DATA NOTE

In silico discovery of terpenoid metabolism in Cannabis sativa
[version 1; peer review: 2 approved, 1 approved with reservations]

Luca Massimino
Molecular Oncology Unit, San Gerardo Hospital, Monza, Italy

Abstract
Due to their efficacy, cannabis based therapies are currently being prescribed for the treatment of many different medical conditions. Interestingly, treatments based on the use of cannabis flowers or their derivatives have been shown to be very effective, while therapies based on drugs containing THC alone lack therapeutic value and lead to increased side effects, likely resulting from the absence of other pivotal entourage compounds found in the Phyto-complex. Among these compounds are terpenoids, which are not produced exclusively by cannabis plants, so other plant species must share many of the enzymes involved in their metabolism. In the present work, 23,630 transcripts from the canSat3 reference transcriptome were scanned for evolutionarily conserved protein domains and annotated in accordance with their predicted molecular functions. A total of 215 evolutionarily conserved genes encoding enzymes presumably involved in terpenoid metabolism are described, together with their expression profiles in different cannabis plant tissues at different developmental stages. The resource presented here will aid future investigations on terpenoid metabolism in Cannabis sativa.

Keywords
terpenoid metabolism, cannabis sativa, in silico gene expression, evolutionarily conserved genes

This article is included in the Data: Use and Reuse collection.
Introduction

Due to its astonishing efficacy\(^1\), nowadays cannabis is prescribed by physicians for the treatment of neurological, psychiatric, immunological, cardiovascular, gastrointestinal, and oncological conditions\(^2–7\). Although therapies based on the use of cannabis flowers or their derivatives are recognized to be very effective, treatments centered on drugs containing \(\Delta^2\)-tetrahydrocannabinol (THC) alone lack efficacy and lead to increased side effects\(^8–9\). This discrepancy seems to result from the absence of the synergistic effects of additional pivotal compounds found in the Phyto-complex, the so-called entourage effect\(^9\). Among these molecules are other cannabinoids and terpenoids, which are thought to play major roles in the modulation of THC\(^10\).

Terpenes are small hydrocarbon (isoprenoid) molecules classified as either as monoterpenes, sesquiterpenes, diterpenes or carotenoids, depending on the number of isoprene units (C2) used to synthesize them. Terpenoids are small lipids derived from terpenes, often accompanied by a strong odor useful for the plants to protect themselves against possible predators\(^12\). Terpenoids not only have important functions when working in concert with cannabinoids; they have been widely investigated in many different plant species and are being exploited as anti-fungal, anti-bacterial, anti-oxidant, anti-inflammatory, anti-stress, anti-cancer and analgesic agents\(^15–18\). However, whilst the gene networks controlling the biosynthesis of cannabinoids and their precursors have been extensively studied\(^19–22\), the biosynthetic pathway of terpenoid molecules in Cannabis sativa is only recently being elucidated. Only two genes have been characterized, one encoding (-)-limonene synthase, the other (+)-\(\alpha\)-pinene synthase\(^23\), two enzymes responsible for the conversion of geranyl pyrophosphate into limonene and pinene, respectively\(^21,25\). Remarkably, while cannabinoids are only found in cannabis plants, terpenoids are also produced by a variety of other plant species, so they must share many of the enzymes involved in their metabolism.

In the present work, evolutionarily conserved genes encoding enzymes predicted to be involved in terpenoid metabolism have been identified within the transcripts of the canSat3 reference transcriptome of Cannabis sativa\(^21\). Moreover, by taking advantage of available gene expression data\(^2\), gene expression profiling of these enzymes was performed in cannabis plant tissue at different developmental stages. The data noted presented here will provide researchers with a corollary of candidate genes that will considerably accelerate future investigations on terpenoid metabolism in Cannabis sativa.

Material and methods

Analysis of evolutionarily conserved transcripts

Cannabis sativa transcript sequences (n=23,630) taken from the canSat3 genome assembly (http://genome.ccbr.utoronto.ca/)\(^21\) were annotated with Blast2GO 4.0.7\(^27\) using NCBI blastx and InterProScan databases. Terpene metabolism related genes were selected if found to be present in datasets downstream of the “terpene metabolic process” gene ontology category from the AmiGO 2 repository (GO:0043692), including the carotene metabolic process (GO:0016119), ent-kaurene metabolic process (GO:0033331), ent-pimara-8(14),15-diene metabolic process (GO:1901539), isoprene metabolic process (GO:0043611), miltiradiene metabolic process (GO:0043692), sesquiterpene metabolic process (GO:1903192), monoterpene metabolic process (GO:1901944), monoterpenoid metabolic process (GO:1903192), sesquiterpene metabolic process (GO:0051761), terpene biosynthetic process (GO:0046246), and terpene catabolic process (GO:0046247)\(^23\).

