The identification of SNPs in THCA synthase gene from Pakistani Cannabis

Saadat Ali¹, Mehreen Mufti², Mehwish Khan³*, Ishrat Aziz⁴

¹University of Veterinary and Animal Sciences, Lahore, Pakistan
²Lahore University, Lahore, Pakistan
³School of Biological Sciences, Quaid-e-Azam campus, University of the Punjab, Lahore, Pakistan, 54000
⁴University of the Punjab, Lahore, Pakistan

Received 12th November 2018 / Accepted 2nd May 2019

Abstract. Five different isolates of Pakistani cannabis belonging to varying locations were analyzed for the presence of a reported tetrahydrocannabinolic acid (THCA) synthase marker or gene. The amplification of the marker (12 kb) from the five isolates confirmed them to be drug-type since the association of the marker with drug-type cannabis plants had already been established in other reports. The sequence analysis of the THCA synthase marker revealed two single nucleotide polymorphisms (SNPs) (i.e. A851→T851 and A883→C883) specific to Pakistani drug-type cannabis. Furthermore, the predicted protein sequence of the isolated sequences also showed two amino acid substitutions (D284→V284 and T295→P295) corresponding to the identified SNPs. However, the homology based three dimensional models of the inferred proteins generated via Swiss-Model-an automated online server did not project any changes at the active sites of the enzyme (THCA) due to D→V and T→P substitutions. The two missense mutations uncovered as a result of this study may assist in distinguishing the products of Pakistani cannabis among the smuggled materials.

Keywords: Cannabis, CBGA, FAD, SNPs, THCA synthase

INTRODUCTION

Pakistan is known for being an eminent representative of the world’s oldest civilization “the Indus valley”. In the present research we have highlighted the specificity of Pakistani gene pool of BHANG (Cannabis sativa L.), which happens to be an ancient crop grown in Indus valley, greatly merited for the oil and fiber obtained from it. Moreover, drugs like marijuana and hashish are prepared from the inflorescence and trichomes of the dried hemp. A large number of countries have implemented stringent laws to prevent farming, importation and distribution of such drug producing crops. Unfortunately, the ‘big road’ of drug trafficking passes through Pakistan and with forensics point of view it is very important for Pakistan to have a clear position in this regard so as to facilitate the detection of illegal narcotic substances in the detained material.

Among 113 varying cannabinoids identified in hemp plant, tetrahydrocannabinol (THC) is the chief psychoactive ingredient which activates cannabinoid receptors in the human body. Therefore, the addiction-causing stupefactive cannabis contains enormous amounts of THC and is referred to as “drug-type”. Conversely, the fiber-rich hemp plant possessing negligible or zero levels of THC is termed as “fiber-type”.

* Author for correspondence: Mehwish Khan, School of Biological Sciences, Quaid-e-Azam campus, University of the Punjab, Lahore, Pakistan, 54000. Email - Mehwish_24_khan@hotmail.com.
Since no considerable variations in the morphology of the afore-mentioned categories of the plant have been documented, a number of molecular tools have been utilized for the characterization of cannabis (Quimby et al., 1973; Small and Cronquist, 1976; Rowan and Fairbairn, 1977; Baker et al., 1982) including Inter-Simple Sequence Repeat amplification (ISSR) (Kojoma et al., 2002), Short Tandem Repeat microsatellite (STR) (Hsieh et al., 2003; Alghanim and Almirall, 2003), Amplified Fragment Length Polymorphism (AFLP) (Miller et al., 2003), 5S-rRNA gene spacer region (Miyahara et al., 1998) and Random Amplified Polymorphism (RAPD) (Gillan et al., 1995; Jagadish et al., 1996; Gigliano et al., 1995; de Meijer et al., 2003). The machinery involved in the biosynthesis of Δ-9-tetrahydrocannabinolic acid (THCA) was unraveled in 1995 by Taura et al. In addition, the discovery of THCA synthase gene has paved the way for detailed analysis of the narcotic material in cannabis (Sirikantaramas et al., 2004).

