

SCIENTIFIC REPORTS

OPEN

Assessment of Genetic Diversity and Population Structure in Iranian Cannabis Germplasm

Aboozar Soorni^{1,2}, Reza Fatahi¹, David C. Haak³, Seyed Alireza Salami¹ & Aureliano Bombarely^{1,2} 

Cannabis sativa has a complex history reflected in both selection on naturally occurring compounds and historical trade routes among humans. Iran is a rich resource of natural populations which hold the promise to characterize historical patterns of population structure and genetic diversity within *Cannabis*. Recent advances in high-throughput DNA sequencing technologies have dramatically increased our ability to produce information to the point that it is now feasible to inexpensively obtain population level genotype information at a large scale. In the present investigation, we have explored the use of Genotyping-By-Sequencing (GBS) in Iranian cannabis. We genotyped 98 cannabis samples 36 from Iranian locations and 26 accessions from two germplasm collections. In total, 24,710 high-quality Single Nucleotide Polymorphisms (SNP) were identified. Clustering analysis by Principal Component Analysis (PCA) identified two genetic clusters among Iranian populations and fineSTRUCTURE analysis identified 19 populations with some geographic partitioning. We defined Iranian cannabis in two main groups using the results of the PCA and discovered some strong signal to define some locations as population according to fineSTRUCTURE analyses. However, single nucleotide variant analysis uncovered a relatively moderate level of variation among Iranian cannabis.

Cannabis sativa L. is a dioecious species in the Cannabaceae family¹ with a broad global distribution which is likely the result of human cultivation. Humans have cultivated the plant as a source of fiber, food, medicines, intoxicants and oils for thousands of years^{1,2}. This use and breeding has led to the selection of two distinct types of *C. sativa*, one for fibre and seed (hemp) and one for medicinal use (marijuana). While these types are morphologically similar, they are distinguished by the type and level of cannabinoids produced. Levels of two types of cannabinoids in particular are used to distinguish marijuana and hemp *C. sativa*. First, D-9-tetrahydrocannabinol (THC) is a psychoactive compound³ found in leaves and inflorescences (but not seeds) of juvenile and mature plants. The second compound, cannabidiol (CBD), is an isomer of THC found in all plant tissues, however, this cannabinoid does not activate cannabinoid receptors^{1,4,5}. Marijuana varieties used for drug consumption are characterized by a high THC content, whereas fibre varieties (hemp) produce CBD as the predominant cannabinoid^{6,7}.

Archaeological and palaeobotanical evidence supports the cultivation and use of *Cannabis* since the Neolithic period with subsequent secondary domestication events in geographical regions outside of the accepted native range^{8–15}. For instance, archaeological evidence for the pharmaceutical or shamanistic use of *Cannabis* has been found in cave artifacts that include a large cache of *Cannabis* dating to ca. 700 BCE¹⁶. This long history of use has resulted in a complex biogeographical history for this species. Based on polymorphism in RAPD markers, the Eurasian Steppe region of Central Asia has been recognized as a putative center of origin for *Cannabis*, spreading from there to the Mediterranean as well as Eastern and Central European countries, in particular, Afghanistan and Pakistan¹⁷. However, the genus has also been described as having two centers of diversity, Hindustani and European–Siberian¹⁸. As with other cultivated plants it is difficult to pinpoint the exact place of origin for *C. sativa*. It is likely that *Cannabis* spread to ancient Persia very early, assisted by Aryan and Scythian tribes expanding westward from central Asia. Evidence for this early spread comes from archeological studies of the Scythians, who occupied an area encompassing large swathes of what is now northwest Iran from the 7th century BCE to the 4th century CE, this culture was known to use *Cannabis* for entertainment and spiritual purposes.

¹Department of Horticulture Sciences, Faculty of Agriculture, University of Tehran, Karaj, 31587, Iran. ²Department of Horticulture, Virginia Tech, Blacksburg, VA, 24061, USA. ³Department of Plant Pathology, Physiology, and Weed Science, Virginia Tech, Blacksburg, VA, 24061, USA. Correspondence and requests for materials should be addressed to S.A.S. (email: asalami@ut.ac.ir) or A.B. (email: aurebg@vt.edu)

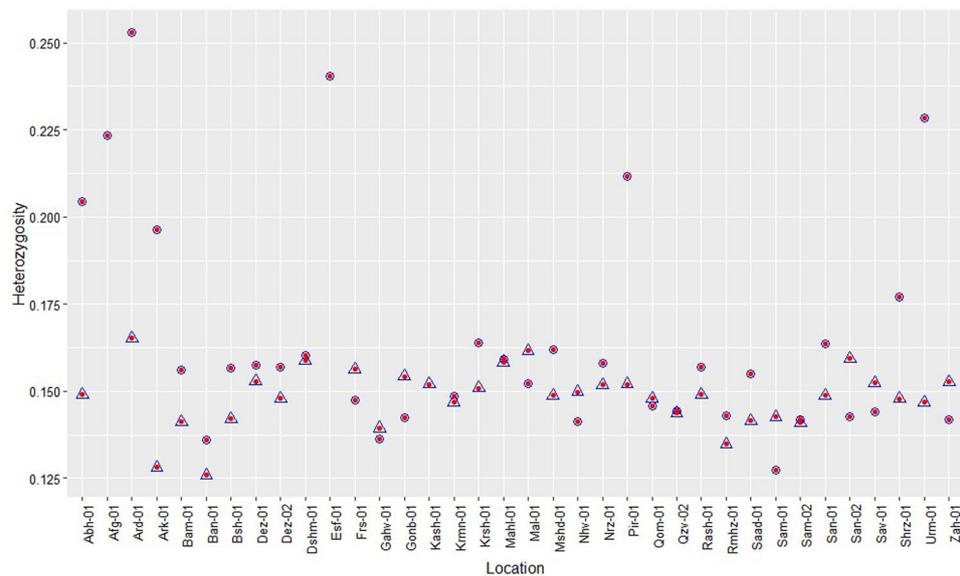


Figure 2. Heterozygosity per location. Triangles represent male samples and circles represent female samples.

