Soler Aleixandre, S.; Gramazio, P.; Figás-Moreno, MDR.; Vilanova Navarro, S.; Rosa-Martínez, E.; Llosa, ER.; Borràs-Palomares, D.... (2017). Genetic structure of Cannabis sativa var. indica cultivars based on genomic SSR (gSSR) markers: implications for breeding and germplasm management. Industrial Crops and Products. 104:171-178. doi:10.1016/j.indcrop.2017.04.043

Genetic structure of *Cannabis sativa* var. *indica* cultivars based on genomic SSR (gSSR) markers: implications for breeding and germplasm management

Salvador Soler[a], Pietro Gramazio[a], Maria R. Figàs[a], Santiago Vilanova[a], Elena Rosa[a], Ernesto R. Llosa[b], Dionís Borràs[a], Mariola Plazas[a], Jaime Prohens[a, *]

[a]Institut de Conservació i Millora de l'Agrodiversitat Valenciana, Universitat Politècnica de València, Camí de Vera 14, 46022 València, Spain

[b]Hemp Trading, S.L.U., Camí del Polio 51, 46469 Beniparrell, Spain

* Corresponding author. Tel: +34 963879424. *E-mail address*: jprohens@btc.upv.es (J. Prohens)

ABSTRACT

Cannabis sativa L. is cultivated for its fiber or seeds (var. *sativa*; hemp), or for its high content in cannabinoids (var. *indica*; marijuana). Knowledge of the genetic structure of *C. sativa* var. *indica* is important for selection and breeding of cultivars with medicinal interest. We used six genomic SSRs (gSSRs) for genotyping 154 individual plants of 20 cultivars of *C. sativa* var. *indica,* plus two cultivars of *C. sativa* var. *sativa*. A very high polymorphism was observed, with an average of 17 alleles and 23.8 genotypes per locus. Expected (H_e) and observed (H_o) heterozygosities were high, with average values of 0.753 and 0.429, respectively. In some cultivars H_e and H_o presented similar values, while in others H_e was considerably higher than H_o suggesting that consanguinity and fixation had taken place during its development. In addition, some cultivars had a reduced number of alleles per locus (in some cases only two) indicating that a genetic bottleneck had taken place during its development. Gene flow (*Nm*) between both botanical varieties was high, with Nm=1.736. The molecular analysis of variance (AMOVA) revealed that only 31.94% of the molecular variation observed was caused by differences among cultivars, while the variation among plants of the same cultivar was of 37.11%, and within individual variation, due to heterozygosity, was of 30.96%. This indicates that a large variation exists within cultivars, which can be exploited for selection, but also complicates germplasm management and regeneration. The population structure analysis identified 14 genetic clusters, with most individuals of a single cultivar clustering together. This analysis, together with UPGMA cluster analysis shows that the two *C. sativa* var. *sativa* cultivars studied are differentiated from *C. sativa* var. *indica,* and that some cultivars of *C. sativa* var. *indica* seem to represent different selections from a common original cultivar. Our results represent the first comprehensive study of intra-

varietal diversity in *C. sativa* var. *indica* and provide information of relevance for selection, breeding, and germplasm conservation, as well as for forensic studies in this crop.

Keywords:
Breeding
Genetic diversity
gSSRs
Heterozygosity
Marijuana
Molecular markers

1. Introduction

Cannabis sativa L. (2n = 2x = 20) is a hypervariable multi-use crop which includes two botanical varieties with different morphologies are recognized: *C. sativa* var. *sativa* (hemp) and *C. sativa* var. *indica* (marijuana) (de Meijer and Keizer, 2012; Salentjin et al., 2015; Small, 2015). While *C. sativa* var. *sativa* is mostly cultivated for its fiber and seeds (Salentijn et al., 2015), *C. sativa* var. *indica* is mainly grown for its high content of cannabinoids, some of which, like the usually predominant tetrahydrocannabinol (THC), have psychotropic activities (Small, 2015). Because these two botanical varieties have been selected for different purposes, the plant morphology is very different (de Meijer and Keizer, 2012) and they are also differentiated at the genetic level (Sawler et al.,

2015). Other than from THC, *C. sativa* var. *indica* also contains other cannabinoids with therapeutic interest, like cannabidiol (CBD) (Small, 2015). In consequence, the development of C. *sativa* var. *indica* cultivars with high contents in CBD and low contents of THC for medical purposes is a current breeding objective; however, its forbidden use as recreational drug in many countries (van Ours, 2012) has hampered genetic diversity studies, a pre-requisite for the application of breeding programs and appropriate germplasm collection and management strategies (Welling et al., 2016).

Cannabis sativa is a essentially dioecious species and therefore is an obligate allogamous species under natural reproduction conditions (Small, 2005), which favours genetic heterogeneity. Sex expression in *C. sativa* is determined by chromosome heteromorphisms, where XY and XX individuals are male and female, respectively (Moliterni et al., 2004). Some cultivars are monoecious despite having an XX chromosomic configuration (Razumova et al., 2016), but have a high degree of allogamy. Because female plants of *C. sativa* var. *indica* have much higher contents in cannabinoids than males (Small, 2015), the production of feminized cultivars, where 100% of the individuals are female, is a common practice (Green, 2005). This is possible thanks to sex reversion, mediated by the application of chemicals such as silver thiosulfate, in inflorescences of genetically female plants (Green, 2005). These functionally male flowers produce 100% of X gametes resulting, after fertilizing the X female gametes, in feminized (XX) cultivars. This alteration of the reproduction, which allows self-pollinations and crossings between genetically female plants, may increase inbreeding and, therefore, have a considerable impact in the genetic structure of *C. sativa* var. *indica* cultivars. Also, breeding methods used to develop new cultivars may have a major impact in the genetic structure of the crops. In this respect, the the most common methods of

developing new *C. sativa* var. *indica* cultivars generally include selection of individuals within heterogenous cultivars, hybridization of different cultivars, or a combination of both (Green, 2005). When the population size used for the development of new cultivars is small, which due to the illegal nature of the crop may be a frequent phenomenon in the development of new *C. sativa* var. *indica* varieties, a genetic bottleneck effect may take place, which reduces diversity and may increase homozygosity due to crossings among related individuals (Sawler et al., 2015; Welling et al., 2016).