In the present study, we used PCR to analyze variations in the cannabis isolates of specific localities (prone for illicit drug, the cannabis usage) of the Punjab, Pakistan. PCR amplification of THCA synthase gene or part of the gene was performed and the nucleotide sequences of the resulting amplicons were studied to determine genetic variations in the gene. Bioinformatic study of the available data confirmed the presence of these novel variations in the putative translated sequences of the gene.

**MATERIALS AND METHODS**

**Plant materials.** We analyzed 5 different localities for their notoriety for Bhang (cannabis) usage. One representative plant of each specific locality i.e. O (Okara), S (Sahiwal), K (Kasur), L1 (Lahore1) and L4 (Lahore 2), has been reported here (the samples are preserved at SBS herbarium under the names SBS-Pak-O-01, SBS-Pak-S-02, SBS-Pak-K-03, SBS-Pak-L1-04, SBS-Pak-L4-05, respectively). All plants used in this study were wild, the effect of environmental differences on plant growth and cannabinoid content is although an unavoidable factor but by collecting plant material from regions of Punjab, Pakistan holding relatively similar edaphic and ecological characteristics the said concern was thought to have addressed.

**DNA isolation and quantification.** Isolation of genomic DNA from 0.2 g of freshly excised 10 weeks old cannabis plant leaves was performed by a slightly improved cetyltrimethylammonium bromide (CTAB) procedure (Gilmore et al., 2003). The quantity and quality of DNA samples were analyzed using agarose gel electrophoresis. Finally, 30 ng/µl concentration of DNA was achieved by diluting the samples using nuclease free water.

**Analysis of the sequencing result.** The amplification of THCA synthase marker linked with drug-type cannabis was carried out using previously published primers (i.e. Forward primer Mh-F, 5'-AAT AAC TCC CAT ATC CAA GCA -3' and reverse primer Mh-R, 5'-AGG ACT CGC ATG ATTAGT TT-3') (Figure 1).

![Figure 1. THCA synthase gene sequence of Cannabis sativa. Primers Mh-F and Mh-R are represented by arrows facing right and left, respectively. Red highlighted sequence is the amplified region](https://link.to/image.png)

In order to perform PCR amplifications, 50 µl of reaction mixtures were prepared using 60 ng of DNA sample, 0.2 mM dNTPs, 10 µM of each primer, 2.5 U of TrueStart™ Hot Start Taq DNA Polymerase (Thermo Scientific™#EP0613), 1.5
mM MgCl₂ and 10X Hot Start Taq buffer. The cycling conditions were programmed as: 15 min of initial denaturation at 95°C, followed by 35 consecutive cycles of denaturation (1 min; 94°C), annealing (45 sec; 55°C) and elongation (1 min; 72°C) each, ending in final elongation at 72°C for 10 min. Electrophoretic analysis of the amplicons was carried out using 1% agarose gel. The product size was 1.2 kb.

**Sequencing.** The amplification products of 1.2 kb were gel extracted using GeneJET Gel Extraction Kit (Thermo Scientific™#K0691). After purification, the amplified DNA segments were ligated with pTZ57R/T vector from InstAalone PCR Cloning Kit (Thermo Scientific™#K1213). The resulting recombinant plasmids were transformed into E. coli (DH5α) cells and cloned using selective media. Finally, the cloned samples were sequenced from the sequencing lab of School of Biological Sciences. Sequence alignment of our representative stocks (O, S, K, L1 and L4) was performed compared with reference sequences of published fiber and drug types (Kojoma et al., 2006) using different software i.e., DNAMAN, DNASTAR and Clustal W. Reproducibility of the results was confirmed by repeating each step at least three times.

