A comprehensive molecular approach to the detection of drug-type *versus* fiber-type hemp varieties

ABSTRACT

The availability of molecular markers able to distinguish drug-type from fiber-type *Cannabis sativa* cultivars would allow fast and cheap analysis of any plant specimen, including seeds and leaves. Several approaches to this issue have been described, using either random amplified polymorphic DNA or single nucleotide polymorphisms. The possibility of using polymorphisms in the genes coding for tetrahydrocannabinol acid synthase or cannabidiolic acid synthase for the design of specific primers has attracted increasing interest. Some studies reported sequencing of these genes from small groups of hemp varieties belonging to both chemotypes, showing the occurrence of specific DNA signatures. However, the effectiveness of the corresponding primers to discriminate among chemotypes has been validated on a limited number of cultivars, or not tested at all. Here we report a thorough *in silico* analysis of available gene sequences for both tetrahydrocannabinol acid and cannabidiolic acid synthases, showing the existence of hypervariable regions at 3' and 5' ends. This notwithstanding, some possible signatures were identified, and 12 putatively specific primer pairs were designed and tested on 16 fiber-type and 11 drug-type varieties. In most cases inconsistent results were obtained, further strengthening the high genetic variability of these genes in hemp germplasm, yet some highly informative polymorphisms were identified. Potentiality and perspectives of this approach are discussed.

Keywords:

Cannabis sativa L.

Fiber hemp

Drug hemp

Chemotype

Single nucleotide polymorphism

Molecular discrimination

Polymerase Chain Reaction

Hemp (*Cannabis sativa* L.) is attracting increasing interest as a sustainable industrial crop for fibers to replace cotton or synthetic materials in a variety of applications, such as in paper, textiles, fabrics, and various construction materials [1]. In the last century hemp cultivation had been substantially discontinued worldwide, mainly because of banning laws adopted to limit the narcotic use of marijuana (*Cannabis indica* L.). Since the two species easily interbreed, making it difficult to distinguish legal from illegal varieties, in many cases the law did not differentiate industrial hemp from psychoactive cannabis [2]. In recent years many countries, comprising the U.S.A. and Italy, reintroduced hemp and legalized its production as an agricultural commodity, leading to a renaissance of this crop [3]. On the other hand, although some countries also authorized fully or in part the medical use of herbal cannabis, in most cases its recreational use remains prohibited [4]. This led to the current jeopardised situation, where extremely different rules govern the conditions under which hemp can be cultivated.

Besides the loss of germplasm well adapted to local pedoclimatic conditions that occurred during prohibition, with the consequent need of new breeding programs, other problems currently limiting the sector revival are unclear agronomic guidance and fertilization recommendations [5], and the unavailability of easy, fast and cheap analytical methods to distinguish between (legal) fiber and (in most cases still prohibited) drug varieties. The two chemotypes differ for their content of δ^9-tetrahydrocannabinol (THC), one of a hundred cannabinoids identified in *Cannabis* spp. Fiber-type cultivars should contain low amounts of this addictive compound, and higher concentration of cannabidiol (CBD). In the U.S.A., Canada, Switzerland and Asia, limits for THC vary from 0.3% to 1.0%, whereas in the European Community the legal threshold is as low as 0.2% [4], causing several old industrial varieties to be discontinued. Although some simple immunological methods for THC detection have been proposed [6], reliable cannabinoids quantitation requires complex protocol of extraction and analysis [e.g. 7]. Moreover, the levels of these substances differ significantly among plant tissues [8], being synthesized and accumulated mainly in thricomes in floral organs [9], a fact that hampers the possibility of reliable analysis of other specimens, such as leaves and seeds.

In this view, the availability of molecular markers discriminating drug-type from fiber-type hemp varieties would be of great interest. Several studies have been undertaken with this aim. Inter simple sequence repeats analysis was applied to 9 fiber-type and 23 drug-type varieties, and principal component analysis of data was shown to discriminate between the groups. Yet, if unweighted pair-group methods were used, no clear separation was obtained. Fiber-type accessions showed high levels of variation compared to drug-type [10]. Using the Random Amplified Polymorphic DNA method, six random decamers were reported to distinguish chemotypes.

However, a complex cluster analysis of data was required, and only 5 fiber-type and 10 drug-type varieties were considered [11].

