Phytochemistry xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Phytochemistry

journal homepage: www.elsevier.com/locate/phytochem

Sequence heterogeneity of cannabidiolic- and tetrahydrocannabinolic acid-synthase in *Cannabis sativa* L. and its relationship with chemical phenotype

Chiara Onofri [a], Etienne P.M. de Meijer [b], Giuseppe Mandolino [a,*]

[a] *Consiglio per la ricerca in agricoltura e l'analisi dell'economia agraria, Centro di Ricerca per le Colture Industriali, via di Corticella 133, 40128 Bologna, Italy*
[b] *GW Pharmaceuticals PLC, Ground Floor South Wing, Kingsgate House, Newbury Road, Andover SP10 4DU, United Kingdom*

ARTICLE INFO

Article history:
Received 2 October 2014
Received in revised form 21 March 2015
Available online xxxx

Keywords:
Cannabis sativa L.
Cannabaceae
Chemotypes
SNPs
Cannabinoids
THCA-synthase
CBDA-synthase
Gene expression

ABSTRACT

Sequence variants of THCA- and CBDA-synthases were isolated from different *Cannabis sativa* L. strains expressing various wild-type and mutant chemical phenotypes (chemotypes). Expressed and complete sequences were obtained from mature inflorescences. Each strain was shown to have a different specificity and/or ability to convert the precursor CBGA into CBDA and/or THCA type products. The comparison of the expressed sequences led to the identification of different mutations, all of them due to SNPs. These SNPs were found to relate to the cannabinoid composition of the inflorescence at maturity and are therefore proposed to have a functional significance. The amount of variation was found to be higher within the CBDAS sequence family than in the THCAS family, suggesting a more recent evolution of THCA-forming enzymes from the CBDAS group. We therefore consider CBDAS as the ancestral type of these synthases.

1. Introduction

1.1. Biochemistry of THCAS and CBDAS

Cannabinoids are terpenophenolic secondary metabolites, produced by all *Cannabis sativa* L. plants in the sessile and stalked trichomes (Happyana et al., 2013). Trichomes are particularly abundant on the inflorescences of the plant, present in lower number on leaves, petioles and stems, and absent on roots and seeds. As a consequence, these latter organs do not contain cannabinoids.

The steps involved in the biosynthesis of the different cannabinoids from the common precursor have been largely elucidated by Taura et al. (1995), Fellermeier and Zenk (1998), Fellermeier et al. (2001). According to this widely accepted pathway, cannabigerolic acid (CBGA) is the first cannabinoid, formed through the condensation of a phenolic moiety (e.g. olivetolic or divarinic acid, Gagne et al., 2012) with the terpenoid component geranyl pyrophosphate. CBGA and its alkyl homolog are considered as the common precursors of all the main cannabinoids produced through an enzyme activity by the plant: i.e. the alkyl homologs of delta 9-tetrahydrocannabinolic acid (THCA), cannabidiolic acid (CBDA) and cannabichromenic acid (CBCA). The most common cannabinoids have a pentyl side chain, but propyl homologs can also occur in vivo (de Zeeuw et al., 1972). Methyl-cannabinoids are known too, though are only present occasionally and in very small amounts. All the CBGA alkyl-homologs can be used as substrate and transformed by plant extracts containing the different cannabinoid synthases in vitro, although the efficiency of conversion was reported to be different for each homolog (Shoyama et al., 1984).

The different synthases catalyzing the oxidocyclization of CBGA into THCA, CBDA or CBCA (and their alkyl homologs), have been characterized in recent years. THCA- and CBDA-synthases share many similarities in their biochemical properties (Taura et al., 1995, 1996), such as the mass (they are monomeric proteins, both 74 kDa as native proteins), p*I*, $v_{\max}$ and K_m for CBGA. They are both soluble enzymes, and once their amino-acid sequence was determined (GenBank accession numbers E55107 and E33090), it was found that they are 84% identical. Both have a 28-amino-acids putative signal peptide that is dissociated in the mature, secreted

Abbreviations: THCA, tetrahydrocannabinolic acid; THCAS, tetrahydrocannabinolic acid synthase; CBDA, cannabidiolic acid; CBDAS, cannabidiolic acid synthase; CBGA, cannabigerolic acid; CBCA, cannabichromenic acid; CBCAS, cannabichromenic acid synthase.

[*] Corresponding author. Tel.: +39 051 6316825; fax: +39 051 374857.
E-mail address: giuseppe.mandolino@entecra.it (G. Mandolino).

http://dx.doi.org/10.1016/j.phytochem.2015.03.006
0031-9422/© 2015 Elsevier Ltd. All rights reserved.

protein (Sirikantaramas et al., 2004, 2005), and a FAD-binding domain (Taura et al., 2007). These findings were consistent with the secretory pathway for these enzymes which are thought to be released into the glandular trichome's cavity, the site of cannabinoid biosynthesis (Kim and Mahlberg, 1997; Mahlberg and Kim, 2004). The tertiary structure of THCA-synthase was recently resolved and amino-acid positions putatively involved in FAD and substrate binding were identified by X-ray crystallography to a 2.75 Å resolution and also by mutational analysis (Shoyama et al., 2012).

Another feature common to THCA- and CBDA-synthases is the presence of a domain showing high homology with the berberine-bridge enzyme involved in the alkaloid biosynthesis of *Eschscholtzia californica*. Both *Cannabis* synthases and the BBE require molecular oxygen for their activity, and form hydrogen peroxide (Sirikantaramas et al., 2004) during the cyclization of the substrate. The elucidation of THCA-synthase tertiary structure also provided hints as to the amino-acid residues involved in catalytic activity and in the recognition of the carboxyl group of CBGA (Shoyama et al., 2012). The similarity of THCA- and CBDA-synthase with BBE is confirmed by comparison of their sequences against the entire GenBank database, where only other THCA synthases, CBDA synthases and BBE are retrieved as the most similar sequences.

1.2. Genetics and genomics of THCAS and CBDAS

It has been previously proposed that the genes coding for the functional THCA- and CBDA-synthase (indicated as B_T and B_D) were allelic and codominant, and that the CBDA/THCA ratio of the heterozygous plants was invariably close to unity, due to the inherent kinetic properties of the two synthases simultaneously present in the B_D/B_T genotypes. This has been demonstrated by genetic analysis in THCA vs. CBDA scatter plots of heterozygous plants (de Meijer et al., 2003). It has also been previously shown that, when examining the chemotype of F_1 hybrids, the slope of the linear regression line through the individual CBDA/THCA ratios can show meaningful variations and that these variations are fully inherited in the inbred F_2 progeny of the heterozygous F_1 plants. It was suggested that these variations could be directly related to a differential efficiency in transforming CBGA between the two co-existing synthases (de Meijer et al., 2003). This difference could be due to either (a) sequence variations at the B_T and/or B_D loci (i.e. to the existence of an allelic series, or of different sequences for one or both the synthases), or (b) expression polymorphism of regulatory elements in *cis* or *trans* to the structural loci.

Following these works, the genetics behind the main known *Cannabis* chemotypes, has been largely elucidated (de Meijer et al., 2003; Mandolino et al., 2003; de Meijer and Hammond, 2005; de Meijer et al., 2009a,b). In the current model, THCA- and CBDA-synthases are coded for by two alleles, B_T and B_D, but it is hypothesized that allelic and/or non-allelic variations may exist and explain the chemotype variation present in *Cannabis* germplasm. It has been proposed that CBGA-accumulating plants, for example, carry defective B_D and/or B_T alleles coding for inactive or minimally active synthases. According to this view, any *Cannabis* plants that accumulate a reduced THCA or CBDA proportion and consistently above-background levels of CBGA, are presumed to be endowed with partially-functional THCA- or CBDA-synthases. The proposed model for these mutants implies the simultaneous presence of the end product(s) of the pathway (THCA or CBDA and their respective alkyl homologs) and the partial accumulation of the precursor CBGA and its alkyl homologs due to inefficient synthases (de Meijer and Hammond, 2005; Mandolino et al., 2003).

Recently, two drafts of the genome sequence of *C. sativa* L. have been published (www.medicinalgenomics.com; van Bakel et al., 2011), implemented by extensive transcriptome sequencing in different organs and strains. The publicly available database (http://genome.ccbr.utoronto.ca/) increased the number of known THCA- and CBDA-synthase gene sequences. It also became clear from this and other works (Kojoma et al., 2006) that there are many THCA- and CBDA-synthase-related pseudogenes in the *Cannabis* genome with several degrees of variation compared with the functional, chemotype-determining ones. Sequence variation was also observed within the putatively functional genes of both enzymes although, under the conditions in which the transcriptome was sequenced, the chemotypes expressed by the plants (the drug strain Purple Kush and the oil seed variety Finola), were not fully specified.

The availability of genomic sequences related to the enzymes involved in the determination of the chemotypes allowed the development of a number of sequence-based markers (Kojoma et al., 2006; Pacifico et al., 2006; Mandolino, 2007; Staginnus et al., 2014) able to discriminate *Cannabis* plants as producing THCA, CBDA or both. Such markers prove to be powerful tools for forensic purposes and the rejection of THCA-containing plants in fiber hemp breeding. Unlike chemical analyses, these markers can demonstrate the presence of THCAS and CBDAS sequences in tissues or plants where they are not expressed, e.g. in seeds and roots or in cannabinoid-free plants.

