Vigilada Mineducación

Genomic annotation and gene expression analysis related to oil production of the non-model crop *Plukenetia volubilis*

Andrés Felipe Florián Cruz

Trabajo de grado

Asesor:

Simón Villanueva Corrales

UNIVERSIDAD EAFIT

Escuela de ciencias

Biología

Medellín

2021

Genomic annotation and gene expression analysis related to oil production of the non-model crop *Plukenetia volubilis*

Andres Felipe Florian[1] and Simón Villanueva-Corrales[2]

[1]Department of Biological sciences, EAFIT University,
Medellin, Colombia
afflorianc@eafit.edu.co
[2]Department of Bioinformatics, Institute of Botany, Czech Academy of Sciences,
Průhonice, Czech Republic
svillanu@eafit.edu.co

November 25, 2021

Abstract

Plukenetia volubilis L. (also known as Sacha inchi) is a crop from the family *Euphorbiaceae* of considerable economic interest given the nutritional properties of its seeds: high content of edible oils, protein and tocopherols. Sacha inchi's seed oil is characterized by a predominant proportion of polyunsaturated fatty acids (PUFAs), which is favorable from a health perspective. In this work, the genome annotation of a previously generated draft-genome assembly was performed. Later, genome-guided transcriptome analyses, including differential expression and co-expression network analyses, were carried out to identify potential regulators of FA and TAG biosynthesis. A total of 51757 genes models were predicted, from which 47531 were functionally annotated. Differential expression dynamics of genes related to FA and TAG biosynthesis is described. From the co-expression network analysis, important putative regulators of FA and TAG biosynthesis were found. In particular, WRI1, FUS3, and LEC1 stand out given their regulatory roles. The identification of regulatory genes involved in FA and TAG biosynthesis pathways will provide useful resources for further research and efforts in the genetic improvement of Sacha Inchi.

1 Introduction

To date more than 370000 plant species have been discovered, from those, only 20 provide most of the world's food, with three prominent crops such as wheat, maize, and rice accounting for 60 % of calories and 56 % of protein consumed from plants by humans [1]. Many edible plant species receive little attention in research, often crops that are used in local or regional scales. These underused crops have the potential to contribute to food security, climate change mitigation and nutrition improvement [1]. This is possible through reduced carbon footprint, enhancements of plant resilience and diet diversification.

Plukenetia volubilis L.(Euphorbiaceae), also known as Sacha inchi(Inca peanut) is a perennial oil seed crop native to the tropical region of south america where it grows at altitudes from 200 to 1200 m.a.s.l [2]. This species is of economic interest due to the nutritional properties of its seeds: a high content of edible oils (∼50 % of dry weight), a high content of protein (∼27 % of dry weight), tocopherols (class of chemicals that display antioxidant activity), among others [3–5].

Sacha inchi's oilseed composition is characterized by a predominant proportion of polyunsaturated fatty acids (PUFAs) that account for ca. 77-88% of total fatty acid content [1, 4], followed by monounsaturated fatty acids (MUFAs) and saturated fatty acids that account for ca. 8-13 % and ca. 7-9

% of total fatty acid content respectively [1]. The PUFA fraction consists primarily of essential fatty acids like α-linolenic acid (18:3 cis-Δ9,12,15; ALA) and linoleic acid (18:2 cis-Δ9,12; LA) accounting for ca. 35-50 % and ca. 33-41 % of total PUFA content, respectvively [3, 4, 6].

As low proportions of Omega-3/Omega-6 fatty acids (a characteristic of Western diets) are related to cardiovascular disease, cancer, and inflammation [7], Sacha inchi seeds, with a high proportion of Omega-3/Omega-6 presents as a healthy source of oils to complement the diet of developed and/or developing countries. For these reasons, understanding the processes behind Sacha inchi's oil production is of great interest, as it can offer targets for the genetic enhancement of this crop and a model of oil production in plants.

Several transcriptomic studies of Sacha inchi have been done to date, in particular regarding floral sex diferentiation [8] and oil production pathways [3, 6, 9]. In the later, expression profiles from different developmental stages were analyzed to identify key genes in FA and TAG biosynthesis pathways. Genes such as FAD2, FAD3, SAD2 are commonly posed as highly related to rapid oil production/accumulation stages [3, 6, 9], but further studies are still needed to fully understand this process. Little attention have been given to the regulators of FA and TAG biosynthesis pathways, which may be of special interest as targets of gnomic modification as they can have effects over many genes of a particular pathway. It is also worth mentioning that those studies were carried out with a *de-novo* transcriptome assembly strategy to identify transcripts, which may limit the precision of their analyses due to the lack of a reference genome. It has been demonstrated that using a genome-guided transcriptome assembly strategy usually generates longer contigs [10, 11] and a more complete set of genes [12] that are not dependent of the sample.

To date, EAFIT has advanced its own genome project on Sacha inchi intending to better understand its oil production and to search targets for genetic enhancement. This work aims to advance this project with the functional annotation of the existing draft genome of Sacha inchi and with a genome-guided transcriptomic analysis (which may improve precision compared to previous *de novo* transcriptomic analysis) of oil production pathways, with a focus on the regulation of these pathways.

2 Methods

2.1 Genome assembly from previous work

The assembled draft genome and repeat annotation from [13] served as the basis of this work.

In short, [13] sequenced, assembled and explored the characteristics of *P. volubilis* genome. It was established that *P. volubilis* is an allotetraploid species, meaning that *P. volubilis* is the product of a hybridization event that occurred at some point in evolutionary history. Hence, *P. volubilis* genome is composed of both paternal and maternal ancestral subgenomes. The subgenomes present a degree of divergence, which partly explains the high heterozygosity of the genome. A summary of the main characteristics of the genome is given in Table 1.