Since single nucleotide polymorphism (SNP) assay would allow simpler and faster discrimination, in their pioneer work [12] Kojoma and co-workers studied the occurrence of SNPs in the genomic DNA sequence for δ^9-tetrahydrocannabinolic acid synthase (THCAS), the enzyme channeling the intermediate cannabigerolic acid toward THC synthesis [13]. By sequencing *THCAS* from 6 drug-type and 7 fiber-type varieties, 37 major substitutions were detected in the alignment of the deduced amino acid sequences, and a specific PCR marker for the drug-type strains was identified [12]. The same approach was used with another small group of 12 drug-type and 4 fiber-type Moroccan isolates, confirming the occurrence of a significant variability in the *THCAS* sequence, and showing some possible diagnostic SNPs [14]. Four polymorphisms within a 399 bp fragment among those described by Kojoma and co-workers were used to distinguish putatively active and inactive THCAS. It was claimed that this SNP assay was able to differentiate chemotypes when used to screen a hundred hemp varieties, where non-drug plants were found to be homozygous at the four sites, while drug plants were either homozygous or heterozygous [15], but no confirmative data were reported. Moreover, THC content in hemp floral tissues does not seem to depend on the presence of active *vs* inactive THCAS forms, but on the competition for cannabigerolic acid among THCAS and other two enzymes catalyzing its oxidocyclization, cannabidiolic acid synthase (CBDAS) and cannabichromenic acid synthase [16]. CBDAS, which directs the precursor into the branch of the biosynthetic pathway leading to CBD, shows a surprising level of homology to THCAS, with about 80-85% identity in a 550-amino acid overlap [17]. Also based on the results of a genetic analysis showing an approximately 1:2:1 segregation of chemotypes in a cross of drug-type *vs* fiber-type cultivars [18], it was therefore hypothesized that THCAS and CBDAS were allelic and co-dominant [17]. A PCR-based marker identified in two segregating populations was linked to either the THC-predominant or the THC-intermediate chemotype (*i.e.*, the presence of at least one *THCAS* allele). When the co-dominant marker system was applied to a larger number of samples in commercially available material, it apparently discriminated most of 25 fiber-type from 12 drug-type cultivars [19]. The analysis of 10 drug-type and 8 fiber-type accessions found out that in several cases more than one expressed sequence for THCAS and CBDAS was present, each showing an ORF allowing translation into an entire protein. The transcription rate of the different sequences was not correlated with the proportion of THCA or CBDA in the cannabinoid fraction. The comparison of the expressed sequences led to the identification of different SNPs in both alleles, which were found to relate to the cannabinoid composition of the inflorescence, and were thus proposed to have a functional significance [20]. Nevertheless, none among these polymorphisms was shared exclusively by the accessions belonging to a given chemotype. The evaluation of an F2 population resulting from the cross of marijuana (Skunk #1) and hemp (Carmen) varieties pointed out the presence in the former of two CBDAS nonfunctional homologs, leading to conclude that plants that are homozygous for functional CBDAS lack the capacity to accumulate THC [21]. This notwithstanding, neither the general occurrence of *CBDAS* containing premature stop codons or frame shift mutations in drug-type cultivars, nor the possible existence of consequently distinctive DNA signatures to be used for SNPs analysis were investigated.

All these studies were in part superseded by recent sequencing data providing conclusive evidence that the THCAS and CBDAS scaffolds are not allelic, but at separate loci, though adjacent on the same chromosome [16]. On this basis, *CBDAS* and *THCAS* from 11 drug-type and 10 fiber-type hemp varieties were sequenced, allowing the identification of multiple genetic markers that discriminated between the two chemotypes. In particular, four functional SNPs that were hypothesized to induce decreased THCAS activity in the fiber-type plants, and a deletion in the CBDAS gene possibly resulting in loss of function of the enzyme in the drug-type varieties were reported [22]. This work was remarkable in that the identification of markers was based on both THCAS and CBDAS, yet neither an *in silico* analysis on other available sequences, nor a confirmative PCR assay was performed.

In summary, several SNPs potentially representing a signature for hemp chemotypes have been proposed, but in most cases their ability to discriminate between drug-type and fiber-type varieties has been evaluated only on the same cultivars whose sequencing had represented the basis for primer design (yielding obviously positive results), or has not been tested at all. Moreover, a large scale single molecule sequencing of *THCAS* showed the occurrence of a highest genetic heterogeneity in drug varieties, both for gene copy number and sequence variation [23], questioning the possibility of chemotype prediction based on simple molecular markers. To address this point, we analysed all *CBDAS* and *THCAS* sequences described in previous papers [12-14, 17, 19-23] and available in public databases. Some putatively discriminating primers were designed, and verified on a group of 27 mostly unsequenced commercial hemp varieties of both chemotypes.

1. Materials and methods

1.1. In silico analysis

Sequences for *C. sativa THCAS* and *CBDAS* that had been described in previous studies [12-14, 17, 19-23] were retrieved from GenBank (https://www.ncbi.nlm.nih.gov/genbank/), and other sequences available in databanks were found using BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) analysis [24], with default settings. The accession numbers of the 145 THCAS and 38

CBDAS genes considered are reported in Supplementary Table S1. Sequences were aligned using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) [24]. The resulting Neighbour-joining tree was visualized with T-REX [25].

1.2. Primer design

Primers were designed with the PRISE2 software [26] with the following settings: primer length 18-24 bp, amplicon size 200-500 bp, melting temperature 48-58 °C, GC content 40-60%.

1.3. Plant samples

Seeds of fiber-type hemp varieties (#01-16) were purchased on the European market. Seeds of drug-type varieties (#17-27) were obtained from Sensi Seeds (Oudezijds Achterburgwal 131, 1012DE Amsterdam, The Netherlands).

1.4. Isolation of DNA

DNA was isolated from a single seed using the REDExtract-N-Amp™ Plant PCR Kit (Sigma XNAP). The seed coat was crushed with tweezers in a 0.2-mL thin wall PCR tube, and the material was resuspended in 0.1 mL of extraction solution. The solubilization was allowed to proceed for 10 min at 95°C, then samples were brought back to room temperature and immediately diluted with 0.1 mL of neutralizing solution. DNA concentration was not quantified, and extracts were stored at 4 ± 1°C. Just before the analysis, samples were centrifuged 30 s at 12,000 *g*.

