University of Murcia

Omics Data Analysis

Cannabis sativa assembly and Gene discovery

Authors
Alberto Gila Navarro
José Luis Sánchez Fernández

May 2020

Table of contents

1 Introduction

(-)-Δ-9-trans-(6aR,10aR)-tetrahydro-cannabinol (Δ-9-THC). *Cannabis sativa* is a domesticated plant that has been used as a source of fibre (including textile production), oil seeds and has medicinal properties, that have been studied in the last decades[1]. *Cannabis sativa* has been altered by intensive breeding, ending in a wide range of morphological and chemical compositions [2]. The main cannabinoids are novel secondary metabolites discovered in *Cannabis sativa*. Dozens of cannabinoids have been isolated from different parts of the plant (specially flowers)[4].

Cannabidiol (CBD) is a non-psychoactive cannabinoid, naturally found in strains at different expression levels [2]. THC has been studied in the treatment of synthoms in multiple sclerosis [3]. Cannabinoids have been researched for treatment of pain such as neuropathic pain and intractable cancer pain. Cannabidiol shows neuroprotective effects such as THC, inhibits glutamate neurotoxicity, and displays great antioxidant activity [5].

The biosynthesis of cannabinoids occurs mostly in glandular trichomes of female flowers [6]. Synthetic cannabinoids have been discovered in last years. The presence of endogenous cannabinoid molecules and receptors in human body have prompted the use of the term phytocannabinoids to describe these compounds [7].

Taking a look at cannabinoid biosynthesis pathway at Metacyc, it is suggested that the pathway is not fully understood and there are gaps in the route [8]. Olivetolate and geranylpyrophosphate condensation is the first step in the route, but the origin of olivetolic acid remains unknown [6]. After the condensation, olivetolic acid and geranylpyrophosphate prenylation is made by a transferase, producing the intermediate cannabigerolate. The ciclation of that molecule could be performed in different ways, obtaining different cannabinoids such as THC or CBD, in their acid form. A decarboxylation to form the active molecules is required, which is a non-enzymatic reaction [9] The last step (decarboxylation) of this pathway occurs spontaneously in a non-enzymatic reaction during storage or smoking but is also found in vivo [10].

2 Objectives

The objective of this project is to identify and study, via homology, genes related to the production of cannabinoid metabolites in the *Cannabis sativa* genome. To accomplish this, we downloaded from the public repositories of

NCBI SRA and Genome DNA-seq reads in fastq format as well as already assembled genomes. Then we tried several pipelines, listed here to identify candidate genes:

(a) *De novo* assembly using a De Bruijn graph algorithm plus prediction and annotation of genes.

(b) Variant calling using a reference genome of high THC-producing strain of *Cannabis* plus filtering of interesting variants to identify candidate genes.

(c) Like (b) but using a more recent and better assembled genome and double the amount of reads to map.

(d) Like (a) but using an OLC assembly algorithm.

3 Methods

The pipelines were performed in public american and european Galaxy servers.

3.1 Pipeline (a)

The reads dataset under the SRA accesion number ERR3850862 was used, which consist of Illumina pair-end reads of around 150 bp. A long read dataset of the same study, ERR3850930, which contains Oxford Nanopore reads of up to 1Mbp of length was also used.

3.1.1 Fastq quality control

First quality control of all three datasets using FastQC was performed. According to the report of said program, it was deemed necessary to trim the short reads from the 3' end using FastQ Trimmer by Column, and so their length was capped at 125 bp maximum. The long reads, given the intrinsic low quality associated with the technology of nanopores that produce them, were left unprocessed. The reads were preprocessed in the same way in every pipeline.

3.1.2 De Buijn graph assembly

Then the ABySS assembly program was used, that uses a De Bruijn graph algorithm, giving the Illumina reads as input to create the contigs and the Nanopore ones to create the scaffolds.
Unfortunately, the execution failed several times due to memory issues, so we could not continue with this pipeline.