For this paper, the variability of the expressed sequences of THCA- and CBDA-synthases was studied, and mutations putatively relevant for both THCA- and CBDA-synthase functions were identified. Sequence-based markers were used to genotype eighteen clones and inbred lines with different geographical and domestication backgrounds at the *B* locus (de Meijer et al., 2003). The different synthases putatively involved in the determination of chemotype were sequenced. The variations found at the level of specific amino-acid substitutions were considered in relation to the cannabinoid profile to thereby deduce the evolutionary relationship of these two synthase families.

2. Results

2.1. Chemotypes and locus B genotypes

Chemotypes of the accessions examined are presented in Table 3. The total cannabinoid content in the dry inflorescences ranged from 0.05% to 15.6% w/w. Individual cannabinoids are presented as neutral, decarboxylated molecules since they transform under the high temperature conditions of GC analysis (e.g. THCA → THC). Substrate for both CBDAS and THCAS is represented by the decarboxylated alkyl homologs of CBGA, i.e. CBGV plus CBG. The proportion of residual substrate was considered a measure for enzyme functionality and ranged from 0.18% to 99.99% of the total cannabinoid fraction. Analogously, the THCAS product is represented by THCV + THC and the CBDAS product by CBDV + CBD. The results of the amplification of DNA with the multiplex, three-primer marker system described by Pacifico et al. (2006) are summarized in the last column of Table 3. All the plants identified as predominantly THC(V) or CBD(V) were correctly genotyped as homozygous B_T or homozygous B_D, thereby confirming the complete association of this marker with the *B* locus for the two most common chemotypes. In three cases, (lines 2009.27.44.2.19.1, 2005.42.11.59 and 2005.16.(2 + 12 + 24).24), CBG(V) was the main cannabinoid. The *B1180/1192* marker identified the first accession as B_T/B_T and the latter two as B_D/B_D. In the B_T/B_T group, the proportion of the end-product (THC + THCV) ranged from 11.05% to 98.58%. In the B_D/B_D group (CBD + CBDV) occupied 0.01–90.23% of the total cannabinoid fraction.

Besides the CBG(V) decarboxylated precursor and the THCAS and CBDAS decarboxylated conversion products, Table 3 also lists

Table 1
The accessions analyzed. Provenances refer to the original material from which clones or inbred lines were selected.

Accession	Provenance	Description
55.22.7.2.2.	Thailand	S3 inbred line from marijuana landrace
55.24.4.34.7.	South India	S3 inbred line from marijuana landrace
94.5.2.30.1.2.4	Turkey	S4 inbred line from fiber landrace
2009.23.25.15.14.4	Germany	S2 inbred line from fiber landrace
2009.27.44.2.19	Malawi	S3 and S4 inbred lines from marijuana landrace
2005.42.11.59	Mixed origin	Complex hybrid clone. Mutant chemotype derived from Italian fiber strain
2005.16.(2 + 12 + 24).24	Mixed origin	Complex hybrid clone. Mutant chemotype derived from Ukrainian fiber strain
55.28.1.4.12.1	U.S.A.	S3 inbred line from marijuana strain
2010.24a.35/T	Morocco, Zoumi region	S1 inbred line from hashish landrace
2010.24a.35/C	Morocco, Zoumi region	S1 inbred line from hashish landrace
M110	U.S.A.	Clone selected from marijuana strain Haze
M124	Afghanistan	Clone selected from hashish landrace
M128	Afghanistan	Clone selected from hashish landrace
55.14.4.27.2.6.4.1	Afghanistan	S5 inbred line from hashish landrace
98.2.21.19.8.7.1	China, Yunnan region	S4 inbred line from complex hybrid
2010.29	China	Seedling from fiber/seed landrace
K310	Northern Russia	S4 inbred line from fiber/seed landrace
Ermo	Italy/Russia	Fiber cv.

proportions of cannabichromene (CBC) and its propyl homolog CBCV, cannabigerol monomethylether (CBGM) and cannabinol (CBN). The generally minor constituents CBC(V) and CBGM are other decarboxylated conversion products of CBG(V)A and their production is genetically independent from THC(V)A and CBD(V)A production (de Meijer et al., 2009a; Shoyama et al., 1970). CBN results from the non-enzymatic oxidation of THC. CBN proportions could therefore be added to THC proportions but as in our samples they were practically negligible we chose to ignore them and list them separately.

The partial lack of association between marker and chemotype can be explained by the assumption that plants of the less common, CBG(V) predominant chemotypes were endowed with partially, or totally, defective variant THCA- or CBDA-synthase sequences similar enough to the normal ones to be amplified by the allele-specific primers *B1180/1192* (de Meijer and Hammond, 2005). If this hypothesis were true, it should be possible to recover a number of variant expressed sequences and to establish correlations between specific mutations and the enzyme functionality.

2.2. Sequence heterogeneity within THCA-synthases

All the cannabinoid synthase sequences were obtained from mRNA extracted from the mature inflorescences of the plants listed in Table 1; therefore they must be considered as fully expressed at maturity in the floral organs and likely to represent transcripts which significantly contribute to the chemical phenotype.

cDNA obtained from mRNA was amplified with both THCAS- and CBDAS-specific primers covering the entire length of the coding sequence. Amplicons for THCAS were only obtained when the THCAS-specific primers were used on plants genotyped by the marker *B1180/1192* as B_T/B_T, as were CBDAS amplicons exclusively obtained from B_D/B_D-genotyped plants. Besides, only complete gene sequences constituting open reading frames able to be translated into putatively functional enzymes were considered in this work; all THCAS and CBDAS sequences discussed here are full-length (1635 bp) and code for 545 and 544 amino-acid proteins, respectively.

The SNPs found in the different THCAS sequences are listed in Fig. 1 with the THCA synthase represented by GenBank accession E33090 used as the reference.

Nine different transcribed THCAS sequences were identified. Four accessions transcribed a single sequence, while in the other four accessions, more than one single sequence was found in cDNA (Fig. 1).

By Clustal W2 cluster analysis (UPGMA clustering method, Fig. 3a), it was possible to group these sequences into 4 families, one of which, group 1, comprised 4 members (1/1–1/4) only differing by a SNP, another (group 2) comprising three members differing from the reference sequence for more than one SNP, and the last two groups (3 and 4) each comprising a single member. Overall, the nine sequences differed for SNPs in 12 positions of the gene (as compared with the E33090 GenBank sequence); however, only 6 of these SNPs determined an amino-acid change in the protein (bottom row in Fig. 1); three of these variations were concentrated between positions 236 and 265 of the protein. These sequence variations in the analyzed accessions resulted in seven variant THCAS proteins, one of which was identical to that of E33090.

Within B_T/B_T germplasm, an important variation in the THCA-synthase gene sequence was observed in the *Cannabis* strain 2009.27.44.2.19, which produced only a minor proportion of THCA, and accumulated CBG(V)A as the main cannabinoid (Fig. 1 and Table 3). In sequence 1/3 (Fig. 1), a single SNP, expressed exclusively in this accession, was detected at position 706 (G → C) but this single mutation implied a change in the first nucleotide of the triplet corresponding to amino-acid position 236 (Glu → Gln). This specific mutation has not been found in

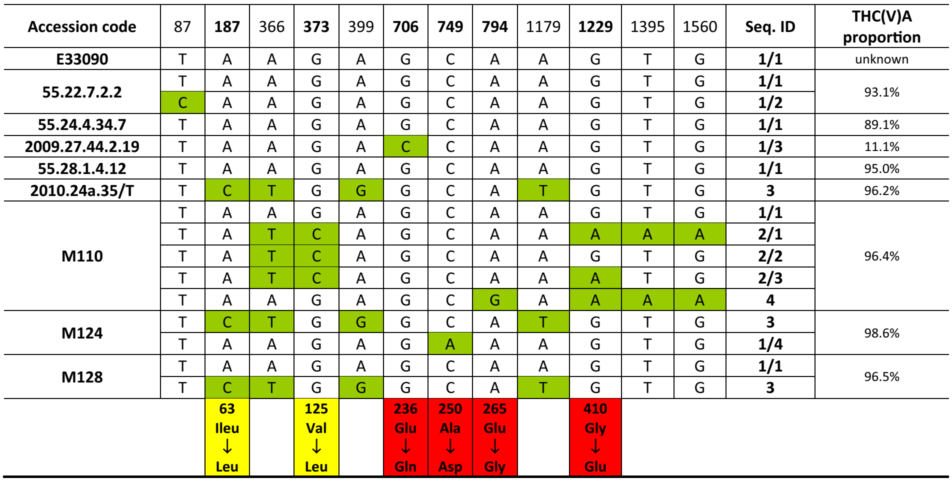
Table 2
Primers used for genotyping, sequencing and qPCR analysis of the THCAS and CBDAS genes.

Genotyping	Primer FW	Primer REV
THCA-synthase	AAGAAAGTTGGCTTGCAG	TTAGGACTCGCATGATTAGTTTTTC
CBDA-synthase	AAGAAAGTTGGCTTGCAG	ATCCAGTTTAGATGCTTTTCGT
Sequencing	Primer FW	Primer REV
THCA-synthase	ATGAATTGCTCAGCATTTTCCTT	ATGATGATGCGGTGGAAGA
CBDA-synthase	ATGAAGTGCTCAACATTCTCC	TTAATGACGATGCCGTGGTT
Real time PCR	Primer FW	Primer REV
THCA-synthase	CAGCAATTCCATTCCCTCAT	TTAGGACTCGCATGATTAGTTTTTC
CBDA-synthase	CAGCAATTCCATTCCCTCAT	ATCCAGTTTAGATGCTTTTCGT
Tubulin	GGCGCTGAGTTGATCGATTC	GTATGGTTCCACAACTGTGTC

Table 3
Chemotypes and marker phenotypes of the sequenced plant materials. Accessions are grouped by their marker phenotype and within groups sorted by an increasing proportion of the central precursors CBGV and CBG.

