DATA NOTE

In silico gene expression profiling in *Cannabis sativa* [version 1; referees: 2 approved]

Luca Massimino

Molecular Oncology Unit, San Gerardo Hospital, Monza, Italy

v1 **First published:** 23 Jan 2017, **6**:69 (doi: 10.12688/f1000research.10631.1)
Latest published: 23 Jan 2017, **6**:69 (doi: 10.12688/f1000research.10631.1)

Abstract

The cannabis plant and its active ingredients (i.e., cannabinoids and terpenoids) have been socially stigmatized for half a century. Luckily, with more than 430,000 published scientific papers and about 600 ongoing and completed clinical trials, nowadays cannabis is employed for the treatment of many different medical conditions. Nevertheless, even if a large amount of high-throughput functional genomic data exists, most researchers feature a strong background in molecular biology but lack advanced bioinformatics skills. In this work, publicly available gene expression datasets have been analyzed giving rise to a total of 40,224 gene expression profiles taken from cannabis plant tissue at different developmental stages. The resource presented here will provide researchers with a starting point for future investigations with *Cannabis sativa*.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 23 Jan 2017	✓ report	✓ report

1 **Gea Guerriero**, Luxembourg Institute of Science and Technology (LIST) Luxembourg

2 **Sergio Esposito**, University of Naples Federico II Italy

Discuss this article

Comments (0)

Corresponding author: Luca Massimino (luca.massimino@unimib.it)

How to cite this article: Massimino L. ***In silico* gene expression profiling in *Cannabis sativa* [version 1; referees: 2 approved]** *F1000Research* 2017, **6**:69 (doi: 10.12688/f1000research.10631.1)

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

Competing interests: No competing interests were disclosed.

First published: 23 Jan 2017, **6**:69 (doi: 10.12688/f1000research.10631.1)

Introduction

The cannabis plant has been used for medical purposes for centuries, before being socially stigmatized for the last half century[1]. Nevertheless, more than 430,000 published scientific papers exist, with about 25,600 works published in 2016 (https://scholar.google.com/). In addition, there are about 600 ongoing and completed clinical trials involving cannabis (https://www.clinicaltrials.gov/).

The endocannabinoid system is involved in virtually every biological function[2], so it is not surprising that cannabis is being used to treat neurological[3], psychiatric[4], immunological[5], cardiovascular[6], gastrointestinal[7], and oncological[8] conditions.

Today, a large amount of high-throughput functional genomic data exists. Nonetheless, even in the era of 'omics, the great majority of researchers feature a strong background in molecular biology but lack advanced bioinformatics skills[9].

In the present work, publicly available gene expression data taken from cannabis plant tissue at different developmental stages (shoot, root, stem, young and mature leaf, early-, mid- and mature-stage flower) have been analyzed, giving rise to 40,224 gene expression profiles. Moreover, the expression patterns of 23 cannabinoid pathway related genes are described. The data note provided here will aid future studies by providing researchers with a powerful resource for future investigations.

Material and methods

Gene expression analysis

Gene expression datasets were downloaded from the NCBI SRA directory[10] (https://www.ncbi.nlm.nih.gov/sra/) with accession numbers SRP006678 and SRP008673. Raw sequences were mapped to the canSat3 reference genome[11] with TopHat2 v2.1.0[12]. Gene counts and relative transcript levels were obtained with Cufflinks v2.2.1.0[13], and submitted to NCBI GEO (https://www.ncbi.nlm.nih.gov/geo/) with accession number GSE93201. Cannabinoid related genes were found within the canSat3 transcripts with the Cannabis genome browser BLAT web tool[11] (http://genome.ccbr.utoronto.ca/cgi-bin/hgBlat?command=start). Gene expression heatmaps and unsupervised hierarchical clustering were carried out with GENE-E[14].

