing stable *Cannabis* lines capable of producing the full range of potentially therapeutic cannabinoids is important for the research and medical communities, which currently lack access to diverse high-quality material in the USA (Nutt *et al.*, 2013).

In this article we extended the initial *Cannabis* genome study (van Bakel *et al.*, 2011), by re-mapping WGS and GBS sequence reads to the existing PK draft scaffolds, to understand diversity and evolutionary relationships among the major lineages. Although hybridization of cultivated varieties (Clarke and Merlin, 2013) and human transport of seeds across the globe was hypothesized to have obscured much of the ancestral genetic signal (Small, 2015b), we found significant evidence for apparent ancestral signals in genomic data derived largely from modern cultivated varieties (Figures 2, 3, and 4). Re-analysis of previously published GBS data (Sawler *et al.*, 2015) provides additional limited evidence for a fourth group (Figures S3 and S4). Interestingly, unique cannabinoid and terpenoid profiles were associated with three of the genetic groups, lending support to their validity, despite the limitations of our sampling scheme. Overall, we hope the publically available data and analyses from this study will facilitate continued research on the history of this controversial plant and the development of the agricultural and therapeutic potential of *Cannabis*.

Acknowledgments

The authors thank Ben Holmes of Centennial Seeds; Devin Liles, Carter Casad, and Jan Cole of The Farm; Ashley Edwards of Ward, Colorado; Jake Salazar of MMJ America; Kevin McKernan of Medicinal Genomics; Ashley and Matt Rheingold of Headquarters; David Salama; Ezra Huscher; and Nico Escondido, and Bob Sievers for providing the DNA samples or sequence information. They thank Reggie

Gaudino of Steep Hill for the advice and assistance with the chemical data.

Funding

This project was supported by donations to the University of Colorado Foundation gift fund 13401977-Fin8 to NCK. Funding for open access of this article was provided by Medicinal Genomics.

Author contributions

RCL, DV, KHW, and NCK designed the project. CJS, MJG, KW, DV, and RCL collected samples. KW generated DNA sequencing libraries. RCL, NCK, and SBT performed bioinformatics analyses. KdC, DPL, and TCR generated chemical data. DV performed chemical data analyses. RCL, DV, and NCK wrote the article.

References

Abrams, D. I., Vizoso, H. P., Shade, S. B., Jay, C., Kelly, M. E., and Benowitz, N. L. 2007. Vaporization as a smokeless cannabis delivery system: a pilot study. *Clin. Pharmacol. Ther.* **82**: 572–578.

Anderson, L. C. 1980. Leaf Variation Among cannabis species from a controlled garden. *Bot. Museum Leafl.* **28**: 61–69.

Bauer, R., Salo-Ahen, K., and Bauer, O. 2008. CB receptor ligands from plants. *Curr. Top. Med. Chem.* **8**: 173–186.

Berry, E. M. and Mechoulam, R. 2002. Tetrahydrocannabinol and endocannabinoids in feeding and appetite. *Pharmacol. Ther.* **95**: 185–190.

Bolger, A. M., Lohse, M., and Usadel, B. 2014. Genome analysis trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bonnie, R. J. and Whitebread, C. H. 1970. The forbidden fruit and the tree of knowledge: an inquiry into the legal history of American marijuana prohibition. *Virginia Law Rev* **56**: 971–1203.

Borrelli, F., Pagano, E., Romano, B., Panzera, S., Maiello, F., Coppola, D., Petrocellis, L., De Buono, L., Orlando, P., and Izzo, A. A. 2014. Colon carcinogenesis is inhibited by the TRPM8 antagonist cannabigerol, a *Cannabis* derived non-psychotropic cannabinoid. *Carcinogenesis* **35**: 2787–2797.

Cherniak, L. 1982. *The Great Books of Cannabis vol. I, Book II*. Damele Publishing, Oakland, CA.

Clarke, R. C. and Merlin, M. D. 2013. *Cannabis Evolution and Ethnobotany,* University of California Press, Berkeley, CA.