(A/C, A/T, C/G or G/T) accounted for 37.3%. The ratio of transitions to transversions is consistent with other studies in various species^{36–39}. Tajima's D was calculated for the filtered SNPs with a mean of -0.18 (range, -2.16 to 3.70) across all samples (Fig. S2) and a mean of 0.007 (range, -1.95 to 3.55) for the 68 samples originally from Iran. Tajima's D is a summary statistic often used for identifying selective sweeps from genomic data, where values are 0 for neutral variation, positive when an excess of rare polymorphism indicates positive selection, and negative with an excess of high-frequency variants, which indicates balancing selection⁴⁰. The distribution of Tajima's D among Iranian cannabis samples suggests that balancing selection likely shaped genetic structure across these populations (Fig. S2). This pattern is common among groups that experience heterozygote advantage, wherein rare alleles are retained at low frequencies. Average heterozygosity was estimated at 0.15 across 68 samples originally from Iran and an Afghanistan. This estimate of heterozygosity is similar to that found by Sawler *et al.*³⁰ for marijuana type accessions. Samples Ard-01-F and Esf-01-F from Ardabil and Isfahan states showed the highest number of heterozygous sites (Fig. 2, Table S1).

Population differentiation resulting from genetic structure was estimated using F_{ST} . For the Iranian samples, the minimum F_{ST} was -0.058 , calculated between Saad and Esf locations, and the maximum F_{ST} was 0.26 for Gahv vs Ard, locations that are separated by 434 km (Fig. 1, Table S2). Low values indicate that genetic diversity is higher within individuals from these locations than between locations, a pattern consistent with gene flow between populations. F_{ST} estimates above 0 indicate a reduction in genetic exchange between population with a value of 1 indicating complete isolation. Across all individuals the maximum F_{ST} , 0.425, was estimated between non-Iranian samples 883049_vs_CAN37. Sample 883049 (from kompolti Sargászárú) has been identified as a fiber cultivar⁴¹. CAN37 was previously described as hemp type and originating in France, however, Sawler *et al.*³⁰ found that it was a distinct outlier and was more closely associated with marijuana and speculated that it could be a mislabeled sample. We also estimated genetic differentiation among marijuana and hemp accessions and Iranian samples and found a larger F_{ST} across hemp 0.086 than marijuana 0.039.

Nei's genetic distance⁴² was evaluated on 13,325 SNPs that were identified across 209 samples (all data, including that from Sawler *et al.*³⁰) as another metric of genetic relationships among types and collections. Nei's genetic distance values ranged from 0.00496 to 0.01932 and largely reflected the DAPC analysis. Similar to Sawler *et al.*³⁰, hemp showed the least genetic distance followed by germplasm collections from CGN and IPK. Marijuana and Iranian cannabis clustered together with genetic distances of 0.00496 and 0.00921, respectively, while the genetic distance between Iranian collections and hemp was estimated at 0.01469 (Fig. 3). Overall, these results suggest that Iranian collections are more genetically similar to marijuana collections than hemp.

Gender, Drug and Non-Drug. To identify of DNA markers associated with gender for rapid/early identification of male and female plants, we examined allele frequency differences between female and male samples at the same position, in a modified bulked segregant analysis. It is important to note that neither of the reference genomes used in this study were from a male plant. Our approach failed to identify sex specific alleles at high frequency outside of the sex determining region.

Previous analyses have shown that marijuana and fibre types differ across the genome and not just at specific loci. Our approach failed to identify positions with significant deviations in allele frequency among 19,345 SNPs between types. Sawler *et al.*³⁰ reported a highest allele frequency of 0.82 in hemp and 0 in marijuana for a single polymorphism. Our reanalysis of these data identified 9 SNPs with allele frequencies of 1 for hemp and 0 for marijuana and 92 SNPs with allele frequency 0 for hemp and 1 for marijuana. All positions and their frequencies are supplied in Table S3.

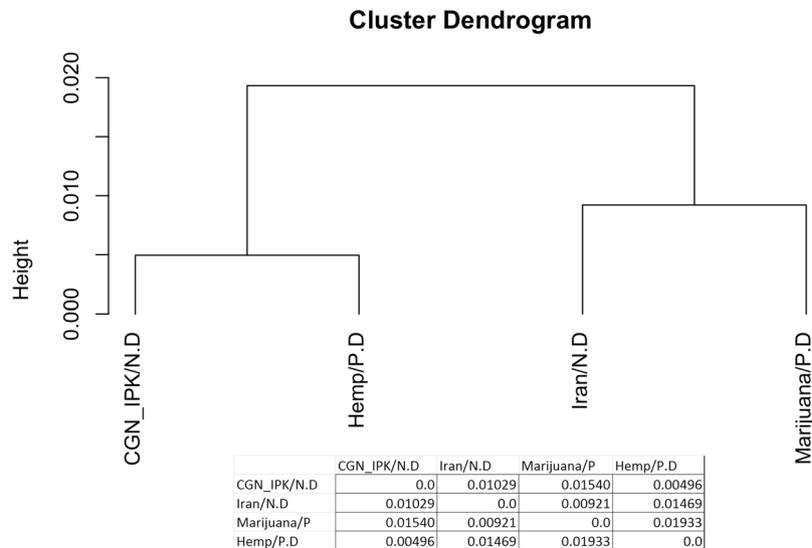


Figure 3. The dendrogram generated from Nei's genetic distance.