Line or clone	Total cannabinoid content[a]	Cannabinoid composition (% w/w of total cannabinoid fraction)										Marker phenotype
		CBDV	CBD	CBCV	CBC	THCV	THC	CBGV	CBG	CBGM	CBN	
M124	6.57		0.19		0.47	0.43	98.15		0.18	0.39	0.19	B_T/B_T
2010.24a.35/T	6.42		0.26		3.10	0.48	95.73		0.43			B_T/B_T
M128	4.62		0.14		2.41	0.79	95.72		0.66		0.27	B_T/B_T
M110	6.70				2.67	1.89	94.53		0.72		0.19	B_T/B_T
55.28.1.4.12	9.65		0.32		1.95	0.37	94.60		2.76			B_T/B_T
55.22.7.2.2	13.04				1.82		93.05		3.74	0.87	0.52	B_T/B_T
55.24.4.34.7	5.39				0.19	0.10	88.95		10.24		0.52	B_T/B_T
2009.27.44.2.19	7.56			1.77	2.97	6.70	4.35	20.15	64.06			B_T/B_T
2010.29	4.08	2.16	88.07		3.11		4.45		1.25	0.96		B_D/B_D
2010.24a.35/C	4.39		87.89		5.34		4.00		1.89	0.87		B_D/B_D
K310	4.27	6.88	83.46	0.18	2.89		3.40		2.16	1.04		B_D/B_D
55.14.4.27.2.6.4	3.85	5.59	83.13	0.18	3.06	0.18	4.16		2.95	0.75		B_D/B_D
2009.23.25.15.14	6.57		83.72		5.77		3.46		3.55	3.51		B_D/B_D
98.2.21.19.8.7	3.43		77.33		4.96		4.88		12.26	0.57		B_D/B_D
94.5.2.30.1.2	5.86		66.78		8.20		4.81		20.22			B_D/B_D
Ermo	0.05		35.21		7.69		5.03		49.03	3.04		B_D/B_D
2005.42.11.59	15.63		6.27		0.55			0.28	92.89			B_D/B_D
2005.16.(2 + 12 + 24).24	6.04		0.01						99.99			B_D/B_D

[a] The total cannabinoid content in the dry mature inflorescence.



Accession code	87	**187**	366	**373**	399	**706**	**749**	**794**	1179	**1229**	1395	1560	Seq. ID	THC(V)A proportion
E33090	T	A	A	G	A	G	C	A	A	G	T	G	**1/1**	unknown
55.22.7.2.2	T	A	A	G	A	G	C	A	A	G	T	G	**1/1**	93.1%
	C	A	A	G	A	G	C	A	A	G	T	G	**1/2**	
55.24.4.34.7	T	A	A	G	A	G	C	A	A	G	T	G	**1/1**	89.1%
2009.27.44.2.19	T	A	A	G	A	C	C	A	A	G	T	G	**1/3**	11.1%
55.28.1.4.12	T	A	A	G	A	G	C	A	A	G	T	G	**1/1**	95.0%
2010.24a.35/T	T	C	T	G	G	G	C	A	T	G	T	G	**3**	96.2%
M110	T	A	A	G	A	G	C	A	A	G	T	G	**1/1**	96.4%
	T	A	T	C	A	G	C	A	A	A	A	A	**2/1**	
	T	A	T	C	A	G	C	A	A	G	T	G	**2/2**	
	T	A	T	C	A	G	C	A	A	A	T	G	**2/3**	
	T	A	A	G	A	G	C	G	A	A	A	A	**4**	
M124	T	C	T	G	G	G	C	A	T	G	T	G	**3**	98.6%
	T	A	A	G	A	G	A	A	A	G	T	G	**1/4**	
M128	T	A	A	G	A	G	C	A	A	G	T	G	**1/1**	96.5%
	T	C	T	G	G	G	C	A	T	G	T	G	**3**	
		63 Ileu ↓ Leu		125 Val ↓ Leu		236 Glu ↓ Gln	250 Ala ↓ Asp	265 Glu ↓ Gly		410 Gly ↓ Glu				

Fig. 1. SNPs in THCAS gene sequences expressed in the *Cannabis* germplasm. All SNPs are defined in comparison with the GenBank accession Nr. E33090. SNPs involving an amino-acid change in the translated protein are indicated in bold, and the type of change is indicated in the bottom row with the corresponding position in the protein sequence (in yellow, amino-acids of the same chemical type, in red amino-acids chemically different). SNP positions are identified by green color and by different sequence ID numbers; in the last column the proportion of the end-product THC(V)A in the total cannabinoid fraction accumulated by each accession is given (see also Table 3).

any of the other sequenced THCA synthases, or in any THCA synthase sequence already present in the genomic databases. It is likely to be responsible for a defective THCAS which is largely unable to convert CBG(V)A into THC(V)A, thus causing CBG(V)A to accumulate. It is proposed here that this mutation represents a new B_T allele, called B_{T0}.

The functionality of the other synthases can be deduced only in those genotypes in which a single THCAS sequence was found. On this basis, sequences 1/1 (85.6% THCA accumulated in genotype 55.24.4.34.7, and over 94% in genotype 55.28.1.4.12) and 3 (above 95% in genotype 2010.24a.35/T) can be considered to code for functional THCAS. The same holds true for sequence 1/2, where the SNP found was not translated in a variant amino-acid compared with E33090 and THCAS 1/1.

The other genotypes transcribed more sequences, and this simultaneous expression in high-THCA producing genotypes, prevents judgment of the individual contributions of the variant synthases 1/4, 2/1, 2/2, 2/3 and 4 in converting the CBGA substrate.

No expressed THCA sequences were recovered from cDNA obtained from B_D/B_D genotypes, suggesting either their absence, or their presence in exceedingly low amounts.

2.3. Sequence heterogeneity within CBDA-synthases

In the accessions identified by the marker *B1180/1192* as B_D/B_D genotypes (putatively CBDA-prevalent), a total of 12 complete transcribed sequences, different from GenBank E55107 were identified. These 12 sequences can be grouped by cluster analysis in two main families, 5 and 6, consisting of 4 and 8 members respectively (Fig. 3b).

Only one sequence, CBDAS 6/2, differed from the reference by a single SNP (Fig. 2); this SNP, a T in position 105 instead of an A, was found to be present in all sequences obtained. All other sequences diverged from E55107 for up to 11 SNPs. A total of 18 SNP positions were found in the CBDAS expressed sequences, only 9 of which implied an amino-acid change in the protein.

Accession code	105	**221**	**428**	**503**	516	**587**	1032	1035	1095	**1123**	1278	1341	**1420**	**1426**	**1465**	1584	**1616**	1617	Seq.ID	CBD(V)A proportion
E55107	G	C	A	A	G	A	G	C	C	G	A	T	A	C	G	C	T	T		unknown
94.5.2.30.1.2.4	T	G	A	G	T	G	G	T	C	C	T	C	C	C	G	T	T	T	**5/1**	66.8%
2009.23.25.15.14.4	T	G	A	G	T	G	G	C	C	G	T	C	C	C	G	T	T	T	**5/2**	83.7%
2005.42.11.59	T	C	A	A	G	A	G	C	C	G	A	T	A	T	G	C	T	T	**6/1**	6.3%
2005.16.(2+..).24	T	G	A	G	T	G	G	C	C	G	T	C	C	C	A	T	T	T	**5/3**	0.01%
55.14.4.27.2.6.4	T	G	A	G	T	G	G	C	C	G	T	C	C	C	G	T	T	T	**5/2**	88.7%
	T	G	A	G	T	G	G	C	C	G	T	C	C	C	G	T	T	A	**5/4**	
	T	C	A	A	G	A	G	C	C	G	A	T	A	C	G	C	T	T	**6/2**	
98.2.21.19.8.7	T	C	A	A	G	A	A	C	T	G	T	T	A	C	G	C	T	T	**6/5**	77.3%
	T	C	A	A	G	A	A	C	T	G	T	T	A	C	G	C	T	A	**6/6**	
	T	C	A	A	G	A	A	C	T	G	T	T	A	C	G	C	A	A	**6/7**	
2010.24a.35/C	T	C	A	A	G	A	G	C	C	G	A	T	A	C	G	C	T	T	**6/2**	87.9%
	T	C	A	A	G	A	G	C	C	G	A	T	A	C	G	C	T	A	**6/3**	
2010.29	T	C	A	A	G	A	G	C	C	G	A	T	A	C	G	C	T	T	**6/2**	90.2%
	T	C	G	A	G	A	G	C	C	G	T	T	A	C	G	C	T	T	**6/8**	
K310	T	G	A	G	T	G	G	C	C	G	T	C	C	C	G	T	T	T	**5/2**	90.3%
	T	G	A	G	T	G	G	C	C	G	T	C	C	C	G	T	T	A	**5/4**	
	T	G	A	G	T	G	G	C	C	G	T	C	C	C	A	T	T	T	**5/3**	
Ermo	T	C	A	A	G	A	G	C	C	G	A	T	A	T	G	C	A	A	**6/4**	35.2%
		74 **Thr** **↓** **Ser**	**143** **His** **↓** **Arg**	**168** **Asn** **↓** **Ser**		**196** **Asn** **↓** **Ser**				**375** **Gly** **↓** **Arg**			**474** **Lys** **↓** **Gln**	**476** **Pro** **↓** **Ser**	**489** **Gly** **↓** **Arg**		**539** **Leu** **↓** **Gln**			

Fig. 2. SNPs in CBDAS gene sequences expressed in the *Cannabis* germplasm. All SNPs are defined in comparison with the GenBank accession Nr. E55107. SNPs involving an amino-acid change in the translated protein are indicated in bold, and the type of change is indicated in the last line with the corresponding position in the protein sequence (yellow, amino-acids of the same chemical type, red amino-acids chemically different). SNP positions are identified by green color and by different sequence ID numbers; in the last column the proportion of the end-product CBD(V)A in the total cannabinoid fraction accumulated by each accession is given (see also Table 3).