Many studies have been performed with different types of molecular markers to evaluate the diversity among accessions in *C. sativa*, revealing that the crop is genetically very diverse (e.g., Shirley et al., 2013; Gao et al., 2014; Sawler et al., 2015). Amazingly, there are very few studies evaluating the intra-varietal diversity of *C. sativa*, which is very important for selection, breeding and developing uniform cultivars. Gilmore et al. (2003) used five simple sequence repeat (SSR) markers to evaluate the molecular diversity of six *C. sativa* var. *sativa* cultivars and nine *C. sativa* var. *indica* germplasm accessions and they found considerable variation among plants of individual accessions. However, only three of the *C. sativa* var. *indica* cultivars had four or more (up to eight) individuals, which is a limited number of cultivars to draw general conclusions. Shirley et al. (2013) using a single SSR marker found several genetic profiles in DNA extracted from individual seeds of each of several *C. sativa* var. *indica* accessions, with an important level of heterozygous loci. However, this study consisted of a single marker and no diversity statistics were presented. Furthermore, as whole seeds were used this means that a mixture of the maternal (endosperm) and zygote (embryo) genomes were surveyed, and the results may not represent a single individual, but a mixture of maternal and zygotic genetic material.

Among the genetic markers used to evaluate genetic diversity in *Cannabis*, SSRs (also called short tandem repeats, STS) have proved to be highly polymorphic (Gilmore et al., 2003; Howard et al., 2008; Allgeier et al., 2011; Shirley et al., 2013; Gao et al., 2014; Soler et al., 2016). *Cannabis sativa* genomic SSRs, which in general are more polymorphic than expressed sequence tag SSRs (EST-SSRs) (Eujayl et al., 2011; Muñoz-Falcón et al., 2011), can be obtained from previously published reports (Gilmore et al., 2003; Soler et al., 2016) or developed *de novo* from the available genome sequence of *C. sativa* (van Bakel et al., 2011).

In this paper, we use gSSR markers to evaluate the inter- and intra-accession genetic diversity of a set of *C. sativa* var. *indica* cultivars. This will provide information of relevance on the extent of intra-varietal genetic diversity, including the level of heterozygosis, as well as on the diversity and relationships of the evaluated varities. The information on the genetic structure will have important implications for breeding new cultivars and for conservation of germplasm, as well as for forensic analyses.

2. Material and Methods

2.1. Plant material

Twenty feminized (i.e., 100% of the seed give female individuals) cultivars of *C. sativa* var. *indica* were provided by Hemp Trading (Beniparrell, Spain) for the study (Table 1). These feminized cultivars were obtained by the breeders by sex reversion of genetically female plants using chemical agents (Green, 2005), which allowed inter-crossing of genetically female individuals which give an offspring of 100% of female

individuals. Some of the cultivars used are selected derivatives of the crossings between different plants of other cultivars, so that the pedigree is indicated in the cultivar name (e.g., 'Black Domina × Black Domina') (Table 1). In addition to these, two *C. sativa* var. *sativa* cultivars were used as controls.

For each cultivar, seeds were germinated in Petri dishes with moistened hydrophylic cotton covered by a layer of filter paper (Soler et al., 2016). Once seeds had germinated, between four and eight plants per cultivar depending on availabilty, totalling 154 plants (on average 7 plants per cultivar), were transferred to seedling trays containing commercial growing substrate. When plantlets had four to six true leafs, young leaf samples of individual plants were taken for DNA extraction. Permission to cultivate the plantlets was obtained from the Agencia Española del Medicamento. After DNA extraction, plantlets were cut and the plant remains were disposed off by delivering them to the Spanish National Police for destruction.

2.2. DNA extraction and SSR characterization

DNA was extracted from young leaf tissue using the CTAB0 protocol, as reported by Soler et al. (2016). DNA concentrations and purity were measured with a NanoDrop ND-1000 spectrophotometer (Nano Drop Technologies, Wilmington, Delaware, USA) and DNA integrity with a 0.8% agarose gel electrophoresis. Initially, 29 dinucleotide genomic simple sequence repeat (gSSR) markers were developed *de novo* from the publicly available nuclear genome sequence of *C. sativa* var. *indica* cultivar 'Purple Kush' (van Bakel et al., 2011). Of these 23 amplified successfully, and after some preliminary tests, six of them were selected because of their polymorphism and good

amplification results in all cultivars (Table 2). The PCR reactions were performed as described in Soler et al. (2016). Basically, the PCR reaction mixture contained 1 × PCR buffer, 1.5 mM $MgCl_2$, 0.04 units of *Taq* DNA polymerase, 0.2 mM dNTPs, 0.05 μM forward primer, 0.25 μM reverse primer, 0.2 μM M13 fluorescent labeled primer and 10 ng of template DNA and H_2O to make a total volume of 10 mL. The PCR amplifications program was as follows: an initial step at 94ºC for 3 min, 35 cycles of 94ºC for 30 s, 58ºC for 45 s, 72ºC for 1 min, and a final step at 72ºC for 10 min. SSR alleles were separated on an ABI Prism 3100 (Applied Biosystems, Carlsbad, California, USA) automatic sequencer using GeneScan 3.7 (Applied Biosystems) software and sized using GeneScan 500 LIZ (Thermo Fisher Scientific, Waltham, Massachussets, USA) molecular size standards with Genotyper 3.7 (Applied Biosystems) software.

2.3. Data analysis

The following diversity statistics were calculated for each gSSR marker: number of alleles, allele range, major allele frequency, number of genotypes, expected (H_e) and observed (H_o) heterozygosities (Nei, 1973), fixation index (F_{is}) (Wright, 1965), and polymorphic information content (PIC) (Botstein et al., 1980). For each cultivar, the H_e, H_o and highest number of alleles per locus found among the six gSSR loci investigated, was determined. Gene flow (*Nm*) between *C. sativa* var. *indica* and *C. sativa* var. *sativa* was estimated according to Wright (1951) using the formula $Nm = [(1/F_{ST})-1]/4$, where F_{ST} is the fixation index between both botanical varieties. In order to investigate the between and within genetic differentiation of the cultivars in study as well as the heterozygosity/homozygosity of each SSR locus in each individual, an analysis of

molecular variance (AMOVA) was performed using GenAlEx version 6.5 (Peakall and Smouse, 2012). Population structure was estimated using a model-based Bayesian structure implemented in the software STRUCTURE version 2.3.4 (Pritchard et al., 2000). The analysis was carried out using a burning period of 500,000 iterations and a run length of 750,000 MCMC replications. A continuous series of clusters (*K*), from 1 to 22 was tested in 20 independent runs. No prior knowledge of the cultivar of origin was introduced. The most informative *K* value was identified using the ΔK statistic (Evanno et al., 2005), using the STRUCTURE HARVESTER version 0.6.94 software (Earl and vonHoldt, 2012). Based on maximum membership probability of each individual plant, they were assigned to the corresponding groups as described by Remington et al. (2001). An unrooted neighbor-joining phenogram based on Nei et al. (1983) genetic distance was built using Powermarker version 3.25 software (Liu and Muse, 2005) to graphically represent the relationships among cultivars. Branch support on the phenogram was tested by bootstrap analysis with 1,000 replications using the PHYLIP version 3.67 software (Felsenstein, 2007). Bootstrap values of 50% or higher were used to indicate support for the phenogram topology at a node (Highton, 1993).