Gene expression analysis

Gene expression profiles from cannabis plant tissue at different developmental stages were downloaded from the NCBI GEO repository (https://www.ncbi.nlm.nih.gov/geo/). Gene expression heatmaps and unsupervised hierarchical clustering were performed with GENE-E 3.0.213\(^29\).
To identify the genes presumably playing a role in terpenoid metabolism, annotated transcripts were filtered for gene ontology (GO) categories involved in terpene biosynthesis and catabolism using the AmiGO 2 reference database. A total of 288 transcripts representing 215 different genes were predicted to be involved in the metabolism of bisabolene, cadinene, carotene, copaene, ent-kaurene, farnesol, geraniol, germacrene, lycopene, limonene, myrcene, phytoene, pinene, squalene, and others (Supplementary table 1). Functional characterization of this subset confirmed an enrichment for GO categories involved in different terpene biosynthetic and catabolic processes (Figure 2).

Terpenoids are produced by several plant species and in several types of plant tissue as defense against predators. Similar to other

![Figure 2](image-url)

**Figure 2.** Gene ontology analysis of putative terpenoid metabolism related transcripts. Functional enrichment analysis of putative terpenoid metabolism related transcripts taken from the canSat3 reference genome. Enrichment for terpene biosynthesis and catabolism is shown for Biological Process (A) and Molecular function (B) Gene ontology categories.
biological compounds, their abundance directly correlates with the expression levels of the enzymes involved in their metabolism. To this end, gene expression analysis of genes likely to be involved in terpenoid metabolism was performed using previously published datasets (Figure 3; Supplementary table 2). Notably, unsupervised hierarchical clustering identified four gene clusters. Cluster 1 genes display high expression in roots and stems; cluster 3 genes in hemp flowers; cluster 4 genes in leaves and flowers. Cluster 2 genes

Figure 3. Gene expression analysis of terpenoid metabolism related genes. Heatmap showing relative expression values (log2 RPKM) of putative terpenoid metabolism related genes from cannabis plant tissue taken at different developmental stages (shoot, root, stem, young and mature leaf, early-, mid- and mature-stage flower). Five gene clusters were defined in accordance to unsupervised hierarchical clustering.
were constitutively expressed in all tissues. These results highlight which enzymes are expressed by specific tissues and will provide a strong rationale for further investigations on the molecular basis of terpenoid metabolism.

**Discussion**

The active principles inside plants have been exploited by humans for centuries, with *Cannabis sativa* being one of the oldest ever used for medicinal purposes. Surprisingly, contrary to whole plant extracts, medicinal products containing exclusively THC have been found to lack efficacy and lead to unbearable side effects. These results arise from the fact that these products lack other important co-factors typically found in the Phyto-complex, such as terpenoids and other cannabinoids that contribute to the synergistic effects seen with whole plant extracts.

While genes involved in cannabinoid biosynthesis have been widely investigated, the gene network controlling terpenoid metabolism is only recently being elucidated, with genes encoding (-)-limonene synthase and (±)-α-pinene synthase being the only two characterized. To this end, *Cannabis sativa* transcripts have been scanned for evolutionarily conserved protein domains and annotated according to their presumptive molecular function. As a result, 215 evolutionarily conserved genes were predicted to be involved in terpenoid metabolism. Furthermore, *in silico* gene expression profiling of these enzymes in cannabis plant tissue at different developmental stages highlighted different gene clusters with peculiar expression patterns. For instance, cluster 3 genes displayed high expression specifically in hemp flowers, which could be of great interest as different cannabis strains harbor different entourage effects.

Since the current cannabis reference transcriptome is still at preliminary stages, it is very likely that false negatives have caused important transcripts to still be missing. For example, the two genes encoding for (-)-limonene synthase and (±)-α-pinene synthase align on the same transcript predicted to encode for Myrcene synthase (PK25781.1 in Supplementary table 1), and therefore cannot be discriminated. Unfortunately, to overcome this issue at whole genome level we need the complete version of the reference transcriptome to be available. Until that time, researchers are forced to validate single transcripts with classic low-throughput technology, such as molecular cloning followed by Sanger sequencing.

Nevertheless, the data presented here will ease future investigations on terpenoid metabolism in *Cannabis sativa* by providing researchers with a collection of candidate genes. For instance, one of these genes was predicted to encode for β-bisabolene synthase (PK05069.1 in Supplementary table 1). Bisabolene is being used as an antimicrobial agent, as well as a biofuel. However, prior to this report nothing was known about the gene network controlling its metabolism in *Cannabis sativa*. As soon as future studies will integrate gene expression data with chemical analysis, the complete molecular scenario underlying terpenoid metabolism will be revealed.

**Data availability**

Processed gene expression data can be found in the NCBI GEO repository (https://www.ncbi.nlm.nih.gov/geo/) with accession number GSE93201.

**Competing interests**

No competing interests were disclosed.

**Grant information**

The author(s) declared that no grants were involved in supporting this work.

**Supplementary material**

**Supplementary table 1. Evolutionarily conserved terpenoid metabolism transcripts**

List of putative terpenoid metabolism genes obtained with Blast2GO.

Click here to access the data.

**Supplementary table 2. Terpenoid metabolism gene profiling in different tissues and developmental stages**

Gene expression matrix of predicted terpenoid metabolism genes. Expression units are expressed in RPKM.