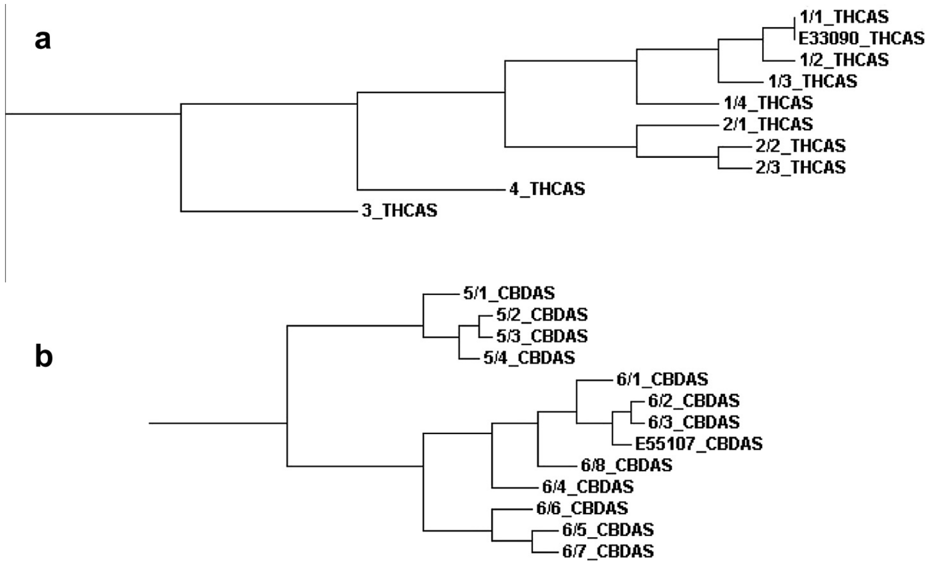


Fig. 3. Similarity within THCA- and CBDA-synthases. The similarity relationships between the sequences obtained from the accessions listed in Table 1 with the reference sequences for THCAS (a – E33090, Taura et al., 2000) and CBDAS (b – E55107, Yoshikai et al., 2001). Codes for the sequences were attributed on the basis of their clustering.

Five out of twelve accessions studied revealed a single CBDAS sequence expressed, so in these accessions the CBDAS sequence identified can be associated with a specific chemotype. Only in the case of accession 2009.23.25.15.14.4, was CBDA the strongly predominant cannabinoid (89.3%). In this case, the associated sequence, 5/2, can be considered as coding for a fully functional CBDAS, despite four amino-acid changes ((Thr → Ser in position 74, Asn → Ser in position 168 and 196 and Lys → Gln in position 474) compared with E55107 CBDAS. Consequently the functionality of the synthase can be considered unaffected by these changes in the protein, at least if they are simultaneously present.

Accession 94.5.2.30.1.2.4 accumulated on average 65.4% of its cannabinoid as CBDA; it is a relatively low proportion compared with other B_D/B_D accessions. This accession only transcribed CBDAS sequence 5/1, with 11 SNPs and five changes in the protein sequence compared with the reference. Four of these changes were shared with the previously described sequence 5/2, but an additional SNP in position 1123 (found exclusively in this strain) caused a change Gly → Arg in position 375 of the enzyme; therefore it can be hypothesized that such change causes the translation of a slightly less efficient CBDA synthase. It should be noted that this accession was the CBDA parent of the heterozygous 99.4 F_1 progeny that had been previously examined (de Meijer et al., 2003), and had a THCA/CBDA ratio significantly higher than the other F_1s; it had been suggested that this deviant ratio with its stable inheritance could reflect the activity of synthase variants; the mutation identified in CBDAS seems to confirm this hypothesis. It is proposed to define this new B_D allele as B_{Dw} (w for *weak*).

Two more accessions, despite being identified by the marker as B_D/B_D, accumulated low or extremely low amounts of CBDA: 2005.42.11.59 (92.9% CBGA and only 6.3% CBDA) and 2005.16.(2 + 12 + 24).24 (99.9% CBGA and only a trace of CBDA). In both cases, only a single transcribed CBDAS was found; in the first case (sequence 6/1 in Fig. 2) only two SNPs were present, one of which (T for C in position 1426 of the gene) induced a Pro → Ser change in position 476 of the enzyme. In the second accession, which accumulated CBGA exclusively, 10 SNPs were found (sequence 5/3), nine of which were common to the previously discussed sequence 5/1. Additionally, in the sequence 5/3, an A for G change in 1465 of the gene resulted in a quite dramatic Gly → Arg change in position 489. As a result, a probably completely non-functional CBDAS is translated, leading to the accumulation of the unconverted CBGA precursor only. These two important variants in the CBDAS group represent B_{D0} mutations as postulated by de Meijer and Hammond (2005). The sequence 6/1 encoding a minimally functional CBDAS is referred to as allele B_{D0^1}. The sequence 5/3 coding for a completely non functional CBDAS is referred to as allele B_{D0^2}.

In the other accessions with CBDA as the prevalent cannabinoid, more than one transcribed sequence was found, and therefore once again it is difficult to deduce the functionality of the translated enzymes for which they code. CBDA proportions varied from 77% to 88% in the presence of two or three different sequences (Fig. 2). Accession 55.14.4.27.2.6.4 transcribed three sequences, including the fully functional 5/2 (already discussed) and the sequence 6/2 coding for a CBDAS the same as the GenBank E55107. Accession 98.2.21.19.8.7 transcribed three sequences, two of which (6/5 and 6/6) coded for enzymes identical to the E55107. Accessions 2010.24a.35/C and 2010.29 each also transcribed two sequences, one coding for a protein identical to the reference E55107. Accession K310 transcribed three sequences, one of which, 5/2, was already known as coding for a fully functional CBDAS. In all these cases, high proportions of CBDA accumulated in the inflorescences.

In cv. Ermo, with a very low cannabinoid content but a relatively high CBGA proportion, only one major transcribed sequence was found: 6/4 (Fig. 2) with 4 SNPs and two variant amino-acids. One of which (Pro → Ser in position 476) already found in the CBGA prevalent accession 2005.42.11.59, genotyped as B_{D0^1}/B_{D0^1}. The other change was a Leu → Gln in position 539. Although this change had already been found in a CBDAS, its functionality could not be assessed due to the simultaneous presence of other synthases.

2.4. qPCR of THCA and CBDA synthase genes

The primers used for qPCR analyses of THCA- and CBDA-synthase genes were designed to avoid the SNPs identified during the sequencing of the cDNA isolated from the different accessions. Therefore, these data account for a cumulative transcription level of all the identified sequences. The relative expression data, along with the relative proportion in the cannabinoid fraction of THCA and CBDA, and the accessions to which the expression refers, are shown in Fig. 4.

The expressions of all THCAS-related sequences within each accession are shown in Fig. 4a, calculated as levels of transcripts relative to accession M124: the one with the highest proportion of THCA in its cannabinoid fraction. There was no apparent correlation between the level of transcription of the THCA-synthase gene(s) and the proportion of THCA actually accumulated in the cannabinoid fraction of the inflorescences. Sequence 1/3 (the only one transcribed in accession 2009.27.44.2.19), coding for a largely defective THCAS, was transcribed at levels only slightly lower than the sequence 3, which codes for the fully functional, unique THCAS of accession 2010.24a.35/T.

For the quantification of CBDAS sequences, the transcription abundances were expressed as relative to the level of one high-CBDA accession, 2010.24a.35/C, and are shown in Fig. 4b. As with the THCAS genes, no strict correlation between observed CBDA proportion putatively determined by each sequence and transcription rate of the synthase genes is observed. However, the sequence 5/3, coding for the completely inactive CBDAS (allele B_{D0^2}) in accession 2005.16.(2 + 12 + 24).24 is transcribed at very low levels compared with the other sequences. The 6/1 sequence (representing the B_{D0^1} allele), coding for the other largely defective CBDAS, and the sequence 5/1, coding for a poorly functional CBDAS, were transcribed at rates comparable with sequences also found in high-CBDA accessions.

The sequence 6/4, the only expressed at significant levels in the low cannabinoid content cv. Ermo (Fig. 2), was also found to be expressed at comparatively low levels.

Overall, these gene expression data confirm that the conversion of CBGA in THCA and CBDA is not transcriptionally regulated in most cases.

When qPCR primers specific for CBDAS are used to amplify cDNA in B_T/B_T accessions, no signal whatsoever was observed; yet, if THCAS-specific primers are used on cDNA from B_D/B_D accessions, some amplified product occasionally occurs, although at very high *Ct* values (>36; data not shown). Therefore, it seems that very low-abundance transcripts very similar to THCAS, and transcribed at exceedingly low levels, might be present in accessions producing mostly CBDA or CBGA. A possible explanation may be that the primers we used for THCAS are also able to amplify sequences transcribed from DNA corresponding to Kojoma's sequences, reportedly only present in some fiber type accessions (Kojoma et al., 2006). These "obscure" THCAS-like sequences appear to co-segregate with CBDAS.