3. Results

The six gSSR markers were polymorphic, yielding a total of 102 alleles, with an average of 17 alleles per locus, which resulted in an average of 23.8 genotypes per SSR locus (Table 3). However, considerable differences were found in the number of alleles and genotypes among gSSR markers, with a minimum of 8 alleles (CGS18 locus) and 12 genotypes (CSG01 locus) and a maximum of 30 alleles and 41 genotypes (CSG12 locus).

As expected for a diploid, for every individual, a maximum of two alleles were found per locus. The allele size varied among gSSR loci in number and range of repeats. In this respect, amplicons of locus CSG13 had the smaller sizes, not overlapping with the ones of markers CSG01, CSG05, CSG18, and CSG24 (Table 3). The major allele frequency ranged from 0.149 (CSG12) to 0.625 (CSG13), with an average value of 0.371 (Table 3). Expected heterozygosity (H_e) was higher than 0.5 for all loci, with an average value of 0.753 and a range between 0.567 (CSG18) and 0.919 (CSG12). On the other had, observed heterozygosity (H_o) was lower than 0.5 for all loci, with an average value of 0.293 and a range between 0.049 (CSG01) and 0.429 (CSG13). For all markers H_o was lower than H_e, although in some cases, like CSG01, the difference was very marked, while in others, like CSG13 the differences were much smaller (Table 3). This resulted in large differences among loci in the fixation index, with values ranging from 0.251 (CSG13) to 0.936 (CSG01). The PIC values ranged between 0.544 for CSG13 and 0.914 for CSG12, with an average value of 0.727 (Table 3).

The observed heterozygosity (H_o) for individual cultivars ranged between 0.150 in 'BL-57' and 0.602 in 'Critical' (Figure 1), with an average value of 0.293, while the expected heterozygosity (H_e) varied between 0.174 for 'NST-75-75' and 0.617 for 'Futura', and had an average value of 0.406. The averages for the observed and expected heterozygosities of the two *C. sativa* var. *sativa* cultivars 'Finola' and 'Futura' (H_o=0.405 and H_e=0.603), were higher than those of the 20 *C. sativa* var. *indica* cultivars (H_o=0.282 and H_e=0.387). Three cultivars ('Critical', 'Hindu Kust' and 'Finola') had more than 50% of loci in heterozygous state ($H_o \geq 0.5$) (Figure 1). In some cultivars, like 'BL-57', 'Deimos', 'Purple', 'Kali Mist × Kali Mist', or 'Futura', the H_o values were considerably lower than those of H_e; in others, like 'NST-75-75', 'BL-26', 'Hindu Kust × Hindu Kust',

231 ‘Hindu Kust’ or ‘Critical’, both values were very similar. When considering the
232 maximum number of alleles per locus found among the six gSSR loci, there were two
233 cultivars with a maximum of two alleles per locus, nine with a maximum of three alleles,
234 six with a maximum of four alleles, two with a maximum of five, two with maximum of
235 six alleles, and one with a maximum of seven alleles per locus (Figure 2). The fixation
236 index (F_{ST}) and gene flow (*Nm*) values between *C. sativa* var. indica and *C. sativa* var.
237 *sativa* were F_{ST}=1.736 *Nm*=1.736.

238 The molecular analysis of variance (AMOVA) revealed that the variation among
239 individuals within cultivar accounts for a greater proportion of the molecular variance
240 (37.11%) than the variation among cultivars (31.94%) (Table 4). Within individual
241 variation due to heterozygosity accounted for 30.96% of the total molecular variance.
242 Analogous results were obtained when AMOVA was performed only on *C. sativa* var.
243 *indica* cultivars, as one third (33.34%) of the variation was detected between cultivars and
244 36.07% among individuals within cultivar, while 30.59% was due to heterozygosity
245 (Table 4).

246 The ΔK statistic had a maximum peak at K=14, indicating the presence of 14
247 clusters in the panel of 154 plants from 22 cultivars of *C. sativa*. In general, most of the
248 plants of a given cultivar belonged to the same cluster (Figure 3), although in some cases,
249 admixture of individuals from different clusters were found in the same cultivar. Most of
250 the individuals of the two *C. sativa* var. *sativa* varieties belonged to a single genetic
251 cluster, which was unique to *C. sativa* var. *sativa* individuals. Also, most of the
252 individuals of some different varieties of *C. sativa* var. *indica* belonged to the same
253 genetic cluster. This was the case of cultivars having a same origin, like ‘BL26’ and ‘BL-
254 58’, ‘NST-75-75’ and ‘NST-75-76’, or ‘Deimos-75-3’, ‘Deimos 75-4’ and ‘Deimos 75-5’

(Figure 3). In some instances, the majority of individuals from cultivars of different origins, like ‘Black Domina × Black Domina’ and ‘Kali Mist × Kali Mist’, or ‘Hindu Kush’ and ‘NST-75-78’ belonged to the same genetic cluster. Amazingly, in some cases, cultivars having a same origin clustered in different genetic groups, like ‘Deimos’ on one hand and ‘Deimos 75-3’, ‘Deimos-75-4’ and ‘Deimos-75-5’ on the other or ‘BL-26’ and ‘BL-58’ on one hand and ‘BL-57’ on the other.

The multivariate UPGMA cluster analysis performed with the 22 *C. sativa* cultivars showed that the two *C. sativa* var. *sativa* cultivars (‘Finola’ and ‘Futura’) cluster together, being differentiated from the *C. sativa* var. *indica* cultivars with a high bootstrap value (Figure 4). The analysis reveals that some groups of cultivars with similar names, like ‘BL-26’, ‘BL-57’ and ‘BL-58’ on one side, ‘Deimos-75-3’, ‘Deimos-75-4’ and ‘Deimos-75-5’ on another, or ‘NST-75-75’, ‘NST-75-76’ and ‘NST-75-75’ on another, cluster together in single branches, generally supported by high bootstrap values. On the other hand, some cultivars that have similar names, like ‘Deimos’ on one side and the ‘Deimos-75-3’, ‘Deimos-75-4’ and ‘Deimos-75-5’ group on the other, or ‘Hindu Kust’ and ‘Hindu Kust × Hindu Kust’, are found in separate branches (Figure 4).