Results

The *Cannabis sativa* reference genome and transcriptome have been published, although data analysis is still at the preliminary stages[11]. In other words, we know what the presumptive genes are, but we do not know the chromosomes they are located in, nor their molecular functions. Given that this high-throughput gene expression data is publicly available, expression analysis of these yet unidentified genes can be performed. To this end, public repositories have been surveyed for transcriptional profiling datasets derived from *Cannabis sativa*. In total, 31 RNA-seq datasets derived from one hemp and two different psychoactive strains (NCBI SRA accession numbers: SRP006678 and SRP008673) of *Cannabis sativa* shoot, root, stem, young and mature leaf, early-, mid- and mature-stage flower have been analyzed. Unsupervised hierarchical clustering of gene expression values revealed six clusters of genes with specific tissue/stage expression (Figure 1). Cluster 1 genes display high expression levels in shoots, mature leaves, and flowers; cluster 2 genes in leaves and flowers; cluster 3 genes in roots and stems; cluster 4 genes in roots, stems, and flowers; cluster 5 genes in hemp flowers and cluster 6 genes in shoots, roots, stems, and flowers.

Genes involved in the biosynthesis of cannabinoids and their precursors have been shown to be overexpressed in flowers[15]. To validate gene expression profiling, cannabinoid, hexanoate, 2-C-methyl-D-erythritol 4-phosphate (*MEP*) and geranyl diphosphate (*GPP)* pathway genes[11,16], together with the olivetol synthase (*OLS*) gene[17,18], the (-)-limonene terpene synthase (*TPS*) gene[19] and the polyketide synthase (*PKS*) gene[20], have been analyzed. As expected, most of these genes were overexpressed in flowers, although many of the genes also displayed high expression in other tissues (Figure 2; Supplementary table 1). Interestingly, virtually all of them were highly expressed in the shoot.

Discussion

Today, cannabis and its derivatives are successfully employed for treatment of a large number of different pathological conditions[3,5–8]. Each year, more articles related to cannabis are published, with about 25,600 studies published in 2016 (https://scholar.google.com/). Remarkably, only 3% of these papers (13,300 out of 432,000) also take genomics into consideration, with very few of them directly relating to the genomics of cannabis. This could be due to the fact that, for obvious reasons, most researchers still lack advanced bioinformatics skills and are therefore limited in their research[9].

To this end, a total of 40,224 gene expression profiles taken from cannabis plant tissue at different developmental stages were obtained by exploiting common bioinformatics pipelines[13]. Moreover, expression profiles of the genes belonging to the cannabinoid pathway[11,16–20] are provided.

Even if these data are preliminary, some observations can already be made. For instance, virtually all genes found to be highly expressed in flowers (Figure 1, cluster 1 and Figure 2) also displayed high expression in the shoot. Having had only one sample at this specific developmental stage, these results could be derived from technical issues rather than differences in gene expression. However, not all transcripts (57%) were found to be overexpressed in the shoot, thus pointing toward the possible specificity of these changes. If this is confirmed, it may provide researchers with the possibility to study the molecular function of flower specific genes directly in sprouting plants, without having to wait for the plant to fully bloom.

Cannabis sativa is a versatile plant - it is being used for medical as well as for industrial purposes[21,22]. For this reason, cutting-edge genomics technology is currently being applied either to ameliorate specific phenotypes, or for breeding purposes[22–27]. Cluster 5 genes (Figure 1) seem of great interest in this regard, as they are visibly overexpressed specifically in non-psychoactive cannabis flowers. These genes could be downregulated in hemp in order to create new strains high in cannabidiol (CBD), but with the proper *entourage* effect commonly found in the psychoactive counterparts[28]. On the other hand, hemp specific genes could be

