

RESEARCH ARTICLE

Classification of cannabis strains in the Canadian market with discriminant analysis of principal components using genome-wide single nucleotide polymorphisms

Dan Jin^{1,2}, Philippe Henry^{3,4}, Jacqueline Shan², Jie Chen^{1,5*}

1 Department of Biomedical Engineering, University of Alberta, Edmonton, Alberta, Canada, **2** PBG BioPharma Inc., Leduc, Alberta, Canada, **3** Egret Bioscience Ltd., West Kelowna, British Columbia, Canada, **4** Lighthouse Genomics Inc., Salt Spring Island, British Columbia, Canada, **5** Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada

* jc65@ualberta.ca



OPEN ACCESS

Citation: Jin D, Henry P, Shan J, Chen J (2021) Classification of cannabis strains in the Canadian market with discriminant analysis of principal components using genome-wide single nucleotide polymorphisms. *PLoS ONE* 16(6): e0253387. <https://doi.org/10.1371/journal.pone.0253387>

Editor: Tzen-Yuh Chiang, National Cheng Kung University, TAIWAN

Received: November 10, 2020

Accepted: June 3, 2021

Published: June 28, 2021

Copyright: © 2021 Jin et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: Raw genome sequencing data for 23 strains are available from the NCBI with BioProject PRJNA683613.

Funding: PBG BioPharma Inc. (<https://pbgbiopharma.com/>) provided funding support in the form of salaries for authors DJ and JS. PBG BioPharma Inc. also provided financial support for genome sequencing and travel expenses. JS is the founder and CEO of PBG BioPharma Inc., and reviewed the manuscript. Labs-Mart Inc. (<http://labs-mart.ca/>) provided chemical standards and

Abstract

The cannabis community typically uses the terms “Sativa” and “Indica” to characterize drug strains with high tetrahydrocannabinol (THC) levels. Due to large scale, extensive, and unrecorded hybridization in the past 40 years, this vernacular naming convention has become unreliable and inadequate for identifying or selecting strains for clinical research and medicinal production. Additionally, cannabidiol (CBD) dominant strains and balanced strains (or intermediate strains, which have intermediate levels of THC and CBD), are not included in the current classification studies despite the increasing research interest in the therapeutic potential of CBD. This paper is the first in a series of studies proposing that a new classification system be established based on genome-wide variation and supplemented by data on secondary metabolites and morphological characteristics. This study performed a whole-genome sequencing of 23 cannabis strains marketed in Canada, aligned sequences to a reference genome, and, after filtering for minor allele frequency of 10%, identified 137,858 single nucleotide polymorphisms (SNPs). Discriminant analysis of principal components (DAPC) was applied to these SNPs and further identified 344 structural SNPs, which classified individual strains into five chemotype-aligned groups: one CBD dominant, one balanced, and three THC dominant clusters. These structural SNPs were all multi-allelic and were predominantly tri-allelic (339/344). The largest portion of these SNPs (37%) occurred on the same chromosome containing genes for CBD acid synthases (CBDAS) and THC acid synthases (THCAS). The remainder (63%) were located on the other nine chromosomes. These results showed that the genetic differences between modern cannabis strains were at a whole-genome level and not limited to THC or CBD production. These SNPs contained enough genetic variation for classifying individual strains into corresponding chemotypes. In an effort to elucidate the confused genetic backgrounds of commercially available cannabis strains, this classification attempt investigated the utility of DAPC for classifying modern cannabis strains and for identifying structural SNPs.

instrumentation support for chemical testing, but did not have any additional role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Egret Bioscience Ltd. (<https://egret.bio>) and Lighthouse Genomics Inc. (<https://lighthousegenomics.com/>) provided support in the form of salaries for author PH, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: The funder provided support in the form of salaries for authors DJ, PH, and JS. This does not alter our adherence to PLOS ONE policies on sharing data and materials.

Introduction

Cannabis has a complex breeding history. Whether its botanical classification is monotypic (*sativa*) or polytypic (*sativa* and *indica*) remains controversial [1]. Since the 1980s, breeding for high psychoactive THC content has occurred very aggressively in North America [2]. Nearly all drug-type cannabis currently cultivated in the USA, Canada, and Europe are hybridized, resulting in thousands of strains [3]. Recent genetic studies focused on validating the vernacular classification of “Sativa” and “Indica” [4–7]. However, this terminology is inadequate for identifying or selecting strains for clinical research and medicinal production due to the misuse of the botanical nomenclature, extensive cross-breeding, and unreliable labelling during unrecorded hybridization [2]. One genetic study found that the reported ancestry percentage of “Sativa” vs. “Indica” for 81 drug stains is only moderately correlated with the calculated genetic structure ($r^2 = 0.36$) [5]. In addition, CBD dominant strains and balanced strains (THC \approx CBD), which have gained increasing attention due to CBD’s use as a therapeutic [8–12], have been omitted in recent classification studies.

Cannabis has a diploid genome ($2n = 20$) with nine autosomal chromosomes and one pair of sex chromosomes [13]. The length of the haploid genome size is 818 Mbp for females and 843 Mbp for males [14]. An SNP is a variation of a single nucleotide at a specific position in the genome, and it is useful for understanding the genetic basis of diversity among populations [15]. SNPs are usually bi-allelic, with two alleles observed in the population [16]. Multiallelic SNPs have more than one alternative allele for that locus. Tri-allelic SNPs, which have three nucleotide substitution-based alleles at the same position, are relatively rare but are being considered of great relevance in epidemiological studies [17], in disaster victim identification using mixed and/or degraded DNA samples [18], and in animals pedigree accuracy studies [19]. Tri-allelic SNPs are reported to have a higher power of discrimination than bi-allelic SNPs requiring fewer markers and lowering costs [18, 20]. However, tri-allelic SNPs have been excluded in cannabis population structural analysis in the current literature [6, 21].

Cannabis classification studies that employ SNPs generally used partial genome information with few or no overlap sequences between datasets [22]. Whole-genome sequencing is used less often in the literature, but is preferable despite its higher cost because it enables comparison of genome datasets from different sources [22]. It also provides comprehensive genetic information [22], as studies showed that differences between fiber- and drug-type cannabis are at a genome-wide level and not necessarily limited to genes involved in THC production [5]. The recent release of the 10-chromosome map of the cannabis genome [23–27] may improve the understanding of the genetic architecture, identify a superior set of SNPs associated with interesting traits, and reduce future targeted genotyping costs by using fewer but more accurate SNPs [28].

Several approaches are now available for the analysis of population genetic structure. One of these approaches is the DAPC, which is a multivariate clustering method that combines the merits of both principal component analysis (PCA) and discriminant analysis (DA) [7, 29–31]. PCA is a multivariate analysis that can be applied to large datasets to reduce dimensions, but does not provide a group assessment, which is essential for investigating genetic structures of biological populations [32]. DA achieves the best classification of individuals into pre-defined groups by maximizing between-group variation and minimizing within-group variation, but the number of variables (alleles) needs to be fewer than the number of observations (individuals), which is generally not the case for SNP data [29]. DAPC first uses PCA to transform raw data (genome-wide identified SNPs) into principal components (PC), which are mutually orthogonal linear combinations of the original variables. This ensures that variables submitted to DA are perfectly uncorrelated and that there are fewer variables than number of individuals.