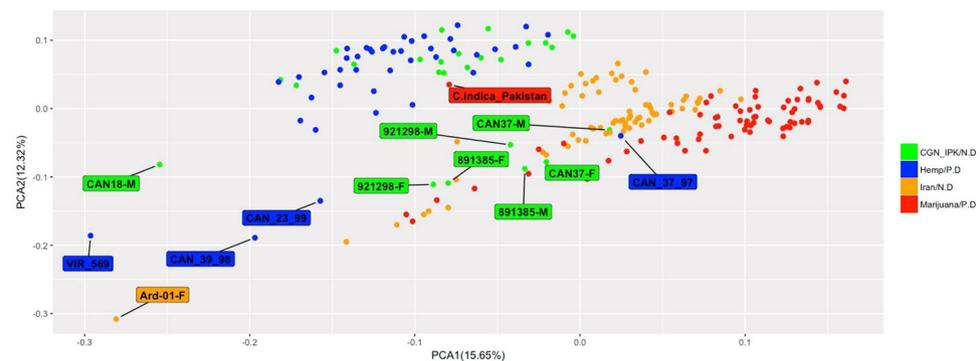


Figure 4. Principle components analysis of 95 samples from Iranian collection, 43 hemp and 71 marijuana samples using 13,325 SNPs. Hemp samples are colored blue and marijuana samples are colored red. Iranian samples divided to two groups, Original from Iran (orange) and come from germplasm collections of CGN and IPK (green) (N.D stand for New Data and P.D stand for previously analyzed data).

Population structure. An initial analysis of population structure was performed using individual-based principal component analysis (PCA). PCA using data from Iranian collections, CGN (A fiber germplasm collection except for one accession, 891385 which known as a drug cannabis)⁴³ and IPK (A hemp germplasm collection), and *C. sativa* GBS data from Sawler *et al.*³⁰ (Fig. 4) revealed two main clusters supporting the reported split between marijuana and hemp accessions³⁰ and revealed that Iranian collections tend to cluster with marijuana accessions. This plot revealed two nonconforming individuals (CAN18-M and Ard-01-F) that failed to group with the two main clusters. Previous outliers from Sawler *et al.*³⁰ were suggested to be sample error or misclassification (hemp vs. marijuana), our data suggests that CAN individuals (CAN37 from our collection and CAN23_99, CAN39_98 and CAN_37/97 from Sawler *et al.*³⁰) are more genetically similar to marijuana type accessions. To further elucidate genetic clustering identified by PCA, we performed a Discriminant Analysis of Principal Components (DAPC)⁴⁴. Consistent with fineSTRUCTURE analysis (Fig. S5), DAPC identified 4 distinct clusters (Fig. 5A). Visualisation of DAPC results using the first 22 principal components clearly clusters, marijuana, hemp, germplasm collections, and Iranian collections (Fig. 5B).

PCA within the Iranian collection identified two primary clusters (Fig. 6) separated along principal component 2, representing 7.8% of variance. This clustering separated accessions from Sanandaj, Samen, Ramhormoz, Gahwareh, Gonabad, Baneh, Arak and Saadat Shahr (Iran's western margin states) and the rest of Iranian accessions. These inferences were also largely consistent with results from a fastSTRUCTURE analysis. Notably, fastSTRUCTURE identifies 2 genetic clusters within Iranian cannabis (Fig. S4). Gene flow estimates between these clusters, identified via MIGRATE-N^{45,46}, indicates an asymmetric sharing of alleles (Table S4) between clusters. This pattern is consistent with reduced gene flow from cluster 1 which includes 18 samples (Fig. 6) such as Rmhz, Gonb, Gahv, Ark, Sam, Nhv, San-01, Ban and Saad-01-M and cluster 2 with all other samples. Genetic clustering