Clarke, R. C., and Merlin, M. D. 2015. Letter to the editor: Small, Ernest. 2015. Evolution and classification of *Cannabis sativa* (Marijuana, Hemp) in relation to human utilization. *Bot. Rev.* **81**: 189–294.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., and Durbin, R. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Datwyler, S.L. and Weiblen, G.D. 2006. Genetic variation in hemp and marijuana (*Cannabis sativa* L.) according to amplified fragment length polymorphisms. *J. Forensic Sci.* **51**: 371–375.

de Lamarck, Jean-Baptiste de Monet. 1783. *Encyclopédie méthodique ou par ordre de matières: botanique*. Vol. 1. Panckoucke.

de Meijer, E. P. M., Bagatta, M., Carboni, A., Crucitti, P., Moliterni, V. M. C., Ranalli, P., and Mandolino, G. 2003. The inheritance of chemical phenotype in *Cannabis sativa* L. *Genetics* **346**: 335–346.

de Meijer, E. P. M. and Hammond, K. 2016. The inheritance of chemical phenotype in *Cannabis sativa* L. (V): regulation of the propyl-/pentyl cannabinoid ratio, completion of a genetic model. *Euphytica* **210**: 291–307.

de Meijer, E. P. M. and Hammond, K. M. 2005. The inheritance of chemical phenotype in *Cannabis sativa* L. (II): cannabigerol predominant plants. *Euphytica* **145**: 189–198.

de Meijer, E. P. M., Hammond, K. M., and Micheler, M. 2008. The inheritance of chemical phenotype in *Cannabissativa* L. (III): variation in cannabichromene proportion. *Euphytica* **165**: 293–311.

de Meijer, E.P.M., Van der Kamp, H.J. and Van Eeuwijk, F.A. 1992. Characterization of *Cannabis* accessions with regard to cannabinoid content in relation to other plant characters. *Euphytica* **62**: 187–200.

De Petrocellis, L., Ligresti, A., Moriello, A. S., Allarà, M., Bisogno, T., Petrosino, S., Stott, C. G., and Di Marzo, V. 2011. Effects of cannabinoids and cannabinoid-enriched *Cannabis* extracts on TRP channels and endocannabinoid metabolic enzymes. *Br. J. Pharmacol.* **163**: 1479–1494.

Devinsky, O., Cilio, M. R., Cross, H., Fernandez-Ruiz, J., French, J., Hill, C., Katz, R., Di Marzo, V., Jutras-Aswad, D., Notcutt, W. G., Martinez-Orgado, J., Robson, P. J., Rohrback, B. G., Thiele, E., Whalley, B., and Friedman, D. 2014. Cannabidiol: pharmacology and potential therapeutic role in epilepsy and other neuropsychiatric disorders. *Epilepsia* **55**: 791–802.

Di Marzo, V., Bifulco, M., and De Petrocellis, L. 2004a. The endocannabinoid system and its therapeutic exploitation. *Nat. Rev. Drug Discov.* **3**: 771–784.

Divashuk, M. G., Alexandrov, O. S., Razumova, O. V, Kirov, I. V, and Karlov, G. I. 2014. Molecular cytogenetic characterization of the dioecious *Cannabis sativa* with an XY chromosome sex determination system. *PLoS ONE* **9**: e85118.

Duchesne, P. and Turgeon, J. 2012. FLOCK Provides reliable solutions to the "Number of Populations" problem. *J. Hered.* **103**: 734–743.

ElSohly, M. A. and Slade, D. 2005. Chemical constituents of marijuana: the complex mixture of natural cannabinoids. *Life Sci.* **78**: 539–548.

Evanno, G., Regnaut, S., and Goudet, J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**: 2611–2620.

Fellermeier, M., Eisenreich, W., Bacher, A., and Zenk, M. H. 2001. Biosynthesis of cannabinoids. *Eur. J. Biochem.* **268**: 1596–1604.

Gao, C., Xin, P., Cheng, C., Tang, Q., Chen, P., Wang, C., Zang, G., and Zhao, L. 2014. Diversity analysis in *Cannabis sativa* based on large-scale development of expressed sequence tag-derived simple sequence repeat markers. *PLoS ONE* **9**: e110638.

Glanzman, A. 2015. Discover Himalaya's Outlawed Marijuana Fields. *Time*. http://time.com/3736616/discover-himalayas-illegal-marijuana-fields/ (Accessed January 1, 2016)

Gross, B. L. and Zhao, Z. 2014. Archaeological and genetic insights into the origins of domesticated rice. *Proc. Natl. Acad. Sci. U. S. A.* **111**: 6190–6197.