Click here to access the data.

**References**


Meirong Jia
Roy J. Carver Department of Biochemistry, Biophysics & Molecular Biology, Iowa State University, Ames, IA, USA

The bioinformatics study here has reported 215 genes that are putatively involved in terpenoid biosynthesis in Cannabis sativa using available genome and transcriptome resources. This knowledge would predictably accelerate the terpenoid investigations in this specific species, or even broader. The research approaches are clear and sound. The writing is generally logical and organized. So I would recommend the publication of this work as a data note.

I would suggest the author to make the following minor changes.

1. The author might wish to clarify throughout the text as to the original “transcriptome” (or “genome”) resources they have used for initial study. For instance, adding the specific link of that file name.
2. In the caption of figure 3, the author claimed “five gene clusters were defined” while in the text, it was “four gene clusters”, the author needs to correct this. Also, it is confused here that some genes clustered are expressed in “Finola” because from the previous text in “introduction” “para 3”, it was said the study was conducted “within the transcripts of the canSat3 reference transcriptome of Cannabis sativa21”. Please clarify this.
3. The 215 gene candidates for terpenoids have been further predicted to express distinctively in tissues and developmental stages. While it is not easy to specifically check the expression pattern of each gene from the currently provided data. It would be great if possible that the author could add this piece of information, which would be valuable for future characterization of these genes. An example study showing the expression pattern of terpene genes would be found in the paper by Wang, et al. 20161.

References

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Partly

**Competing Interests:** No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 09 June 2017

https://doi.org/10.5256/f1000research.11622.r23373

© 2017 Page J et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

 Jonathan E Page  
 Botany Department, University of British Columbia, Vancouver, BC, Canada

 Judith K. Booth  
 Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada

The present study is appropriate as a data note, and may provide a useful shortcut for future studies on terpenoid biosynthesis in cannabis. The author identifies transcripts putatively involved in terpenoid biosynthesis, including processes of both primary and specialized metabolism. The rationale is well explained, and the methods are technically sound.

Major changes:

- More information is required in the Materials and Methods for the study to be replicable. Two versions of the canSat3 transcriptome are available, and the authors should specify whether they used the full or representative transcript set. The authors must also provide the NCBI accession numbers for the GEO dataset they used.
- There is an unsupported statement in the results section, which reads: “Similar to other biological compounds, their abundance directly correlates with the expression levels of the enzymes involved in their metabolism.” In a complex system, this is not obvious and requires textual support.

Minor changes:
The Introduction generally provides a good rationale for the study. However, the definitions provided for ‘terpene’ and ‘terpenoid’ in the second sentence of the second paragraph do not reflect the definitions used in reference 12, nor are they the traditional definitions for those terms. References 24 and 25 address enzymes characterized in other organisms, and so do not support the statement regarding their biosynthetic activities in cannabis. That information is provided in reference 23. The final statement of the second introductory paragraph is unsupported and requires citation. Figure 2 caption states that the transcripts were taken from the reference genome, but the Materials and Methods indicate that a transcriptome was used. Figure 3 caption insufficiently explains x axis labels. The caption should define ‘PK’ and ‘Finola’. Otherwise, this figure shows an interesting and valuable result.

Readers should note the specific products of terpene synthases and other enzymes in specialized metabolism are currently difficult to predict using in silico methods. A recent paper (Booth et al., 2017) includes biochemical and phylogenetic analysis of some of the candidate genes highlighted here, and may be of interest to readers.

References

Is the rationale for creating the dataset(s) clearly described?
Partly

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Plant metabolism

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.
Akan Das  
Department of Applied Biology, University of Science and Technology, Meghalaya (USTM), Ri-Bhoi, Meghalaya, India

In the present work, 23,630 transcripts were scanned for evolutionarily conserved protein domains and annotated using Gene Ontology analysis. A total of 215 evolutionarily conserved genes encoding enzymes presumably involved in terpenoid metabolism are described on the basis of Gene Expression profiles of NCBI GEO repository. The identification of candidate genes on terpenoid metabolism in Cannabis sativa will ease the future investigations on terpenoid metabolism pathways.

It is a small piece of work on bioinformatics analysis of transcripts which present useful information to the plant researcher. Hence, the article can be recommended to be published as a Data Note. I suggest the author make a minor change in the title to be "In silico discovery of terpenoid metabolism associated transcripts in Cannabis sativa" and that the discussion, rather than describing general things, should be modified to discuss the findings in the work.

Is the rationale for creating the dataset(s) clearly described?  
Yes

Are the protocols appropriate and is the work technically sound?  
Yes

Are sufficient details of methods and materials provided to allow replication by others?  
Yes

Are the datasets clearly presented in a useable and accessible format?  
Yes

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

• Your article is published within days, with no editorial bias
• You can publish traditional articles, null/negative results, case reports, data notes and more
• The peer review process is transparent and collaborative
• Your article is indexed in PubMed after passing peer review
• Dedicated customer support at every stage