Then, linear discriminant functions, which are synthetic variables of linear combinations of these SNPs, are constructed to maximize inter-cluster differences and minimize intra-cluster variation [29]. By combining the advantages of PCA and DA, DAPC can identify groups, assign individuals to groups, visualize between-population differentiation, and identify individual alleles that have contributed to population structuring.

The objectives of this study are to:

1. investigate whether modern cannabis strains can be classified and differentiated at the whole-genome level, and
2. investigate the chromosomal location and putative functions of identified structural SNPs.

This study is a part of an integrated cannabis strain classification project utilizing genetic, chemical, and morphological profiles, wherein plants were grown in a commercial greenhouse under the same condition.

Materials and methods

DNA extraction and whole genome sequencing

This study included 23 commercially available cannabis strains, and the research was carried out under a cannabis research license issued by Health Canada. Where possible, the reported ancestry (“Sativa”, “Indica”, or “Sativa-dominant” and “Indica-dominant”) was obtained from the licensed producer providing the strain or from an online strain database (<https://www.leafly.ca>) (Table 1). Each strain was analyzed for chemical composition using methods established in our previous study [33] and labelled as “THC dominant”, “balanced”, and “CBD dominant”. DNA was extracted from 100 mg of fresh leaves for each strain using a Qiagen DNeasy Plant Mini Kit (QIAGEN, Canada). DNA concentrations were determined using a Qubit Fluorometer (Thermo Fisher Scientific, US). DNA integrity was tested by agarose gel electrophoresis. Library construction and sequencing were performed by BGI (USA) using DNBseq™ sequencing technology to a depth of 30x. DNBseq™ is a high-throughput sequencing solution, where DNA is fragmented into 100–300 bp and made into DNA nanoballs (DNB™), which are continuous DNA molecule with multiple head-to-tail copies of the same DNA fragment by linear isothermal rolling-circle replication. They are loaded onto high-density sequencing templates and sequenced by combinatorial probe-anchor synthesis (cPAS), where fluorescently tagged nucleotides complete for addition to the growing chain. After the addition of each nucleotide, high-resolution digital imaging is carried out where the DNB clusters are excited by a light source and a characteristic fluorescent signal is emitted. Hundreds of and thousands of clusters are sequenced in a massively parallel process. The emission wavelength, along with the signal intensity, determines the base call and the number of the cycles determines the length of the read. Sequence reads were then aligned to the reference genome assembly ASM23057v4 of a drug type strain Purple Kush (PK) in the NCBI BioProject database under accession number PRJNA73819 [34] using Burrows-Wheeler Alignment (BWA) tool [35]. New assignments of chromosomes numbers (1–10) were used as in ASM23057v5 [36]. The first step of SNP calling is marking duplications in BAM format files, and selected duplications are included in SNP calling by GATK (Genome Analysis Toolkit) (<https://www.broadinstitute.org/gatk/>). Local realignment around inDels is performed to avoid the bias of SNP calling, and the variation sites around inDel are identified as SNPs. A total of 235,334 SNPs was identified, including 225,046 bi-allelic and 10,288 multiallelic SNPs. After filtering for SNPs with no missingness by locus and a minor allele frequency less than 10% using

Table 1. Strain information of 23 strains and preassigned clusters by DAPC.

Strain number	Strain name	Chemotypes	Clusters (W-SNPs)	Clusters (I-SNPs)	"Sativa" or "Indica"
1	Lemon Garlic OG	1-Balanced	C1	C4	"Indica" dominant
2	Royal Medic	2-Balanced	C3	C2	"Sativa" dominant
3	Blue Hawaiian	3-CBD	C3	C1	"Sativa" dominant
4	Kandy Kush	4-CBD	C3	C1	"Sativa" dominant
5	Special	5-CBD	C3	C1	Not provided
6	NN	6-CBD	C3	C1	Not provided
7	Dance World	7-Balanced	C3	C2	"Sativa" dominant
8	Treat	8-CBD	C3	C1	Not provided
9	High	9-Balanced	C3	C2	Not provided
10	CB7	10-CBD	C3	C1	Not provided
11	33°	11-THC	C1	C4	Not provided
12	Banana Cake	12-THC	C2	C5	"Indica" dominant
13	Bananium	13-THC	C3	C3	"Indica" dominant
14	Burmese Blueberry	14-THC	C2	C5	"Indica" dominant
15	Divine Banana	15-THC	C2	C4	"Indica" dominant
16	Granddaddy Purple	16-THC	C2	C5	"Indica" dominant
17	Lemon Love	17-THC	C1	C5	"Indica" dominant
18	Lemon Sorbet	18-THC	C1	C4	"Indica" dominant
19	MeatHead	19-THC	C2	C5	"Indica" dominant
20	Nanitra	20-THC	C1	C4	"Indica" dominant
21	Platinum Jelly Punch	21-THC	C1	C4	"Indica" dominant
22	SBSK2 (Lemon Thai)	22-THC	C3	C3	50/50 hybrid
23	Super sherbet	23-THC	C1	C4	"Indica" dominant

*The column of clusters W-SNPs was obtained using the whole set of 137,858 filtered SNPs. The column of clusters I-SNPs was obtained using 344 structural SNPs.

<https://doi.org/10.1371/journal.pone.0253387.t001>

VCFtools, 137,858 SNPs, including 128,810 bi-allelic and 9,048 multiallelic SNPs, remained for analysis.

Analysis of population structure and identification of structural SNPs

The population structure in this work was analyzed by DAPC using the *adegenet* package [37] in R software [38]. First, the *find.clusters* function ran successive K-means [39] for a range of k values (where the number of clusters $k = K$), and identified the optimal number of clusters by comparing the Bayesian Information Criterion (BIC) [40] of the corresponding models. After groups were assigned, a cross-validation function (*xvalDapc*) was used to determine the optimal number of PCs to avoid over-sacrificing information or over-fitting in the subsequent DAPC. In cross-validation, the data were divided into a training set (90% of the data) and a validation set (10% of the data) by default. DAPC was carried out on the training set and the accuracy of predicting the membership of individuals in the validation set was used to identify the number of PCs. The sampling and DAPC were repeated 30 times by default at each level of PC retention. After assigning individuals to clusters, DA was carried out on the retained PCs and contributions of the alleles to each discriminant function were stored. An SNPZIP analysis (*snpzip*) in R was then used to provide objective delineation between structural and non-structural SNPs, as identified by DAPC, to determine which SNPs contribute significantly to the between-population structure [41].

First, the whole set of 137,858 SNPs were applied to DAPC to identify SNPs that contributed most to the identified clusters. DAPC was carried out again using the identified SNPs to validate their differentiation efficiency by confirming the separation of the 23 strains into their preassigned clusters. A short sequence (about 600 nt) around each one of these identified SNP was searched using the BLAST software (<https://blast.ncbi.nlm.nih.gov>) against *Cannabis sativa* Annotation Release 100 [42]. In addition to DAPC, other clustering methods, including PCA, neighbor-joining (NJ) tree [43], and hierarchical dendrogram using Ward's minimum variance method [44], were also employed to assess the robustness of the final inferred clusters. PCA and NJ tree were plotted using R. The hierarchical dendrogram was plotted using JMP 14.0.0.