1.5. PCR

PCRs were carried out in a final volume of 15 µL containing 7.5 µL of 2X REDExtract-N-Amp™ PCR ReadyMix™ (Sigma R4775), 0.5 µM each of forward and reverse primers and 3 µL of template DNA. Unless specified otherwise, PCR conditions were 95 °C for 3 min and 35 cycles of 94 °C for 1 min, 49 °C for 30 s, and 72 °C for 1 min, with a final extension at 72 °C for 5 min. Samples were then brought to 4 °C and immediately analysed.

1.6. Analysis of PCR products

Amplification products were separated by electrophoresis on 1% agarose gels run at 44 V, stained with a 1:5000 dilution of the fluorescent dye Nancy-520 (Sigma 01494), and visualized under blue light. Images were acquired with a Gel Doc 2000 system and the Quantity One software (Bio-Rad). Semi-quantitative evaluation of amplification products was obtained by measurement of band intensities using the ImageJ software [27].

2. Results

2.1. In silico analysis of THCAS and CBDAS sequences from fiber-type and drug-type hemp varieties

A total of 38 *CBDAS* and 145 *THCAS* sequences for *Cannabis sativa* were retrieved from public databases. For 36 and 67 of them information about the chemotype of the plant from which the gene had been cloned was also available in the literature, respectively (Supplementary Table S1). The sequences were aligned (Supplementary Fig. S1), showing a relatively large, highly conserved central region between approximately residues 400 and 750, with 69% identity over a 346 bp stretch. On the contrary, at both 5' and 3' ends no identities were found. A Neighbour-joining tree generated from the aligned sequences (Fig. 1) clearly disclosed two distinct clusters, as expected. A first, more homogeneous group contained all *CBDAS* but one, the second one included all *THCAS* as well as accession number AB292683.1, which had been annotated as a "*CBDAS* homologue". Interestingly, within each cluster two subgroups were evident. In the case of *CBDAS*, although at a low phylogenetic distance, the two subsections perfectly resolved all genes from drug-type varieties from those cloned in fiber-type varieties. For *THCAS*, two clades at a more remarkable genetic distance were found. Also in this case most sequences of a given chemotype clustered together: in the first clade 23 genes belonging to fiber-type cultivars were present, along with the drug-type accession numbers KJ469379.1 and JQ437490.1, whereas in the second clade 40 genes belonging to the drug-type cultivars were present, plus the fiber-type accession numbers KJ469381.1 and KJ469383.1 (Supplementary Fig. S2). Results thus suggest that, following the duplication of an ancestral cannabinoid acid synthase using cannabigerolic acid as substrate, the two paralogs evolved separately leading to a different enzymatic product from the same precursor, *i.e.* to THCAS or CBDAS activity. In both paralogs a divergent evolution has thereafter occurred that most likely caused a prevalent activity of either enzyme, leading to the alternative accumulation of THC or CBD.

2.2. Design of putatively specific primers for either gene and chemotype

The obtained picture was consistent with the possibility to find genetic signatures that would allow chemotype discrimination. To identify specific primers, at first an overall analysis with all *CBDAS* and *THCAS* sequences was carried out. Both genes were considered in order to rule out the possibility that a primer couple able to distinguish fiber-type from drug-type *CBDAS* would cause amplification of a corresponding sequence in *THCAS*, and *vice-versa*. In other terms, a clade was used as target cluster, and the other clade of the same gene and both clades of the other gene were defined as non-target sequences, in all four possible combinations. However, because of the high genetic variability and the large number of sequences, no discriminating primers were recognized in such a way using the PRISE2 software. To overcome this difficulty, the analysis was repeated using only 8 sequences, two for each group: fiber-type (MG996405.1 and AB212830.1) and drug-type (KJ469378.1 and JQ437491.1) *THCAS*, and fiber-type (MG996434.1 and KP970859.1) and drug-type (KJ469376.1 and KJ469375.1) *CBDAS*, respectively. With this approach, 12 putatively specific primer pairs were identified (Table 1). The presence of primer target sequences in most -if not all- genes of the same cluster was then verified, as well as the occurrence of at least 2-3 mismatches in the sequences of the other 3 clusters (Supplementary Fig. S3A-L). The position of each primer pair within the two aligned genes is shown in Supplementary Fig. S4.