2.5. Comparison with other THCAS and CBDAS sequences

The expressed sequences were also compared by cluster analysis (Clustal W2-Phylogeny) with other sequences, especially with genomic sequences deposited in the GenBank and with expressed sequences identified in the genome sequencing and RNA-Seq analysis by van Bakel et al. (2011) or those kindly provided by F. Licausi (PlantLab, Scuola Superiore Sant'Anna, Pisa). Due to the existence of several pseudogenes, we decided to compare only complete sequences from other sources, with 544 or 545-amino-acid open reading frames.

The results are shown in Fig. 5. The sequences are clearly separated by UPGMA analysis into two main clusters, corresponding to CBDAS and THCAS sequences. In the CBDAS cluster, the sequences obtained in this work and coded as 5, form a minor cluster. This is distinct from the sequences of family 6 which group with most of the other known CBDAS sequences: two NCBI complete ones (the reference E55107, and AB292682), and one sequence from RNA-Seq data provided by F. Licausi (ifc0_CBDS_152). Hence, most of the CBDAS sequences identified in this work have not been previously reported and underline the huge amount of sequence variation accumulated in the CBDAS gene family. A minor CBDAS cluster was formed by Taura's sequences AB292683 and AB292684 and had more than 150 SNPs compared with the E55107 reference. These are indicated as coding for non-functional synthases (Taura et al., 2007). Sequence scaffold 39155 (from cv. Finola, van Bakel et al., 2011) is the most dissimilar.

The main THCAS cluster includes all our sequences, the reference sequence NCBI33090, 4 sequences obtained by RNA-Seq and provided by F. Licausi, two THCAS from NCBI, (AB212838 and

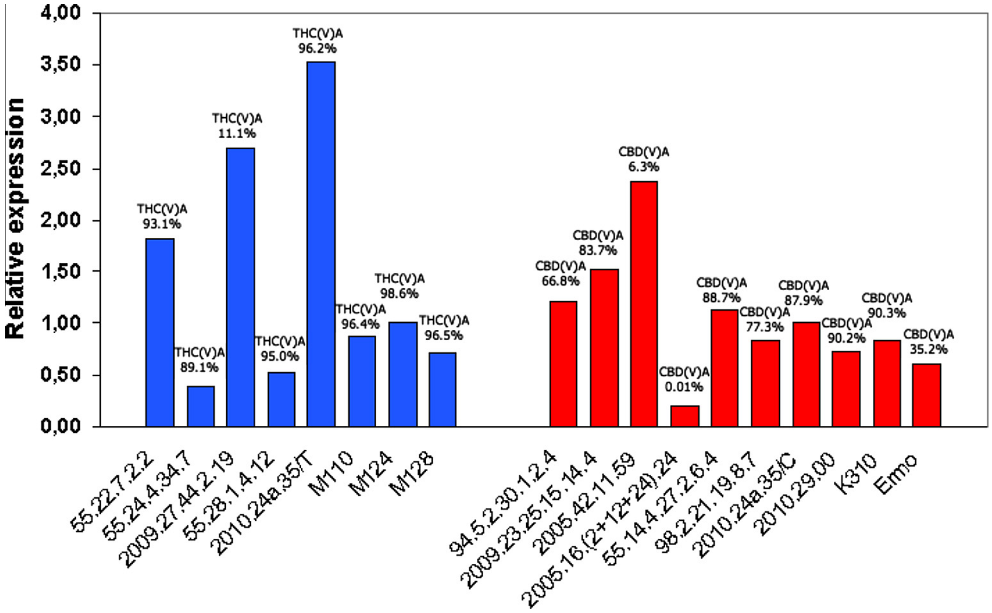


Fig. 4. Gene expression of THCA- and CBDA-synthases. Relative expression of THCA (blue bars)- and CBDA (red bars)-synthases in the different accessions tested. Primers used for qPCR were designed so to amplify all the variant THCAS found or present in the database. THCAS transcription in accession M124 and CBDAS transcription in accession 2010.24a.35/C, with the highest THCA and CBDA proportions in the cannabinoid fraction, were taken as calibrators for the respective quantification of transcripts in the other accessions. Above the bars, the proportions of THC(V)A and CBD(V)A in the cannabinoid fraction of the accessions are indicated. Codes along X-axis refer to the accession codes, as listed in Table 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

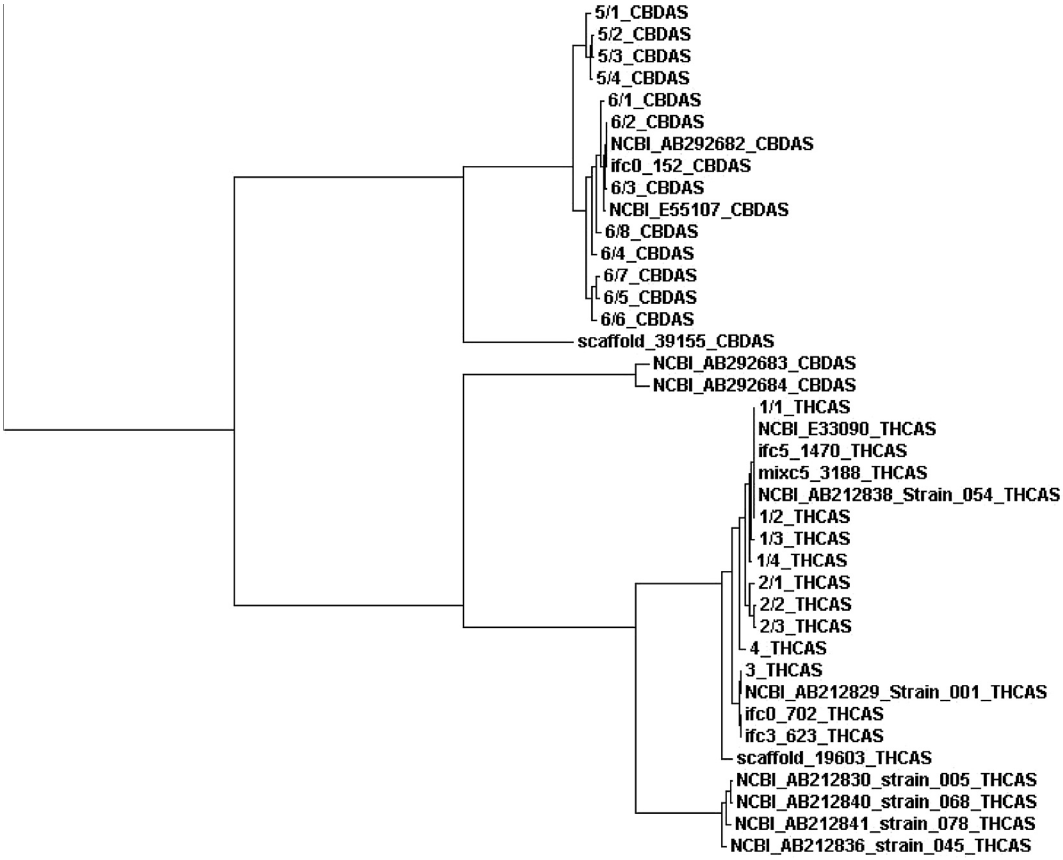


Fig. 5. Cluster analysis of synthase sequences. Cluster analysis (UPGMA, Clustal Phylogeny tool) of the cannabinoid synthases present in the databases or sequenced during this work. Sequences AB212829, AB212830, AB212836, AB212838, AB212840 and AB212841 from Kojoma et al., 2006; sequences AB292682, AB292683 and AB292684 are from Taura et al. (2007); sequences infc0_152, infc5_1470, mixc5_3188, infc0_702 and infc3_623 are THCAS or CBDAS sequences provided by Dr. F. Licausi, Dept. Plant Biology, Sant'Anna School, Pisa (Italy); sequences indicated as scaffolds are from van Bakel et al. (2011), downloaded at http://genome.ccbr.utoronto.ca/.

AB212829), and two genomic sequences obtained from drug strains (Kojoma et al., 2006). Once again, the THCAS complete sequence obtained from Purple Kush by van Bakel et al. (2011) is the most different (scaffold 19603 in Fig. 5). Besides, a separate cluster of NCBI sequences, formed by AB212830/36/40/41 and having a high number of SNPs (63 or 64 SNPs) has been identified by Kojoma et al. (2006) from *Cannabis* CBDA-predominant or mixed THCA/CBDA strains. It is hypothesized that these THCAS sequences, assumed to be completely non-transcribed by Kojoma et al. (2006), might instead be responsible for the low-level qPCR signals in some B_D/B_D accessions observed when using THCAS specific primers.

The complete THCAS sequences obtained by RNA-Seq by van Bakel et al. (2011) and Licausi et al. (personal communication) are different by only 5 SNPs that do not coincide with those identified in this study and with others previously deposited in Genbank, However, our sequence 1/1 and Kojoma's sequence AB212838 proved identical to the reference, E33090. Our sequence 3, found in two high-THCA accessions, was identical to Kojoma's sequence AB212829 (Fig. 5).