Results and discussions

Discriminant analysis of principal components using 137,858 SNPs

As indicated by the elbow in the curve of BIC values as a function of k in Fig 1(a), the optimal number of identified clusters was three, corresponding to the lowest BIC values. The number of PCs retained for DAPC analysis was four, as calculated by cross-validation in Fig 1(b), where it had 100% predictive success, and 0% associated root mean squared error (RMSE). In this study, the number of PCs associated with the highest mean success was also associated with the lowest MSE, which made it easier to choose the number of PCs to retain. For the subsequent DAPC analysis, four PCs and two discriminant functions were retained. The DAPC

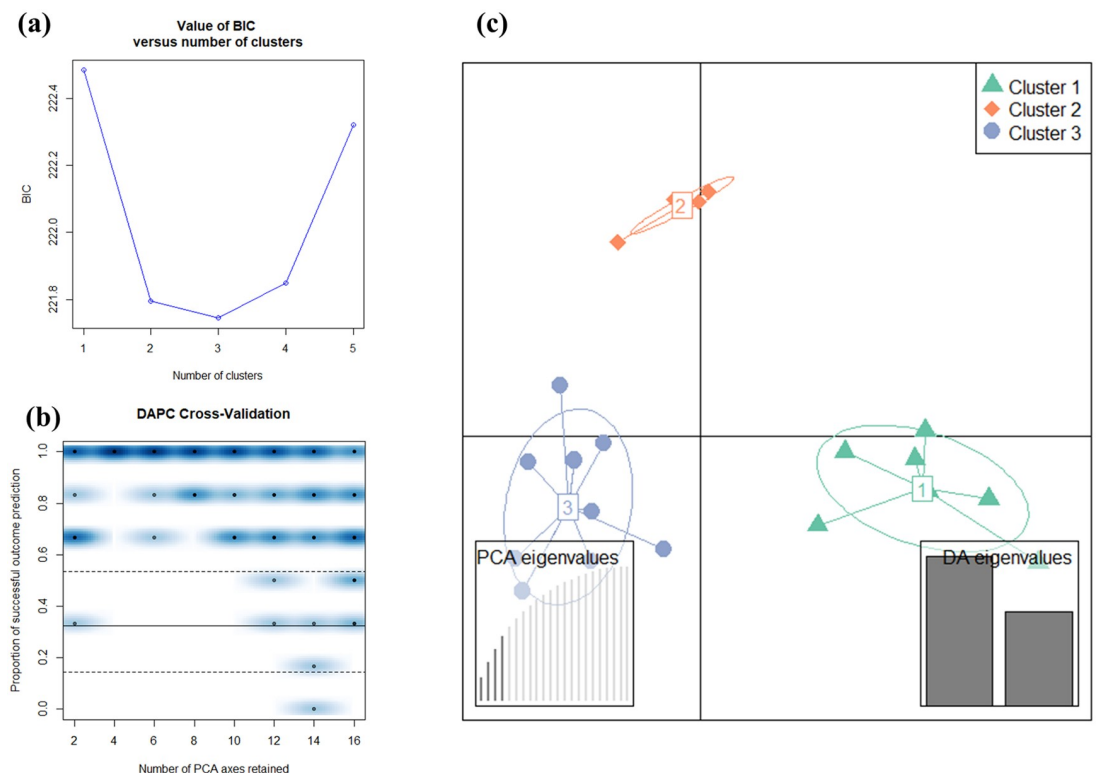


Fig 1. DAPC for 23 cannabis genotypes. (a) The x-axis is the number of clusters k and the y-axis is the corresponding value of BIC. (b) The plot of DAPC cross-validation. The x-axis is the number of PCA axes retained for DAPC, and the y-axis is the proportion of successful outcome prediction. Individual replicates appear as points, and the density of those points in different regions of the plot is displayed in blue. (c) DAPC plot for 23 cannabis genotypes along two linear discriminants (LD 1 and LD 2).

<https://doi.org/10.1371/journal.pone.0253387.g001>

plot of 23 cannabis genotypes is shown in Fig 1(c). The grouping assignment for individual strains by DAPC is listed in Table 1 (as W-SNPs). C1 is a THC dominant cluster and includes six THC dominant strains (11, 17, 18, 20, 21, and 23-THC) and one balanced strain (1-balanced). C2 is another THC dominant cluster and includes five THC dominant strains (12, 14, 15, 16, and 19-THC). C3 is a cluster dominated by CBD dominant and the balanced strains which includes six CBD dominant strains (3, 4, 5, 6, 8, and 10-CBD), three balanced strains (2, 7, and 9-balanced), and two THC dominant strains (13 and 22-THC). While C2 is closer to C3 and is more distant to C1, C1 and C3 are clearly separated along linear discriminant 1 (LD1). While C1 and C3 are roughly at the same level with respect to linear discriminant 2 (LD2), C2 is separated from both. PCA was also carried out on the same set of SNPs and results are shown in S1 Fig. Twenty-three cannabis strains are plotted along pair-wise PCs of the first 4 PCs, which account for 18.4%, 11.5%, 9.5%, and 8.7% of the total variance, respectively. Similarly, the first PC suggests the existence of a relatively compact CBD & balanced clade on the left side of the plot and a more dispersed THC dominant clade on the right side of the plot. Balanced strains share a closer gene pool with CBD dominant strains, while the THC gene pool is more dispersed. Because THC is psychoactive and its potency can be readily assessed through consumption, selection for increasing THC content started early and widely for recreational purposes by traditional breeding [45]. In contrast, CBD is non-psychoactive and must be analyzed in a laboratory for potency, and therefore breeding for high CBD concentrations began later [45]. A complete genome assembly implied that CBD dominant varieties were generated by integrating hemp-type CBD acid synthase gene clusters into a background of drug-type cannabis to elevate CBDA production [24]. These balanced strains may have been created by crossing purebred THC dominant types with CBD dominant types [46]. Therefore, there may be a relatively limited selection of CBD dominant strains for breeding balanced strains.

Discriminant analysis of principal components using 344 structural SNPs

DAPC was repeated using identified 344 structural SNPs. The optimal number of identified clusters was five, corresponding to the lowest BIC values (Fig 2(a)). Two PCs were retained for the following DAPC analysis in Fig 2(b), where it had 98.9% predictive success and 0.04% RMSE. For the subsequent DAPC analysis, two PCs and two discriminant functions were retained. The grouping assignment for individual strains by DAPC is listed in Table 1 (as I-SNPs). Within the five clusters (Fig 2(c)), C1 is a CBD dominant cluster that includes six strains (3, 4, 5, 6, 8, and 10-CBD), C2 includes three balanced strains (2, 7, and 9-balanced), and C3, C4, and C5 are THC dominant clusters that include two (13 and 22-THC), seven (1-balanced, 11, 15, 18, 20, 21, 23-THC), and five (12, 14, 16, 17, and 19-THC) strains, respectively.