2.3. Validation of the identified SNPs as chemotype markers

The presence of a conserved sequence in all genes in a cluster, coupled with the presence of some mismatches in the genes of the other clusters, should allow molecular discrimination. However, available *THCAS* and *CBDAS* sequences represent only a minimal part of hemp germplasm, thus the possibility exists that with cultivars not included in the above sequence analysis the designed primer pairs would not result into the expected amplification pattern. To investigate this aspect, DNA was extracted from single seeds of a set of 16 fiber-type and 11 drug-type varieties (Table 2). Among them, for only 3 (namely Carmagnola selezionata, Futura 75 and Uso 31) and 1 (Northern Lights) *THCAS* and *CBDAS* sequences were available, respectively [22]. PCR was carried out with the 12 putatively specific primer pairs, *plus* a thirteenth pair (T) designed to amplify a region that had been found highly conserved in both *THCAS* and *CBDAS* (Supplementary Fig. S3T), to be used as a positive control. DNA was not quantified, and low stringency conditions and high number of cycles were used to ensure amplification. Three different seeds for each genotype were separately extracted and analysed, and consistent results were obtained without exceptions. Results, summarized in Table 2, pointed out a heterogenous picture. With primers pair T, despite conservation in all known sequences, amplification was not obtained in one case (Fedora 17). Moreover, even though the sequence was expected to be present in both *THCAS* and *CBDAS*, the amplicon quantity obtained (expressed as percent band intensity of the more abundant amplification product among those obtained with the 13 primer pairs) varies greatly among genotypes, from 45 to 87% (not shown). Conversely, when using primer pair B that had been designed to amplify *CBDAS* from fiber- and not from drug-type varieties, the most abundant amplification product was obtained in all cases but two. Considering the other primer pairs, in most cases results differed significantly from the expected amplification patterns. With primer pair L, designed to amplify *THCAS* from drug-type varieties, amplification was obtained also with 15 out of 16 fiber-type varieties. With pairs D and F, designed to amplify *CBDAS* from drug-type varieties, a consistent result was obtained for target varieties, but amplification occurred also with several cultivars of the other chemotype. This notwithstanding, consistent results were obtained with primer pair A, designed to amplify *CBDAS* from fiber-type varieties, and pair K, designed to amplify *THCAS* from drug-type varieties. In both theses cases all target sequences were identified, while non-target sequences did not result in amplicon formation, with only 2 exceptions out of 16 for primer pair K (Fig. 2).

3. Discussion

The availability of molecular markers to distinguish legal from illegal hemp varieties would represent a very attractive result, greatly facilitating the resurgence of this crop as an agricultural commodity worldwide. Because the two main chemotypes are characterized by the mutually exclusive accumulation of CBD and THC, most studies focused on the properties of the enzymes channeling the common precursor cannabigerolic acid into either the biosynthetic branch leading to these cannabinoids, namely CBDAS and THCAS. However, the attainment of this goal was initially hampered by the erroneous assumption that *THCAS* and *CBDAS* were allelic and co-dominant [17-21], a hypothesis that misdirected the attention toward differences in sequence between these two genes. A few years ago sequencing data provided on the contrary conclusive evidence that the THCAS and CBDAS scaffolds are at separate loci, though adjacent on the same chromosome [16]. This prompted a more recent study in which *CBDAS* and *THCAS* from 11 drug-type and 10 fiber-type hemp varieties were sequenced in parallel, allowing the identification of some SNPs that were hypothesized as signatures for decreased THCAS activity in fiber-type plants, and deletions possibly resulting in CBDAS loss-of-function in drug-type plants [22]. In fact, also a few other *CBDAS* previously isolated from drug-type varieties had been found to contain mutations causing frameshifts or premature stop codons [21]. Nevertheless, the presence of these possible molecular markers was investigated only in those 21 varieties, and the actual ability of the corresponding primers to discriminate between chemotypes was not assessed experimentally. The rationale of the present work was, therefore, to re-evaluate all the information previously made available under a correct perspective, where *CBDAS* and *THCAS* are two distinct genes present in both drug-type and fiber-type plants.

A wide set of 38 *CBDAS* and 145 *THCAS* sequences, for 36 and 67 of which the chemotype was known, was considered. A Neighbour-joining tree generated from the aligned sequences not only confirmed the divergence between the two genes, suggesting that accession number AB292683.1 should be re-annotated as *THCAS*, but clearly showed the presence within each cluster of two subgroups in which almost all sequences of a given chemotype co-clustered. In the case of *CBDAS*, for which a lower number of sequences were available, the two clades perfectly resolved drug-type from fiber-type varieties. For *THCAS*, with a double number of informative sequences available, the two clades were at a remarkably higher genetic distance, but four accessions clustered in the wrong subgroup. However, the two fiber-type accessions (KJ469381.1 and KJ469383.1) that clustered with the drug-type varieties were present in a lateral branch of the clade, near the node that divides the two chemotypes. The two drug-type accessions (KJ469379.1 and JQ437490.1) that clustered with all the other fiber-type varieties were on the contrary scattered within the clade. This inconsistency may depend on a wrong classification of their chemotype, or on the presence of drug-type alleles in fiber-type varieties. Indeed, recent results suggested that synthases for the cannabinoid pathway are highly duplicated, and that hemp plants probably express the paralogs of these genes differently in specific tissues. Gene copy number was also found to at least partially explain variation in cannabinoid content [28]. Whatever the reasons for these few exceptions, the whole picture strongly suggested that in most cases drug-type hemp varieties contain both *THCAS* and *CBDAS* forms that are dissimilar from those in fiber-type cultivars, and that these differences could be used to identify SNPs potentially able to discriminate the chemotypes.