3.2 Pipeline (b)

3.2.1 Alignment with BWA-MEM

The processed reads were aligned to a *Cannabis sativa* Purple Kush cultivar as the reference genome, which differs from our sample in the production of cannabinoids. The quality of the output was filtered using BAM/SAM Filter, with a minimum MAPQ quality score threshold of 50.

3.2.2 Variant calling

After the quality control, variant calling was made with bcftools mpileup and default parameters. A *Cannabis sativa* Purple Kush cultivar was the reference genome. Bcf stats was used after the variant calling to check results. In order to improve the quality of the variants, those with quality equal to zero were discarded using SnpSift Filter. Finally, another bcf stats was performed to check the variants.

3.2.3 Variant selection

We designed a strategy based on filtering variants according to their positions in the genome. We wanted to select those variants that were in a gene of interest, specifically those found in chromosome 9, because it contains the locus of the main enzymes of cannabinoid biosynthesis pathway. The chromosome 9 was extracted from the whole genome using a Perl script called Chromosome extractor.

In order to obtain the location of the genes, a protein prediction was made with Augustus tool. The training set was *Arabidopsis thaliana*. Then, Augustus was run with the chromosome 9 previosly described. Further analysis was performed with InterproScan to obtain all proteins in our chromosome 9. Then, a script (Protein Extractor 2) was performed to obtain the ID of each protein and its GO terms. Another script compared those GO terms with those belonging to enzymes of the cannabinoid biosynthesis pathway.

Unfortunately no variants of quality greater than 0 were found, so we were not able to get any results from this pipeline. The variant selection ended with a list of protein with GO terms and the genomic coordinates, but we could not use them.

3.3 Pipeline (c)

This pipeline was performed in the same was as b), but with the next changes:

1. The processed reads were aligned to a better assembled genome which GenBank assembly accession is GCA_900626175.2, which was used in the variant calling as well.

2. Another pair of short reads were aligned to the reference genome and then merged to the another alignment in order to improve the quality of the alignment.

Unfortunately, there was no results after filtering the variants by quality score.

3.4 Pipeline(d)

We finally decided to give a try to the method of assembly that the authors of the study where the reads come from used. We were reluctant at first because it is an OLC (Overlap-Layout-Consensus) algorithm, which is considerably more inefficient than a De Bruijn algorithm and mostly out of reach for this project given the computational resources available at Galaxy servers. Nevertheless, we tried using only one fifth of the data, the ERR3850930 reads instead of all the Nanopore fastq dataset that the authors employed, and we finally managed to complete the execution of the necessary programs.

3.4.1 OLC assembly

First the aligner Minmap2 was used in the O phase to align every read against each other, giving the same fastq dataset as reference and target.Afterwards, the fast assembler miniasm was executed using as input the original reads and the Minimap2 alignment in paf format. This constitutes the L phase. Next, for the C phase, Minimap2 was run again to align the original reads to the preliminary assembly of miniasm formatted to fasta. Following this, the Racon program was used to create the consensus assembly, giving as input the original reads in fasta format, the preliminary miniasm assembly in SAM

format as well; and the second alignment of Minimap2 in SAM format.

3.4.2 Assembly correction

Additionally, as the authors did, the definitive assembly was further corrected using pilon. To do this step, first an alignment of the short Illumina reads over the consensus algorithm was executed using BWA-MEM. This alignment in BAM format together with the consensus assembly were given as input to pilon. pilon was executed with VCF output disabled and it was specified that the organism is diploid and that the alignment used to correct the assmebly is made of pair-end reads. The option to filter stray reads, that is, not properly mapped reads, was also activated. Unfortunately the program execution did not manage to end after two days running in the server due to memory issues, so we could not use the improvements of pilon.

3.4.3 Assembly quality control

The quality of the final assembly was assessed with Quast and it was compared to the most recent assembly of *Cannabis sativa* found in NCBI Genome, GCA_900626175.1 cs10.