These multiallelic SNPs were also subjected to PCA, NJ tree, and hierarchical clustering analysis. In Fig 3, the 23 cannabis strains are plotted along PC1 and PC2, which account for 44.5% and 10.0% of the total variance, respectively. The proportions of explained variance are higher compared to the previous PCA results (18.4% and 11.5%) obtained using the whole set of SNPs. CBD dominant cluster C1 and balanced cluster C2 are on the left side of the scatter plot ($PC1 < 0$) and the THC dominant clusters C3, C4, and C5 are on the right side of the scatter plot ($PC1 > 0$). Notably, six CBD dominant strains are separated from three balanced strains, while they were previously combined in the analysis using the whole set of SNPs. In addition, two THC dominant strains 13-THC and 22-THC are separated from the CBD and balanced cluster, and instead placed closer to other THC dominant strains. Strain 1-balanced is closer to THC dominant strain regardless of whether the whole set of SNPs or 344 identified SNPs were used.

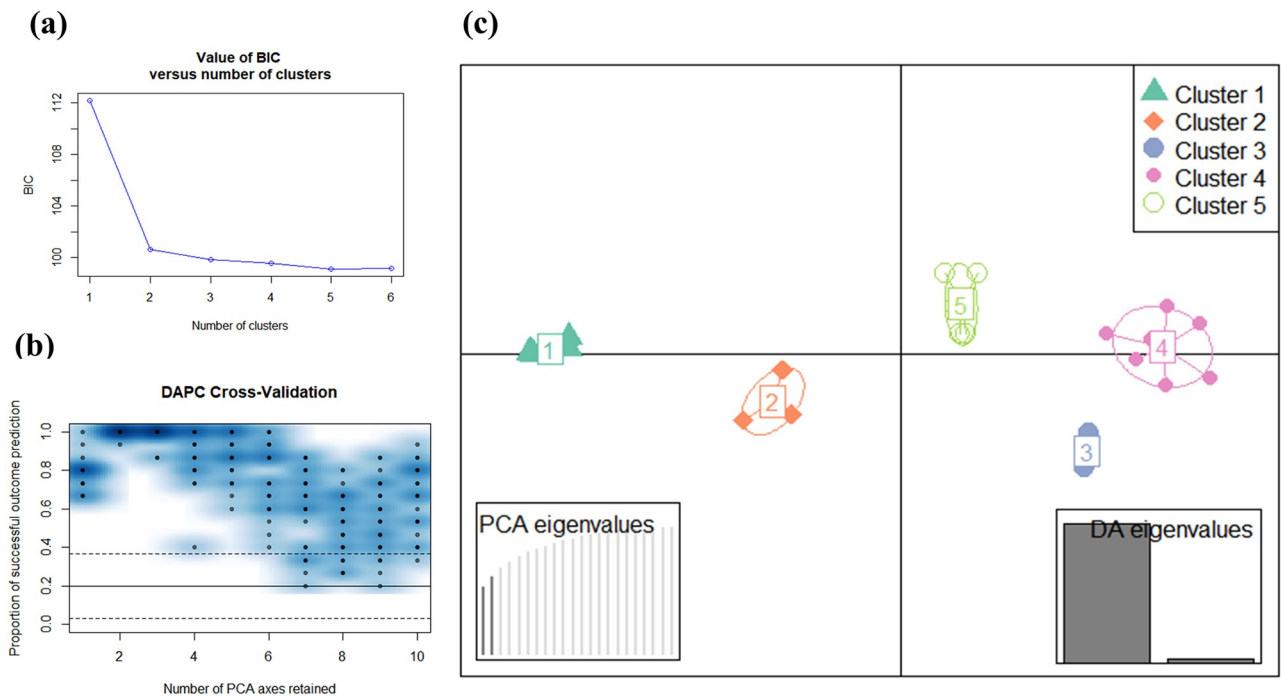


Fig 2. DAPC of 23 cannabis genotypes using 344 multiallelic structural SNPs. Clusters indicated as C1, C2, C3, C4, and C5 corresponds to the I-SNPs in Table 1.

<https://doi.org/10.1371/journal.pone.0253387.g002>

The genetic structure from NJ-tree and hierarchical clustering using the 344 multiallelic are displayed in Fig 4, mostly congruent with that of DAPC. In the NJ-tree, all six CBD dominant strains are clustered together, with three balanced strains clustered closer on the same branch (Fig 4(a)). Most THC dominant strains are also clustered adjacent to strains within their own clusters. The dendrogram using hierarchical clustering by Ward's method reveals two major groups, where one group is comprised of CBD dominant & balanced strains, and the other of THC dominant strains (Fig 4(b)). They are further separated into five subclusters, where CBD dominant and balanced clusters are consistent with the DAPC grouping results, and several THC dominant strains clustered differently. Two strains, 15-THC and 18-THC, were assigned to C4 using DAPC but are assigned closer to C5 in the dendrogram. Two other strains, 14-THC and 16-THC, were assigned to C5 in DAPC but are assigned closer to C3 in the dendrogram. The clustering results are congruent between DAPC and hierarchical clustering with an assignment agreement rate of 83% (19/23).

Allele frequencies for 344 multiallelic SNPs in three chemotypes

DAPC identified 344 highly contributing SNPs (S1 Table). All the structural SNPs are multiallelic, among which 98.5% (339/344) are tri-allelic and the remainder 1.5% (5/344) are tetra-allelic. The dendrogram of 23 strains using hierarchical clustering based on the allele counts in the 344 structural SNPs (S2 Table) separated the strains into CBD dominant, balanced, and THC dominant strains, mostly corresponding to the grouping results of DAPC (Fig 5). The allele frequency was calculated by dividing the counts of that allele for all strains within the targeted group by the sum of the counts for each allele for that SNP within the targeted group. Allele frequencies of the structural SNPs were calculated for three major branches, each corresponding one of three chemotypes. (S1 Table). If 1-balanced strain was assigned to the THC

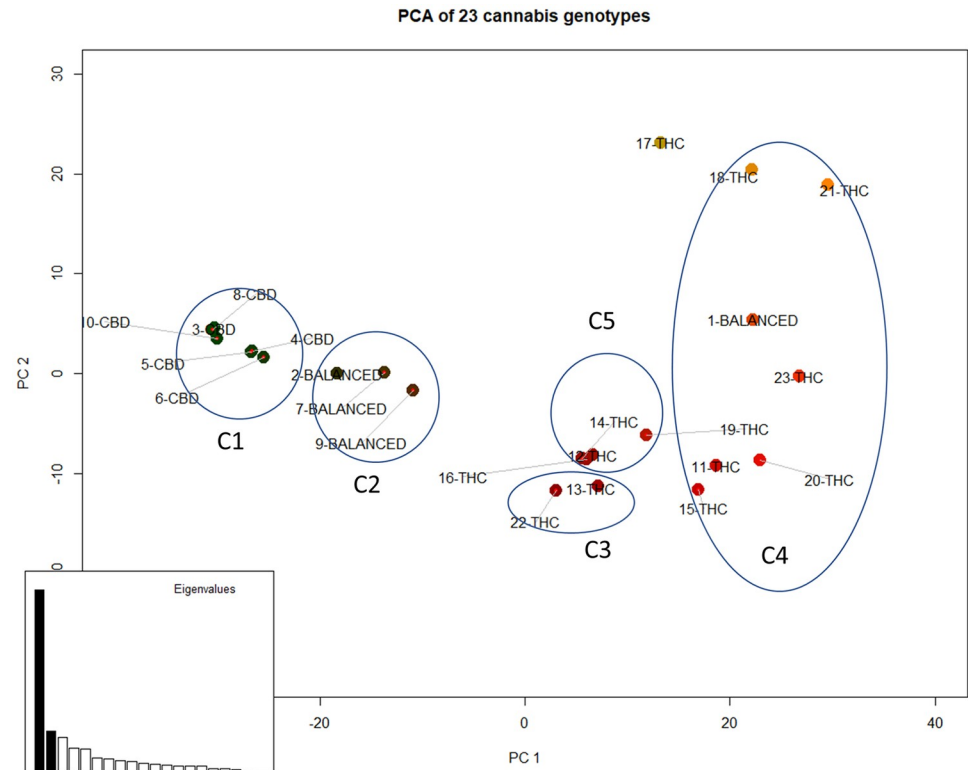


Fig 3. Scatter plot of 23 cannabis strains on PC1 & PC2 using 344 structural SNPs. Clusters indicated as C1, C2, C3, C4 and C5 correspond to I-SNPs in Table 1.