**Bioinformatic analysis.** Putative protein sequences of isolated THCA synthase genes (O, S, K, L1 and L4) as well as those of reported Drug-type (DT) and Fiber-type (FT) THC-synthase genes were predicted using an online computational translation program called ExPASy-translate tool. Multiple sequence alignment of hypothetical protein sequences was performed with the help of DNASTAR software. An online protein-modeling software Swiss-Model was utilized to facilitate three-dimensional homology modeling of the predicted proteins based on X-ray structures of the given template (3VTE). Since, complete sequences of DT, FT and partial sequences of the isolated samples were available; in order to model a complete structure, the missing flanking nucleotides were added to each of these marker sequences from a known sequence (accession # AB212834). The detailed secondary and 3D structures of the modeled proteins were analyzed with software PDBeSum, UCSF chimera 1.11.2 and Discovery Studio V3.1.

**RESULTS AND DISCUSSION**

**Sequence analysis of THCA synthase marker in Pakistani cannabis.** Kojoma et al. (2006) analyzed cannabinoids (THCA and CBDA) by HPLC in mature leaves of *C. sativa* strains. On the basis of THCA content, cannabis strains have been divided into two groups: drug type (with high THCA content) and fiber type (with low or no THCA content). Furthermore, they were also successful in identification of a THCA synthase sequence-based DNA marker exclusively found in drug type strains. In case of the present study, the specific THCA-synthase marker (1.2 kb) was retrieved from all the DNA samples of Pakistani cannabis (O, S, K, L1 and L4); thereby confirming all five to be drug-type in nature. The PCR products were sequenced and accessioned effectively (LC378399, LC378398, LC378397, LN998182 and LC378400).

In the previous studies it has been noted that there exist some significant single nucleotide polymorphisms, (a total number of 62 SNPs) between the drug type and fiber type THCA synthase gene sequences (Kojoma et al., 2006). Similarly, multiple sequence alignment of the marker sequences understudy and the well-recognized THCA synthase gene sequences belonging to exotic cannabis plants (Czech, Iran, Mexico, Japan, France, Germany Hungary, Italy and Poland), also unraveled the presence of those landmark SNPs in our isolates which are specific to drug-type strains (Figure 2). However, two nucleotide substitutions were exclusive to Pakistani isolates, i.e. A at positions 851 and A at position 883 were replaced with T and C, respectively (Figure 2). The identified markers were about 99.8% identical to the drug-type THCA synthase genes and 95.8% to that of fiber-type strains. Interestingly, the same percentage of homology (approximately 96%) occurs between the markers obtained from Pakistani isolates and the fiber type THCA synthase gene sequences representing varying populations of the world. Moreover, the sequences of the individual markers obtained from the five cannabis isolates of Pakistani origin were 100% identical to one and other (O, S, K, L1 and L4). The phylogenetic dendrogram constructed on the basis of multiple sequence alignment of all the sequences revealed
that the sequences diverged into two distinct
groups at the genetic distance of 2.1. Group I
consisted of all the previously documented drug
type THCA synthase genes as well as markers of
Pakistani isolated, while group II comprised of all
the known fiber type THCA synthase genes
(Figure 3). All these findings reestablished the
association of the observed markers with THCA
synthase gene of drug type cannabis.

**Figure 2.** Alignment of the ‘drug type’ THCA synthase marker sequences isolated from Pakistani cannabis with the reported coding region of THCA synthase gene belonging to different areas of the world. ‘DT’ represents reported ‘drug type’; ‘FT’ indicates reported ‘fiber type’ gene sequences; and sequences from K to S represent THCA marker sequences of Pakistani cannabis. The red circles highlight the two SNPs (A→T and A→C) in the sequence of Pakistani cannabis.

**Figure 3.** Phylogenetic tree based on multiple sequence alignment of isolated and reported THCA synthase gene sequence. The two main groups diverged at 2.1 genetic distance; all the known ‘fiber type’ related sequences (FT) constitute one cluster; while the ‘drug type’ related sequences (DT) make a separate cluster including the THCA marker sequences isolated from Pakistani cannabis (O, S, L4, L1 & K).