Table 1. Summary of *P. volubilis* draft genome assembly from [13].

Parameter	Value
Genome size (Mb)	594
Estimated haploid genome size (Mb)	334
Heterozygosity (%)	11.6
GC content (%)	29.4
Number of scaffolds(# sequences)	13713
N50 (Kb)	112
Complete BUSCOs (%)	96.4

Given the above characteristics, the draft genome presents some redundancy and thus is not an haploid representation.

2.1.1 Repeat sequences annotation

Briefly, the EDTA (v1.6.4) pipeline was used, and thus all its dependencies [14–21], to annotate the repetitive content of the MaSuRCA assembly, with options *sensitive*, *anno* and *evaluate* [13].

2.2 Genome annotation

An illustration of the pipeline used in the "Gene model prediction" and "Functional annotation of genes" sections, is given in Figure S1.

2.2.1 Gene model prediction

The strategy used for gene prediction was based on homology of RNA-seq data (Our libraries and external RNA-seq data from other *P. volubilis* studies [3, 6, 8], protein data from closely related species(*Euphorbiaceae* family) available at the time in the NCBI database (*Ricinus communis* [22], *Jatropha curca* [23], *Manihot esculenta* [24], *Hevea brasiliensis* [25]) and *de novo* predictions.

First, our RNA-seq data were aligned to *P. volubilis* genome with Magicblast [26]for subsequent transcriptome reconstruction with StringTie2 [27]. This transcriptome was used as input along with the protein data for BRAKER2 [28–38](which uses Augustus [39]and Genemark [40]for gene prediction) to generate an initial prediction. This prediction was later passed to MAKER [41] (which uses Augustus, Genemark and SNAP [42]) along with the transcriptome, the protein data and annotations mapped with Liftoff [43]from closely related species (Ricinus communis [22], Jatropha curca [44], Manihot esculenta [24], Hevea brasiliensis [25]). Four rounds of MAKER runs were made to optimize gene models. Two additional rounds were made using a transcriptome extended with external RNA-seq data, which was generated as described previously.

2.2.2 Functional annotation of genes

Functional annotation was carried out with the help of two pipelines: MANTIS [45] (wich internally compares against eggNOG5 [46], TIGRfams [47], Kofams [48], Pfam [49]databases) and Taxonomy Oriented Annotation (TOA) [50](which internally compared against Dicots PLAZA 4.0, Gymno PLAZA 4.0 [51], NCBI Refseq plant and NCBI blast nr databases). KEGG Automatic Annotation Server (KAAS) [52]with bidirectional hit rate was also used to complement the annotation and identification of oil related production genes.

2.2.3 Ortholog analysis

To identify orthologs between *P. volubilis* and *Arabidopsis thaliana* an ortholog analysis with OrthoFinder [53] was run. In order to make a robust analysis recomendations from [54] were followed. For that reason 9 sets of proteins from different species related to *P. volubilis* were used to run OrthoFinder: *Arabidopsis thaliana* (TAIR10) [55], *Populus alba* [56], *Salix suchowensis* [57], *Mercurialis annua* [58], *Hevea brasiliensis* [59], *Manihot esculenta* (NCBI assembly accession: GCF_001659605.2), *Jatropha curcas* [60], *Ricinus communis* [61] and *P. volubilis*(from this work).

2.3 Transcriptomic experiment

The "proprietary" RNA-seq data used in this work comes from a previous transcriptome analysis [62]. The sections "Plant material extraction" and "Construction of RNA Libraries (TruseqTM, Illumina)" were taken from that work.

2.3.1 Plant material extraction

Collections were made at the San Luis farm in Antioquia located 675 m.a.s.l (N6 ° 01'30.7 "; W074 ° 58'06.1"). Samples of three different trees that had ideal agronomic characteristics were collected. Developing leaf, flower, and seed samples in stages E2, E3, E4, and E5. Each of the fruits were characterized according to their size and color.

2.3.2 Construction of RNA Libraries (TruseqTM, Illumina)

RNA extractions were made from leaf, flower, and developing seed at stages E2, E3, E4, and E5. A total of fourteen samples were purified using the "GeneJet Plant RNA Purification Kit" following the manufacturer's recommendations. The RNA was quantified by spectrophotometry (260/280 nm) in a NanoDrop ND-2000 UV-Vis Spectrophotometer equipment (NanoDrop Technologies, USA) and it was analyzed in 1 % agarose gel, visualized under ultraviolet light by staining with Gel Red (0.5 μg / ml). The samples were stored at -80 oC until they were sent for sequencing. 2 ug of total RNA were used for each sample analyzed. To determine the integrity of the total RNA, the Bioanalyzer 2100 equipment, Agilent Technologies®, was used. Samples that showed an integrity number greater than or equal to 8 were used to make the libraries. The entire process of library construction and sequencing was carried out by the High Throughput Sequencing Facility (HTSF) of the University of North Carolina in Chapel Hill, United States. Sequencing was performed using the second generation Illumina HiSeq2500 platform, obtaining paired readings of 50 bases.

2.3.3 Transcriptome analysis

For the analysis of RNA-seq data, the NASA GeneLab RNA-Seq Consensus Pipeline [63] was adapted to our datasets. In short, RNA-seq reads were aligned to Sacha inchi's draft genome(the final gene prediction was also used in this step) with STAR [64], then those alignments were passed to RSEM [65]to calculate gene counts, and finally DESeq2 [66]was used for the differential expression analysis. Genes were treated as differentially expressed if they had an adjusted p-value (Bonferroni correction) less than or equal to 0.05.