3.4.4 Gene prediction and annotation

After the assembly was done, the program Augustus with the training set of *Arabidopsis thaliana* was run in order to predict genes in the genome. The training set of *A.thaliana* was chosen because it is by far the best annotated plant genome available.
Then the annotation program Interproscan was tried to be executed with the predicted sequences, but to no avail, since it did not manage to end the run in the Galaxy after two days of execution. As an alternative approach, the NCBI Blastp was executed twice restricting the search to *Arabidopsis thaliana* and using exclusively one database at a time to look for matches. The databases used were RefSeq and SwissProt. In order to make up for the capability of Interproscan to annotate with GO terms the predicted proteins of Augustus, the Ensembl BioMart Plant tool was used instead, utilizing the *Arabidopsis thaliana* genome as database. The best matches present in the Blastp report for every predicted protein were used as input for BioMart as a filter and the GO terms associated were selected as annotation output. To extract the accessions numbers of the blastp reports, we used a perl script,

named BlastParser.pl which takes as input a Blastp report in text format returns a two-column text file with the query ID and the best match ID. It can be accessed in the linked GitHub account

3.4.5 Homology study

In order to do this analysis, we made use of two more perl scripts which can also be found in the GitHub account. First, we selected the GO terms of interest based on the ones associated with the known enzymes of the cannabinoid biosynthesis pathway. The metabolic database Biocyc was consulted and later the GO database to extract all the terms related to each enzyme. The script GO_gapdExtractor.pl was then employed to group all the GO terms of the different gapd files in a single and simple text file containing only the proteins' accession numbers and below one GO term per line.
We then extracted the candidate genes for homology study based on the associated GO terms using a simple perl script named ProteinSelector3.pl. The script takes as input the output file from GO_gapdExtractor.pl and the results of BioMart Plant in text format. It then selects any protein that has a GO term in common with one of the ones related to the proteins of the biosynthetic pathway. The output returned by said script is a text file which contains tab separated fields for the protein ID, the unique GO terms it has in common with the cannabinoids' biosynthetic enzymes, the number of these GO terms, and how many GO terms it shares with each enzyme in particular.
After the execution of this script, the output was sorted with a very simple bash script which implements a loop to get the two proteins with most GO terms in common with each enzyme of the pathway. It can also be found in the GitHub account under the name SuperSorter.sh. The selection was done twice, once per database used in the Blastp. The selected proteins were investigated looking the availbale information in the UniProt database. To investigate the ones coming from RefSeq, BioMart Plant was used again to change the identifiers from RefSeq to UniProt.

3.5 External links

Below are the links to the galaxy servers were we tried the different pipelines. This one for pipelines (a) and (d) and this one for pipelines (b) and (c). The fastq trimmer was shared between servers in order to not have to load the short reads twice. The link to the GitHub account is this. And these are the links to the workflows of each pipeline, a, b, c d

4 Results

Only the (d) pipeline gave some results, and therefore are the only ones presented here. Pipelines (b) and (c) did manage to run to completion but no results were obtained.

4.1 Preprocessing

Due to the low quality of ending nucleotides of each read, preprocessing was necessary to improve the quality.

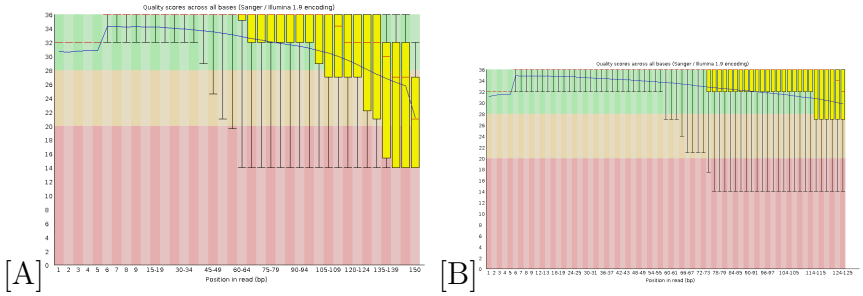

Figure 1: Graphics produced by FASTQC tool. A) Per base sequence quality before preprocessing. B) Per base sequence quality after preprocessing.