Three primer pairs were actually identified for each subgroup whose complementary sequences are present in virtually all target varieties, while showing a significant number of SNPs in most sequences belonging to the other three, non-target subgroups. Some other primers, potentially discriminating the two chemotypes within a gene, were discarded because they would have amplified also the other gene. Several other putative signatures for a chemotype were found within the two genes, but at this stage of the research only the 12 primer pairs described in Table 1 were considered. When these primers were used to analyse DNA extracted from a number of mostly unsequenced hemp varieties, the patterns obtained were in general strikingly different from the expected ones, even in the case of the 4 varieties for which *CBDAS* and *THCAS* sequences had been considered in *in silico* analysis. Such inconsistent results may depend in part on the fact that no attempts were made to optimize PCR conditions. We aimed at assessing the potentiality of a very simple protocol in which DNA is extracted from a specimen as small as a single seed, and an end-point PCR is performed even without template DNA quantification. Also the amount of the amplification products was roughly estimated by image analysis following gel visualization. Of course an increase of stringency with the adoption of higher annealing temperatures, the normalization of the concentration of the template in the reaction mixture and/or the use of more sophisticated and truly quantitative PCR techniques could improve discrimination of drug-type from fiber-type varieties, and overcome some of the inconsistencies found.

This notwithstanding, the dataset herein described allows drawing some conclusions. A few genetic signatures in *CBDAS* and *THCAS* sequences are indeed present that may contribute to distinguish hemp chemotypes. Despite the basic protocol adopted, in the case of primer pairs A, C and K, consistent results (*i.e.* amplification in target genotypes and lack of amplification in non-target genotypes) were obtained in 27, 26 and 25 out of 27 varieties tested, respectively. On the other hand, data confirmed previous studies showing an extreme genetic variability in hemp germplasm concerning these genes. Coupled with the possible occurrence of multiple gene copy numbers [28] allowing the presence of drug alleles in fiber varieties and *vice-versa*, this implies that the identification of a single SNP able in all cases to discriminate the two chemotypes is unlikely. Therefore, a slightly different approach could be pursued, in which some other similarly informative SNPs would be further identified, and used together to genotype a large number of hemp varieties. The availability of this primer panel, and the building of a detailed database of the corresponding amplification patterns in as many hemp cultivars as possible, would be useful not only for chemotype DNA barcoding, but also for varietal identification, another essential application for breeding programs and seed patent protection. Work is currently in progress with this aim.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/....

References

[1] C. Schluttenhofer, L.Yuan, Challenges towards revitalizing hemp: A multifaceted crop. Trends Plant Sci. 22 (2017): 917–929.

[2] J. Fike, Industrial hemp: Renewed opportunities for an ancient crop, Crit. Rev. Plant Sci. 35 (2016): 406–424.

[3] M. Tyler, J. Shepherd, D. Olson, W. Snell, S. Proper, S. Thornsbury, Economic viability of industrial hemp in the United States: A review of state pilot programs, EIB-217 (2020), U.S. Department of Agriculture, Economic Research Service.

[4] R. Abuhasira, L. Shbiro, Y. Landschaft, Medical use of cannabis and cannabinoids containing products - Regulations in Europe and North America. Eur. J. Intern. Med. 49 (2018): 2–6.

[5] I. Adesina, A. Bhowmik, H. Sharma, A. Shahbazi, A review on the current state of knowledge of growing conditions, agronomic soil health practices and utilities of hemp in the United States. Agriculture 10 (2020): 129.

[6] D.W. Lachenmeier, S.G Walch, Analysis and toxicological evaluation of Cannabinoids in hemp food products-a review. Electr. J. Environ. Agric. Food Chem. 4 (2005): 812–826.

[7] M. Hädener, S, König, W.Weinmann, Quantitative determination of CBD and THC and their acid precursors in confiscated cannabis samples by HPLC-DAD. Forensic Sci. Int. 299 (2019): 142–150.

[8] T.J. Raharjo, I. Widjaja, S. Roytrakul, R. Verpoorte, Comparative proteomics of *Cannabis sativa* plant tissues. J. Biomol. Tech. 15 (2004): 97–106.

[9] M.D. Marks, L. Tian, J.P. Wenger, S,N, Omburo, W. Soto-Fuentes, J. He, D.R. Gang, G.D. Weiblen, R.A. Dixon Identification of candidate genes affecting δ^9-tetrahydrocannabinol biosynthesis in *Cannabis sativa*. J. Exp. Bot. 60 (2009): 3715–3726.

[10] E.E. Hakki, S.A. Kayis, E. Pinarkara, A. Sag, Inter simple sequence repeats separate efficiently hemp from marijuana (*Cannabis sativa* L.). Electron. J. Biotech. 10 (2007): 570–581.

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

[12] M. Kojoma, H. Seki, S. Yoshida, T. Muranaka, DNA polymorphisms in the tetrahydro-cannabinolic acid (THCA) synthase gene in "drug-type" and "fiber-type" *Cannabis sativa* L. Forensic Sci. Int. 159 (2006): 132–140.

[13] S. Sirikantaramas, S. Morimoto, Y. Shoyama, Y. Ishikawa, Y. Wada, Y. Shoyama, F. Taura. The gene controlling marijuana psychoactivity: molecular cloning and heterologous expression of δ^1-tetrahydrocannabinolic acid synthase from *Cannabis sativa* L. J. Biol. Chem. 279 (2004): 39767–39774.

[14] M.A. El Alaoui, M. Melloul, S.M. Udupa, E.E. Hakki, H. Stambouli, A. El Bouri, S.A. Amine, A. Soulaymani, E. El Fahime, Study of Moroccan *Cannabis sativa* DNA polymorphism in the THCA synthase gene from seized Moroccan cannabis resin (Hashish). J. Plant Biol. Res, 5 (2016): 1–11.