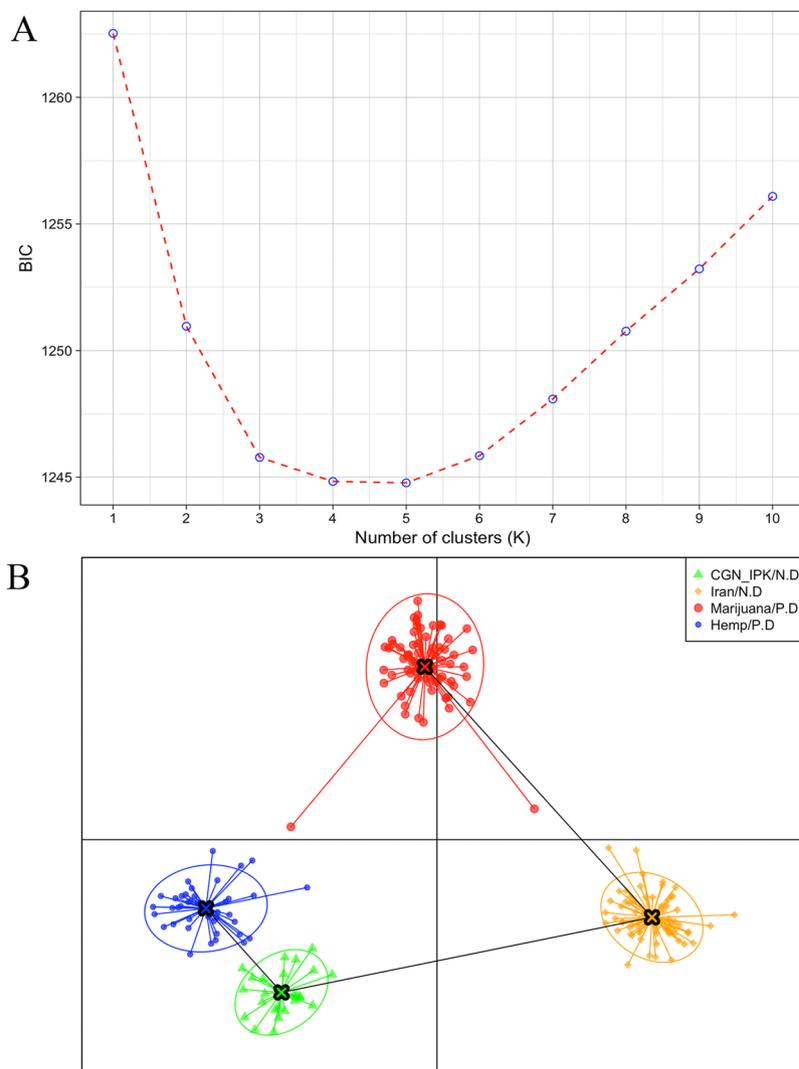


Figure 5. Discriminant analysis of principal components (DAPC) results. (A) The optimal number of clusters (K) as determined by ‘k-means’. The graph shows a clear decrease of BIC until k = 4 clusters to be the most likely value of K, after which BIC increases. (B) Scatterplot based on the DAPC output for four assigned genetic clusters, each indicated by different colours. Dots represent different individuals.

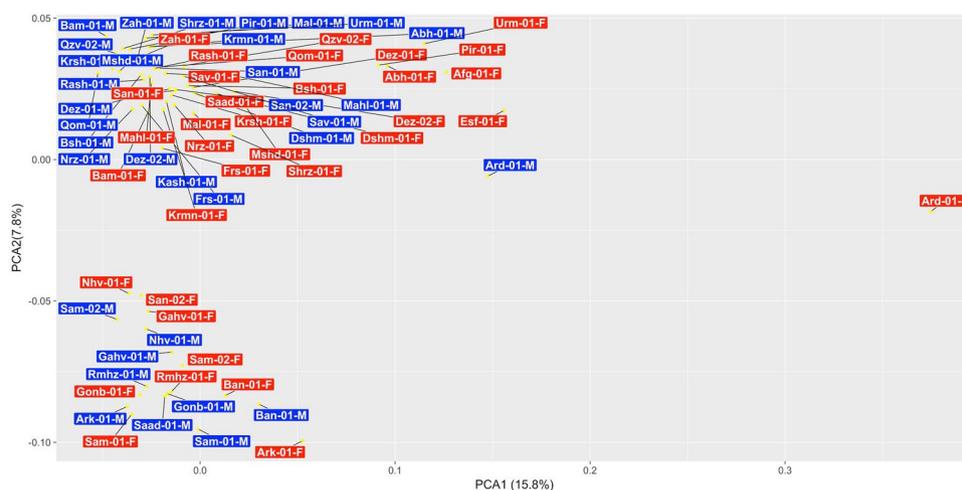


Figure 6. Individual-based principal components analysis for 35 Iranian regions and Afghanistan using 29,647 SNPs. Male plants are colored blue and female plants are colored red.

with fineSTRUCTURE identified genetic structure among individuals from the same region (Fig. S4, Table S5). According to these results we can define distinct genetic clusters for locations Neyriz, Piranshahr, Gahwareh, Arak, Urmia and Abhar. Analysis of Molecular Variance (AMOVA), as implemented in Arlequin⁴⁷, on the two and 19 genetic clusters obtained by fastSTRUCTURE and fineSTRUCTURE respectively, found that variation within populations was very high (93.09% and 95.74%) compared to between population estimates (1.38% and 1.02%; Table S6). This pattern is consistent with perennial dioecious plants wherein the majority of variation is harbored within populations⁴⁸. Together these suggest that Iranian cannabis populations tend to share more DNA with geographically proximate populations where may have genomes made up of mixtures of inferred source populations, while our simulation incorporated drift between locations, but not admixture.

Discussion

Cannabis, both marijuana and fibre types, is a globally important plant, driving a multi-billion dollar industry. Unraveling the population genomic parameters of natural populations can help identify sources of genetic diversity, as well as describing patterns of domestication for this widely used plant. In this study, we have found that natural populations of *Cannabis* in Iran are more closely related to marijuana than hemp, and that these populations harbor unique pools of genetic diversity. Taken together these data support the hypothesis that reduced diversity across fibre types suggests that hemp cultivars are derived from marijuana³⁰.

Population analyses among all accessions sampled defined 4 distinct genetic clusters (Figs 3,4 and 5). These analyses support previous findings (Sawler *et al.*³⁰) that marijuana and hemp are differentiated and identify Iranian collections as genetically more similar, yet distinct from, marijuana. This evidence provides support for the hypothesis that Iranian cannabis harbors unique genetic diversity and may represent a distinct genetic lineage of marijuana. Heterozygosity indicates levels of genetic diversity within populations, and has also been used to estimate genetic distance between populations^{49,50}. Consistent with genetic diversity levels in the present study, previous estimates of heterozygosity across diverse marker types (e.g., SNP, SSR, AFLP) typically identify higher levels of heterozygosity in hemp compared to marijuana^{30,51–53}. However, it should be noted that one study found lower levels of heterozygosity in hemp varieties across 195 samples and 2894 SNPs²⁹. It has been suggested that this may result from limited hemp sample representation in the collection²⁹. Heterozygosity estimates within our Iranian collection were similar to those found by Sawler *et al.*³⁰ for marijuana type accessions. If, as we surmise, Iranian cannabis are marijuana accessions, then these accessions likely represent remnants of cultivated germplasm from the other regions, possibly through migration of *Cannabis* from neighboring countries like Afghanistan and Pakistan into Iran. These results demonstrate that Iran is a public repository of marijuana genetic diversity; however, the loss of this unique germplasm is of great concern as there are no breeding programs and growing *Cannabis* is associated with strict legal penalties.