4.2 OLC assembly

The OLC pipeline involving Minimap2, miniasm, Minimap2 again and Racon had a runtime of approximately 6 hours. The process that took the longest to complete was Racon. As explained in the Methods sections, the correction phase of pilon was not able to finish. The results of the assembly are summarized in the table below. The whole table can be found in the Galaxy history, which includes more details about unaligned contigs and misassembled ones:

Metric	Value
Contigs	88
Contigs $1000bp$	85
Total length	1977764 bp
Total length $1000bp$	1975490 bp
Reference length	875732047 bp
Largest contig	412867
GC %	35.84
GC % reference	33.41
N50	36368 bp
N75	18983 bp
L50	10
L75	27

Table 1: Assembly metrics.

As can be seen, the assembled genome is far from complete. The reference has less than 0.5% of the total length of the latest *Cannabis sativa* genome. This very bad result can be attributed to the usage of only one Nanopore reads dataset for the assembly pipeline, due to limited computational resources availability. Furthermore, the assembly is highly fragmented, which can be inferred through the N50 and L50 values. The length of the contig that makes half of the assembly length is two orders of magnitude below the total assembled genome. Which means that there plenty of contigs of even smaller length. This is corroborated with the L50 value, which indicate that half of the assembled base pairs are distributed in only 10 contigs, the other half therefore is made up of 77 contigs.
The genome browser Icarus can be accessed in the Galaxy history to visualize this assembly and the differences between it and the reference genome used. Since the result is really bad and the assembly very small, it was not feasible to select only certain contigs to predict and annotate genes based on their similarity with the reference genome assembled chromosomes. Ideally the contigs aligned over chromosomes 9, 1 and 3 would have been selected. Chromosome 9 because it contains the locus for one critical enzyme of the cannabinoid biosynthesis pathway, and the other two becuase they encompass several QTL regions linked to the production of cannainoid metabolites. Therefore the whole assembly was given as input to the Augustus predictor in order to continue with the workflow.

4.3 Gene prediction and annotation

The prediction of Augustus, using the training dataset of *Arabidopsis thaliana* gave a total of 493 sequences. The Blastp report, restricted to *Arabidopsis thaliana* and using the RefSeq database contained a total of 234 proteins for whom matches were found. The Blastp report using the SwissProt database had 235 proteins, a similar number. After extracting the GO terms with BioMart Plant and filtering using the script ProteinSelector3.pl, the total amount of proteins deemed candidate to further investigate was reduced to 86 in the case of the ones coming from RefSeq and 119 in the one coming from SwissP

4.4 Homology study

The following table summarizes the proteins chosen for further analysis:

Enzyme Name:AN	**#Total GOs shared**	**2 top proteins**	**#GOs shared with top protein**	**Database**
AAE1:F4HUK6	22	Q8VZF1 Q84P23	4	SwissProt
CBDS:A6P6V9	0	Q8VZF1 Q84P23	0	SwissProt
OAC:I6WU39	89	Q9SP32 Q9M039	2	SwissProt
OLS:B1Q2B6	49	Q8GYW8 O80467	3	SwissProt
THCA:Q8GTB6	66	Q9SU40 O03042	4	SwissProt
AAE1:F4HUK6	15	NP_201143.1 NP_188316.1	4	RefSeq
CBDS:A6P6V9	0	NP_201143.1 NP_188316.1	0	RefSeq
OAC:I6WU39	62	NP_568232.1 NP_680598.2	2	RefSeq
OLS:B1Q2B6	42	NP_180087.1 NP_001329809.1	3	RefSeq
THCA:Q8GTB6	41	NP_051067.1 NP_001329809	3	RefSeq

Table 2: Summary of gene selection process.

It is worth noting that no protein was found that shared at least one GO term with the CBDS enzyme. The possible reasons will be discussed in the next section. Now we proceed to discuss each protein in detail and its possibles links to the syntehsis of cannabinoids.