4. Discussion

The high values obtained for molecular diversity statistics confirm the wide genetic diversity of *C. sativa* (Piluzza et al., 2013; Gao et al., 2014; Sawler et al., 2015) and the hypervariability of gSSR in *Cannabis* (Gilmore et al., 2003). In this respect, these latter authors found a similar number of alleles per locus (15.7 alelles/locus) in a sample of 93 individual plants from 15 cultivars of *C. sativa* genotyped with gSSRs (Gilmore et

al., 2003). On the other hand, Gao et al. (2014) using EST-SSRs for the characterization of 115 cultivars of *C. sativa* found a much lower number of alleles per locus (2.87). This confirms that, as in other crops (Eujayl et al., 2001; Muñoz-Falcón et al., 2011), gSSRs are more informative than EST-SSRs for genetic fingerprinting and diversity studies. The large number of alleles found per locus is probably caused by the fact that new SSR alleles can be produced at a high rate (Kalia et al., 2011), and that in the particular case of gSSRs it is assumed that most of them are neutral and so new allelic variants are not eliminated by selection. Also, the natural strictly allogamous reproductive system of *Cannabis sativa* (Small, 2005) would have contributed to the maintenance of large number of alleles. In all cases, with our dinucleotide gSSR markers we detected a maximum of two alleles per locus in individual plants, confirming the diploidy of *C. sativa* (van Bakel et al., 2011).

When comparing the two *C. sativa* var. *sativa* cultivars with the 20 *C. sativa* var. *indica*, some differences were found in heterozygosity. In this respect, the higher observed and expected heterozygosity values found in the former were probably caused by the higher inbreeding in *C. sativa* var. *indica* cultivars that reduces diversity through genetic drift in small populations and increases homozygosis. In the particular case of the new cultivars obtained by selection within a breeding stock, in general low values of observed heterozygosity have been observed compared to older prominent cultivars. Also, the fact that by means of chemical treatments it is possible to reverse sex in genetical female plants (Green, 2005) makes possible performing selfings of selected *C. sativa* var. *indica* plants. In fact, for some cultivars, a maximum of two alleles are found in the population, suggesting that these cultivars could come from the selfing of a single selected founder plant, resulting in a genetic bottleneck. In addition, selfing in *C. sativa*

induces inbreeding depression for vigour and plant size (Small, 2015), which may be detrimental in hemp production (Salentijn et al., 2015), but much less, and even may be favoured, in marijuana (Green, 2005). In any case, the lower heterozygosity in *C. sativa* var. *indica* with respect to *C. sativa* var. *indica* has not been a general phenomenon and the range in the observed heterozygosity has been very variable, reflecting differences in the breeding processes. In this respect, in some cultivars the expected heterozygosity is much higher than the observed one, revealing fixation probably due to inbreeding, while in others both values are similar. Despite the differences in genetic structure between *C. sativa* var. *indica* and *C. sativa* var. *sativa*, the estimate of fixation index (F_{ST}) among them is very low, while the gene flow (*Nm*) is high, considerably higher than *Nm*=0.5, indicating that gene flow between both botanical varieties is a major determinant of their genetic structure (Wright, 1951).

The high intra-varietal diversity observed has several implications of interest for breeding and conservation of germplasm (Salentijn et al., 2015; Weiling et al., 2016). On one hand this high diversity can be exploited for selection, and in fact many cultivars of *C. sativa* var. *indica* have been produced by selection within variable landraces or cultivars (Green, 2005). However, it also difficults producing genetically homogeneous cultivars, which can be needed for cultivars used for the medical industry, in which a high uniformity is needed (Potter, 2013). In these cases, the use of methods based on inbreeding that increase uniformity or the use of clonal propagation may be needed (Green, 2005; Lata et al., 2009). Regarding conservation of germplasm, the large intra-varietal diversity will require a high number of plants to be used in the multiplication of accesions to avoid loss of alleles at low frequencies and genetic drift (Weiling et al., 2016). Finally, the large intra-varietal diversity for gSSRs also suggests that other types of

markers may be needed to discriminate precisely between hemp and drug types. In this respect, Rotherham and Harbison (2011) suggested the use of a single single nucleotide polymorphism (SNP) from the tetrahydrocannabinolic acid synthase gene for distinguishing both types of *Cannabis* plants.

The partition of molecular variation reveals that only around one third of the molecular variation is attributable to differences among *C. sativa* var. *indica* cultivars. Gilmore et al. (2003) in a study of six cultivars of *C. sativa* var. *sativa* and five of *C. sativa* var. *indica* found similar results. However, this study contained mostly individuals of *C. sativa* var. *sativa* and therefore these results mostly represent *C. sativa* var. *sativa* intra-varietal variation. Datwyler and Weiblen (2006), using dominant amplified fragment length polymorphism (AFLP) markers in four *C. sativa* var. *indica* cultivars found that inter-individual variation accounted for 51.9% of the molecular variance. Anyway, due to the different nature of markers used (AFLPs) and low number of variaties used, the results may not be comparable to ours. Piluzza et al. (2013) also performed an AMOVA with AFLP data of 19 cultivars of *Cannabis* using two bulks of ten plants per cultivar, which obviously reduces the estimate of intra-accession diversity. Even under these conditions the intra-varietal component of variation was of 26%. All these results indicate that much of the molecular variation observed in *C. sativa* is attributable to differences within cultivars. This has important consequences for studies of relationships among *C. sativa* materials based on a single plant per cultivar, as depending on the plant used the results and interpretations may be very variable. Our study also reveals that a large proportion of the molecular variation observed is within-individual variation due to heterozygosis.

In our case, in which several plants per cultivar were used, the genetic structure analysis reveals a large number of genetic clusters (14) compared to other allogamous crops (Vigouroux et al., 2008; Richards et al., 2009; Hu et al., 2015), although most of the individuals of a single cultivar generally belong to the same genetic cluster. Nonetheless, the genetic structure analysis also shows that some cultivars present a high degree of genetic admixture, with individuals belonging to different genetic clusters, suggesting genetic flow and that these cultivars may not be stable from a genetic point of view. The genetic structure and UPGMA cluster analyses confirm the genetic differentiation of *C. sativa* var. *sativa* from *C. sativa* var. *indica* (Sawler et al., 2015). Both analysis also show that some cultivars with similar names cluster together, suggesting a common origin. These results confirm the utility of gSSRs for establishing relations among *C. sativa* cultivars, at least when several plants per cultivar are included in the study.

In conclusion, our study reveals that gSSRs are highly informative for studying the genetic structure, diversity and relationships in *C. sativa* var. *indica*. Despite evidence of inbreeding and genetic bottlenecks in some *C. sativa* var. *indica* cultivars, a large variation exists within these materials, which has important implications for selection, breeding and germplasm management, as well as for forensic applications.

References

Allgeier, L., Hemenway, J., Shirley, N., LaNier, T., Coyle, H.M. 2011. Field testing of collection cards for *Cannabis sativa* samples with a single hexanucleotide DNA marker. J. Forensic Sci. 56, 1245-1249.