2.4 Co-expression network

2.4.1 Construction with WGCNA

The RNA-seq data used for the network construction included all the available data from [3, 6, 8] (this data was also used in the Gene prediction step), plus data from [4] (this data was not used in the Gene prediction step as we got access to that data well after the gene prediction step). A total of 55 RNA-seq libraries were used for network construction, spanning tissues from seeds at different development stages, fruit, male and female flower, stem, leaves, shoot apex and root (see Table S6).

To run the R package WGCNA (the package behind the network construction), an appropriate gene count normalization is needed as the results of the coexpression network are dependent on correlations across genes calculated from these counts. Based on the results of [67], read counts were normalized

with the weighted trimmed mean of M-values (TMM) method, with the help of edgeR [68]. The genes used to construct the network were the ones that had a total sum of gene counts higher than 50 across libraries and a coefficient of variation across all libraries greater than 0.5. This last choice was inspired by [69] and is used to reduce the effect of wrong correlations across the network. In total, 18769 genes were used for network construction.

A signed adjacency network was constructed with a soft-power threshold of 12 (important parameter for network construction in WGCNA), based on the scale-free topology criterion (See supplementary figure S3). This criterion is recommended for the selection of the soft-power threshold parameter [70, 71]. Then an adjacency matrix was transformed into a topological overlap measure (TOM) matrix to estimate gene-wise similarities [70, 72].

2.4.2 Co-expression module detection

From the TOM matrix, a TOM-based dissimilarity matrix [70] was created and from it, a dendrogram was constructed, from the dendogram the "cutreeDynamic" function of the "dynamicTreeCut" package was used to detect co-expression modules with an automatic tree height of 0.99 . Then modules with a correlation greater than 0.75 were merged to get the final coexpression modules from which subsequent analyses were made. This process was done following default parameters for co-expression module detection [71].

2.4.3 GO enrichment analysis of coexpression modules

The "Functional prediction" tool from the "Ortholo-gous MAtrix" project (OMA) was used to generate GO annotations for all the P. volubilis genes. The "topGO" R package [73] was used to estimate GO enrichment per module. Fischer exact test was used to estimate enrichment significance, and the algorithm "elim" was selected to manage the GO graph topology.

3 Results and discussion

3.1 Genome annotation

MAKER, the principal tool used for gene prediction, offers the option to iteratively predict genes based on results of previous runs, using previous gene models and training gene predictors on those gene models. This feature improves the accuracy of the predicted gene models and the quality of downstream analysis. Three iterations were performed based on proprietary mRNA and proteins and annotations from related species (see Methods). These iterations showed an increase in quality as suggested by Benchmarking Universal Single-Copy Orthologs (BUSCO) scores (Tables 3 and 2). Iterations 4 and 5 included external mRNA libraries, which had several advantages: longer lengths (90 to 150 bp in external libraries vs 50 bp in our libraries), higher sequencing depth (ca. 26 to 65 M of reads per external library vs ca. 16 M of reads per proprietary library), and increased transcript diversity (additional tissues included). These last two iterations showed significant improvement in quality according to BUSCO scores (Tables 3 and 2).

Table 2. Gene prediction by iteation summary of Sacha inchi draft genome. ncRNA prediction was done only in the last iteration.

Prediction	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Gene models number	58787	51214	50935	51786	51757
Number of mRNAs	59690	51757	51455	56090	58134
Average gene length (bp)	2303	3782	3832	3950	4199
Mean exon length (bp)	278	228	227	229	237

Average exon per mRNA	4.1	7.4	7.5	7.8	8.1
Mean intron length (bp)	380	326	325	343	357
tRNAs[a]	-	-	-	-	535
rRNAs[b,c]	-	-	-	-	572
snRNAs[b]	-	-	-	-	667
miRNAs[b]	-	-	-	-	667

[a] tRNA genes were predicted with tRNAscan-SE [74]within MAKER.
[b] The Rfam database [75]and the Infernal software [76]were used to predict rRNA, snRNA, and miRNA genes.
[c] rRNA subunits were identified by RNAmmer [77].

Table 3. BUSCO score by iteration. A constant growth and reduction of complete and missing orthologs, respectively, suggests an increase in the prediction quality of the genome. Iterations 2 and 3 had slight variations in the orthologous numbers that are not well reflected in percentages.

Category BUSCOs	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5
Complete (C)	84.26%	89.42%	89.42%	89.60%	93.12%
Complete single-copy (S)	77.39%	84.14%	84.14%	78.59%	77.90%
Complete duplicated (D)	6.88%	5.29%	5.29%	11.01%	15.22%
Fragmented (F)	8.21%	5.37%	5.37%	4.90%	3.18%
Missing (M)	7.52%	5.20%	5.20%	5.50%	3.70%

Overall in the final iteration, 51757 genes were predicted with 58314 mRNAs associated. These genes have an average gene length of 4199 bp, average exon length of 237 bp, average exon number per mRNA of 8.1, and an average intron length of 357 (Table 2). BUSCO analysis revealed that 2166 complete BUSCO sequences (93.12 %) could be identified from the eudicots dataset, indicating that the predicted gene set was highly complete (Table 3). When comparing the number of genes detected with other reported genome annotations of the *Euphorbiaceae* family, we found that *P. volubilis* has the highest number of genes, followed by *H. brasiliensis* and *E. lathyris* (Table 4). The high heterozygosity in the genome of *P. volubilis* [13] may in part be responsible for this high gene count. During genome assembly, heterozygotic loci of a given gene might be hard to resolve onto a single haploid representation if the divergence between the alleles is high enough. Thus, the different alleles from the same locus might be represented in the final genome assembly, resulting in multiple copies of genes for those "inflated" loci.