**Prediction of THCA protein sequence.** After the discovery of two single nucleotide substitutions in THCA synthase fragment of Pakistani cannabis, it was imperative to find out if these mutations are merely silent or they do have a considerable impact on the sequence, structure and function of the translated protein (Raamsdonk et al., 2001). The task at hand was achieved by translating and analyzing the sequences using software designated for each purpose. Since the marker sequences (O, S, K, L1 and L4) being studied are 100% identical to one another, only one isolate L1 was selected as a representative sequence for Pakistani gene pool. Likewise, one representative of each known drug type and fiber type THCA synthase gene was picked to perform the comparative study of the predicted amino acid sequences. The original number of the amino acids constituting THCA synthase is 545; however, none of the translated proteins except that of reported drug type and

Identification of SNPs in THCA synthase gene in Cannabis

fber type gave complete sequence of THCA synthase (409 amino acids) as the amplified fragments were partial gene sequences (Sirikantaramas et al., 2004). Twenty-seven of the amino acid substitutions detected formerly in fiber type plants were absent in Pakistani hemp, hence reinstating the assumption that the obtained isolates were in fact drug type (Kojoma et al., 2006). It was also noticed that the two nucleotide substitutions specific to Pakistani drug type cannabis encoded amino acids different from that of published THCA synthase sequences i.e., valine (V) was replaced with aspartic acid (D) at position 284 and threonine (T) at position 295 was substituted by proline (P) (Figure 4). Thus, the two mutations are considered non-synonymous which means that the change in a single amino acid is a direct result of a single nucleotide substitution (Kumar et al., 2009).

In order to ascertain whether the identified SNPs in our isolated marker lied in the conserved domains of the protein or not, the sequences were analyzed using NCBI’s conserved domain (CDD) search (Marchler-Bauer et al., 2014). Consequently, two main domains were recognized in the THCA synthase protein, i.e. FAD-binding domain and berberine & berberine like domain spanning amino acids 81 to 218 and 480 to 538, respectively (Figure 5) (Daniel et al., 2017). Similar domains were discerned in the predicted THCA synthase protein of Pakistani cannabis (Figure 5). Based on this outcome, it became clear that the two mutations (Val284 and Pro295) indicated in the current study were not located inside the highly conserved regions of the protein sequence.

Figure 4. Conserved domain database search for one representative of predicted THCA synthase protein sequence of known drug-type, known fiber-type and Pakistani cannabis (L1) each. ‘FAD binding domain’ is indicated in drug type and Pakistani cannabis proteins starting from position 81 and ending at position 218.

Comparison of predicted 3D structures of the obtained sequences. The resultant models (DT, FT and L1) were assessed for their authenticity and quality by comparing them with the template (3VTE) structure. Each of the modeled targets, i.e. DT (Accession # AB212834), FT (Accession # AB212830) and L1 (Accession # LN998182) are presented in the current study (Figure 6, Fig. 3a, 4a, and 5a; Supplementary Material). According to the results, the error values for residues turned out to be quite insignificant as determined by ERRAT; the quality factor for the generated structures (93.5 to 93.7) was very much analogous to that of the template (92.81). Similar quality factor of ERRAT was noted by El Alaoui et al. (2014) in 3VTE (Table 1). In addition, the z-score of the THCA synthase proteins under scrutiny laid further emphasis on the proximity of these models with the scientifically proven X-ray and NMR structures of similar sized proteins (Biasini et al., 2014). The structures understudy had z-score values (-10.51 to -10.28) which were
Identification of SNPs in THCA synthase gene in Cannabis

The residues at the active-site of the computed THCA synthase models were determined adequately; His114 and Cys176 seemed to play the most predominant role in FAD binding (Figure 8), while Tyr175, Tyr417 and Tyr484 were supposedly the main participants in the binding of CBGA (Figure 9). Shoyama et al. (2012) and El Alaoui et al. (2014) had reached the same conclusion regarding the catalysis of THCA.

Figure 5: Predicted 3D structure of the putative THCA synthase protein L1 (accession # LN998182) identified in Pakistani cannabis, modeled using Swiss Model and observed with UCSF Chimera. The blue molecule shows the ligand FAD; the green amino acids (Val284 & Pro295) are the substitutions observed only in Pakistani cannabis; and the yellow residues indicate the key amino acids at active site of the enzyme i.e. Tyr175, Cys176, Tyr484 & His114.