3. Discussion

In a previous paper (de Meijer et al., 2003), it was demonstrated that the value of the CBDA/THCA ratio in an F_1 cross between homozygous pure-THCA and pure-CBDA parents, was an inherited trait and differed significantly, depending on the F_1 considered. This fact suggested a possible differential efficiency of one or both synthases – when both present in the heterozygous plants – to compete in the conversion of the common precursor. Such slightly different abilities in oxydocyclizing CBGA was in turn tentatively ascribed to the occurrence of an allelic series at the *B* locus which was responsible for differential catalytic efficiencies of the encoded synthases. The complete sequences of both CBDA- and THCA-synthases of the *Cannabis* strains used to obtain the variant F_1s were here determined, and a specific variation in the sequence of the CBDA synthase of strain 94.5.2.30.1.2.4 (sequence 5/1) was actually found. Potentially, this variation has a detrimental impact on the efficiency of the enzyme, thereby determining the lower accumulation of CBDA. The allele coding for this specific sequence was indicated as B_{DW}.

It was previously proposed (de Meijer and Hammond, 2005) that *Cannabis* chemotypes characterized by a strong CBGA accumulation could be due to defective synthases unable, or less able, to convert the precursor CBGA into the end-products THCA or CBDA; therefore, it should be possible to find such defective alleles at both the B_D and B_T loci postulated in the genetic model proposed by de Meijer et al. (2003). So the convergence toward the same phenotype, i.e. CBGA accumulation, was predicted to occur through different mutations of different alleles. One such defective allele, B_{D0^1}, had already been genetically characterized by de Meijer and Hammond (2005). In the current work, it was confirmed that this allele expresses a unique CBDA-synthase sequence (6/1, Fig. 2) with substitutions in the amino-acid sequence that can explain an inefficient CBGA conversion. In addition, two defective alleles (B_{D0^2} and B_{T0}, sequences 5/3 and 1/3 in Figs. 1 and 2) were found and sequenced; their characteristics give further credence to the hypothesis that CBGA chemotypes are caused by minimally functional CBDA- or THCA-synthases. It should be stressed that in all these CBGA-accumulating accessions, only one expressed sequence was retrieved, thus making practically certain that it is determining the phenotype.

The positions of the aminoacid substitutions in the deduced tertiary structure of the CBGA-accumulating synthases are shown in Fig. 6, where they are compared with the wild-type structures. If the phenotypic effects represented by the different accessions can be traced back to sequence variation in CBDAS and THCAS, inferences on the amino-acid positions crucial for the complete enzymatic function can be obtained. B_{D0^1} and B_{D0^2} alleles code for different sequences (6/1 and 5/3), with only one SNP (which is not relevant for protein function; see Fig. 2) in common. These two mutations then are very different, the first characterized by a single C → T transition, leading to a Pro → Ser change in position 476 of the protein, and the second by a number of SNPs, 5 of which lead to amino-acid changes in the enzyme primary structure and are scattered throughout the different protein domains. Among these variations, the most crucial (and the only one unique to this mutant) appears to be the mutation occurring in position 489 (Gly → Arg). The B_{T0} allele is represented by sequence 1/3 with a single SNP causing the change of an acidic aminoacid (Glu) to a basic one (Gln) in position 236 of the enzyme. These variations are not in the primer binding regions and do not affect the identification of the different accessions by the B_T- or B_D-specific markers, but they appear likely to affect, by different mechanisms, the functionality of the encoded enzymes leading to a CBGA-accumulating chemotype (Fournier et al., 1986; Mandolino and Carboni, 2004). The weak but still functional synthase coded for by the B_{Dw} allele discussed above and represented by sequence 5/1, is also characterized by a unique mutation Gly → Arg in position 375 of the enzyme, and results in a partial accumulation of CBGA due to the partly impaired activity of the CBDA-synthase. Both B_{Dw} and B_{D0^2} CBDAS, with impaired function, appear to have evolved from the fully functional CBDAS coded for by sequence 5/2, sharing with it 4 of their 5 aminoacids variations (Figs. 2 and 6).

It should be noted that no variation was found in the first 28 amino-acids of the protein representing the trans-membrane domain of the synthases. All variant proteins found, when analyzed by PSORT 6.4 software, were equally described as putatively directed to the secretory route with very high probability (>90%, data not shown). Protein characteristics as indicated by Sable analysis of the secondary structure of the variant synthases (http://sable.cchmc.org/), and by the protein parameters predicted by ExPASy (www.expasy.org) are not dramatically different for the variant synthases compared with the reference ones (data not shown), and the same holds true for the calculated tertiary structure (Fig. 6). Furthermore, no SNPs were found in the CBDAS amino-acidic residues indicated to be putative glycosylation sites of the mature protein or in the flavin-binding consensus sequence by Taura et al. (2007). As for the THCAS identified, none of the observed mutations in the enzyme primary sequences corresponds to the mutations induced in the THCA-synthase sequence by Shoyama et al. (2012). According to these authors, enzyme activity was severely reduced by a Tyr to Phe change in position 417 and by Glu to Gln change in position 442, while it was completely shut off by a His to Ala (292) and by a Tyr to Phe change in 482. In our group of B_T/B_T accessions the only functionally crucial mutation appeared to be the sequence 1/3 of allele $B_{T0,}$ located in the subdomain Ib as identified by X-ray crystallography by Shoyama et al. (2012), involving the Glu to Gln change already discussed (236). It should be noted though that our B_{T0} mutant is not a complete loss-of-function one, as the accession carrying it still accumulates some of its total cannabinoid fraction as THC(V)A (ca 10%). Also, none of the 12 residues of THCA-synthase reported by Shoyama et al. (2012) to bind to FAD, either covalently or by H bonds, was involved in the B_{T0} mutation described here.

In their tertiary structure analysis, Shoyama et al. (2012) examined THCAS and neglected CBDAS. We compared their induced mutations in THCAS (Tyr417Phe and Glu442Gln for a partial loss

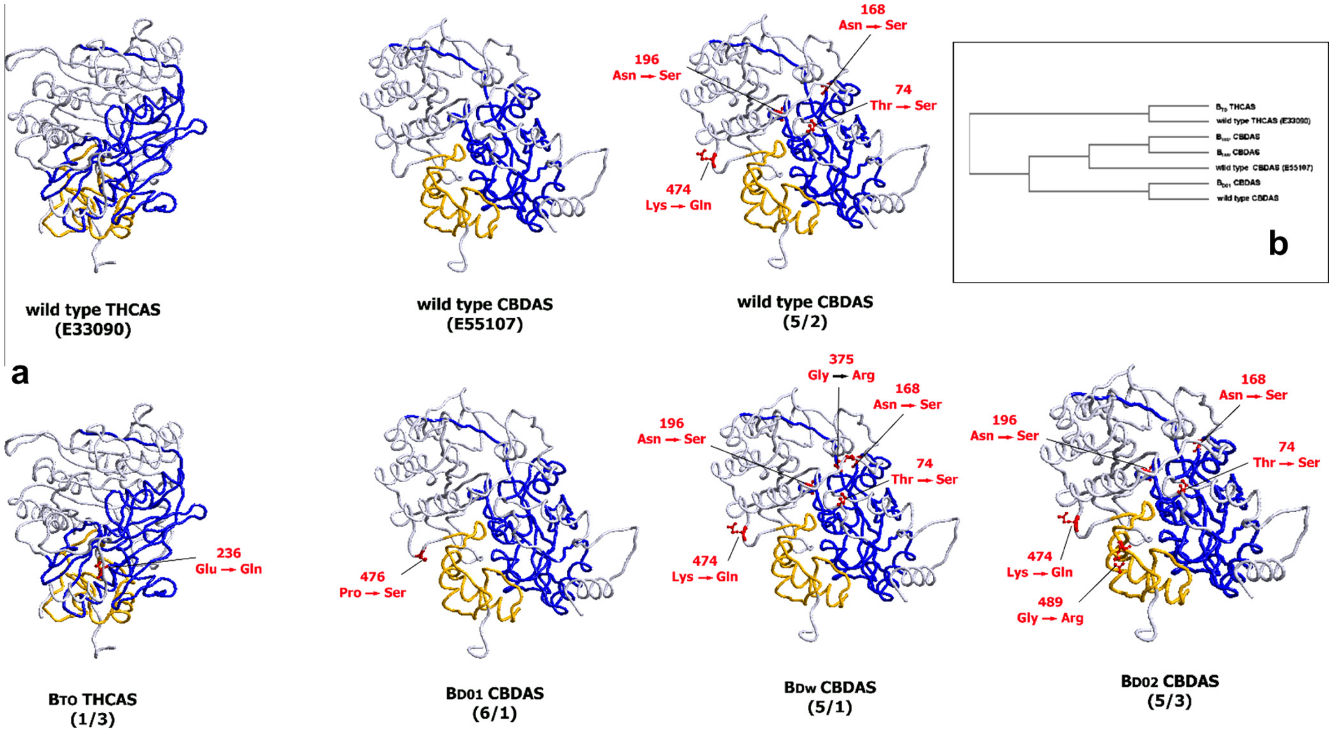


Fig. 6. Models of the tertiary structure of THCA synthases (left) and CBDA synthases (right) with the indication of the aminoacid changes for the main wild types and mutants discussed in the text. The berberine-bridge enzyme-like domain is indicated in yellow and the FAD-binding domain in blue. This view of the THCAS and CBDAS structures was chosen to optimize visibility of the position of the variant aminoacids within the tertiary structure. In the inset (b), the dendrogram shows the relationship between the structures. Codes of the respective sequences, as specified in Figs. 1 and 2, are in brackets.

of function, His292Ala and Tyr484Phe for a complete knock-out) with those putatively crucial in our synthases (Pro → Ser in position 476 in the B_{D0^1} mutation, Gly → Arg in position 489 in the B_{D0^2} mutation, and Glu → Gln in position 236 in the B_{T0} mutation). It is evident that mutations greatly affecting enzyme activity are likely to fall in two areas: the first directly, or indirectly, involved in FAD binding and the second close to the catalytic site in the berberine bridge-like domain of the enzyme (Fig. 6).