<https://doi.org/10.1371/journal.pone.0253387.g003>

dominant group as indicated by DAPC for allele frequency calculation, there are 87% (300/344) SNPs in CBD dominant clusters, 46% (157/344) SNPs in balanced clusters, and 11% (39/344) SNPs in THC dominant clusters that have one allele with allele frequencies > 80% (S1 Table). Among them, 140 SNPs shared same alleles with allele frequencies > 80% in CBD dominant strains (140/300) and balanced strains (140/157), which further indicated that CBD dominant strains and balanced strains closely share a gene pool. There are 38 SNPs that have one allele present in CBD dominant strains with allele frequencies > 80% and are not detected in THC dominant strains. There are 322 SNPs whose alleles that are present in THC dominant strains but were not detected in CBD dominant strains.

If the 1-balanced strain is assigned to the balanced group for allele frequency calculation, there are 87% (300/344) SNPs in CBD dominant clusters, 10% (36/344) SNPs in balanced clusters, and 13% (44/344) SNPs in THC dominant clusters that have one allele with allele frequencies > 80% (S2 Table). Among them, 32 SNPs shared same alleles with allele frequencies > 80% in CBD dominant strains (32/300) and balanced strains (32/36). There are 38 SNPs that have one allele present in CBD dominant strains with allele frequencies > 80% and are not detected in THC dominant strains. There are 321 SNPs whose alleles are present in THC dominant strains but were not detected in CBD dominant strains. Assigning the 1-balanced strain to the balanced group added more genetic diversity to the balanced group, and the effect of adding or deleting this strain for the THC dominant group in terms of allele frequency is small and can be neglected.

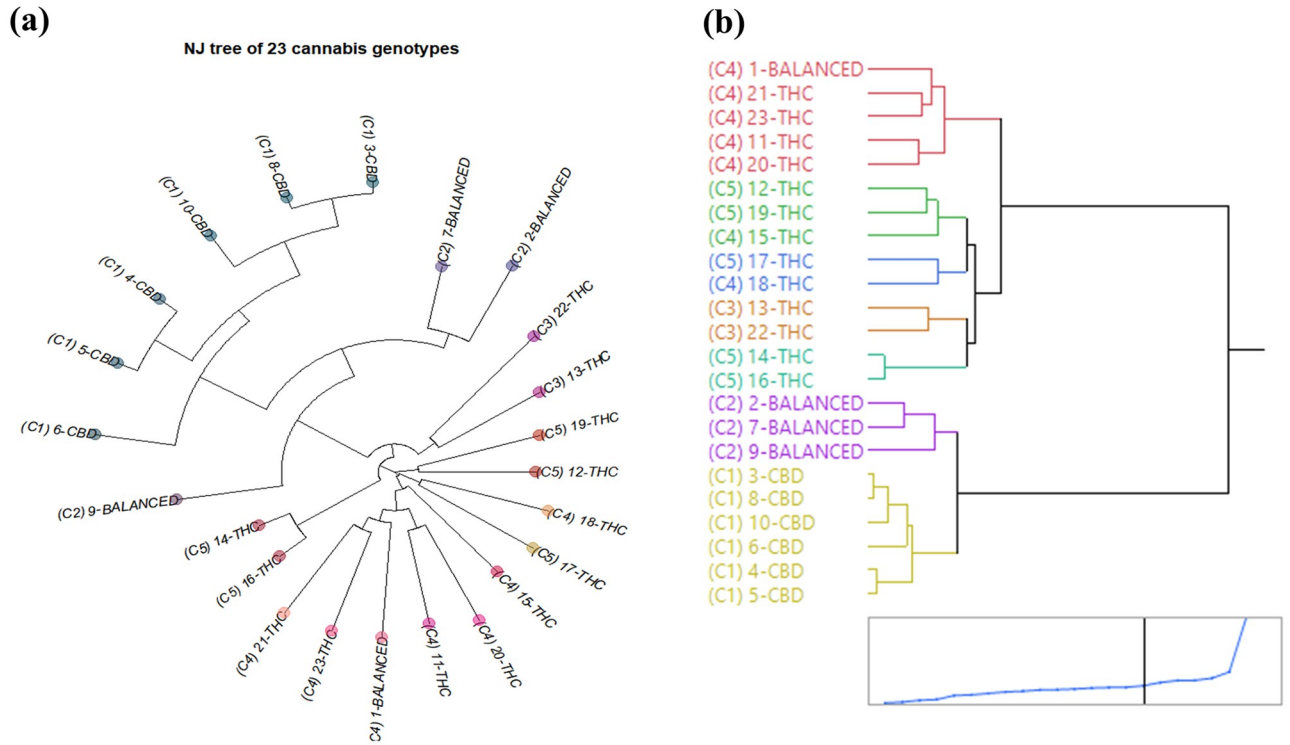


Fig 4. NJ-tree and hierarchical clustering using the 344 multiallelic SNPs (a) NJ-tree and (b) The dendrogram using hierarchical clustering by Ward’s method for 23 cannabis genotypes. Clusters indicated as C1, C2, C3, C4, and C5 corresponds to I-SNPs in Table 1.