[15] D. Rotherham, S.A. Harbison, Differentiation of drug and non-drug Cannabis using a single nucleotide polymorphism (SNP) assay. Forensic Sci. Int. 207 (2011): 193–197.

[16] K.U. Laverty, J.M. Stout, M.J. Sullivan, H. Shah, N. Gill, L. Holbrook, G. Deikus, R. Sebra, T.R. Hughes, J.E. Page, H. van Bakel, A physical and genetic map of *Cannabis sativa* identifies extensive rearrangements at the THC/CBD acid synthase loci. Genome Res. 29 (2019): 146-156.

[17] F. Taura, S. Sirikantaramas, Y. Shoyama, K. Yoshikai, Y. Shoyama, S. Morimoto, Cannabidiolic-acid synthase, the chemotype-determining enzyme in the fiber-type *Cannabis sativa*. FEBS Lett. 581 (2007): 2929–2934.

[18] E.P. de Meijer, M. Bagatta, A. Carboni, P. Crucitti, V.M. Moliterni, P. Ranalli, G. Mandolino, The inheritance of chemical phenotype in *Cannabis sativa* L. Genetics 163 (2003): 335–346.

[19] C. Staginnus, S. Zörntlein, E.de Meijer, A PCR marker linked to a THCA synthase polymorphism is a reliable tool to discriminate potentially THC-rich plants of *Cannabis sativa* L. J. Forensic Sci. 59 (2014): 919–926.

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

[21] G.D. Weiblen, J.P. Wenger, K.J. Craft, M.A. El Sohly, Z. Mehmedic, E.L. Treiber, M.D. Marks, Gene duplication and divergence affecting drug content in *Cannabis sativa*. New Phytol. 208 (2015): 1241–1250.

[22] F. Cascini, A. Farcomeni, D. Migliorini, L. Baldassarri, I. Boschi, S. Martello, S. Amaducci, L. Lucini, J. Bernardi, Highly predictive genetic markers distinguish drug-type from fiber-type *Cannabis sativa* L. Plants (Basel) 8 (2019): 496.

[23] K.J. Mc Kernan, Y. Helbert, V. Tadigotla, S. Mc Laughlin, J. Spangler, L. Zhang, D. Smith, Single molecule sequencing of THCA synthase reveals copy number variation in modern drug-type *Cannabis sativa* L. bioRxiv (2015): doi: https://doi.org/10.1101/028654.

[24] F. Madeira, Y.M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A.R.N. Tivey, S.C. Potter, R.D. Finn, R. Lopez, The EMBL–EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47 (2019): W636–W641

[25] A. Boc, Diallo, B.Alpha, V. Makarenkov, T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucl. Acids Res. 40 (2012): W573–W579.

[26] Y.T. Huang, J.I. Yang, M. Chrobak, J. Borneman, PRISE2: software for designing sequence-selective PCR primers and probes. BMC Bioinformatics 15 (2014): 317.

[27] C.A.Schneider, W.S. Rasband, K.W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis, Nature Methods 9 (2012): 671–675.

[28] D. Vergara, E.L. Huscher, K.G. Keepers, R.M. Givens, C.G. Cizek, A. Torres, R. Gaudino, N.C. Kane, Gene copy number is associated with phytochemistry in *Cannabis sativa*. AoB Plants 11 (2019): plz074.

Table 1

Putatively specific primers to amplify hemp *THCAS* and *CBDAS* sequences. The expected formation of amplicons is indicated.

Pair	Primer, forward	Primer, reverse	size	*CBDAS*		*THCAS*	
				fiber cluster	drug cluster	fiber cluster	drug cluster
A	GAATCTGTATTTGTCCAAA	AAGGAGTCATGAAGTTAT	235	☑			
B	CACTATTCTATGCTCCAAGAAA	GTAGACTTTGGGACAGCA	480	☑			
C	AGAATCTGTATTTGTCCAAA	TTCCTATATCAAGGTCTCTA	294	☑			
D	CACTATTCTATGTCCAAGAAAA	AAGTGTGCATCAACGATATT	341		☑		
E	CCAATGTAACAAATCTAAA	TTGAATGCATGTTTCTCA	277		☑		
F	CAAGGCACTATTCTATGTC	AGACTTTGTTGGGACAGC	485		☑		
G	GCTATAGTAGACTTGAGAAA	TGAGTCGTGAGCATTAAA	494			☑	
H	TCAAAGTAGATATTCATAGCCAAA	AAGTGAGTCGTGAGCATTAAA	466			☑	
I	TAAGAAACTAATACCTGAAA	ACATAAGGAGTTGTGAAA	255			☑	
J	TACATGGTTACTTCTCTTCAA	AATTGGTTTCTTAACATAGTCTAA	251				☑
K	ACTCACTTCATAACAAAGAA	ACTTAATTGAGAAAGCCGT	283				☑
L	TATTATTGATGCACACTTAGT	AAAATTTACAACACCACTG	472				☑
T	TTGGAGAAGTTTATTATTGG	ACTAGACTATCCACTCCACCA	510	☑	☑	☑	☑

Table 2

Amplification patterns obtained with the putatively specific primers using DNA template from various fiber-type and drug-type hemp varieties.