Table 4. Number or gene models per species(currently published genomes) in the spurge family (Euphorbiaceae)

Species	Number of gene models	Reference
Plukenetia volubilis	51757	-
Hevea brasiliensis	44187	[59]
Euphorbia lathyris	36342	[78]
Manihot esculenta	34483	[79]
Ricinus communis	30066	[80]
Vernicia fordii	28422	[81]
Jatropha curcas	27619	[82]

Among the predicted genes, 47531 (91.83 %) were functionally annotated. MANTIS, TOA, and KAAS allowed the annotation of 90.82 % , 85,78 % and 17.34 % of all genes, respectively.

Additionally, several types of non-coding RNAs were identified, including 667 microRNA genes, 535 tRNA genes, 572 rRNA genes, and 667 small nucleolar RNA (snoRNA) genes (Table 2).

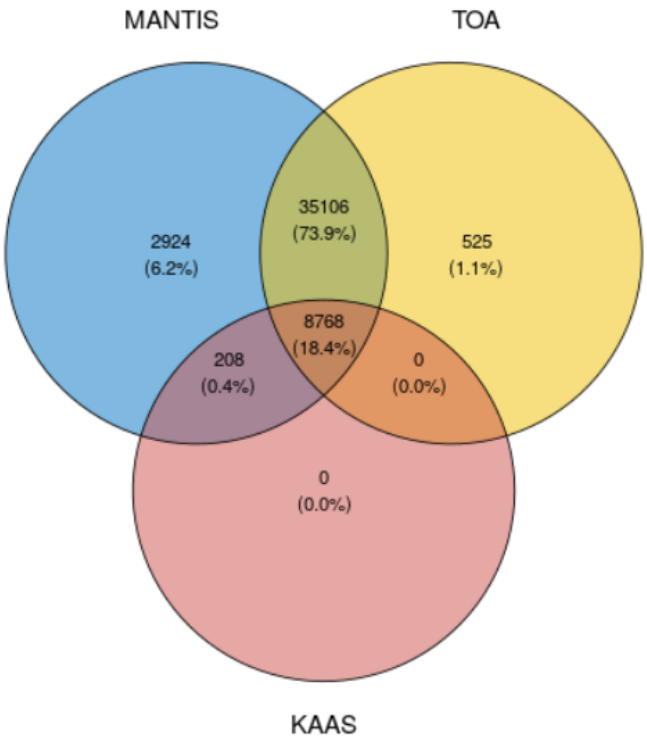

Figure 1. Annotation overlap of TOA, MANTIS and KAAS functional annotation pipelines of Sacha inchi gene models.

3.2 Identification of ALA biosynthesis related genes

Enzymes related to Pyruvate metabolism, ALA biosynthesis, TAG biosynthesis, and FA catabolism were searched in the functional annotations of MANTIS, TOA and KAAS. A total of 37 enzymes across 197 genes were identified (Table 5). In a separate analysis with orthofinder 27 putative ortholog regulators of FA and TAG biosynthesis were found (orthologs with *A. thaliana*) (Table 6).

Next, a short description of the role the identified genes have in their respective pathway is provided:

3.2.1 Acetyl-CoA production

The biosynthesis of ALA occurs in two steps: (1) FA elongation and (2) desaturation (illustrated in Figure 3). The acyl chain elongation step uses acetyl-CoA as starting substrate. The acetyl-CoA may come from the conversion of pyruvate via Pyruvate dehydrogenase complex (PDHC, which includes four subunits: α-PDH, β-PDH, LPD and DHLAT) [3, 4, 6]. The pyruvate in turn, may come from the conversion of Phosphoenolpyruvate via Pyruvate kinase (PK). The flux of carbon from Phosphoenolpyruvate might be redirected to protein biosynthesis by the action of Phosphoenolpyruvate carboxykinase (PEPC) or Phosphoenolpyruvate carboxylase (PPC) (see Figure 2) [81].

Previously, ACC and PEPC have been proposed as key enzymes driving the carbon flux to either oil or protein biosynthesis in the seed. In particular, Zhang *et. al.* [81] suggested that the ratio between the number of ACC genes to PEPC genes is possibly related to the relative contributions of oil and protein to seed mass, where more ACC genes than PEPC genes would yield higher oil contents in the seed. When we compared the ratio ACC/PEPC numbers with the reported seed oil contents across seven species (Table S1, Figure S2), we found that Sacha inchi follows the proposed correlation. Sacha presents a similar number of ACC and PEPC genes to *Vernicia fordii* and relatively similar seed oil contents (ca. 50 % and 60 % for Sacha inchi and Tung tree, respectively) [81]. Nonetheless, this correlation should be handled with care as other factors may affect the carbon flux for oil biosynthesis