<https://doi.org/10.1371/journal.pone.0253387.g004>

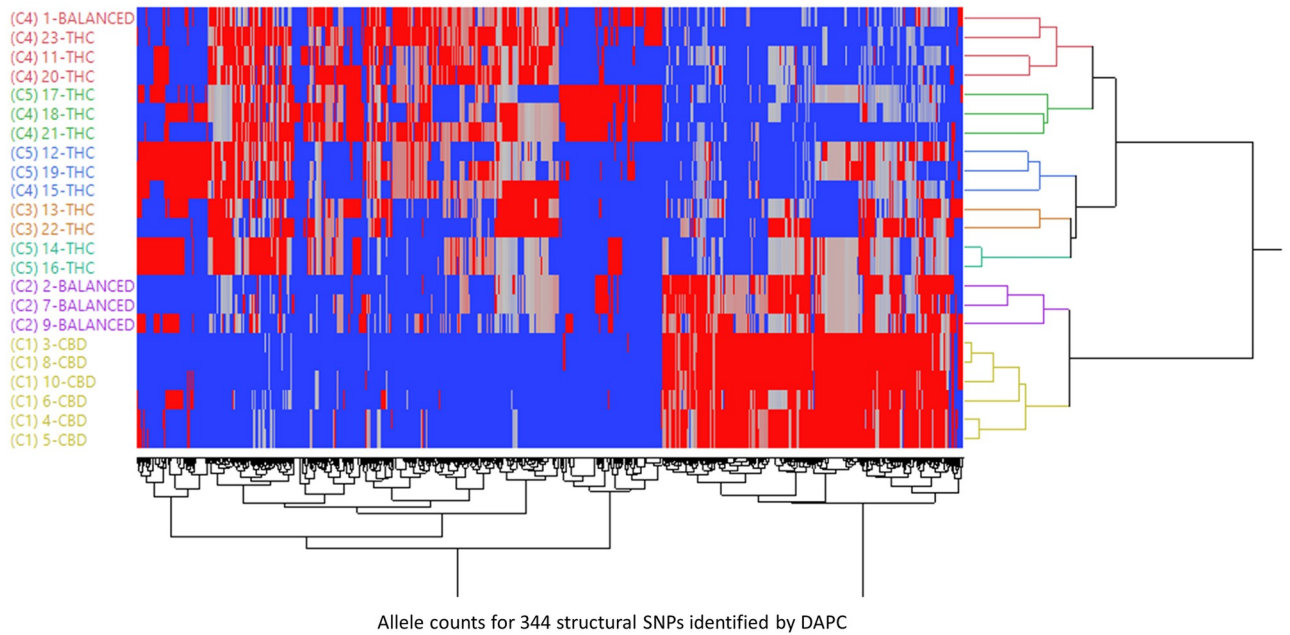


Fig 5. Hierarchical clustering of 23 strains based on the allele counts for 344 structural SNPs identified by DAPC.

<https://doi.org/10.1371/journal.pone.0253387.g005>

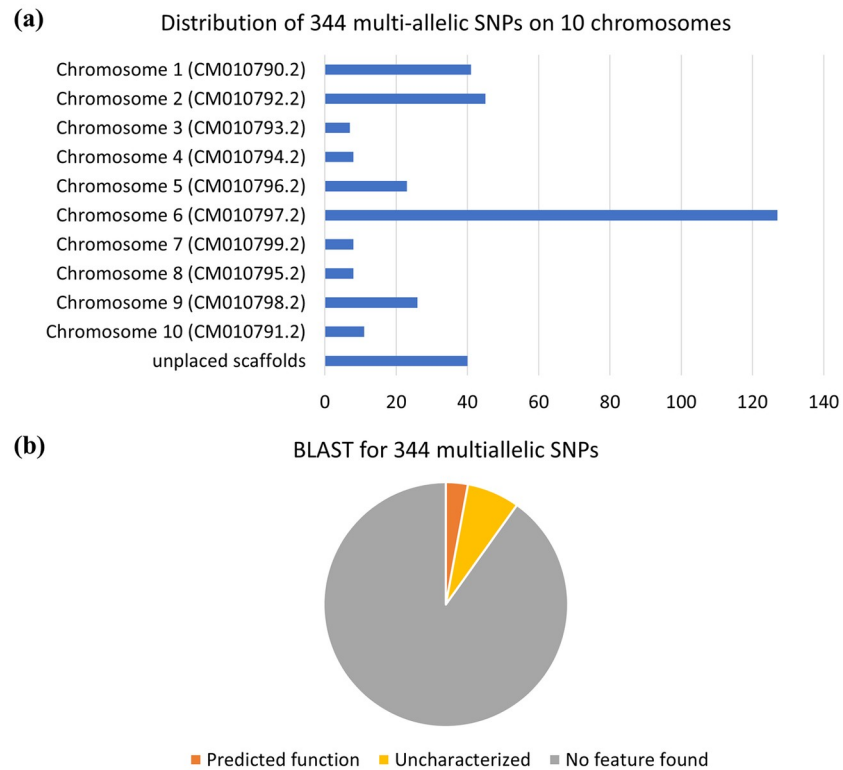


Fig 6. Features of 344 multiallelic SNPs (a) Distribution of structural SNPs on chromosome 1–10 and unplaced scaffolds. (b) BLAST results for structural SNPs against a fully annotated genome.

<https://doi.org/10.1371/journal.pone.0253387.g006>

BLAST analysis of 344 multiallelic SNPs

These 344 SNPs were spread across all 10 chromosomes (Fig 6(a)), indicating that commercially available cannabis strains in North America are significantly differentiated at a genome-wide level. The number of identified SNPs ranged from 7 to 127 on each genome, with 37% of the genetic variation occurring (127 SNPs) on chromosome 6, where CBDAS and THCAS are located [13]. The rest SNPs were spread over the remaining nine chromosomes. All ten chromosomes have genes related to the biochemical pathways of secondary metabolites, including cannabinoids, monoterpenes, and sesquiterpenes [13, 24, 47–51]. BLAST results showed that 90% (310/344) of these structural SNPs had no feature, 7% (24/344) are uncharacterized loci with unknown functions, and 3% (10/344) are predicted for certain functions (Fig 6(b)).

Conclusions

Although the cannabis industry is rapidly advancing after the relaxation of legal restrictions in North America, the increasing number of THC dominant strains, CBD dominant strains, and balanced strains only adds confusion to the currently poorly understood genetic background of the thousands of varieties already in existence. Although there were only 23 strains included in this study, they covered the three typical chemotypes of cannabis strains currently available in the market. Leveraging as much genetic variation as possible using whole-genome sequencing, we identified 344 multiallelic SNPs that were used to investigate the genetic structure of 23 cannabis genotypes using DAPC, PCA, NJ tree, and hierarchical clustering, which provided consistent observations and groupings despite the differences in algorithms. The clustering results revealed that these 23 strains could be separated into five clusters, with one cluster

containing six CBD dominant strains, another cluster containing three balanced strains, and the remaining three clusters containing 13 THC dominant strains and one balanced strain. CBD dominant strains and the balanced strains are closer genetically. This may be attributed to how medical interest in breeding for non-psychoactive, CBD-elevated strains (CBD dominant and balanced strains) has only recently been in vogue, resulting in an overlapping and less diverse gene pool for CBD dominant and balanced strains compared to the longer breeding history for THC strains. Some alleles are only present in CBD dominant strains or in THC dominant strains. More alleles present in balanced strains are shared with CBD dominant strains. One third of these structural SNPs are located on the chromosome containing THCAS and CBDAS. The remaining SNPs are located on the other nine chromosomes. An area of potential investigation is how the identified structural SNPs are associated with the production of other cannabinoids, mono- and sesquiterpenes, flavonoids, other compounds, or morphological characteristics.

Since the late 20th century, genetic methodologies have been developed for separating industrial hemp from drug-type cannabis for forensic purposes, thus differentiating CBD dominant and THC dominant strains [52–56]. For the past 20 years, with the extensive hybridization of THC dominant strains, many classification studies have focused on separating “Sativa” and “Indica” strains and many have suggested abolishing this vernacular [5–7]. The genotyping results of this study indicate that modern, extensively hybridized strains can still be separated using genome-wide information. As a powerful multivariate approach that investigates population structures based solely on genetic information, DAPC separated strains into clusters aligned with their chemotypes. Additionally, DAPC has the potential to sort the disordered genetic background of thousands of THC dominant strains by identifying the number of genetic clusters within THC dominant strains, describing clusters by interpreting group memberships, and identifying the contributing SNPs that have the potential to be used as genetic markers for strain classification and identification. This would require a concerted effort from the cannabis industry by contributing whole genome sequence data to public databases and by building a common taxonomy based on genomics. Optimally, the identified genetic markers can be used as genomic fingerprints in combination with chemical fingerprints and morphological characteristics for strain identification. These markers can be leveraged for strain selection in clinical trials and for manufacturing cannabis-based products and medicines.