373 Botstein, D., White, R.I., Skolnich, M., Davis, R.W. 1980. Construction of a genetic
374 linkage map in man using restriction fragment length polymorphism. Am. J. Genet.
375 32, 324-331.
376 Datwyler, S.L., Weiblen, G.D. 2006. Genetic variation in hemp and marijuana (*Cannabis*
377 *sativa* L.) according to amplified fragment length polymorphisms. J. Forensic Sci. 51,
378 371-375.
379 de Meijer, E.P.M., Keizer, L.C.P. 2012. Patterns of diversity in *Cannabis*. Genet. Res.
380 Crop. Evol. 43, 41-52.
381 Earl, D.A., vonHoldt, B.M. 2012. STRUCTURE HARVESTER: a website and program
382 for visualizing STRUCTURE output and implementing the Evanno method. Cons.
383 Genet. Resour. 4, 359-361.Eujayl, I., Sorrells, M., Baum, M., Wolters, P., Powell, W.
384 2001. Assessment of genotypic variation among cultivated durum wheat based on
385 EST-SSRS and genomic SSRS. Euphytica 119, 39-43.
386 Eujayl, I., Sorrells, M., Baum, M., Wolters, P., Powell, W. 2001. Assessment of
387 genotypic variation among cultivated durum wheat based on EST-SSRS and genomic
388 SSRS. Euphytica 119, 39-43.
389 Evanno, G., Regnaut, S., Goudet, J. 2005. Detecting the number of clusters of individuals
390 using the software STRUCTURE: a simulation study. Mol. Ecol. 14, 2611-2620.
391 Felsenstein, J. 2007. PHYLIP (Phylogeny Inference Package) version 3.67. Department
392 of Genome Sciences, University of Washington, Seattle, WA, USA.
393 Gao, C., Xin, P., Cheng, C., Tang, Q., Chen, P., Wang, C., Zang, G., Zhao, L. 2014.
394 Diversity analysis in *Cannabis sativa* based on large-scale development of expressed
395 sequence tag-derived simple sequence repeat markers. PLOS ONE 9, e110638.

396 Gilmore, S., Peakall, R., Robertson, J. 2003. Short tandem repeat (STR) DNA are
397 hypervariable and informative in *Cannabis sativa* L.: implications for forensic
398 investigations. Forensic Sci. Intl. 9, 65-74.
399 Green, G. 2005. The Cannabis breeder's bible. Green Candy Press, San Francisco, USA.
400 Highton, R. 1993. The relationship between the number of loci and the statistical support
401 for the topology of UPGMA trees obtained from genetic distance data. Mol.
402 Phylogenet. Evol. 2, 337-343.
403 Howard, C., Gilmore, S., Robertson, J., Reakall, R. 2008. Developmental validation of a
404 *Cannabis sativa* STR multiplex system for forensic analysis. J. Forensic Sci. 53, 1061-
405 1067.
406 Hu, J., Wang, P., Su, Y., Wang, R., Li, Q., Sun, K. 2015. Microsatellite diversity,
407 population structure, and core collection formation in melon germplasm. Plant Mol.
408 Biol. Rep. 33, 439-447.
409 Kalia, R.K., Rai, M.K., Kalia, S., Singh, R., Dhawan, A.K. 2011. Microsatellite markers:
410 an overview of the recent progress in plants. Euphytica 177, 309-334.
411 Lata, H., Chandra, S., Khan, I.A., ElSohly, M.A. 2009. Propagation through alginate
412 encapsulation of axillary buds of *Cannabis sativa* L. – an important medicinal plant.
413 Physiol. Mol. Biol. Plants 15, 79-86.
414 Liu, K., Muse, S.V. 2005. PowerMarker: integrated analysis environment for genetic data.
415 Bioinformatics 20, 2128-2129.
416 Moliterni, V.M.C., Cativelli, L., Ranalli, P., Mandolino, G. 2004. The sexual
417 differentiation of *Cannabis sativa* L.: a morphological and molecular study. Euphytica
418 140, 95-106.

Muñoz-Falcón, J.E., Vilanova, S., Plazas, M., Prohens, J. 2011. Diversity, relationships, and genetic fingerprinting of the *Listada de Gandía* eggplant landrace using genomic SSRs and EST-SSRs. Sci. Hort. 129, 238-246.

Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA 70, 3321-3323.

Nei, M., Tajima, F., Tateno, Y. 1983. Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. J. Mol. Evol. 19, 153-170.

Peakall, R., Smouse, P.E. 2012. GenAlEx 6.5: genetic analysis in Excel. Population genetic analysis for teaching and research-an update. Bioinformatics 28, 2537-2539.

Piluzza, G., Delogu, G., Cabras, A., Marceddu, S., Bullitta, S. 2013. Differentiation between fiber and drug types of hemp (*Cannabis sativa*) from a collection of wild and domesticated accessions. Genet. Resour. Crop Evol. 60, 2331-2342.

Prichard, J.K., Stephens, M., Donnelly, P. 2000. Inference of population structure from multilocus genotype data. Genetics 155, 945-959.

Potter, D.J. 2013. A review of the cultivation and processing of cannabis (*Cannabis sativa* L.) for production and prescription medicines in the UK. Drug Testing Anal. 6, 31-38.

Razumova, O.V., Alexandrov, O.S., Divashuk, M.G., Sukhorada, T.I., Karlov, G.I. 2016. Molecular cytogenetic analysis of monoecious hemp (*Cannabis sativa* L.) cultivars reveals its karyotype variations and sex chromosomes constitution. Protoplasma 253, 895-901.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M., Buckler, E.S. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98, 11479-11484.

443 Richards, C.M., Volk, G.M., Reilley, A.A., Henk, A.D., Lockwood, D.R., Reeves, P.A.,
444 Forsline, P.L. 2009. Genetic diversity and population structure in *Malus sieversii*, a
445 wild progenitor species of domesticated apple. Tree Genet. Genom. 5, 339-347.
446 Rotherham, D., Harbison, S.A. 2011. Differentiation of drug and non-drug *Cannabis*
447 using a single nucleotide polymorphism (SNP) assay. Forensic Sci. Intl. 207, 193-197.
448 Salentijn, E.M.J., Zhang, Q., Amaducci, S., Yang, M., Trindade, L.M. 2015. New
449 developments in fiber hemp (*Cannabis sativa* L.) breeding. Ind. Crops Prod. 68, 32-
450 41.
451 Sawler, J., Stout, J.M., Gardner, K.M., Hudson, D., Vidmar, J., Butler, L., Page, J.E.,
452 Myles, S. 2015. The genetic structure of marijuana and hemp. PLOS ONE 10,
453 e0133292.
454 Shirley, N., Allgeier, L., LaNier, T., Coyle, H.M. 2013. Analysis of the NMI01 marker
455 for a population database of *Cannabis* seeds. J. Forensic Sci. 58, S176-S182.
456 Small, E. 2015. Evolution and classification of *Cannabis sativa* (marijuana, hemp) in
457 relation to human consumption. Bot. Rev. 81, 189-204.
458 Soler, S., Borràs, D., Vilanova, S., Sifres, A., Andújar, I., Figàs, M.R., Llosa, E.R.,
459 Prohens, J. 2016. Use of embryos extracted from individual *Cannabis sativa* seeds for
460 genetic studies and forensic applications. J. Forensic Sci. 61, 494-500.
461 van Bakel, H., Stout, J.M., Cole, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R., Page,
462 J.E. 2011. The draft genome and transcriptome of *Cannabis sativa*. Genome Biol. 12,
463 R102.
464 van Ours, J.C. 2012. The long and winding road to *Cannabis* legalization. Addiction 107,
465 872-873.