In conclusion, the novel THCAS and CBDAS sequences identified in this work confirm that different mutations can affect the functionality of the cannabinoid synthases and explain most of the different chemotypes observed.

It should be noted that most of the expressed sequences obtained are different from GenBank ones, and that overall, we identified 9 THCAS sequences and 12 CBDAS sequences, only one of which (THCAS sequence 1/1) identical to the genomic sequence E33090 deposited in NCBI. Other complete sequences present in the database, or obtained by other investigators, add further variation to the known cannabinoid synthase genes that can be considered a true gene family. The THCAS determined in this work averaged 2.9 SNPs per sequence, while the CBDAS had 5.7 SNPs per sequence, so a higher mutation rate seems to be a feature of CBDAS. This suggests that, if the hypothesis of a common ancestor for these genes is accepted (Taura et al., 2007), the CBDAS gene may be the ancestral one from which THCAS genes evolved. Alternatively, THCAS mutations could be more deleterious to fitness and thus more constrained than CBDAS mutations. All the nucleotide sequences described here are deposited in GenBank (accession numbers KP970849 to KP970856 for THCAS and KP970857 to KP970868 for CBDAS).

One of the main evolutionary mechanisms giving rise to gene families is duplication and divergence (Kliebenstein et al., 2001; Osbourn, 2010) and given the strong similarity between the two main types of synthases, this mechanism has also been suggested as an explanation for cannabinoid diversity (Taura et al., 2007). Sequencing data (van Bakel et al., 2011 and this paper) indicate the presence of more than one transcribed and potentially translatable gene for each of the two main synthases. The relationships within the entire family of fully sequenced CBDA- and THCA-synthases are shown in Fig. 3. The two clusters of CBDAS and THCAS are generally well separated, but there are sequences that could be considered somehow intermediate between the two main groups. GenBank sequences AB292683 and AB292684 obtained by Taura et al. (2007) are the most distant from both groups. These sequences have over 150 SNPs and 80 amino-acid changes compared with NCBI reference CBDAS sequence E55107, although a putative 545-amino-acids ORF is still conserved. Taura et al. (2007) were unable to associate these sequences with any enzymatic activity in transformed insect cells.

In Fig. 3 a small cluster, distinct from the other THCAS, is formed by the THCAS gene sequences AB212830/36/40/41. It is possible to speculate that these sequences, found by Kojoma et al. (2006) only in some CBDA predominant accessions, could be considered evolutionarily intermediate or side-branches between the CBDAS and the fully functional THCAS. We did not recover THCAS-like sequences from cDNA of B_D/B_D accessions. However, by real time analysis using consensus primers amplifying all THCAS, we found that some pure-CBDA accessions transcribed a THCAS-like sequence ($Ct > 36$) to a very small extent, while no THCA accession showed any detectable transcription of CBDAS-related sequences. The B_T-specific marker failed to find these THCAS-like sequences in CBDA-predominant genomic DNA. This was probably due to two mismatches in the sequence of one of the primers used for the marker which caused the B_T-specific marker's failure to identify these mostly inactive sequences. The other THCAS sequences described by Kojoma et al. (2006) are typical, functional THCAS with a very great similarity to those found both in our accessions and by other authors by RNA-Seq (F. Licausi, personal communication), and clustering along with them.

Consequently, the small but constant amounts of THCA found in CBD-predominant, highly inbred *Cannabis* lines in this work (3–5% of the total cannabinoids, Table 3) may be due to a low-level activity of Kojoma's type THCAS present in CBDA accessions. The opposite, i.e. the synthesis of small amounts of CBDA in inbred THCA-pure lines, does not seem to occur: the observed range of CBDA proportion in THCA-predominant material is below 0.03%. This might be due to the absence of low-activity CBDA synthase in high-THCA material, and is confirmed by our real time PCR data and by the findings of van Bakel et al. (2011) who analyzed the whole Purple Kush transcriptome to find, apart from THCAS sequences, only pseudogenes with homology with CBDAS.

The complexity of the above described frame suggests that CBDA accessions might be the ancestral ones. Events of duplication could have led to a higher CBDAS variation, to the formation of CBDAS pseudogenes and to the rise of new sequences coding for a new enzyme able to oxidocyclize CBGA in a new way, thus producing THCA. The subsequent loss of function of CBDAS sequences led to the appearance of THCA chemotypes, presumably in environments conferring some advantage to a high THCA synthesis (as hypothesized by Pate, 1983). Finally, the long history of *Cannabis* domestication led to further separation and specialization of the sequences.

In the *Cannabis* germplasm analyzed, we found the simultaneous expression of more than two sequences in some cases (Figs 1 and 2); this fact is apparently inconsistent with the strictly monogenic inheritance model proposed by de Meijer et al. (2003). The expression of multiple, related sequences has also been observed by high-throughput RNA-Seq analyses (van Bakel et al., 2011; F. Licausi, personal communications). This is the first report of the existence of different (up to 5 in high-THCA accession M110), complete, expressed sequences coding for variant CBDA- or THCA-synthases. In many cases the duplication of a single ancestral gene sequence leads to several, tightly associated loci, between which recombination cannot be observed by genetic means especially if such sequences keep coding for functional synthases. Therefore, we propose here that the monogenic model (de Meijer et al., 2003) can still be considered valid for all practical purposes (e.g. breeding and genetic analysis), and it remains fully consistent with all genetic dfata available. Nevertheless, it is necessary to become aware of the possible variation and multiplicity occurring at the molecular level in the coding sequences of the functional cannabinoid synthases contributing to the final chemotype.

4. Conclusions

In several *Cannabis* accessions, we found more than one expressed sequence for THCAS and CBDAS, each possessing an ORF allowing the translation into an entire protein. In four accessions, characterized by high levels of CBGA, we identified different mutations in the CBDAS and THCAS sequences that are likely to impair enzyme functionality, thus explaining CBGA accumulation. New allelic codes (B_{Dw}, B_{D0^1}, B_{D0^2}, B_{T0}) were attributed to these sequences that expand the locus *B* allelic series.

The amino-acid substitutions observed in the mutants do not correspond to any critical position in the tertiary structure as described by Shoyama et al. (2012) or to any of the induced mutations of the THCAS expressed in heterologous systems by the same authors; therefore the phenotypic effects of these mutations provide new information on critical amino-acid positions in both THCAS and CBDAS.

The transcription rate of the different sequences was rarely correlated with the proportion of THCA or CBDA in the cannabinoid fraction but it cannot completely be ruled out that besides the mutations in the synthases, regulatory elements in *cis* or *trans*, contribute to the chemotype of the plants carrying them. This agrees with recent data on the lack of correlation between THCAS transcription levels and THC accumulated in the leaves or inflorescences (Cascini et al., 2013).

Finally, the mutation rate found for THCAS and CBDAS, and the comparison of the sequences identified with those present in the databases, allowed to propose a hypothesis on the phylogenetic relationships between these sequences, where the CBDAS is considered the ancestral one, and THCAS sequences arose in CBDA-predominant germplasm by duplication and divergence.

5. Experimental

5.1. Plant material

The accessions (inbred lines, clones, a landrace and a cultivar) used for the sequencing of the cannabinoid synthases are listed in Table 1, where the geographic provenance of the original source population is also indicated. The plant material listed has been selected not only on the basis of the chemotype, but also of geographical distribution and crop use type, intended to comprise maximal sequence heterogeneity and functional variation. All plants tested were diploid females. Per population or inbred line, at least two seedlings, and per clone, at least two cuttings were used for RNA extraction and cDNA sequencing.

5.2. Chemotype assessment

The plants were raised from seed or cuttings, in a computer-controlled greenhouse, under a 24 h photoperiod for 3 weeks, and then moved to a 12 h photoperiod for flower induction for a 9 week generative period. The temperature was 25 °C and the conditions were kept uniform for all plants during the entire growth period. Mature floral samples (ca. 1 g) were collected, dried and treated as described elsewhere for gas-chromatographic determination of cannabinoids (de Meijer et al., 2009a).

5.3. Molecular markers

DNA was extracted from dried leaves of all plants using the Power Plant DNA Isolation kit (Mobio laboratories, Inc.) according to the manufacturer's instruction. For primer design, the GenBank sequences E55107 and E33090 were used (CBDA-synthase and THCA-synthase respectively; Yoshikai et al., 2001; Taura et al., 2000); the markers at the synthase loci were obtained using a multiplex system with three primers described elsewhere (Pacifico et al., 2006), the first designed on a sequence common to both synthases, and the other two respectively THCAS- and CBDAS-sequence specific. This marker has already been described as correctly identifying THCA – (genotype B_T/B_T) and CBDA – (B_D/B_D) prevalent plants, and their hybrids (B_T/B_D; Pacifico et al., 2006; Mandolino, 2007). Fifty nanograms of genomic DNA were amplified using the conditions previously described (Pacifico et al., 2006).

5.4. RNA extraction and cDNA synthesis

RNA was isolated from snap frozen ground inflorescences with the Plant RNeasy Mini Kit (Qiagen) according to manufacturer instruction, starting from 100 mg of pulverized tissue. One microgram of RNA was treated with DNase (Life Technologies) to remove any possible DNA contamination and retro-transcribed into cDNA with the High Capacity cDNA Reverse Transcription Kit (Applied Biosystems).