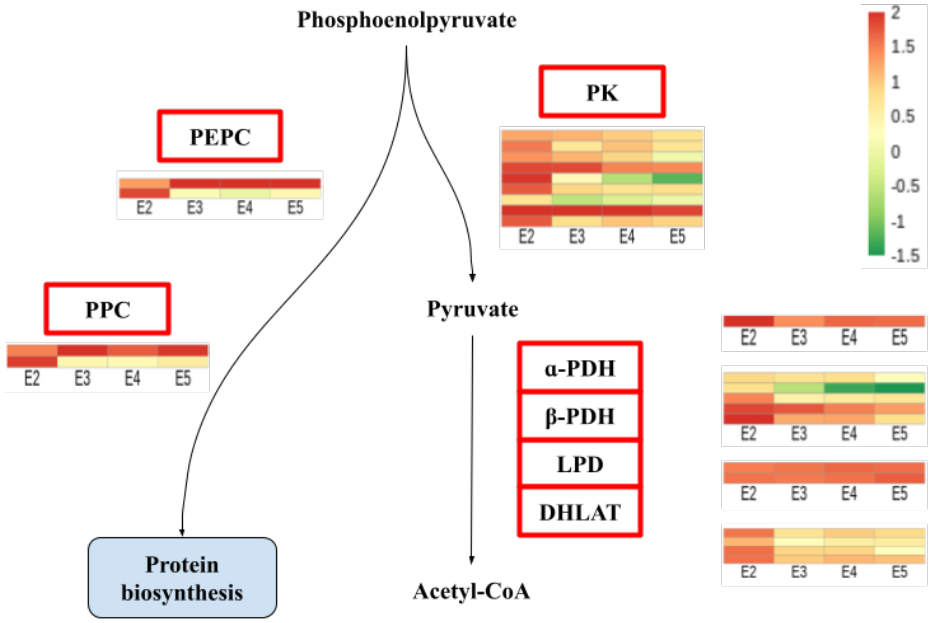

Figure 2. Acetyl-CoA production pathway. In red boxes are enzymes for which a putative gene was found. Heatmaps correspond to log10 TPM(Transcripts per Million) normalized expression in the corresponding genes. Only genes with a TPM > 0 and differentially expressed across any of the following comparisons: E2 v E3, E3 v E4, E4 v E5 were shown.

that are not evident in the analyzed species.

3.2.2 ALA biosynthesis

To better illustrate this process an schematic of this pathway is presented in Figure 3.

FA elongation starts with Acetyl-Co catalyzed by acetyl-CoA carboxylase (ACC) to form malonyl-CoA. Then, malonyl-CoA is converted to malonyl-ACP by Malonyl-CoA ACP Malonyltransferase (MCMT). malonyl-ACP is the primary substrate for the following condensation reactions [3, 4, 6, 81]. Next, Ketoacyl-ACP synthase III (KAS III), Ketoacyl-ACP reductase (KAR), 3-hydroxyacyl-ACP dehydratase (HAD), and Enoyl-ACP reductase (EAR) mediate the production of a C4:0-ACP, which is the substrate for further elongation [3, 4, 6]. Ketoacyl-ACP synthase I (KAS I) is used for the elongation from C4:0-ACP to C14:0-ACP [4, 6]. Ketoacyl-ACP synthase II (KAS II) catalyzes the reaction from C14:0-ACP to C:16:0-ACP, and C:18:0-ACP [4]. SAD-ACP inserts the first desaturation in the carbon chain, FAD2/6 (omega-6 desaturases) inserts the next desaturation, and then FAD3/7/8 (omega-3 desaturases) insert the final desaturation ending up in the formation of ALA. During the last to carbon additions in C16:0-ACP to C18:0-ACP, and the subsequent desaturations up to C18:3-ACP, FATB and FATA may release free fatty acids [3, 4, 6, 81]. These are then activated to CoA esters via Long-chain acyl-CoA synthetase (LACS) and transported to the Acyl-CoA pool in the Endoplasmatic Reticulum (ER) [3, 4, 6, 81].

3.2.3 TAG biosynthesis

TAGs (a composition of FAs and Glycerol) are the main way neutral lipids are stored in oleaginous seeds [4]. Biosynthesis of TAGs (illustrated in Figure 4) occurs in the ER [3, 4, 81].

Dihydroxyacetonephosphate is reduced by Glycerol-3-phosphate dehydrogenase (GPDH) to form glycerol-3-phosphate (G3P) [3, 4]. G3P is acylated by glycerol-3-phosphate acyltransferase (GPAT) to produce 2-lysophosphatidic acid (LPA) [3, 4, 6, 81]. Phosphatidic acid (PA) is synthesized from the acylation