Q8VZF1 This enzyme is the a peroxisomal acetate/butyrate–CoA ligase. It catalyzes the synthesis of acetyl-CoA by linking acetate to the coenzime A using ATP. The acetyl-CoA is a necessary substrate for some of the enzymes

of the cannabinoid biosynthetic pathway, like AAE1, hence this enzyme could be related to the biosyntehsis by yielding the requiered cofactor. It is also capable of linking longer fatty acids to the coenzime A, like butyric acid.

Q84P23 This enzyme is a 4-coumarate-CoA ligase-like that catalyzes the formation of coumaroyl-CoA. Its activity resembles that of AAE1 and and the aforementioned Q8VZF1, since all three link coenzime A by creating tioester bonds between it and a carboxylic group. It is not involved in the synthesis of cannabinoids, but participates in the synthesis of other, polyphenolic metabolites, like such as flavonoids.

Q9SP32 This enzyme is an homolog of the Dicer endoribonuclease, and is involved in the process of post-transcriptional gene silencing and ARN interference. It is involved in the defence against RNA virus but also in the gene RNA-dependent regulation of gene expression. It promotes the flowering process by repressing the flowering locus C. It does not appear to be related to the cannabinoid synthesis in any way except for the fact that these molecules are accumulated in glandular trichomes all over the plant but especially in flowers.

Q9M039 This enzyme is the pyruvate decarboxylase. It transforms pyruvic acid into acetaldehyde. There are some decarboxylation reactions in the last steps of the synthesis of cannabinoids, which according to Biocyc are spontaneous, but perharps they can be catalyzed by some decarboxylase enzyme.

Q8GYW8 This enzyme is the spermidine coumaroyl-CoA acyltransferase. It binds and acyl group of the coumaroyl-CoA to the spermidine, creating bis(coumaroyl)-spermidine molecule and releasing coenzime A, which is needed in the synthesis of cannabinoids.The reaction is also similar to the one catalyzed by OLS, which also forms a bond between a coenzime A-activated acyl group donor and another molecule, releasing coenzime A during the process. Its main function howevers lies within the polyamines metabolism, such as spermidine.

O80467 This enzyme is the spermidine sinapoyl-CoA acyltransferase, which catalyzes a similar reaction as the previously described enzyme and OLS, the formation of an acyl bond between a coenzime A activadted molecule, in this

case sinapoyl-CoA, and another substrate. This enzyme is also involved in the polyamine metabolism.

Q9SU40 This enzyme is the monocopper oxidase-like. Its specificty is not known. There reactions catalyzed by THCA and CBDS both use oxygen as electron acceptor for the reaction and turn it to hydrogen peroxyde. These reactions eventually lead to the production of THC and CBD. There is a third reaction that according to biocyc uses the same substrate as THCA and CBDS but an unknown oxidized molecule as electron acceptor. Perhaps an oxidase similar to this one is involved in its generation.

NP_003042 This enzyme is the long chain of the ribulose biphosphate carboxylase, a part of the RuBisCO enzyme. It is involved in the Calvin cycle and assimilation of carbon dioxyde during the dark phase of photosynthesis as well as the photorespiration pathway. It is not related to the synthesis of cannabinoids directly.

NP_001320124.1 This enzyme is a protein from the peroxidase superfamily. It is located in plasmodesma, cell wall and extracellular region, and it is involved in peroxidation, response to oxidative stress and heme binding.

NP_001329809.1 Trehalose phosphatase/synthase 5 is a glycosyltransferase that catalyses the synthesis of alpha,alpha-1,1-trehalose-6-phosphate from glucose-6-phosphate using a UDP-glucose donor. It is a key enzyme in the trehalose synthesis pathway.

NP_568232.1 This enzyme is a zinc dependent oligopeptidase that may be involved in the degradation of proteasome-generated peptides by hydrolysis. It is up-regulated during senescence and pathogen-infection.