PCA and fastSTRUCTURE analysis of the Iranian collection identified two genetic clusters (Fig. 6, S4) separated along an east west gradient. Further analyses via fineSTRUCTURE showed that some locations are supported as distinct genetic populations (Fig. S3). These observations reveal that Iranian cannabis, despite clear evidence of admixture (likely the result of breeding), harbors distinguishable pools of genetic diversity. The lack of strong population differentiation is unsurprising since, all known cultivars of *Cannabis* are wind-pollinated and highly heterozygous (confirmed by AMOVA, Table S6). Population structure is further complicated by the fact that marijuana cultivars are clonally propagated in order to retain high-levels of THC production. Intentionally growing *Cannabis* plants in Iran is punishable by prison sentence, populations of plants are more likely to have arisen from seed and therefore represent more natural populations. Although Iranian cannabis is not likely a subspecies it does represent a genetically unique variety of marijuana, and thus provides a novel source of genetic material for cultivar development.

In plants, the sex determination system is important for two reasons; first, understanding the role of sex determination in shaping plant evolution, and second, diversity in the mechanisms through which sex is determined. There have been many studies on gender in *Cannabis*, including whether a plant should be classified as female or male, and in addition to the identification of sex chromosomes²¹, some male-specific DNA markers have been identified in *C. sativa*, allowing verification of gender during early developmental stages^{20,22,54}. Sex determination in *Cannabis* is a complex process and can be modified or reversed by environmental factors and chemical treatment^{55,56}. Additionally, male flowers are able to develop on female plants under extreme conditions⁵⁷. Because confirmed sex-associated DNA markers such as MADC2 sometimes fail to discriminate sex phenotype²², we attempted to identify sex associated markers from autosomal regions. While our study generated thousands of differentiating markers, we failed to find sex locus specific SNPs. This is likely because no male reference genome is available and the proportion of coding regions covered by the GBS derived SNPs. Future studies can capitalize on the utility of high-throughput sequencing technologies to look for markers associated with sex-determining loci, in particular coding derived SNPs (e.g., RNA-seq). We were able, however, to identify marijuana and fibre type specific markers through reanalysis of previously published data.

Our conclusions, consistent with previous studies, show that genetic differences between hemp and marijuana accessions are widely distributed across the genome³⁰. Comparative analysis of Purple Kush (marijuana) and Finola (fibre) genomes revealed highly discriminative SNPs that are distributed across the genome and are not restricted to particular loci (e.g., cannabinoid production)³¹. While previous work focused on THC:CBD ratios and the associated B locus (a single locus with two co-dominant alleles)⁴¹, recent work has identified SNPs in THCA and CBDA synthases associated with chemotype variation⁵⁸. Thus, associating SNPs with active and inactive forms of THC and CBD synthases will continue to be a powerful tool for distinguishing *Cannabis* types. In this study, we identified SNPs that appear to be tightly linked to type, and are outside of cannabinoid genes, which should prove useful for future research. More immediately, these markers can be validated for early and rapid identification of marijuana and fibre type plants for current breeding programs.

Materials and Methods

Collection of Genetic Material. Natural populations of *Cannabis* in Iran were identified and seeds were collected for growing in the field in university of Tehran. Sex identities were verified using taxonomic keys. A set of different accessions provided by CGN and IPK and one population from Afghanistan were used for analysis in this study as well (Table S1, Fig. 1). Figure 1 was produced using the R software version 3.1.3 with the packages: Raster version 2.5-8 (<https://cran.r-project.org/web/packages/raster/index.html>) and Ggplot2 version 2.2.1 (<http://ggplot2.tidyverse.org/reference/>)⁵⁹. Additionally Dplyr version 0.5.0 was used to manipulate the data-frames (<https://cran.r-project.org/web/packages/dplyr/index.html>).

DNA Extraction, Library Preparation and Sequencing. DNA was extracted using a Qiagen DNeasy plant mini-kit, from leaf tissue of one female and one male plant from each location. The isolation procedure was carried out according to the manufacturer's guidelines. We performed *in silico* digestion of the *Cannabis* genome sequence with *Pst*I and *Ape*KI to select the best restriction enzyme library preparation. Libraries were prepared using the GBS protocol published by Sonah *et al.* (2013). A 150 ng genomic DNA template was used to prepare the library using the *Ape*KI enzyme. High-throughput was performed on an Illumina HiSeq. 2500, Rapid-run mode, single-end 100 base reads, at Duke Center for Genomics and Computational Biology.

Bioinformatics Analysis. Demultiplexing and Read Filtering. After unzipping fastq.gz files to fastq files by gunzip command, the GBSX package⁶⁰ was used for demultiplexing of reads. Reads were organized into new files with adapter sequences removed, reads were discarded that were, shorter than 50 bases, and trim leading and trailing low quality regions (<Q30) by fastq-mcf, a widely available open source software⁶¹. To elucidate the relationship of Iranian cannabis with marijuana and fibre type accessions, we merged our data with marijuana and hemp data prepared by Sawler *et al.*³⁰ (downloaded from NCBI SRA BioProject: PRJNA285813).