Vigouroux, Y., Claubitz, J.C., Matsuoka, Y., Goodman, M.M., Sánchez, J., Doebley, J. 2008. Population structure and genetic diversity of New World maize races assessed by DNA microsatellites. Amer. J. Bot. 95, 1240-1253.

Welling, M.T., Shapter, T., Rose, T.J., Liu, L., Stanger, R., King, G.J. 2016. A belated green revolution for *Cannabis*: virtual genetic resources to fast-track cultivar identification. Frontiers Plant Sci. 7, 1113.

Wright, S. 1951. The genetical structure of populations. Ann. Eugenics 15, 323-354.Wright, S. 1965. The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19, 395-420.

Table 1

Cultivars of *C. sativa* var. *indica* and *C. sativa* var. *sativa* used for genotyping with SSR markers, including the number of plants genotyped, origin and sexual type.

Cultivar	n	Origin	Sexual type
C. sativa var. *indica*			
‘BL-26’	4	New cultivar derived by selection from the BL breeding stock	Female
‘BL-57’	8	New cultivar derived by selection from the BL breeding stock	Female
‘BL-58’	8	New cultivar derived by selection from the BL breeding stock	Female
‘Black Domina × Black Domina’	8	New cultivar derived by intercrossing of selected individuals of cultivar ‘Black Domina’	Female
‘Critical’	7	Prominent cultivar of unknown pedigree	Female
‘Deimos’	8	Cultivar of unknown pedigree	Female
‘Deimos-75-3’	7	New cultivar derived by selection from cultivar ‘Deimos’	Female
‘Deimos-75-4’	7	New cultivar derived by selection from cultivar ‘Deimos’	Female
‘Deimos-75-5’	8	New cultivar derived by selection from cultivar ‘Deimos’	Female

Gitane	6	Cultivar derived from intercrossing of cultivars ‘Super Silver Haze’ and ‘Critical Chiva’	Female
‘Hindu Kush’	6	Prominent cultivar of unknown pedigree	Female
‘Hindu Kush × Hindu Kush’	8	New cultivar derived by intercrossing of selected individuals of cultivar ‘Hindu Kush’	Female
‘Kali Mist × Kali Mist’	8	New cultivar derived by intercrossing of selected individuals of cultivar ‘Kali Mist’	Female
‘Northern Lights × Northern Lights’	8	New cultivar derived by intercrossing of selected individuals of cultivar ‘Northern Lights’	Female
‘NST-75-75’	6	New cultivar derived by selection from the NST breeding stock	Female
‘NST-75-76’	7	New cultivar derived by selection from the NST breeding stock	Female
‘NST-75-78’	7	New cultivar derived by selection from the NST breeding stock	Female
‘Purple’	7	Prominent cultivar of unknown pedigree	Female
‘Somango × Somango’	7	New cultivar derived by intercrossing of selected individuals of cultivar ‘Somango’	Female

‘White Widow × White Widow’	6	New cultivar derived by intercrossing of selected individuals of cultivar ‘White Widow’	Female
C. sativa var. *sativa*			
Finola	6	Cultivar developed by Finola ky (Finland)	Dioecious
Futura	7	Cultivar developed by FNPC (France)	Monoecious

478

Table 2

Characteristics of the six genomic SSR (gSSR) markers used for molecular characterization of *C. sativa* materials. These SSR markers were obtained from the genome sequence of *C. sativa* var. *sativa* 'Purple Kush' (van Bakel et al., 2011).

SSR marker	Genomic scaffold	Repeat motif	Size (bp)	Primer sequence (5'-3')	Annealing temperature (ºC)
CSG01	309	$(AT)_{86}$	350	F- ACAACCAACCTGGAATCTGC	59.5
				R- TCAATCTGTGTGCTGTGTGC	
CSG05	5874	$(AC)_{50}$	307	F- GCCCATAAGGGGGTTTGTAT	60
				R- CTCACCTCGTCCTTTGATCC	
CSG12	160138	$(AG)_{47}$	144	F-TCCAACACCTTGATGCTCTG	59
				R-GGGCATAAAACCTAACATGAGA	
CSG13	17433	$(AG)_{51}$	189	F-TGGTTTCTACTCCCTCTTACTCG	59.5
				R-TTCAAGCTCCAAATCAAGCA	
CSG18	829	$(AT)_{72}$	236	F-CCGGTGGTTGTGGAGATGAT	59
				R-GATCAGAGTAGAGAGAGGCGA	
CSG24	4142	$(AG)_{53}$	343	F-CAGGGAAGCAATTGGAAAAA	60
				R-GATGAGGAGAAAAGGGCACA	

Table 3

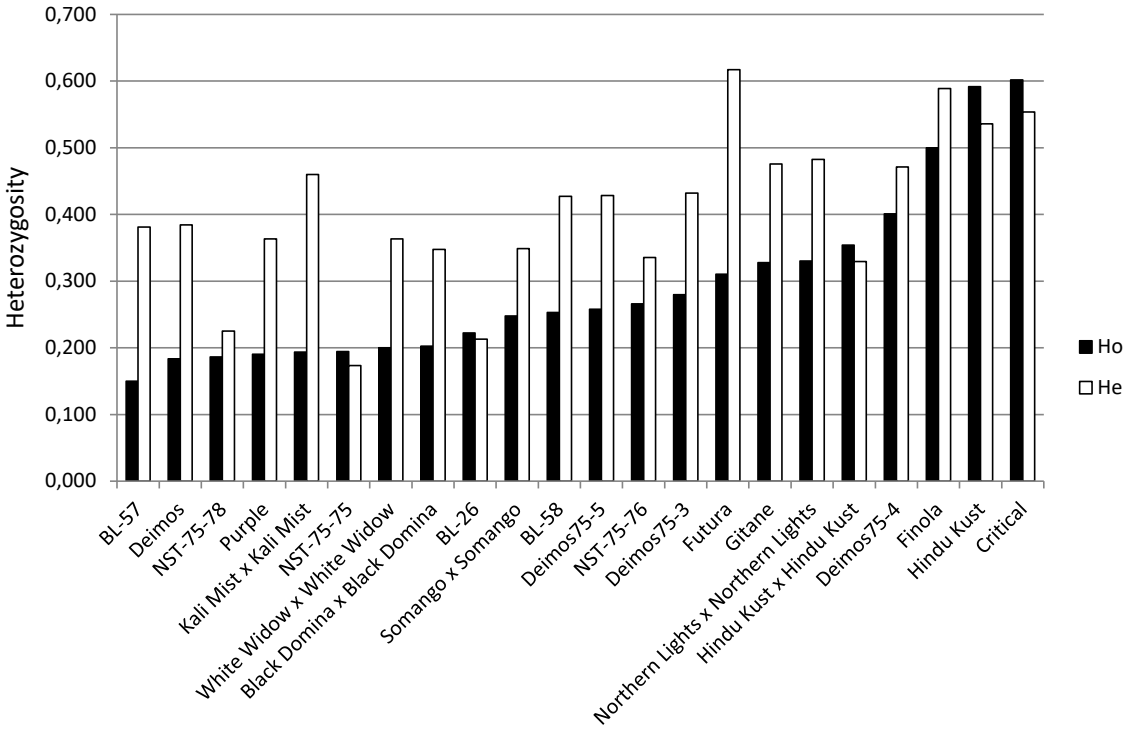
Diversity statistics for the six genomic SSR (gSSR) markers studied in a collection of 22 *C. sativa* accessions.