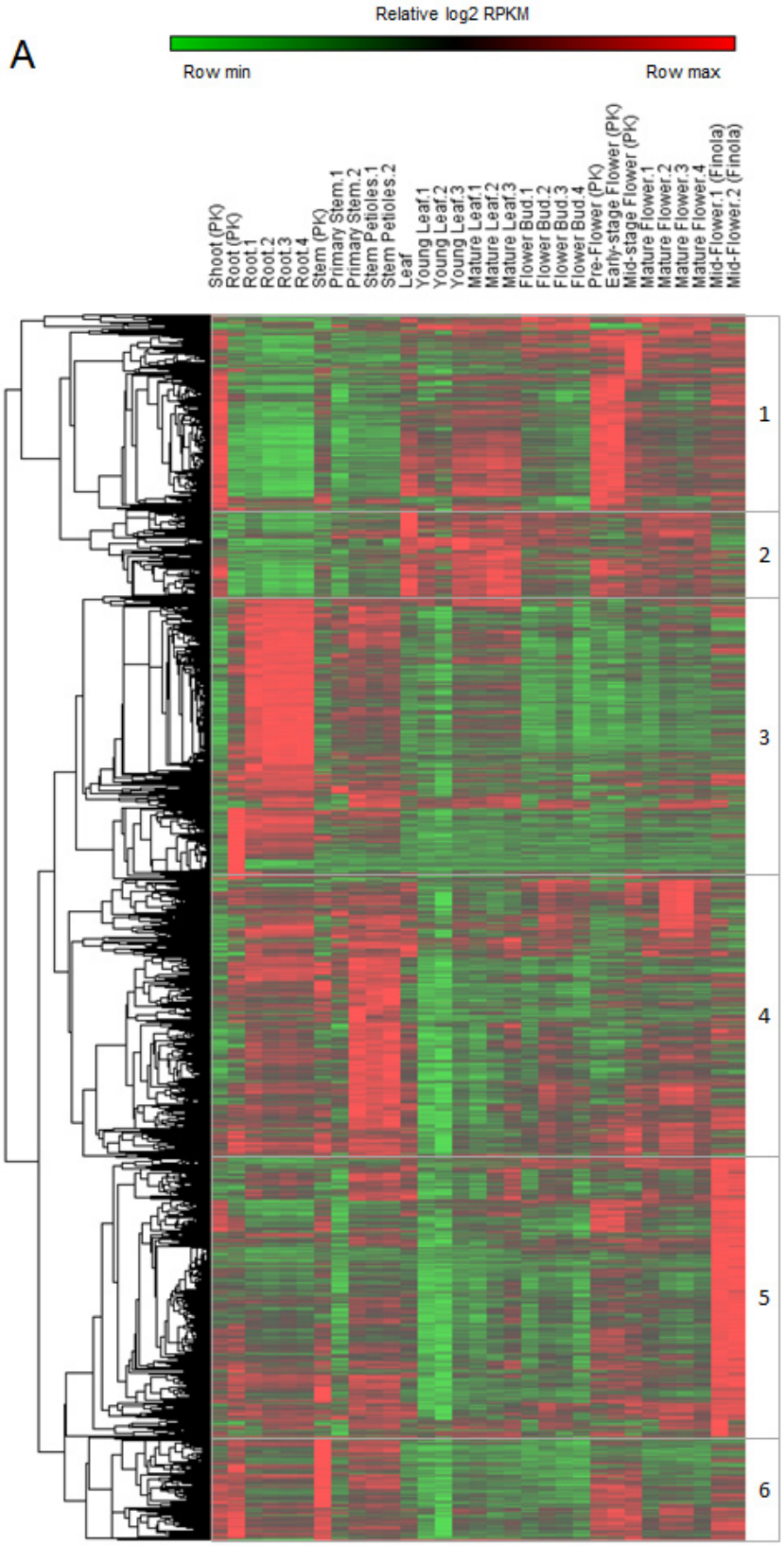

Figure 1. Gene expression profiles taken from cannabis plant tissue at different developmental stages. Heatmap showing relative expression values (log2 RPKM) of the highest expressed genes. Six gene clusters were defined in accordance with the unsupervised hierarchical clustering.