Supporting information

S1 Fig. PCA of 23 strains using whole set of SNPs.

(PDF)

S1 Table. 344 multiallelic SNPs identified by DAPC.

(XLSX)

S2 Table. Allele counts for 344 structural SNPs identified by DAPC.

(XLSX)

Acknowledgments

The authors are grateful to licensed grower Emerald Flower Farm who provided commercial greenhouse to cultivate cannabis. The authors are also grateful to Dr. Limin Wu for assisting DNA extraction, Dr. Jie Zeng for assisting BLAST analysis, and Shengxi Jin for proofreading the manuscript.

Author Contributions

Conceptualization: Dan Jin.

Data curation: Dan Jin.

Formal analysis: Dan Jin, Philippe Henry.

Funding acquisition: Dan Jin, Jacqueline Shan.

Investigation: Dan Jin.

Methodology: Dan Jin, Philippe Henry.

Project administration: Dan Jin.

Resources: Dan Jin.

Software: Dan Jin, Philippe Henry.

Supervision: Jie Chen.

Validation: Dan Jin, Philippe Henry.

Visualization: Dan Jin, Philippe Henry.

Writing – original draft: Dan Jin.

Writing – review & editing: Philippe Henry, Jacqueline Shan, Jie Chen.

References

1. Hillig KW (2005) A systematic investigation of Cannabis, PhD thesis, Indiana University. PhD Thesis
2. McPartland JM (2017) Cannabis sativa and Cannabis indica versus “Sativa” and “Indica”. In: Cannabis sativa L.—Botany and Biotechnology. Springer, pp 101–121
3. McPartland JM, Guy GW (2017) Models of Cannabis Taxonomy, Cultural Bias, and Conflicts between Scientific and Vernacular Names. *The Botanical Review* 4:327–381
4. Knight G, Hansen S, Connor M, Poulsen H, McGovern C, Stacey J (2010) The results of an experimental indoor hydroponic Cannabis growing study, using the ‘Screen of Green’(ScrOG) method—Yield, tetrahydrocannabinol (THC) and DNA analysis. *Forensic science international* 202:36–44 <https://doi.org/10.1016/j.forsciint.2010.04.022> PMID: 20462712
5. Sawler J, Stout JM, Gardner KM, Hudson D, Vidmar J, Butler L, et al. (2015) The genetic structure of marijuana and hemp. *PLoS one* 10:e0133292 <https://doi.org/10.1371/journal.pone.0133292> PMID: 26308334
6. Lynch RC, Vergara D, Tittes S, White K, Schwartz CJ, Gibbs MJ, et al. (2016) Genomic and chemical diversity in Cannabis. *Critical Reviews in Plant Sciences* 35:349–363
7. Henry P (2015) Genome-wide analyses reveal clustering in Cannabis cultivars: the ancient domestication trilogy of a panacea. <https://doi.org/10.7287/peerj.preprints.1553v2>
8. McGuire P, Robson P, Cubala WJ, Vasile D, Morrison PD, Barron R, et al. (2018) Cannabidiol (CBD) as an adjunctive therapy in schizophrenia: a multicenter randomized controlled trial. *Am J Psychiatry* 175:225–231 <https://doi.org/10.1176/appi.ajp.2017.17030325> PMID: 29241357
9. French J, Thiele E, Mazurkiewicz-Beldzinska M, Benbadis S, Marsh E, Joshi C, et al. (2017) Cannabidiol (CBD) significantly reduces drop seizure frequency in Lennox-Gastaut syndrome (LGS): results of a multi-center, randomized, double-blind, placebo controlled trial (GWPCARE4)(S21.001). *Neurology* 88:S21–001
10. Avraham Y, Grigoriadis NC, Poutahidis T, Vorobiev’ L, Magen I, Ilan Y, et al. (2011) Cannabidiol improves brain and liver function in a fulminant hepatic failure-induced model of hepatic encephalopathy in mice. *British journal of pharmacology* 162:1650–1658 <https://doi.org/10.1111/j.1476-5381.2010.01179.x> PMID: 21182490
11. Bloomfield MAP, Green SF, Hindocha C, et al (2020) The effects of acute cannabidiol on cerebral blood flow and its relationship to memory: An arterial spin labelling magnetic resonance imaging study. *Journal of psychopharmacology (Oxford, England)* 269881120936419 <https://doi.org/10.1177/0269881120936419> PMID: 32762272