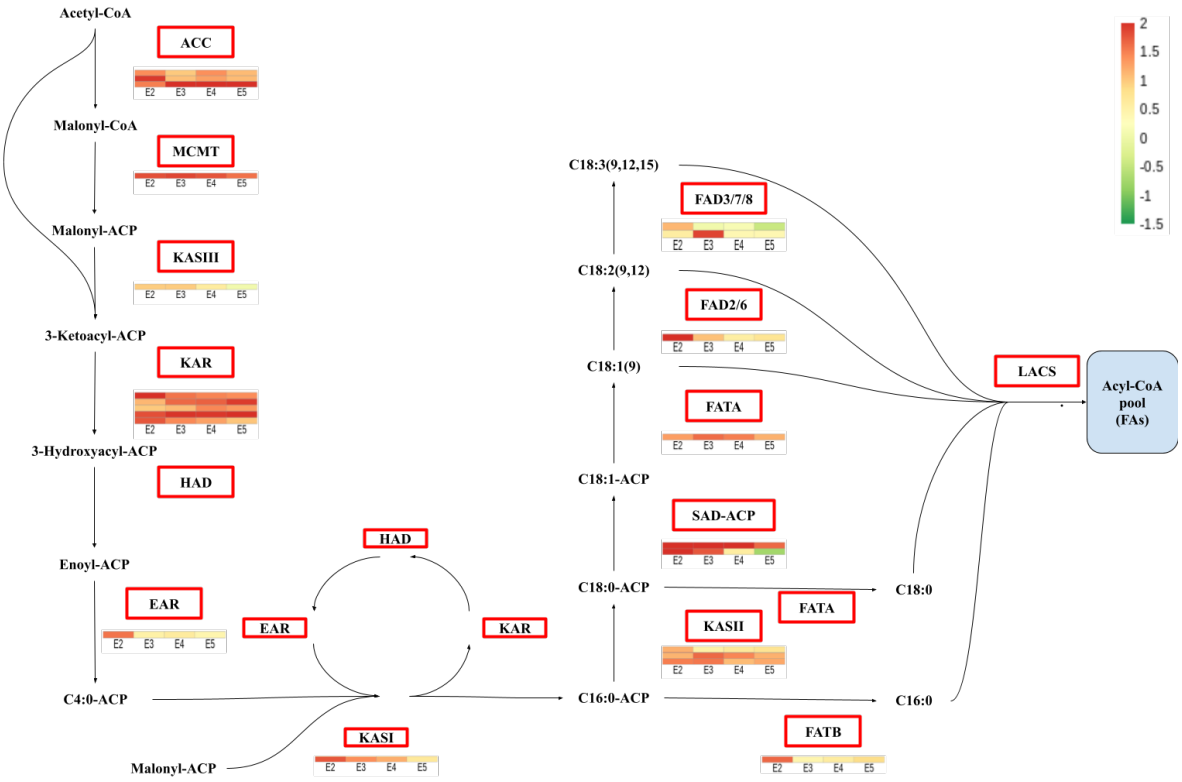

Figure 3. FA biosynthesis pathway. In red boxes are enzymes for which a putative gene was found. Heatmaps correspond to log10 TPM(Transcripts per Million) normalized expression if the corresponding genes. Only genes with a TPM > 0 and differentially expressed across any of the following comparisons: E2 v E3, E3 v E4, E4 v E5 were shown.

of LPA by Lysophosphatidic acid acyltransferase (LPAAT) [3, 4, 6, 81]. PA is transformed into Diacylglycerol (DAG) in a reaction catalyzed by phosphatidate phosphatase (PP) [3, 4, 6, 81]. PP is acylated by Diacylglycerol acyltransferase (DGAT) to form TAG [3, 4, 6, 81]. Alternatively, Phospholipid: diacylglycerol acyltransferase (PDAT) may catalyze the conversion of DAG to TAG using phosphatidylcholine (PC) as an acyl donor [4, 81]. PC production for TAG synthesis from DAG is dependent on the activity of LPCAT that uses Lysophosphatidylcholine (LPC) as a sustrate. DAG may also be produced from PC via DAG-CPT [4, 81]. Free FAs enter the acyl-CoA pool for TAG synthesis via Phospholipase A2 (PLA2) released by the hydrolysis of phospholipids. Once TAGs production is done, layers of phospholipids and protein steroleosins, oleosins and caleosins will surround these TAGs to form structures referred to as oil bodies [4, 81]. Oil accumulation in seeds is usually related with an increase in the number of oil bodies [3].

3.2.4 FA catabolism

The major pathway of FA catabolism is the beta-oxidation pathway (illustrated in Figure 5) and it encompasses the acyl-CoA oxidase (ACX), ketoacyl-CoA thiolase (KAT) and multifunctional protein (MFP2) [6]. The enzyme LACS mediates the initial reactions of FA catabolism by providing the Acyl-CoA pool for the oxidation [6].

3.2.5 Regulators of FA/TAG biosynthesis

The regulation of fatty acid metabolism and oil biosynthesis is well studied and characterized in *Arabidopsis thaliana*. Several orthologs of regulator genes were found in *P. volubilis* annotation. These results are described in Table 6.

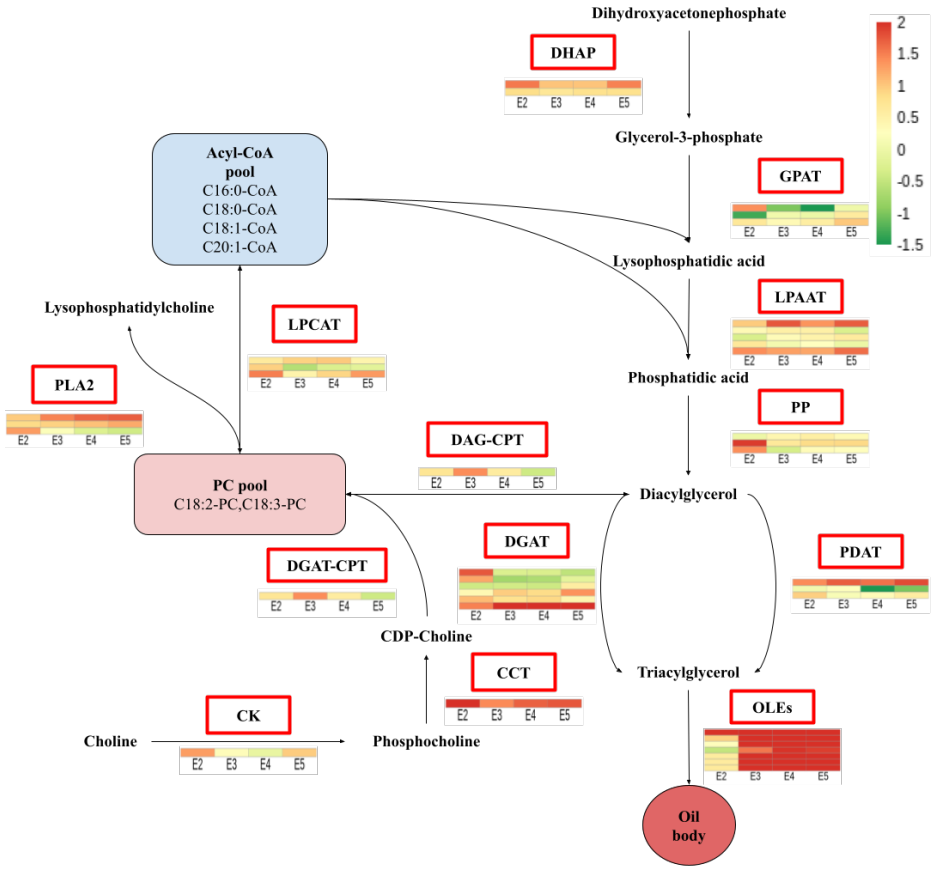

Figure 4. TAG biosynthesis pathway. In red boxes are enzymes for which a putative gene was found. Heatmaps correspond to log10 TPM(Transcripts per Million) normalized expression if the corresponding genes. Only genes with a TPM > 0 and differentially expressed across any of the following comparisons: E2 v E3, E3 v E4, E4 v E5 were shown.