5.5. Sequencing, SNPs analysis and bioinformatics

In order to sequence the fragments corresponding to the complete THCA- and CBDA-synthases, 50 ng of the cDNA of each plant was amplified with specific primers designed to amplify the complete gene sequence of 1635 bp, using the high fidelity Accuprime Pfx DNA Polymerase (Life Technologies). Each amplicon was then purified with Purelink PCR purification kit and cloned into the vector of the Zero Blunt PCR cloning kit (Invitrogen). Five to ten clones per sample were then grown overnight and the extracted transformed vector checked with M13 PCR for fragment insertion. The transformed plasmids were then all sequenced according to the Sanger method from BMR Genomics (Padua, Italy). Only complete sequences, obtained by bacterial clones harboring a full 1635-bp insert were considered. The complete cDNA sequences were aligned with the two sequences of THCA- and CBDA-synthase (GenBank accessions E33090 and E55107), taken as reference. The sequence comparison was carried out by ClustalOmega software (www.ebi.ac.uk), and SNP positions identified upon complete alignment. Relationships between the different sequences were analyzed by Clustal W2-Phylogeny software using the UPGMA clustering method. Protein sequences were obtained with ExPASy Translation Tool (www.expasy.org). Tertiary structures were calculated with the SWISS-MODEL tool (http://swissmodel.expasy.org/) and 3-D structures visualized with software RasTop 2.2.

5.6. Real time analysis

CBDA-synthase and THCA-synthase specific primers for the qPCR quantification of the transcript levels of the different cannabinoid synthase genes were designed in consensus regions of the CBDA- and THCA-synthase cDNA sequences obtained, respectively.

One hundred nanograms of cDNA from the previously described samples were used. qPCR analysis was carried out using a Rotor-Gene 6000 apparatus (Corbett), and the Rotor-Gene SYBR Green PCR master mix (Qiagen), according to the manufacturer's instruction. Tubulin was used as internal reference gene (Marks et al., 2009). The primers used for the detection of THCA- and CBDA-synthase and for the tubulin transcript levels were checked for comparable amplification efficiencies in serial dilutions of the cDNA, in order to analyze the data obtained through the relative quantification method of $2^{-\Delta\Delta Ct}$ (Livak and Schmittgen, 2001). The amplification efficiency was close to 100%. All the primers used in this work, for genotyping, sequencing or qPCR analysis, are listed in Table 2.

Acknowledgments

We are grateful to Dr. Francesco Licausi (Scuola Superiore Sant'Anna, Pisa) for providing his RNA-Seq data on *Cannabis* synthases and for helpful discussions. We also thank Kathy Hammond for her valuable technical assistance and for reviewing the text. The present work was supported in the frame of the project "Agronanotecnologie 2 – Nuove tecnologie molecolari", financed by the Italian Ministry of Agriculture and Forestry.

References

Cascini, F., Passerotti, S., Boschi, I., 2013. Analysis of THCA synthase gene expression in cannabis: a preliminary study by real-time quantitative PCR. Forensic Sci. Int. 231, 208–212.

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

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

de Meijer, E.P.M., Hammond, K.M., Micheler, M., 2009a. The inheritance of the chemical phenotype in *Cannabis sativa* L. (III): variation in cannabichromene proportion. Euphytica 165, 293–311.

de Meijer, E.P.M., Hammond, K.M., Sutton, A., 2009b. The inheritance of the chemical phenotype in *Cannabis sativa* L. (IV): cannabinoid-free plants. Euphytica 168, 95–112.

de Zeeuw, R.A., Wijsbek, J., Breimer, D.D., Vree, T.B., van Ginneken, C.A., van Rossum, J.M., 1972. Cannabinoids with a propyl side chain in *Cannabis*. Occurrence and chromatographic behaviour. Science 175, 778–779.

Fellermeier, M., Zenk, M.H., 1998. Prenylation of olivetolate by a hemp transferase yields cannabigerolic acid, precursor of tetrahydrocannabinol. FEBS Lett. 427, 283–285.

Fellermeier, M., Eisenreich, W., Bacher, A., Zenk, M.H., 2001. Biosynthesis of cannabinoids. Incorporation experiments with ^{13}C-labeled glucoses. Eur. J. Biochem. 268, 1596–1604.

Fournier, G., Richez-Dumanois, C., Duvezin, J., Mathieu, J.-P., Paris, M., 1986. Identification of a new chemotype in *Cannabis sativa*: cannabigerol-dominant plants, biogenetic and agronomic prospects. Planta Med. 53, 277–280.

Gagne, S.J., Stout, J.M., Liu, E., Boubakir, Z., Clark, S.M., Page, J.E., 2012. Identification of olivetolic acid cyclase from *Cannabis sativa* reveals a unique catalytic route to plant polyketides. Proc. Natl. Acad. Sci. U.S.A. 109, 12811–12816.

Happyana, N., Agnolet, S., Muntendam, R., Van Dam, A., Schneider, B., 2013. Analysis of cannabinoids in laser-microdissected trichomes of medicinal *Cannabis sativa* using LCMS and cryogenic NMR. Phytochemistry 87, 51–59.

Kim, E.S., Mahlberg, P., 1997. Immunochemical localization of tetrahydrocannabinol (THC) in cryofixed glandular trichomes of *Cannabis* (Cannabaceae). Am. J. Bot. 84, 336–342.

Kliebenstein, D.J., Lambrix, V.M., Reichelt, M., Gershenzon, J., Mitchell-Olds, T., 2001. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. Plant Cell 13, 681–693.

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

Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real time quantitative PCR and the $2^{-\Delta\Delta Ct}$ method. Methods 25, 402–408.

Mahlberg, P., Kim, E.S., 2004. Accumulation of cannabinoids in glandular trichomes of *Cannabis* (Cannabaceae). J. Ind. Hemp 9, 15–36.

Mandolino, G., 2007. Marker assisted selection and genomics of industrial plants. In: Ranalli, P. (Ed.), In Improvement of Crop Plants for Industrial End Uses. Springer, pp. 59–82.

Mandolino, G., Carboni, A., 2004. Potential of marker-assisted selection in hemp genetic improvement. Euphytica 140, 107–120.

Mandolino, G., Bagatta, M., Carboni, A., Ranalli, P., de Meijer, E.P.M., 2003. Qualitative and quantitative aspects of the inheritance of chemical phenotype in *Cannabis*. J. Ind. Hemp 8 (2), 51–72.

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

Osbourn, A., 2010. Secondary metabolic gene clusters: evolutionary toolkits for chemical innovation. Trends Genet. 26, 449–457.

Pacifico, D., Miselli, F., Micheler, M., Carboni, A., Ranalli, P., Mandolino, G., 2006. Genetics and marker-assisted selection of the chemotype in *Cannabis sativa* L. Mol. Breed. 17, 257–268.

Pate, D.W., 1983. Possible role of ultraviolet radiation in evolution of *Cannabis* chemotypes. Econ. Bot. 37, 396–405.

Shoyama, Y., Yamauchi, T., Nishioka, I., 1970. *Cannabis*. V. Cannabigerolic acid monomethyl ether and cannabinolic acid. Chem. Pharm. Bull. (Tokyo) 18, 1327–1332.

Shoyama, Y., Hirano, H., Nishioka, I., 1984. Biosynthesis of propyl cannabinoid acid and its biosynthetic relationship with pentyl and methyl cannabinoid acids. Phytochemistry 23, 1909–1912.

Shoyama, Y., Tamada, T., Kurihara, K., Takeuchi, A., Taura, F., Arai, S., Blaber, M., Shoyama, Y., Morimoto, S., Kuroki, R., 2012. J. Mol. Biol. 423, 96–105.

Sirikantaramas, S., Morimoto, S., Shoyama, Y., Ishikawa, Y., Wada, Y., Shoyama, Y., Taura, F., 2004. The gene controlling marijuana psychoactivity. J. Biol. Chem. 279, 39767–39774.

Sirikantaramas, S., Taura, F., Tanaka, Y., Ishikawa, Y., Morimoto, S., Shoyama, Y., 2005. Tetrahydrocannabinolic acid synthase, the enzyme controlling marijuana psychoactivity, is secreted into the storage cavity of the glandular trichomes. Plant Cell Physiol. 46, 1578–1582.

Staginnus, C., Zorntlein, S., de Meijer, E.P.M., 2014. A PCR marker linked to a THCA synthase polymorphism is a reliable tool to discriminate potentially THC-rich plants of *Cannabis sativa* L. J. Forensic Sci. http://dx.doi.org/10.1111/1556-4029.12448.

Taura, F., Morimoto, S., Shoyama, Y., 1995. First direct evidence for the mechanism of Δ^1-tetrahydrocannabinolic acid biosynthesis. J. Am. Chem. Soc. 117, 9766–9767.

Taura, F., Morimoto, S., Shoyama, Y., 1996. Purification and characterization of cannabidiolic-acid synthase from *Cannabis sativa* L. J. Biol. Chem. 271, 17411–17416.

Taura, T., Matsushita, H., Morimoto, S., Masayama, Y., 2000. Tetrahydrocannabinolic acid synthase gene. GenBank, E33090.1.

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

van Bakel, H., Stout, J.M., Cote, A.G., Tallon, C.M., Sharpe, A.G., Hughes, T.R., Page, J.E., 2011. Genome Biol. 12, R12.

Yoshikai, K., Taura, T., Morimoto, S., Masayama, Y., 2001. DNA encoding cannabidiolate synthase. GenBank, E55107.1.