NP_680598.2 This enzyme is a homolog of yeast polyadenylation factor Protein 1. It is involved in positive regulation of flower development. Cannabinoids are located in the flower too, so it could have some relation between them.

5 Discussion

The assembly pipelines we tried had no or limited success due mainly to computational resources restraints in the Galaxy server. If we had had the

possibility to use all the available data, we could have obtained a better, less fragmented and more representative assembly. We could also have compared the De Bruijn graph-based ABySS with the OLC pipeline of the authors of the study,pilon correction included. The trials we did with the variant calling pipelines to find relevant genes also failed due to the lack of quality of variants. Also, it would have been easier if a golden reference of *Cannabis sativa* genome had been available.

Even though the assembly was suboptimal, we managed to find genes with a similar functions to those involved in cannabinoid biosynthesis. We think that there is some kind of relation between the proteins NP_680598.2, Q9SP32 and cannabinoid biosynthesis pathway. These proteins are involved in flower development and regulation, and the cannabionoids are mainly located in female flowers. It could be a interesting relationship to further research.

Some of the proteins were related to the pathway at the catalyzed reaction type level only, performing similar types of reactions, such as acyl-CoA transferase activity, decarboxylation or oxidation.

However, this lack of direct association was likely to have happened anyways because the pathway is not very well understood, many enzymes are not known and are poorly annotated, and is difficult to find homologs in *Arabidopsis thaliana* because this pathway belongs to the secondary metabolism and as such is not preserved in every plant, since it is not essential for survival. One way to improve the analyses would have been to use only the GO terms of molecular function and biological process as well as Kegg and Reactome pathway terms of the pathway's enzymes available in Biocyc to filter the Blastp report, thus narrowing the candidate protein group to study.

6 Bibliography

1. Laverty KU, Stout JM, Sullivan MJ, Shah H, Gill N, Holbrook L, Deikus G, Sebra R, Hughes TR, Page JE, van Bakel H.A physical and genetic map of Cannabis sativa identifies extensive rearrangements at the THC/CBD acid synthase loci. *Genome Res* 2019 Jan;29(1):146-156.

2. Van Bakel H, Stout JM, Cote AG, Tallon CM, Sharpe AG, Hughes TR, Page JE. The draft genome and transcriptome of Cannabis sativa. *Genome Biol* 2011 Oct 20;12(10):R102.

3. Lakhan SE, Rowland M. Whole plant cannabis extracts in the treatment of spasticity in multiple sclerosis: a systematic review. *BMC Neurol* 2009 Dec 4;9:59.

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

5. Russo E. Cannabinoids in the management of difficult to treat pain. *Ther Clin Risk Manag* 2008 Feb; 4(1): 245–259.

6. Raharjo TJ, Chang WT, Verberne MC, Peltenburg-Looman AM, Linthorst HJ, Verpoorte R. Cloning and over-expression of a cDNA encoding a polyketide synthase from Cannabis sativa. *Plant Physiol Biochem* 2004 Apr;42(4):291-7.

7. Elsohly MA, Slade D: Chemical constituents of marijuana: The complex mixture of natural cannabinoids. *Life Sci* 2005, 78:539-548.

8. Caspi R, Billington R, Fulcher C, Keseler I, Kothari A, Markus Krummenacker M, Mario Latendresse M, Peter E Midford P, Ong Q, Ong W, Paley S, Subhraveti P, Karp P. The MetaCyc database of metabolic pathways and enzymes, *Nucleic Acids Research* 2018, 46(D1):D633-D639.

9. Morimoto S, Komatsu K, Taura F, Shoyama Y. Purification and characterization of cannabichromenic acid synthase from Cannabis sativa. *Phytochemistry* 1998 Nov;49(6):1525-9 .

10. Baker PB, Taylor BJ, Gough TA. The tetrahydrocannabinol and tetrahydrocannabinolic acid content of cannabis products. *J Pharm Pharmacol* 1981 Jun;33(6):369-72.