A schematic of the dynamic these regulators have is presented in 6 . bZIP67 is reported to have a direct positive regulation of the omega-3 desaturase FAD3 in *A. thaliana* [83] and because its overexpression was associated with a gain-of-function phenotype regarding seed dormancy [84]. WRI1 is often cited as a "master regulator" in the transcriptional regulation of plant triacylglycerol (TAG) biosynthesis especially given that *A. thaliana* loss-of-function mutant (*wri1-1*) shows an 80 % reduction in seed oil content [85, 86]. Related to WRI1, FUS3 has been shown to promote FA biosynthesis by regulation of FA-related genes. LEC1 promotes FA biosynthesis partially through the regulation of WRI1, FUS3, and ABI3 [86, 87].

TT8 can directly inhibit the transcription of LEC1,LEC2, and FUS3 [86, 88]. TT2 participates the in inhibition of fatty acid (FA) biosynthesis in the seed embryo by directly binding to the promoter of FUS3 [89, 90]. TTG1, together with TT2 and TT8, coordinates carbon partitioning driving the flux of carbon into the synthetic pathways of seed coat mucilage and flavonoid pigments, which compete with oil for photosynthates [86, 91]. In particular, TTG1 binds to the GL2 locus downregulating GL2 which is related to carbon allocation to seed coat mucilage [86].

BPM1 interacts with WRI1 mediating its destabilization and degradation via the 26S proteasome [92, 93]. A model suggests that BPM1 competes with 14-3-3 proteins for binding with WRI1. These proteins are related to protein stabilization and activity enhancement of WRI1. Overexpression of 14-3-3 proteins in *Nicotiana benthamiana* leaves leads to increased oil biosynthesis [94].

Mediator complex MED15 physically interacts with WRI1 to form MED15/WRI1 complex, which in turn transcriptionally upregulates WRI1 target genes [95]. However, it is also suggested that MED15 may interact with transcriptional regulators other than WRI1 to regulate the gene expression of WRI1 targets [95].

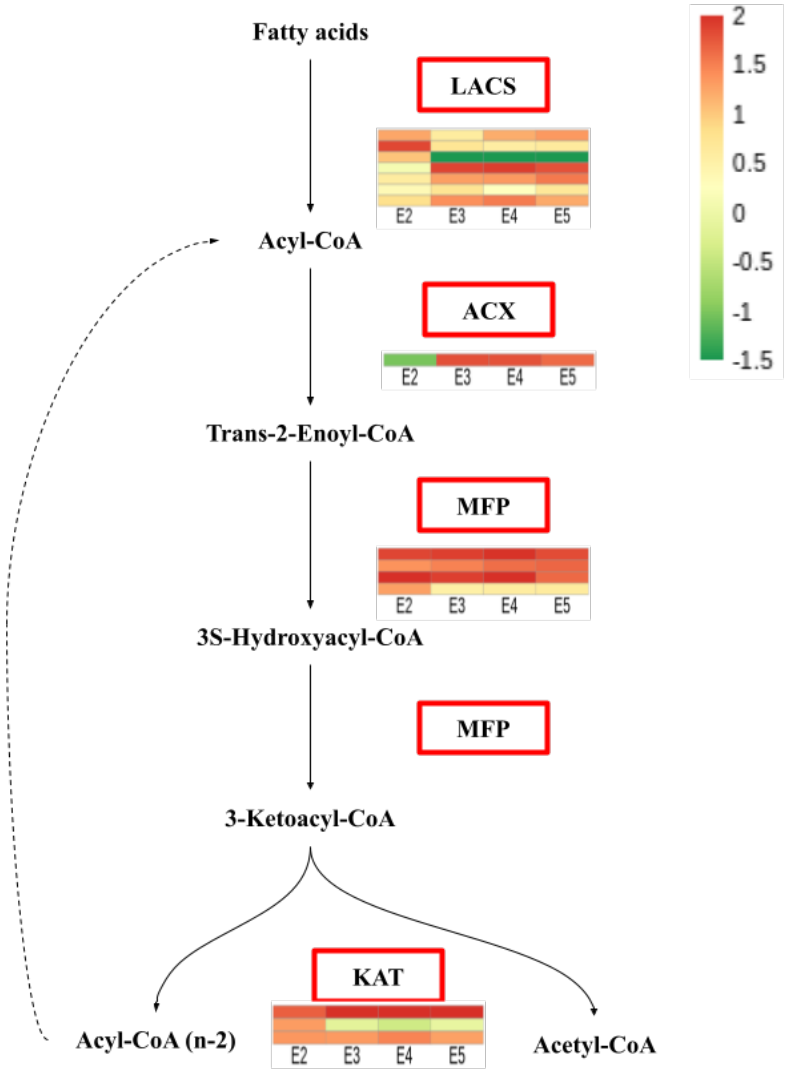

Figure 5. FA catabolism pathway. In red boxes are enzymes for which a putative gene was found. Heatmaps correspond to log10 TPM(Transcripts per Million) normalized expression if the corresponding genes. Only genes with a TPM > 0 and differentially expressed across any of the following comparisons: E2 v E3, E3 v E4, E4 v E5 were shown.