#	Cultivar	Primer pair	A	B	C	D	E	F	G	H	I	J	K	L	T
fiber-type															
01	Carmagnola selezionata		+++	++++	++	++++	–	++++	++++	++++	+	+	++	++++	+++
02	Fedora 17		+	++++	+	–	–	–	–	++++	–	–	–	–	–
03	Felina 32		+++	++++	++	–	–	–	++	++++	+	–	–	+++	+++
04	Fibrol		++++	++++	+++	+++	++	+++	+++	++++	++	+	–	+++	+++
05	Futura 75		+++	++++	++	–	–	–	++	++++	–	–	–	++	+++
06	Jubileu Secuieni		+	+++	–	–	–	–	–	++++	–	–	–	+	++
07	KC Dora		++++	++++	++	–	–	–	++	+++	+	–	–	+++	+++
08	KC Zuzana C1		++++	++++	+++	+++	–	++	++	+++	++	++	–	++++	+++
09	KC Zuzana C2		++++	++++	+++	+	–	–	++++	++++	+	+	–	++++	+++
10	Kompolti		++++	++++	+++	++	–	++	+	++	–	–	–	+++	++
11	Monoica		++++	++++	++	++++	+	++++	+++	++++	+	++	+++	+++	++
12	Silvana		++++	++++	+++	+++	–	+	++	++++	++	+	–	+++	+++
13	Tiborszallasi		++++	++++	+++	++	–	+	+++	++++	+	+	–	++++	+++
14	Tisza		++++	++++	++++	+++	–	+	+++	+++	++	+	–	++++	+++
15	Uso 31		++++	++++	+++	+++	+	++	+++	+++	++	++	–	+++	+++
16	Zenit		++++	+++	+++	++	–	+	+++	++++	+++	+++	–	+++	+++
drug-type															
17	Afghani #1		–	++++	–	++++	+	++++	–	–	–	+++	+++	+++	+++
18	Black Domina		–	++++	–	++++	++	++++	+++	++++	++	++	+++	+++	+++
19	Durban		–	++++	–	++++	–	++++	–	++++	–	+	+	++	++
20	Jack Flash #5		–	++++	–	++++	++	++++	+++	++++	++	+++	++++	++++	+++
21	Jack Herer		–	++++	–	++++	+	++++	–	–	–	++	+++	+++	+++
22	Jamaican Pearl		–	++++	–	++++	++	++++	+++	+++	+	+++	+++	+++	+++
23	Northern Lights		–	++++	–	++++	+	++++	+++	++++	+	+	+++	+++	+++
24	Northern Lights #5 x Haze		–	++++	–	++++	++	++++	+++	+++	++	+++	+++	+++	++++
25	Sensi Skunk		–	++++	–	++++	+	++++	–	–	–	+	++	+++	+++
26	Shiva Skunk		–	++++	–	++++	+	++++	–	–	–	++	+++	+++	+++
27	Silver Haze		–	++++	–	++++	+	++++	+++	++++	++	+++	++++	+++	++

Band intensity was quantified with ImageJ. ++++. +++. ++. + and –: 81-100, 61-80, 41-60, 21-40 and ≤ 20% intensity of the more abundant amplification product obtained with the 13 primer pairs, respectively.