gSSR locus	Number of alleles	Number of genotypes	Allele size range	Major allele frequency	Expected heterozygosity (H_e)	Observed heterozygosity (H_o)	Fixation index (F_{is})	Polymorphic information content (PIC)
CSG01	9	12	218-231	0.332	0.770	0.049	0.936	0.735
CSG05	25	32	198-375	0.243	0.867	0.296	0.659	0.854
CSG12	30	41	144-340	0.149	0.919	0.304	0.669	0.914
CSG13	10	16	114-147	0.625	0.573	0.429	0.251	0.544
CSG18	8	13	198-215	0.593	0.567	0.304	0.464	0.509
CSG24	20	29	318-340	0.286	0.823	0.379	0.539	0.804
Mean	17.0	23.8	-	0.371	0.753	0.293	0.587	0.727

Table 4

Molecular analysis of variance (AMOVA) among cultivars, among individuals and within individuals for the 22 *C. sativa* cultivars evaluated (20 *C. sativa* var. *indica* plus 2 *C. sativa* var. *sativa*), and for the subset of 20 *C. sativa* var. *indica* cultivars. Results are based on six genomic SSR (gSSR) markers.

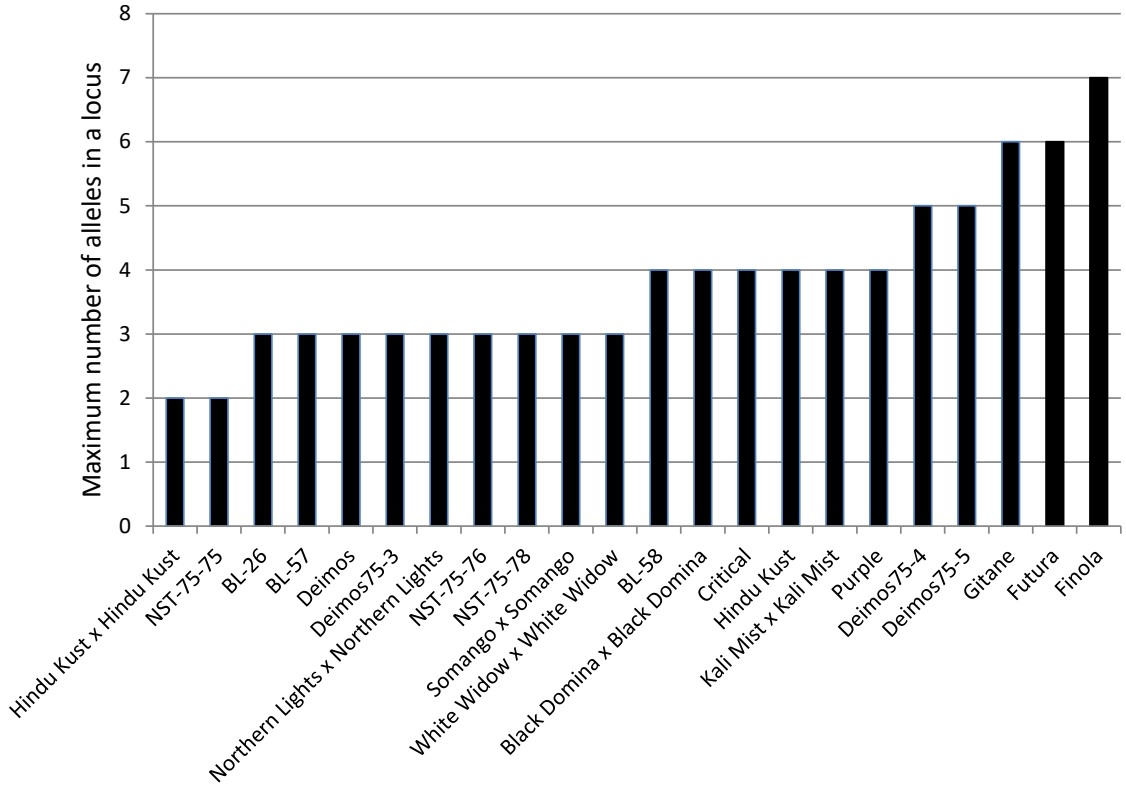
Source of variation	d.f.	S.S.	M.S.	Variance components	Percentage of variation
		C. sativa (n=22)			
Total	307	729.41		2.412	
Among cultivars	21	279.54	13.312	0.770	31.94
Among individuals	132	334.87	2.537	0.895	37.11
Within individuals	154	115.00	0.747	0.747	30.96
		C. sativa var. *indica* (n=20)			
Total	281	653.06		2.365	
Among cultivars	19	257.12	13.53	0.788	33.34
Among individuals	121	293.94	2.429	0.853	36.07
Within individuals	141	102.00	0.723	0.723	30.59



495

496 **Fig. 1.** Observed (H_o; black bars) and expected (H_e, white bars) heterozygosity based on
497 genomic SSR (gSSR) markers for each of the *C. sativa* cultivars evaluated. Cultivars are
498 ordered according to increasing values for the H_o value.