Mapping, SNPs Discovery and filtering. In a high-throughput genotyping workflow, alignment of short reads to a reference genome is the first step after read processing and filtering. BWA⁶² was used to map reads of the individual genotypes to the reference genome with the default parameters. Reads mapped to Purple Kush (canSat3: a special variety of hemp) and Finola (finola1: a special variety of marijuana) *C. sativa* reference genome assembly separately which are known as high and low-THCA producing varieties respectively. The mapping outputs were used for removing unmapped reads to produce BAM files using Samtools⁶³ and only reads mapping to a unique location in the genome were retained. Merging all BAM files into one stream by bamaddrg utility (<https://github.com/ekg/bamaddrg>), sorting and indexing BAM files by Samtools package⁶³ were primary stages for use of FreeBayes⁶⁴ to detect variants. Before running FreeBayes, we estimated the number of markers for each individual by "bedtools genomcov"⁶⁵ and percentage of coverage by dividing marker number times read length by genome size. FreeBayes was run using default parameters. This was performed for or males and females and drug and non-drug types separately to find positions linked to gender and type. Bi-allelic, missingness, quality, and depth were filtered. The aim of the QC on SNPs was to define high quality set of individuals for analysis. Bi-allelic markers were identified by a command-line written in our lab. Then got vcflib freely available (<https://github.com/vcflib/vcflib.git>) packages to filter down the SNPs that had mapping quality <30 and read depth <5. This package can filter each position for each individual. Filtering was initially performed using VCFtools package⁶⁶, VarFilter from BCFtools is freely available (<https://github.com/samtools/BCFtools>) packages. After screening a few markers we found that read depth and quality were not being appropriately filtered for our data set and therefore we opted to use vcflib. To filter missing data we used "-max-missing 1.0" option in VCFtools package⁶⁶. Finally, summary statistics were collected using vcf-stats before and after data filtering.

Scan for Identification of SNPs associated with gender and type. Identification of DNA markers associated with gender and type was carried out based on comparison of SNP allele frequency differences between each group (female-male and marijuana-fibre). To do this, we called SNPs for sample pairs female and male, marijuana and fibre, separately using FreeBayes⁶⁴. After filtering variants for read depth (>5), read mapping quality (>30) and minor allele frequency (>2.5%), we generated allele frequency estimates and compared frequencies at the same position across the genome.

Analysis of population structure. We computed the fixation index (F_{ST}) using VCFtools⁶⁶ among all wise locations in the Iranian collection and also between marijuana and hemp types. Estimation of heterozygosity for each individual was conducted with custom command-line scripts by dividing the number of heterozygous sites by the number of non-missing genotypes. The number of heterozygous sites was counted by vcflib tools. We pursued principal components analysis (PCA) to investigate genetic relationships using a distance matrix obtained by TASSEL version 5⁶⁷. Plotting PCA results was completed via the ggplot2⁵⁹ package in Rstudio version 0.99.902. We also applied discriminant analysis (DA) of principal components⁴⁴ using the adegenet package⁶⁸. Discriminant analysis can ascribe relationships for pre-defined groups without relying on a particular population genetics mode⁴⁴. Files were read using the function read.vcf and converted into geneid objects with the vcfR-2geneid function⁶⁹. In DAPC, data is first transformed using a principal components analysis (PCA) and subsequently the number of genetic clusters was assessed using the find.clusters function. The Bayesian Information Criterion (BIC) was calculated for $K = 1-10$. For k-means clustering, all of the principal components were retained. The K value with the lowest BIC was selected as the optimal number of clusters. DAPC was implemented using the optimized number of principal components as determined by the optim.a.score function. Nei's genetic distance⁴² among populations was calculated using the STAPP package for R⁷⁰ and the resultant dendrogram was drawn using the standard R function plot.hclust. To determine the most probable number of genetic clusters, fast-STRUCTURE⁷¹ was run at $K = 1$ and $K = 10$, with an average of 22600 iterations, using default parameters for the Iranian samples. The analysis at $K = 2$ was performed to test the extent to which the samples reflect two distinct

groups. Other values of K were tested (not shown), but did not provide further optimization or descriptive value. Additionally, the cannabis population structure was investigated using fineSTRUCTURE⁷². To visualize populations, we plotted the output data via the fineSTRUCTURE graphical user interface.

The genetic clusters from fastSTRUCTURE and fineSTRUCTURE were used to estimate gene flow and population size via MIGRATE-N (v. 3.6.11)^{45,46}. In this case, gene flow was estimated between two clusters obtained by fastSTRUCTURE only for Iranian cannabis (69 samples). MIGRATE-N was implemented with following parameters: the Bayesian inference strategy, 1000 for number of recorded steps in chain, a burn-in of 1000 for each chain and a full migration model with two population sizes and two migration rates. The starting values for θ and M were generated initially from Fst, Migrate-n was subsequently run using the resulting θ and M values of the previous run. The runs were conducted on 5 K of markers. Hierarchical analysis of molecular variance (AMOVA) was performed using the Arlequin software package (v. 3.1)⁴⁷. Significance levels for variance components and F-statistics were estimated using 1000 permutations.