Habib, R., Finighan, R., and Davenport, S. 2013. *Testing for Psychoactive Agents*. http://lcb.wa.gov/publications/Marijuana/BOTEC%20reports/1c-Testing-for-Psychoactive-Agents-Final.pdf

Hanuš, L. O., Levy, R., De La Vega, D., Katz, L., Roman, M., and Tomíček, P. 2016. The main cannabinoids content in hashish samples seized in Israel and Czech Republic. *Israel J. Plant Sci.* **63**: 182–190.

Hazekamp, A. and Fischedick, J. T. 2012. *Cannabis* – from cultivar to chemovar. *Drug Test. Anal.* **4**: 660–667.

Hazekamp, A., Peltenburg, A., Verpoorte, R., and Giroud, C. 2005. Chromatographic and spectroscopic data of cannabinoids from *Cannabis sativa* L. *J. Liq. Chromatogr. Relat. Technol.* **28**: 2361–2382.

Hillig, K. W. 2004. A chemotaxonomic analysis of terpenoid variation in *Cannabis*. *Biochem. Syst. Ecol.* **32**: 875–891.

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

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

Holopainen, J. K. and Gershenzon, J. 2010. Multiple stress factors and the emission of plant VOCs. *Trends Plant Sci.* **15**: 176–184.

Huson, D. H. and Bryant, D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**: 254–267.

Izzo, A. A., Capasso, R., Aviello, G., Borrelli, F., Romano, B., Piscitelli, F., Gallo, L., Capasso, F., Orlando, P., and Di Marzo, V. 2012. Inhibitory effect of cannabichromene, a major non-psychotropic cannabinoid extracted from *Cannabis sativa*, on inflammation-induced hypermotility in mice. *Br. J. Pharmacol.* **166**: 1444–1460.

Janischevsky, D. E. 1924. *Cannabis ruderalis. Proc. Saratov* **2**: 14–15.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Li, H. L. 1973. An archaeological and historical account of *Cannabis* in China. *Econ. Bot.* **28**: 437–448.

Linnaeus, C. 1753. *Species Plantarum* (1st ed.). Laurentius Salvius, Stockholm.

Lynch, R. C. 2015. *Genomics of Adaptation and Diversification*. University of Colorado, Boulder. ProQuest Dissertations & Theses A&I. http://gradworks.umi.com/37/04/3704759.html

McPartland, J. M., Duncan, M., Marzo, V. Di, and Pertwee, R. G. 2015. Are cannabidiol and Δ9-tetrahydrocannabivarin negative modulators of the endocannabinoid system? *A systematic review, British J. Pharma.* **172**: 737–753.

McPartland, J. M., Matias, I., Di Marzo, V., and Glass, M. 2006. Evolutionary origins of the endocannabinoid system. *Gene* **370**: 64–74.

Mechoulam, R., and Gaoni, Y. 1967. Recent advances in the chemistry of hashish. *Chemie Org. Naturstoffe* **25**: 175–213.

Mehmedic, Z., Chandra, S., Slade, D., Denham, H., Foster, S., Patel, A. S., Ross, S. A., Khan, I. A., and ElSohly, M. A. 2010. Potency trends of Δ9-THC and other cannabinoids in confiscated *Cannabis* preparations from 1993 to 2008. *J. Forensic Sci.* **55**: 1209–1217.

Mechtler, K., Bailer, J. and De Hueber, K. 2004. Variations of Δ 9-THC content in single plants of hemp varieties. *Ind. Crops Prod.* **19**: 19–24.

Nutt, D. J., King, L. A, and Nichols, D. E. 2013. Effects of Schedule I drug laws on neuroscience research and treatment innovation. **14**: 577–585.

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

Pacher, P. and Mechoulam, R. 2011. Is lipid signaling through cannabinoid 2 receptors part of a protective system? *Prog. Lipid Res.* **50**: 193–211.

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

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

Pickrell, J. K. and Pritchard, J. K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**: e1002967.

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

Pritchard, J. K., Stephens, M., and Donnelly, P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155**: 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**: 559–575.

Radwan, M. M., Ross, S. A., Slade, D., Ahmed, S. A., Zulfiqar, F., and Elsohly, M. A. 2008. Isolation and characterization of new *Cannabis* constituents from a high potency variety. *Planta Med.* **74**: 267–272.

Raj, A., Stephens, M., and Pritchard, J. K. 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics* **197**: 573–589.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. 2009. Reconstructing Indian population history. *Nature* **461**: 489–494.

Rieseberg, L. H., and Willis, J. H. 2007. Plant speciation. *Science* **317**: 910–914.