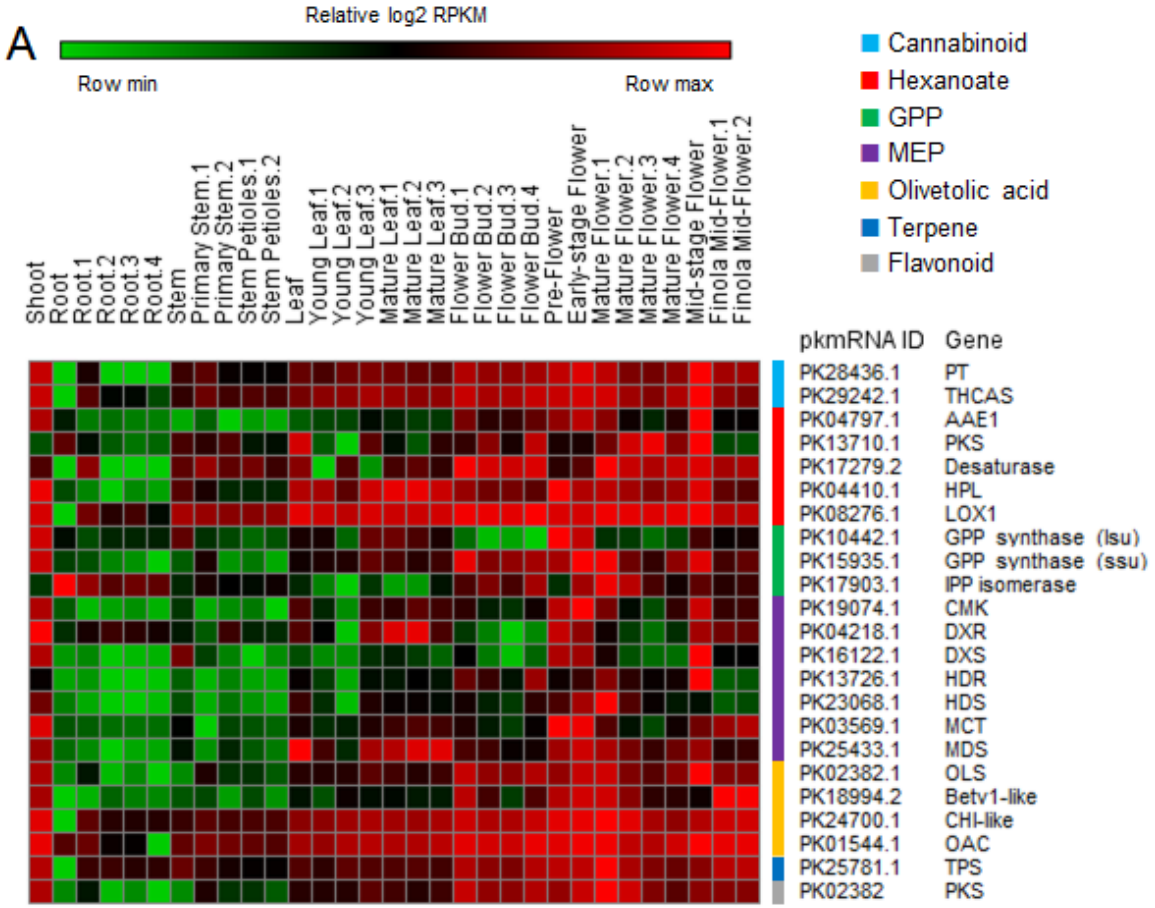

Figure 2. Gene expression analysis of the cannabinoid pathway. Heatmap showing relative expression values (log2 RPKM) of genes belonging to cannabinoid and precursor (hexanoate, GPP, MEP, olivetolic acid) pathways, together with terpene synthase (TPS) and polyketide synthase (PKS).

upregulated in marijuana to produce high fiber/oil containing crops harboring therapeutically valuable active principles within their flowers. One potential candidate is the *Csfad2a* gene which was recently found to be highly expressed only in some hemp strains. Here, high *Csfad2a* expression was correlated with both higher oil content and lower oxidation tendency, eventually leading to the production of a significantly better commercial product[26].

Perhaps the major pitfall of this kind of analysis comes from the fact that although the current cannabis reference genome and transcriptome have been published, data analysis is still at the preliminary stages[11]. Like in other plants, the cannabis genome is highly redundant and difficult to resolve[29]. It is very likely that false negatives have caused important transcripts to still be missing. Nevertheless, these 40,224 gene expression profiles will provide researchers with a valuable resource and important genomic insights for future investigations with *Cannabis sativa*.

Data availability

Raw expression data can be found in the NCBI SRA directory (https://www.ncbi.nlm.nih.gov/sra/) with accession numbers SRP006678 and SRP008673.

Processed 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. Cannabinoid metabolism related gene profiling in different tissues and developmental stages. Gene expression matrix of cannabinoid pathway genes. Expression values are expressed in RPKM.

Click here to access the data.