Figure 6. Secondary structure alignment of the modeled Pakistani protein ‘L1’ with the modeled known ‘drug type’ and ‘fiber type’ THCA synthase as well as template ‘3VTE’. The dark blue color highlights the preserved amino acids in all sequences; orange bands represent α-helices; and β-sheets are given as blue pointers.
Identification of SNPs in THCA synthase gene in Cannabis

Table 1. General attributes of the modelled protein structures and RMSD values of the models aligned against template 3VTE.

<table>
<thead>
<tr>
<th>Protein</th>
<th>ID</th>
<th>Z-score</th>
<th>Verify %</th>
<th>ERRAT</th>
<th>RMSD A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template THCA synthase</td>
<td>PDB: 3VTE</td>
<td>-9.26</td>
<td>95.23</td>
<td>92.813</td>
<td>-</td>
</tr>
<tr>
<td>Drug-type THCA synthase (DT)</td>
<td>Accession: AB212834</td>
<td>-10.51</td>
<td>95.56</td>
<td>93.491</td>
<td>0.147</td>
</tr>
<tr>
<td>Fiber-type THCA synthase (FT)</td>
<td>Accession: AB212830</td>
<td>-10.28</td>
<td>97.88</td>
<td>93.700</td>
<td>0.127</td>
</tr>
<tr>
<td>Pakistani THCA synthase (L1)</td>
<td>Accession: LN998182</td>
<td>-10.37</td>
<td>95.56</td>
<td>93.491</td>
<td>0.146</td>
</tr>
</tbody>
</table>

Table 2. Ramachandran plot data computed by RAMPAGE, a good quality model would be expected to have over 90% in the most favoured regions.

<table>
<thead>
<tr>
<th>Protein</th>
<th>ID</th>
<th>Residues in favoured region</th>
<th>Residues in allowed region</th>
<th>Residues in outlier region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template THCA synthase</td>
<td>PDB: 3VTE</td>
<td>92.5%</td>
<td>7.1%</td>
<td>0.4%</td>
</tr>
<tr>
<td>Drug-type THCA synthase (DT)</td>
<td>Accession: AB212834</td>
<td>93.8%</td>
<td>5.2%</td>
<td>1%</td>
</tr>
<tr>
<td>Fiber-type THCA synthase (FT)</td>
<td>Accession: AB212830</td>
<td>93.2%</td>
<td>6.0%</td>
<td>0.8%</td>
</tr>
<tr>
<td>Pakistani THCA synthase (L1)</td>
<td>Accession: LN998182</td>
<td>93.8%</td>
<td>5.2%</td>
<td>1%</td>
</tr>
</tbody>
</table>

CONCLUSION

In this study, we isolated a DNA marker linked with THCA synthase of drug type cannabis from 5 isolates of Pakistan. The DNA sequence of the obtained marker was compared with the reported sequences of the enzyme (THCA synthase) belonging to different regions of the world. It was found that the identified sequences were highly identical (99.8%) to the sequence of drug type THCA synthase with the exception of only 2 nucleotides which were specific to only Pakistani cannabis. Moreover, the protein sequence...
prediction of the sequences showed that the 2 variations also got translated into two amino acid substitutions. However, the changes in amino acid sequence did not apparently affect the structure of the protein, nor did it influence its affinity to the substrate (CBGA) as revealed by the homology modeling of the putative protein sequences. The analysis of FAD/CBGA binding pocket also showed the presence of similar residues at active site as indicated in earlier studies. The detection of two nucleotide mutations as well as two amino acid substitutions in ‘drug type’ Pakistani cannabis will help point to the identification of cannabis plant being illegally transported from Pakistan to other countries and will greatly aid forensic studies.

ACKNOWLEDGEMENT

We are grateful to all the members of Plant Biotechnology lab of School of biological sciences for helping us out in every way possible.

REFERENCES