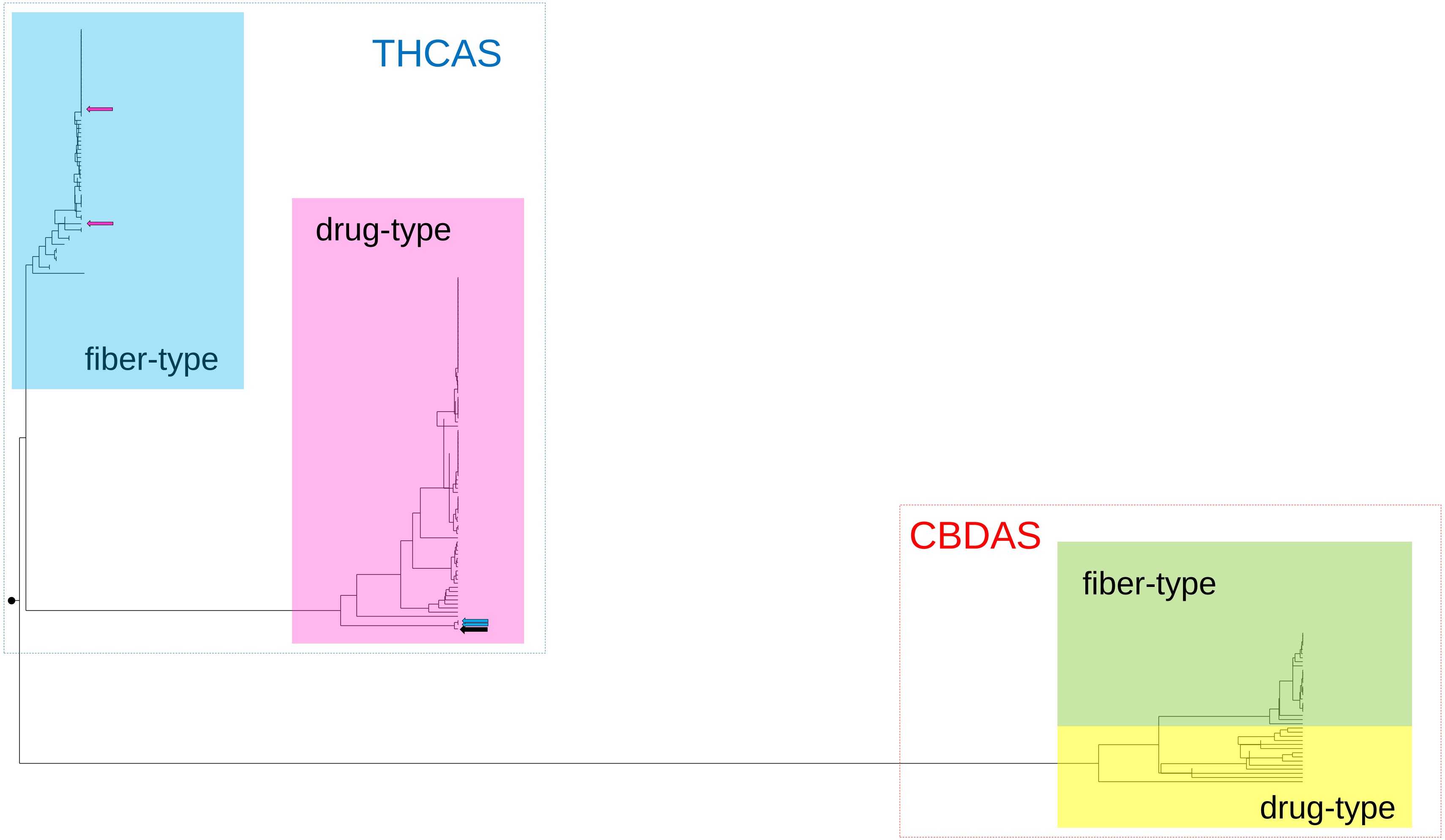

Fig. 1. Neighbour-joining tree generated from the aligned sequences of all hemp *THCAS* and *CBDAS* available from public databases. The black arrow points at the only sequence (accession number AB292683.1) clustering with *THCAS* but annotated as "*CBDAS* homologue". The clade containing CBDAS genes from fiber-type varieties are emphasized in green shading, whereas that clustering drug-type accessions is shaded in yellow. In the case of *THCAS*, the presence of most fiber-type varieties is emphasized in cyan, whereas that of drug-type accessions is shaded in hot pink. Cyan and hot pink arrows show the fiber-type and the drug-type accessions that cluster in the opposite clade, respectively. A more analytical picture with all accession numbers and the position of the sequences cloned from plants of known chemotype is provided as Supplementary Fig. S2.

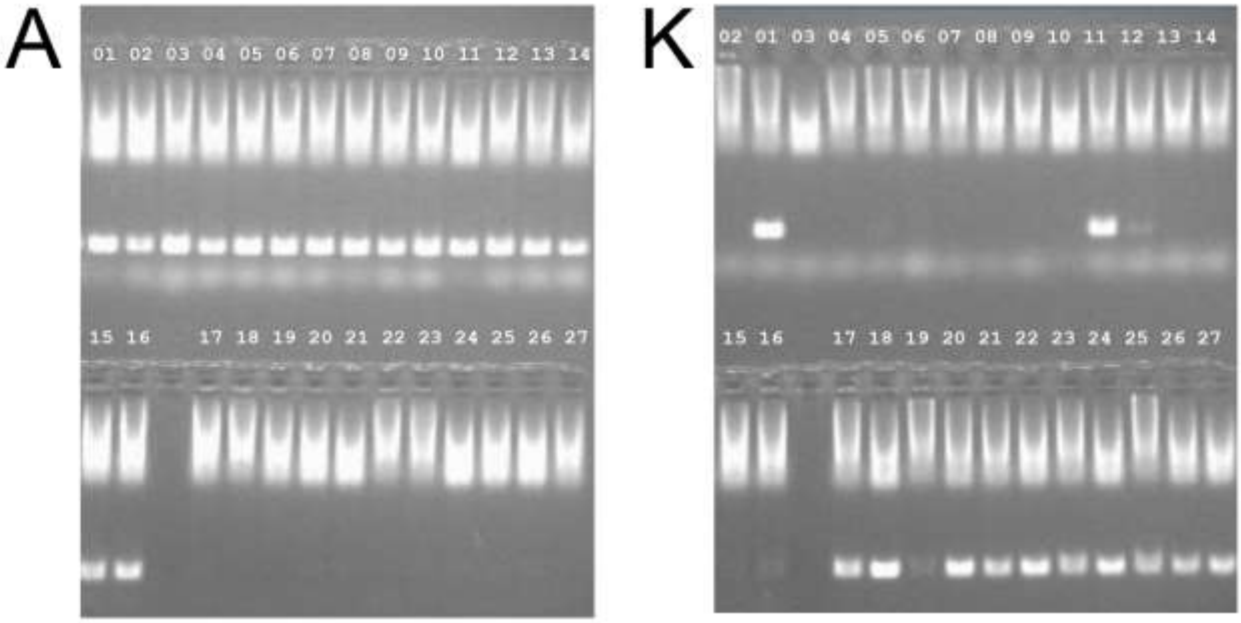

Fig. 2. The patterns of DNA amplicons obtained with primer pairs A and K using template DNA from fiber-type (01-16, as listed in Table 2) or drug-type (17-27) hemp varieties.