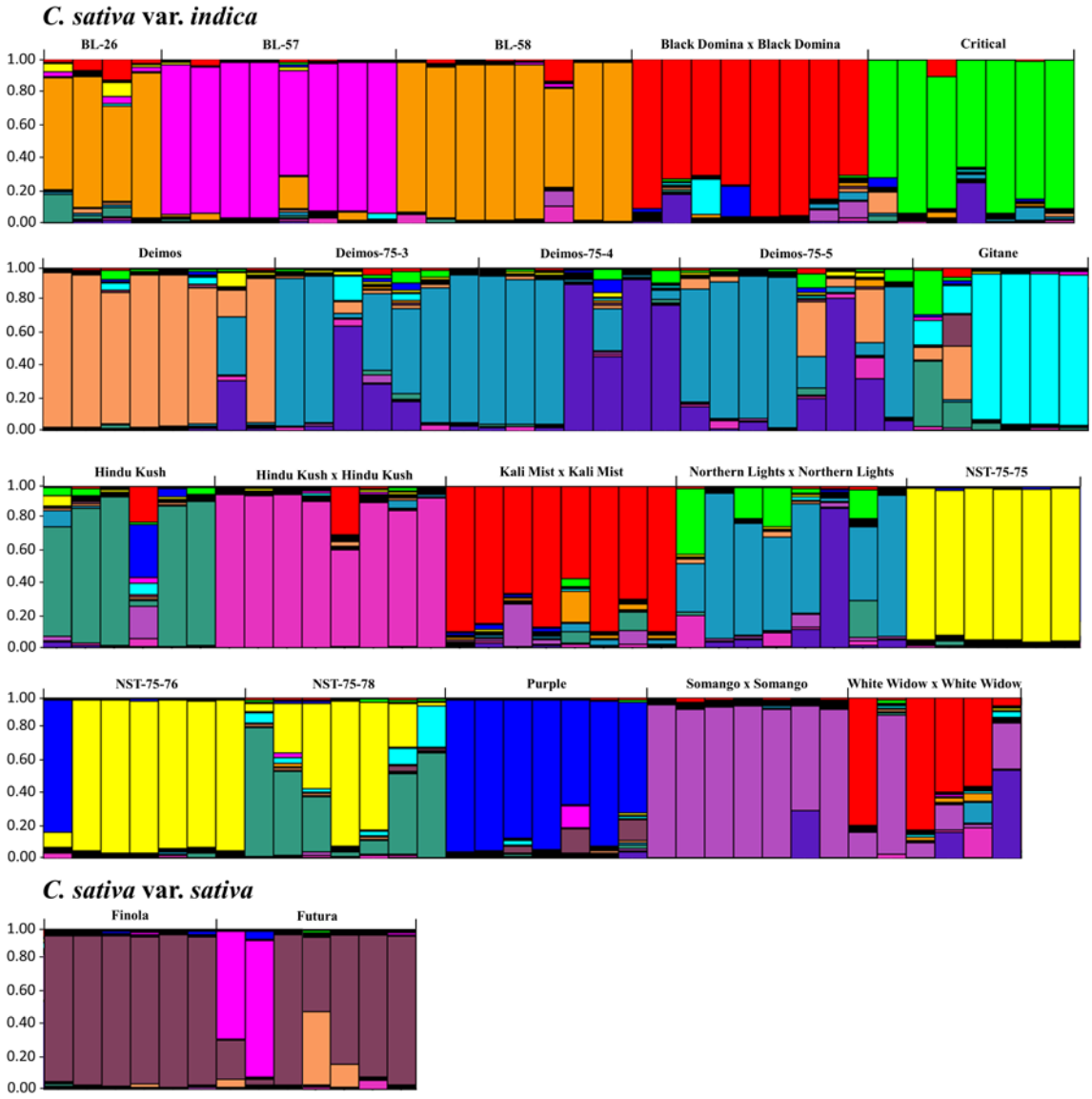
499



500

501 **Fig. 2.** Maximum number of alleles found in a single genomic SSR (gSSR) locus among
502 the six gSSR loci scored for each of the *C. sativa* cultivars evaluated. Cultivars are
503 ordered according to increasing values of number of alleles.

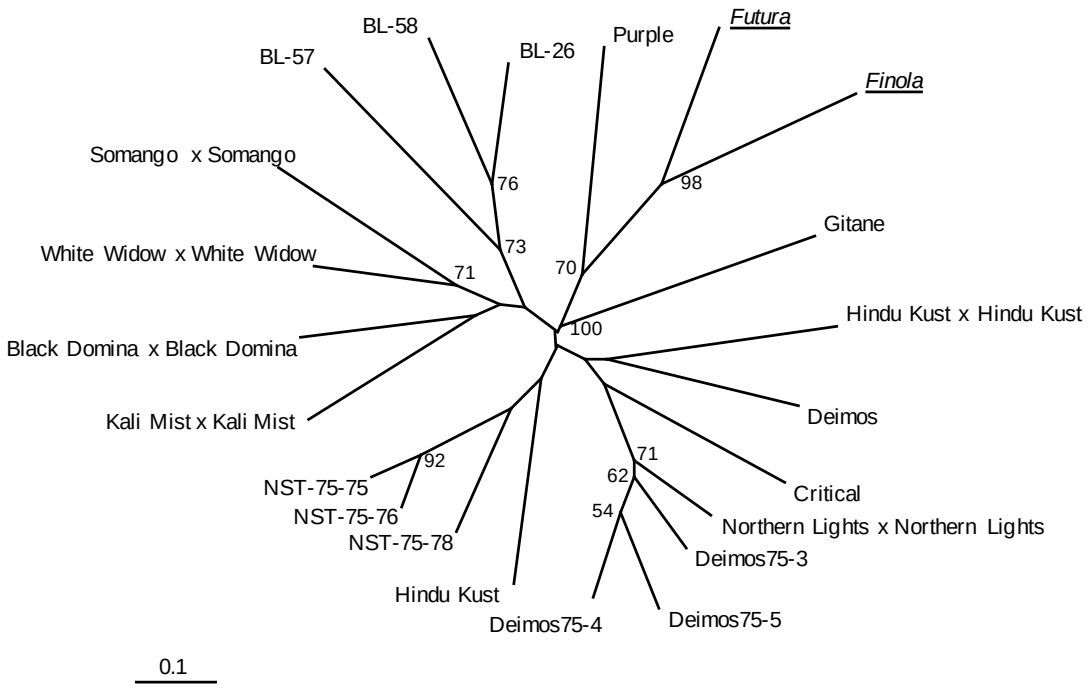
504



505

506 **Fig. 3.** Estimated population structure for 154 individual plants of 20 *Cannabis sativa* var.
507 *indica* and two *C. sativa* var. *sativa* cultivars based on a number of clusters (*K*) of *K*=14.
508 Each individual plant is represented by a vertical bar, which is partitioned into coloured
509 segments that provide an estimate of the membership fraction in each of the 14 genetic
510 clusters.

511



512

513 **Fig. 4.** Unrooted neighbor-joining phenogram of 22 *C. sativa* cultivars based on six
514 polymorphic genomic SSR (gSSR) markers. Phenetic relationships among cultivars were
515 derived from Nei et al. (1983) genetic distances. Bootstrap values (based on 1,000
516 replications; expressed in percentage) greater than 50% are indicated at the corresponding
517 nodes. *Cannabis sativa* var. *indica* cultivars are represented in normal font, while *C.*
518 *sativa* var. *sativa* in italics and underlined.