Russo, E. B. 2007. History of *Cannabis* and its preparations in saga, science, and sobriquet. *Chem. Biodivers.* **4**: 1614–1648.

Russo, E. B. 2011. Taming THC: potential *Cannabis* synergy and phytocannabinoid-terpenoid entourage effects. *Br. J. Pharmacol.* **163**: 1344–1364.

Rustichelli, C., Ferioli, V., Vezzalini, F., Rossi, M. C., and Gamberini, G. 1996. Simultaneous separation and

identification of hashish constituents by coupled liquid chromatography-mass spectrometry (HPLC-MS). *Chromatographia* **43**: 129–134.

Rustichelli, C., Ferioli, V., Baraldi, M., Zanoli, P. and Gamberini, G. 1998. Analysis of cannabinoids in fiber hemp plant varieties (*Cannabis sativa* L.) by high-performance liquid chromatography. *Chromatographia* **48**: 215–222.

Sakamoto, K., Akiyama, Y., Fukui, K., Kamada, H., and Satoh, S. 1998. Characterization; Genome sizes and morphology of sex chromosomes in hemp (*Cannabis sativa* L.). *Cytologia.* **63**: 459–464.

Sawler, J., Stout, J. M., Gardner, K. M., Hudson, D., Vidmar, J., Butler, L., Page, J. E., and Myles, S. 2015. The genetic structure of marijuana and hemp. *PLoS ONE* **10**: e0133292.

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

Small, E. 2015a. Evolution and classification of *Cannabis sativa* (Marijuana, Hemp) in relation to human utilization. *Bot. Rev.* **81**: 189–294.

Small, E. 2015b. Response to the erroneous critique of my *Cannabis* monograph by R. C. Clarke and M.D. Merlin. *Bot. Rev.* **81**: 306–316.

Small, E. and Cronquist, A. 1976. A practical and natural taxonomy for *Cannabis*. *Taxon* **25**: 405–435.

Small, E., Pocock, T., and Cavers, P. 2003. The biology of Canadian weeds. 119. *Cannabis sativa* L. *Can. J. Plant Sci.* **83**: 217–237.

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

Swift, W., Wong, A., Li, K. M., Arnold, J. C., and McGregor, I. S. 2013. Analysis of *Cannabis* seizures in NSW, Australia: *Cannabis* potency and cannabinoid profile. *PLoS ONE* **8**: 1–9.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30**: 2725–2729.

Tramèr, M. R., Carroll, D., Campbell, F. A., Reynolds, D. J., Moore, R. A., and McQuay, H. J. 2001. Cannabinoids for control of chemotherapy induced nausea and vomiting: quantitative systematic review. *BMJ* **323**: 16–21.

van Bakel, H., Stout, J. M., Cote, A. G., Tallon, C. M., Sharpe, A. G., Hughes, T. R., and Page, J. E. 2011. The draft genome and transcriptome of *Cannabis sativa*. *Genome Biol.* **12**: R102.

Vavilov, N. I. and Bukinich, D. D. 1929. *Zemledel'cheskii Afganistan [Agricultural Afghanistan]*. Leningrad: Vse soyusnyi institut prikladnoi botaniki i novykh kul'tur pri SNK SSSR.

Vergara, D., Baker, H., Clancy, K., Keepers, K.G., Mendieta, J. P., Pauli, C.S., Tittes, S.B., White, K.H., and Kane, N.C. 2017. Genetic and genomic tools for *Cannabis sativa*. *Crit. Rev. Plant Sci.* **x**: xx–xx.

Volkow, N. D., Baler, R. D., Compton, W. M., and Weiss, S. R. B. 2014. Adverse health effects of marijuana use. *N. Engl. J. Med.* **370**: 2219–2227.

Wade, D. T., Makela, P. M., House, H., Bateman, C., and Robson, P. 2006. Long-term use of a *Cannabis*-based medicine in the treatment of spasticity and other symptoms in multiple sclerosis. *Mult. Scler.* **12**: 639–645.

Weiblen, G. D., Wenger, J. P., Craft, K. J., ElSohly, M. A., Mehmedic, Z., Treiber, E. L., and Marks, M. D. 2015. Gene duplication and divergence affecting drug content in *Cannabis sativa*. *New Phytol.* **208**: 2141–1250.

Zogopoulos, P., Vasileiou, I., Patsouris, E., and Theocharis, S. E. 2013. The role of endocannabinoids in pain modulation. *Fundam. Clin. Pharmacol.* **27**: 64–80.