References

- Small, E. & Cronquist, A. A practical and natural taxonomy for *Cannabis*. *Taxon*. **25**, 405–435 (1976).
- Adams, I. R. & Martin, B. R. *Cannabis*: pharmacology and toxicology in animals and humans. *Addiction*. **91**, 1585–1614 (1996).
- Gaoni, Y. & Mechoulam, R. Isolation, structure and partial synthesis of an active constituent of hashish. *Journal of the American Chemical Society*. **86**, 1646–1647 (1964).
- Siniscalco Gigliano, G. *Cannabis sativa* L. botanical problems and molecular approaches in forensic investigations. *Forensic Science Review*. **13**, 1–17 (2001).
- Taura, F. *et al.* Cannabidiolic-acid synthase, the chemotype-determining enzyme in the fiber-type *Cannabis sativa*. *Federation of European Biochemical Societies*. **581**, 2929–2934 (2007).
- Broseus, J., Anglada, F. & Esseiva, P. The differentiation of fibre- and drug type Cannabis seedlings by gas chromatography/mass spectrometry and chemometric tools. *Forensic Science International*. **200**, 87–92 (2010).
- Hillig, K. Genetic evidence for speciation in *Cannabis* (Cannabaceae). *Genetic Resources and Crop Evolution*. **52**, 161–180 (2005).
- Bradshaw, R. H. W., Coxon, P., Greig, J. R. A. & Hall, A. R. New fossil evidence for the past cultivation and processing of hemp (*Cannabis sativa* L.) in Eastern England. *New Phytologist*. **89**, 503–510 (1981).
- Duvall, C. S. Drug laws, bioprospecting and the agricultural heritage of *Cannabis* in Africa. *Space Polity*. **20**, 10–25 (2016).
- Herbig, C. & Sirocko, F. Palaeobotanical evidence for agricultural activities in the Eifel region during the Holocene: plant macro-remain and pollen analyses from sediments of three maar lakes in the Quaternary Westeifel Volcanic Field (Germany, Rheinland-Pfalz). *Vegetation History and Archaeobotany*. **22**, 447–462 (2013).
- Li, H.-L. The origin and use of cannabis in eastern asia linguistic-cultural implications. *Economic Botany*. **28**, 293–301 (1974).
- Murphy, T. M., Ben-Yehuda, N., Taylor, R. E. & Southon, J. R. Hemp in ancient rope and fabric from the Christmas Cave in Israel: talmudic background and DNA sequence identification. *Journal of Archaeological Science*. **38**, 2579–2588 (2011).
- Piluzza, G., Delogu, G., Cabras, A., Marceddu, S. & Bullitta, S. 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 (2013).
- Rivoira, G. Canapa. In: Baldoni R, Giardini L (eds) *Coltivazioni erbacee*. Patron, Bologna (1981).
- Small, E. & Marcus, D. "Hemp: a new crop with new uses for North America," in *Trends in New Crops and New Uses*, eds J. Janick and A. Whipkey (*Alexandria, VA: ASHS Press*), 284–326 (2002).
- Russo, E. B. *et al.* Phytochemical and genetic analyses of ancient cannabis from Central Asia. *Genetic Resources and Crop Evolution*. **59**, 4171–4182 (2008).
- Faeti, V., Mandolino, G. & Ranalli, P. Genetic diversity of *Cannabis sativa* germplasm based on RAPD markers. *Plant Breeding*. **115**, 367–370 (1996).
- Zeven, A. C. & Zhukovskiy, P. M. Cannabidaceae. In: *Dictionary of cultivated plants and their centres of diversity*. 62–63, 129–130 (Centre for Agricultural Publishing and Documentation, Wageningen, The Netherlands, 1975).
- Green, G. *The Cannabis Breeder's Bible*. 15–17 (Green Candy Press, 2005).
- Mandolino, G., Carboni, A., Forapani, S., Faeti, V. & Ranalli, P. Identification of DNA markers linked to the male sex in dioecious hemp (*Cannabis sativa* L.). *Theoretical and Applied Genetics*. **98**, 86–92 (1999).
- Mandolino, G., Carboni, A., Bagatta, M., Moliterni, V. M. C. & Ranalli, P. Occurrence and frequency of putatively Y chromosome linked DNA markers in *Cannabis sativa* L. *Euphytica*. **126**, 211–218 (2002).
- Sakamoto, K., Shmamura, K., Komeda, Y., Kamada, H. & Satoh, S. A male associated DNA sequence in a dioecious plant, *Cannabis sativa* L. *Plant cell physiology*. **36**, 1549–1554 (1995).
- Sakamoto, K. *et al.* RAPD markers encoding retrotransposable elements are linked to the male sex in *Cannabis sativa* L. *Genome*. **48**, 931–936 (2005).
- Gilmore, S. & Peakall, R. Isolation of microsatellite markers in *Cannabis sativa* L. (marijuana). *Molecular Ecology Notes*. **3**, 105–107 (2003).
- Gilmore, S., Peakall, R. & Robertson, J. Short tandem repeat (STR) DNA markers are hypervariable and informative in *Cannabis sativa*: implications for forensic investigations. *Forensic Science International*. **131**, 65–74 (2003).
- Pacifico, D. *et al.* Genetics and marker-assisted selection of the chemotype in *Cannabis sativa* L. *Molecular Breeding*. **17**, 257–268 (2006).
- Alghanim, H. J. & Almirall, J. R. Development of microsatellite markers in *Cannabis sativa* for DNA typing and genetic relatedness analyses. *Analytical and Bioanalytical Chemistry*. **376**, 1225–1233 (2003).
- Hakki, E. E. Inter simple sequence repeats separate efficiently hemp from marijuana (*Cannabis sativa* L.). *Electronic Journal of Biotechnology*. **10**, 4 (2007).
- Lynch, R. C. *et al.* Genomic and chemical diversity in *Cannabis*. *Critical Reviews in Plant Sciences*. **35**, 349–363 (2015).
- Sawler, J. *et al.* The Genetic Structure of Marijuana and Hemp. *PLoS One*. **10**, e0133292 (2015).
- van Bakel, H. *et al.* The draft genome and transcriptome of *Cannabis sativa*. *Genome Biology*. **12** (2011).
- Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One*. **6**, e19379 (2011).
- Deschamps, S., Llaca, V. & May, G. D. Genotyping-by-Sequencing in Plants. *Biology*. **1**, 460–483 (2012).
- Soorni, A., Nazeri, V., Fattahi, R. & Khadivi-Khub, A. DNA fingerprinting of *Leonurus cardiaca* L. germplasm in Iran using amplified fragment length polymorphism and interretrotransposon amplified polymorphism. *Biochemical Systematics and Ecology*. **50**, 438–447 (2013).
- Bailey, T. *et al.* Practical guidelines for the comprehensive analysis of ChIP-seq data. *PLOS Computational Biology*. **9**, e1003326 (2013).
- Batley, J., Barker, G., O'Sullivan, H., Edwards, K. J. & Edwards, D. Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol*. **132**, 84–91 (2003).

37. Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. Molecular basis of base substitution hotspots in *Escherichia coli*. *Nature*. **274**, 775–780 (1978).
38. Pootakham, W. *et al.* Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). *Genomics*. **105**, 288–295 (2015).
39. Shearman, J. R. *et al.* SNP identification from RNA sequencing and linkage map construction of rubber tree for anchoring the draft genome. *PLoS One*. **10**, e0121961 (2015).
40. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. **123**, 585–595 (1989).
41. De Meijer, E. P. M. Variation of *Cannabis* with reference to stem quality for paper pulp production. *Industrial Crops and Products*. **3**, 201–211 (1994).
42. Nei, M. Genetic distance between populations. *The American Naturalist*. **106**, 283–392 (1972).
43. De Meijer, E. P. M. M. Diversity of cannabis. [dissertation]. *Wageningen: Wageningen University*. (1994).
44. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94 (2010).
45. Beerli, P. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics*. **22**, 341–345 (2006).
46. Beerli, P. How to use MIGRATE or why are Marko Chain Monte Carlo Programs Difficult to use? New York: Cambridge University Press. (2009).
47. Excoffier, L., Laval, G. & Schneider, S. Arlequin (version 3.0): an integrated software package for population genetic analysis. *Evolutionary Bioinformatics Online*. **1**, 47–50 (2005).
48. Sheng, Y., Zheng, W., Pei, K. & Ma, K. Genetic Variation Within and Among Populations of a Dominant Desert Tree *Haloxylon ammodendron*(Amaranthaceae) in China. *Annals of Botany*. **96**(2), 245–252 (2005).
49. Chakraborty, R. Relationship between heterozygosity and genetic distance in the three major races of man. *American Journal of Physical Anthropology*. **65**, 249–258 (1984).
50. Guerreiro, J. F., Santos, E. J. M. D. & Santos, S. E. B. D. Effect of average heterozygosity on the genetic distance of several Indian tribes from the Amazon region. *Annals of Human Biology*. **21**, 589–595 (1994).
51. Gao, C. *et al.* Diversity analysis in *Cannabis sativa* based on large-scale development of expressed sequence tag-derived simple sequence repeat markers. *PLoS ONE*. **9**, e110638 (2014).
52. Hu, Z. G. *et al.* Genetic diversity research of hemp (*Cannabis sativa* L) cultivar based on AFLP analysis. *Journal of Plant Genetic Resources*. **13**, 555–561 (2012).
53. Zhang, L. G. *et al.* Analysis of the genetic diversity of Chinese native *Cannabis sativa* cultivars by using ISSR and chromosome markers. *Genetics and Molecular Research*. **13**, 10490–10500 (2014).
54. Törjék, O. *et al.* Novel male-specific molecular markers (MADC5, MADC6) in hemp. *Euphytica*. **127**, 209–218 (2000).
55. Chailakhyan, M. K. Genetic and hormonal regulation of growth, flowering and sex expression in plants. *American Journal of Botany*. **66**(6), 717–736 (1979).
56. Mohan Ram, H. Y. & Sett, R. Sex reversal in the female plants of *Cannabis sativa* by cobalt ions. *Proceedings of the Indian Academy of Sciences*. **88**(4), 303–308 (1979).
57. Clarke, R. K. Hanf: Botanik, Anbau Vermehrung und Züchtung. Aarau, Schweiz: AT Verlag (1997).
58. Onofri, C., de Meijer, E. P. M. & Mandolino, G. Sequence heterogeneity of cannabidiolic- and tetrahydrocannabinolic acid-synthase in *Cannabis sativa* L. and its relationship with chemical phenotype. *Phytochemistry* **116**, 57–68 (2015).
59. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York, 2009).
60. Herten, K., Hestand, M. S., Vermeesch, J. R. & Van Houdt, J. K. GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics*. **16**, 37 (2015).
61. Aronesty, E. Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal* **7**, 1–8 (2013).
62. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*. **26**, 589–595 (2010).
63. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics*. **25**, 2078–2079 (2009).
64. Garrison E. & Marth G. Haplotype-based variant detection from short-read sequencing. <http://arxiv.org/abs/1207.3907> (2012).
65. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**, 841–842 (2010).
66. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics*. **27**, 2156–2158 (2011).
67. Bradbury, P. J. *et al.* TASSEL: Software for association mapping complex traits in diverse samples. *Bioinformatics*. **23**, 2633–2635 (2007).
68. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics*. **24**, 1403–1405 (2008).
69. Knaus, B. J. & Grunwald, N. J. Vcfr: an R package to manipulate and visualize VCF format data. *Molecular Ecology Resources*. Pre print, (2016).
70. Pembleton, L. W., Cogan, N. O. I. & Forster, J. W. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources*. **13**(5), 946–952 (2013).
71. Raj, A., Stephens, M. & Pritchard, J. K. fastSTRUCTURE: Variational inference of population structure in large SNP datasets. *Genetics*. **197**, 573–589 (2014).
72. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genetics*. **8**, e1002453 (2012).

Acknowledgements

The authors are grateful to the Ministry of Science, Research and Technology of Iran as a funding source of this project. Special thanks to Dr. Luisa Trindade of Wageningen UR Plant Breeding and Dr. Bert Visser of Center for Genetic Resources, The Netherlands, and IPK, Germany for providing the seeds.

Author Contributions

A.S., S.A.S. and A.B. participated in the experimental design. A.S., R.F. and S.A.S. participated in the sample collection and DNA extraction. A.S. prepared the libraries. A.S., A.B. and D.H. analyzed the data. All the authors participated in the discussion of the results and writing of the article.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-017-15816-5>.

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017