12. Upton R, Craker L, ElSohly M, Romm A, Russo E, Sexton M (2014) Cannabis inflorescence: cannabis spp.; standards of identity, analysis, and quality control. American Herbal Pharmacopoeia, Scotts Valley, CA <https://doi.org/10.4081/ijfs.2020.8581> PMID: 32913724
13. Kovalchuk I, Pellino M, Rigault P, et al (2020) The genomics of cannabis and its close relatives. *Annu Rev Plant Biol* 71:713–739 <https://doi.org/10.1146/annurev-arplant-081519-040203> PMID: 32155342
14. Sakamoto K, Akiyama Y, Fukui K, Kamada H, Satoh S (1998) Characterization; Genome Sizes and Morphology of Sex Chromosomes in Hemp (*Cannabis sativa* L.). *Cytologia* 63:459–464
15. Shastry BS (2007) SNPs in disease gene mapping, medicinal drug development and evolution. *Journal of Human Genetics* 52:871–880 <https://doi.org/10.1007/s10038-007-0200-z> PMID: 17928948
16. Sachidanandam R, Weissman D, Schmidt SC, et al (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933 <https://doi.org/10.1038/35057149> PMID: 11237013
17. Casci T (2010) SNPs that come in threes. *Nature Reviews Genetics* 11:8–8 <https://doi.org/10.1038/nrg2725> PMID: 20050277
18. Westen AA, Matai AS, Laros JFJ, Meiland HC, Jasper M, de Leeuw WJF, et al. (2009) Tri-allelic SNP markers enable analysis of mixed and degraded DNA samples. *Forensic Science International: Genetics* 3:233–241
19. Kalbfleisch T, Petersen JL, Tait RG Jr., Qiu J, Basnayake V, Hackett PH, et al. (2020) Using triallelic SNPs for determining parentage in North American yak (*Bos grunniens*) and estimating cattle (*B. taurus*) introgression. *F1000Res* 9:1096 <https://doi.org/10.12688/f1000research.25803.2> PMID: 33163159
20. Phillips C, Amigo J, Tillmar AO, et al (2020) A compilation of tri-allelic SNPs from 1000 Genomes and use of the most polymorphic loci for a large-scale human identification panel. *Forensic Science International: Genetics* 46:102232 <https://doi.org/10.1016/j.fsigen.2020.102232> PMID: 31986343
21. Henry P, Khatodia S, Kapoor K, et al (2020) A single nucleotide polymorphism assay sheds light on the extent and distribution of genetic diversity, population structure and functional basis of key traits in cultivated north American cannabis. *Journal of Cannabis Research* 2:26 <https://doi.org/10.1186/s42238-020-00036-y> PMID: 33526123
22. Vergara D, Baker H, Clancy K, Keepers KG, Mendieta JP, Pauli CS, et al. (2016) Genetic and genomic tools for *Cannabis sativa*. *Critical Reviews in Plant Sciences* 35:364–377
23. Lavery KU, Stout JM, Sullivan MJ, et al (2019) A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci. *Genome Res* 29:146–156 <https://doi.org/10.1101/gr.242594.118> PMID: 30409771
24. Grassa CJ, Wenger JP, Dabney C, Poplawski SG, Motley ST, Michael TP, et al. (2018) A complete *Cannabis* chromosome assembly and adaptive admixture for elevated cannabidiol (CBD) content. *bioRxiv* 458083
25. Jenkins C, Orsburn B (2019) Constructing a Draft Map of the Cannabis Proteome. *bioRxiv* 577635
26. Jenkins C, Orsburn B (2019) The Cannabis Multi-Omics Draft Map Project. *bioRxiv* 753400
27. Jenkins C, Orsburn B (2019) The First Publicly Available Annotated Genome for Cannabis plants. *bioRxiv* 786186
28. Benevenuto J, Ferrão LFV, Amadeu RR, Munoz P (2019) How can a high-quality genome assembly help plant breeders? *Gigascience*. <https://doi.org/10.1093/gigascience/giz068> PMID: 31184361
29. Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11:94 <https://doi.org/10.1186/1471-2156-11-94> PMID: 20950446
30. Henry P (2017) Cannabis chemovar classification: terpenes hyper-classes and targeted genetic markers for accurate discrimination of flavours and effects. *PeerJ Preprints* 5:e3307v1
31. Henry P (2018) The genetic basis of the human-cannabis relationship. *bioRxiv* 287938
32. Jin D, Jin S, Yu Y, Lee C, Chen J (2017) Classification of Cannabis Cultivars Marketed in Canada for Medical Purposes by Quantification of Cannabinoids and Terpenes Using HPLC-DAD and GC-MS. *J Anal Bioanal Tech* 8:2
33. Jin D, Dai K, Xie Z, Chen J (2020) Secondary Metabolites Profiled in Cannabis Inflorescences, Leaves, Stem Barks, and Roots for Medicinal Purposes. *Scientific Reports* 10:3309 <https://doi.org/10.1038/s41598-020-60172-6> PMID: 32094454
34. ASM23057v4—Genome—Assembly—NCBI. https://www.ncbi.nlm.nih.gov/assembly/GCA_000230575.4/.
35. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26:589–595 <https://doi.org/10.1093/bioinformatics/btp698> PMID: 20080505

36. ASM23057v5—Genome—Assembly—NCBI. https://www.ncbi.nlm.nih.gov/assembly/GCA_000230575.5.
37. Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405 <https://doi.org/10.1093/bioinformatics/btn129> PMID: 18397895
38. R Core Team (2018) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
39. Liu N, Zhao H (2006) A non-parametric approach to population structure inference using multilocus genotypes. *Human Genomics* 2:353 <https://doi.org/10.1186/1479-7364-2-6-353> PMID: 16848973
40. Konishi S, Kitagawa G (2008) Information Criteria and Statistical Modeling. <https://doi.org/10.1007/978-0-387-71887-3>
41. Deperi SI, Tagliotti ME, Bedogni MC, Manrique-Carpintero NC, Coombs J, Zhang R, et al. (2018) Discriminant analysis of principal components and pedigree assessment of genetic diversity and population structure in a tetraploid potato panel using SNPs. *PLoS One* 13:e0194398 <https://doi.org/10.1371/journal.pone.0194398> PMID: 29547652
42. Cannabis sativa Annotation Report. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Cannabis_sativa/100/.
43. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406–425 <https://doi.org/10.1093/oxfordjournals.molbev.a040454> PMID: 3447015
44. Ward JH Jr (1963) Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* 58:236–244
45. Clarke RC, Merlin MD (2016) Cannabis domestication, breeding history, present-day genetic diversity, and future prospects. *Critical Reviews in Plant Sciences* 35:293–327
46. de Meijer EPM, Bagatta M, Carboni A, Crucitti P, Moliterni VMC, Ranalli P, et al. (2003) The inheritance of chemical phenotype in Cannabis sativa L. *Genetics* 163:335–346 PMID: 12586720
47. Stout JM, Boubakir Z, Ambrose SJ, Purves RW, Page JE (2012) The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in Cannabis sativa trichomes. *The Plant Journal*. <https://doi.org/10.1111/j.1365-313x.2012.04949.x> PMID: 22353623
48. Gagne SJ, Stout JM, Liu E, Boubakir Z, Clark SM, Page JE (2012) Identification of olivetolic acid cyclase from Cannabis sativa reveals a unique catalytic route to plant polyketides. *Proceedings of the National Academy of Sciences* 109:12811–12816 <https://doi.org/10.1073/pnas.1200330109> PMID: 22802619
49. Van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, et al. (2011) The draft genome and transcriptome of Cannabis sativa. *Genome biology* 12:R102 <https://doi.org/10.1186/gb-2011-12-10-r102> PMID: 22014239
50. Zager JJ, Lange I, Srividya N, Smith A, Lange BM (2019) Gene Networks Underlying Cannabinoid and Terpenoid Accumulation in Cannabis. *Plant Physiology* 180:1877–1897 <https://doi.org/10.1104/pp.18.01506> PMID: 31138625
51. Booth JK, Page JE, Bohlmann J (2017) Terpene synthases from Cannabis sativa. *PLOS ONE* 12: e0173911 <https://doi.org/10.1371/journal.pone.0173911> PMID: 28355238
52. Piluzza G, Delogu G, Cabras A, Marceddu S, Bullitta S (2013) Differentiation between fiber and drug types of hemp (Cannabis sativa L.) from a collection of wild and domesticated accessions. *Genetic resources and crop evolution* 60:2331–2342
53. Datwyler SL, Weiblen GD (2006) Genetic variation in hemp and marijuana (Cannabis sativa L.) according to amplified fragment length polymorphisms. *J Forensic Sci* 51:371–375 <https://doi.org/10.1111/j.1556-4029.2006.00061.x> PMID: 16566773
54. Gilmore S, Peakall R, Robertson J (2003) Short tandem repeat (STR) DNA markers are hypervariable and informative in Cannabis sativa: implications for forensic investigations. *Forensic Sci Int* 131:65–74 [https://doi.org/10.1016/s0379-0738\(02\)00397-3](https://doi.org/10.1016/s0379-0738(02)00397-3) PMID: 12505473
55. Hakki EE, Kayis SA, Pinarkara E, Sag A (2007) Inter simple sequence repeats separate efficiently hemp from marijuana (Cannabis sativa L.). *Electronic Journal of Biotechnology* 10:570–581
56. Hilyard A, Lewin S, Johnson S, Henry P, Orser C (2019) Application of a Simple Genetic Assay to Discriminate Hemp from Drug-Type Cannabis. In: Cannabis Science Tech. <https://www.cannabissciencetech.com/view/application-simple-genetic-assay-discriminate-hemp-drug-type-cannabis>.