References

1. Pain S: **A potted history.** *Nature.* 2015; **525**(7570): S10–S11.
PubMed Abstract | Publisher Full Text
2. Di Marzo V, Bifulco M, De Petrocellis L: **The endocannabinoid system and its therapeutic exploitation.** *Nat Rev Drug Discov.* 2004; **3**(9): 771–84.
PubMed Abstract | Publisher Full Text
3. Hosking R, Zajicek J: **Pharmacology: Cannabis in neurology--a potted review.** *Nat Rev Neurol.* 2014; **10**(8): 429–30.
PubMed Abstract | Publisher Full Text
4. Curran HV, Freeman TP, Mokrysz C, *et al.*: **Keep off the grass? Cannabis, cognition and addiction.** *Nat Rev Neurosci.* 2016; **17**(5): 293–306.
PubMed Abstract | Publisher Full Text
5. Klein TW: **Cannabinoid-based drugs as anti-inflammatory therapeutics.** *Nat Rev Immunol.* 2005; **5**(5): 400–11.
PubMed Abstract | Publisher Full Text
6. Di Marzo V, Després JP: **CB1 antagonists for obesity--what lessons have we learned from rimonabant?** *Nat Rev Endocrinol.* 2009; **5**(11): 633–8.
PubMed Abstract | Publisher Full Text
7. Gerich ME, Isfort RW, Brimhall B, *et al.*: **Medical marijuana for digestive disorders: high time to prescribe?** *Am J Gastroenterol.* 2015; **110**(2): 208–14.
PubMed Abstract | Publisher Full Text
8. Swami M: **Cannabis and cancer link.** *Nature Reviews Cancer.* 2009; **9**:148.
Publisher Full Text
9. Chang J: **Core services: Reward bioinformaticians.** *Nature.* 2015; **520**(7546): 151–152.
PubMed Abstract | Publisher Full Text
10. Barrett T, Clark K, Gevorgyan R, *et al.*: **BioProject and BioSample databases at NCBI: Facilitating capture and organization of metadata.** *Nucleic Acids Res.* 2012; **40**(Database issue): D57–63.
PubMed Abstract | Publisher Full Text | Free Full Text
11. van Bakel H, Stout JM, Cote AG, *et al.*: **The draft genome and transcriptome of *Cannabis sativa*.** *Genome Biol.* 2011; **12**(10): R102.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Kim D, Pertea G, Trapnell C, *et al.*: **TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol.* 2013; **14**(4): R36.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Trapnell C, Roberts A, Goff L, *et al.*: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc.* 2012; **7**(3): 562–78.
PubMed Abstract | Publisher Full Text | Free Full Text
14. **GENE-E.** Cambridge (MA): The Broad Institute of MIT and Harvard.
Reference Source
15. Sirikantaramas S, Taura F, Tanaka Y, *et al.*: **Tetrahydrocannabinolic acid synthase, the enzyme controlling marijuana psychoactivity, is secreted into the storage cavity of the glandular trichomes.** *Plant Cell Physiol.* 2005; **46**(9): 1578–82.
PubMed Abstract | Publisher Full Text
16. Stout JM, Boubakir Z, Ambrose SJ, *et al.*: **The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in *Cannabis sativa* trichomes.** *Plant J.* 2012; **71**(3): 353–365.
PubMed Abstract | Publisher Full Text
17. Taura F, Tanaka S, Taguchi C, *et al.*: **Characterization of olivetol synthase, a polyketide synthase putatively involved in cannabinoid biosynthetic pathway.** *FEBS Lett.* 2009; **583**(12): 2061–2066.
PubMed Abstract | Publisher Full Text
18. Gagne SJ, Stout JM, Liu E, *et al.*: **Identification of olivetolic acid cyclase from *Cannabis sativa* reveals a unique catalytic route to plant polyketides.** *Proc Natl Acad Sci U S A.* 2012; **109**(31): 12811–6.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Günnewich N, Page JE, Köllner TG, *et al.*: **Functional expression and characterization of trichome- specific (-)-limonene synthase and (+)-α-pinene synthase from *Cannabis sativa*.** *Nat Prod Commun.* 2007; **2**(3): 223–232.
Reference Source
20. Flores-Sanchez IJ, Linthorst HJ, Verpoorte R: ***In silicio* expression analysis of *PKS* genes isolated from *Cannabis sativa* L.** *Genet Mol Biol.* 2010; **33**(4): 703–13.
PubMed Abstract | Publisher Full Text | Free Full Text
21. de Meijer EP, Hammond KM, Sutton A: **The inheritance of chemical phenotype in *Cannabis sativa* L. (IV): cannabinoid-free plants.** *Euphytica.* 2009; **168**(1): 95–112.
Publisher Full Text
22. Salentijn EM, Zhang Q, Amaducci S, *et al.*: **New developments in fiber hemp (*Cannabis sativa* L.) breeding.** *Ind Crops Prod.* 2015; **68**: 32–41.
Publisher Full Text
23. Mandolino G, Carboni A: **Potential of marker-assisted selection in hemp genetic improvement.** *Euphytica.* 2004; **140**(1): 107–120.
Publisher Full Text
24. van den Broeck HC, Maliepaard C, Ebskamp MJM, *et al.*: **Differential expression of genes involved in C_1 metabolism and lignin biosynthesis in wooden core and bast tissues of fibre hemp (*Cannabis sativa* L.).** *Plant Sci.* 2008; **174**(2): 205–220.
Publisher Full Text
25. Guerriero G, Sergeant K, Hausman JF: **Integrated -omics: A powerful approach to understanding the heterogeneous lignification of fibre crops.** *Int J Mol Sci.* 2013; **14**(6): 10958–10978.
PubMed Abstract | Publisher Full Text
26. Bielecka M, Kaminski F, Adams I, *et al.*: **Targeted mutation of Δ12 and Δ15 desaturase genes in hemp produce major alterations in seed fatty acid composition including a high oleic hemp oil.** *Plant Biotechnol J.* 2014; **12**(5): 613–23.
PubMed Abstract | Publisher Full Text
27. Massimino L: **Cannabis growing meets genomics.** *F1000Research.* 2017; **6**: 15.
28. Russo EB, Taming TH: **potential cannabis synergy and phytocannabinoid-terpenoid entourage effects.** *Br J Pharmacol.* 2011; **163**(7): 1344–1364.
PubMed Abstract | Publisher Full Text
29. Türktaş M, Kurtoğlu KY, Dorado G, *et al.*: **Sequencing of plant genomes - A review.** *Turkish J Agric For.* 2015; **39**: 361–376.
Publisher Full Text

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 02 May 2017

doi:10.5256/f1000research.11455.r20301

✔ **Sergio Esposito**
Dipartimento di Biologia, University of Naples Federico II, Napoli, Italy

The bioinformatic approach presented in this study is potentially highly interesting, providing to scientists a useful data set to investigate the different pathways activated in *Cannabis sativa* at different developmental stages, and in different organs.
I would only ask to the Author if some of the gene sets identified could be further divided into sub-clusters: e.g. there is a “green” area in Fig.1 – cluster 4 in leaves and flower buds (about in the second quarter from above). On the opposite, this area is "red" in cluster 5 – first quarter. Could be these results interpreted under the light of the considerations exposed in the discussion?

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

Referee Report 20 February 2017

doi:10.5256/f1000research.11455.r20299

✔ **Gea Guerriero**
Environmental Research and Innovation (ERIN), Luxembourg Institute of Science and Technology (LIST), L-4362 Esch/Alzette, Luxembourg

The present study suits the Data Note format and can be a useful resource for future studies centered on *Cannabis sativa.* I find the bioinformatics approach sound. I have one suggestion for the author. How about enriching Figure 1 with a representation of GO/pathway enrichment analysis for each cluster (for example with ClueGO in Cytoscape; Bindea et al., 2009[1])?

The interest around this multi-purpose crop is increasing and recently transcriptomics data have been published for a fiber variety too (see for example Behr et al., 2016[2]). The approach described in this note can be applied also to other varieties in the future and/or to other tissue types.

References

1. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, Fridman WH, Pagès F, Trajanoski Z, Galon J: ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks.*Bioinformatics*. 2009; **25** (8): 1091-3 PubMed Abstract | Publisher Full Text
2. Behr M, Legay S, Žižková E, Motyka V, Dobrev PI, Hausman JF, Lutts S, Guerriero G: Studying Secondary Growth and Bast Fiber Development: The Hemp Hypocotyl Peeks behind the Wall.*Front Plant Sci*. 2016; **7**: 1733 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

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