KIN10 has been shown to phosphorylate WRI1, leading to WRI1 degradation [92, 96]. The site in WRI1 where phosphorylation occurs is adjacent to the 14-3-3/BPM1 overlapping motif model suggesting a relationship between the two regulatory mechanisms involving proteasome-mediated protein degradation [92, 96]. T6P (Trehalose 6-phosphate) was found to increase WRI1 stabilization and increase fatty acids generation via suppression of KIN10 activity [92, 97].

SPT promotes FA accumulation through upregulation of FA synthesis genes during seed development while inhibiting the expression of seed storage protein (SSP) genes [90, 98].

SHN1 upregulates the expression of numerous genes involved in de novo fatty acid synthesis, wax accumulation, and oil production [86, 99].

WRKY6, a WRKY family transcription factor, was recently demonstrated to regulate seed oil content and FA compositions, suppressing the expression of genes that are involved in FA biosynthesis and modification during seed development [90]. Mutation of this gene increased FA content and C18:3 proportion in seeds of *A. thaliana* [90].

DREB2C, a TF related to response to abiotic stresses, emerged to promote 18:3 production. When

transformed into *A. thaliana* DREB2C could upregulate FAD3, FAD7, and FAD8 [86, 100].

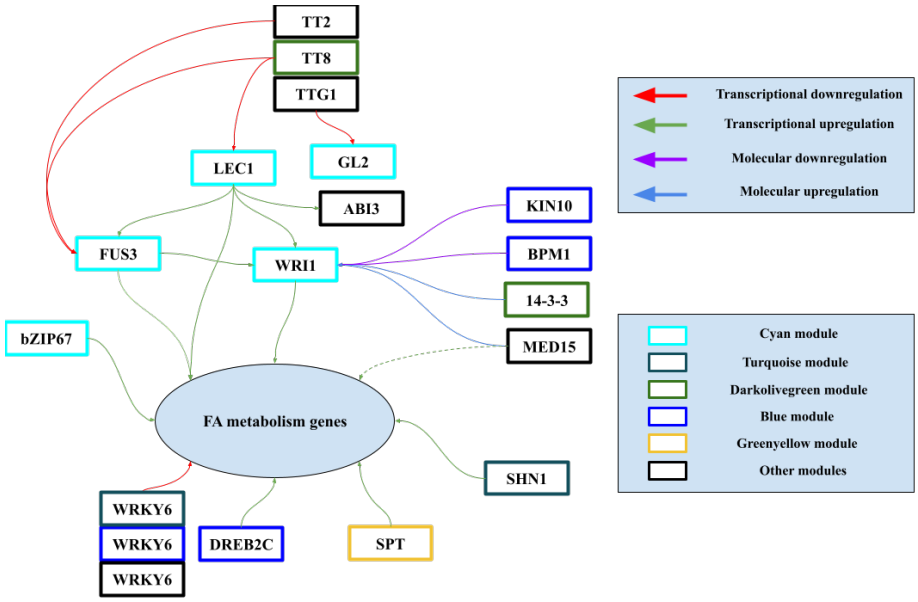

Figure 6. Dynamics of *A. thaliana* regulators of FA and TAG biosynthesis. Molecular upregulation or downregulation refers to interactions at the protein level that not necessarily affect the transcript levels of the target protein. The Boxes representing the regulators are colored depending in the coexpression module they were found. Dotted lines represent indirect regulation.

Table 5. Summary of enzymes involved in ALA metabolism identified by the annotation of the Sacha Inchi transcriptome

Pyruvate metabolism related genes		
Enzyme abreviation	Full name	Number of putative genes
α-PDHC	Pyruvate dehydrogenase α subunit	4
β-PDHC	Pyruvate dehydrogenase β subunit	3
LPD	Dihydrolipoamide dehydrogenase	3
DHLAT	Dihydrolipoamide acetyltransferase	6
PEPC	Phosphoenolpyruvate carboxykinase	3
PPC	Phosphoenolpyruvate carboxylase	4
PK	Pyruvate kinase	8
FA biosynthesis related genes		
ACC	Acetyl-CoA carboxylase	8
MCMT	Malonyl-CoA ACP Malonyltransferase	1
KASIII	Ketoacyl-ACP synthase III	2
KAR	Ketoacyl-ACP reductase	8
HAD	Hidrocyacyl-ACP dehydrase	1
EAR	Enoyl-ACP reductase	4
KASI	Ketoacyl-ACP synthase I	3
KASII	Ketoacyl-ACP synthase II	3
FATB	Acyl-ACP thioesterase B	7
SAD-ACP	Stearoyl-ACP desaturase	5
FATA	Acyl-ACP thioesterase A	1
LACS	Long-chain acyl-CoA synthetase	12
FAD2/6	Fatty acid desaturase 2/6	3
FAD3/7/8	Fatty acid desaturase 3/7/8	3
TAG biosynthesis related genes		
GPDH	Glycerol-3-phosphate dehydrogenase	5
GPAT	Glycerol-3-Phosphate Acyltransferase	14
LPAAT	Lysophosphatidic acid Acyltransferase	8
PP	Phosphatidate Phosphatase	7
DGAT	Diacylglycerol O-Acyltransferase	13
PDAT	Phospholipid:Diacylglycerol Acyltransferase	4
DAG-CPT	CDP-choline-diacylglycerol cholinephosphotransferase	2
PLC	Phospholipase C	6
CK	Choline Kinase	2
CCT	Choline-Phosphate Cytidylyltransferase	1
PLA2	Phospholipase A2	3
LPCAT	1-Acylglycerol-3-Phosphocholine Acyltransferase	5
OLE	Oleosin	10
FA catabolism related genes		
MFP2	Multifunctional protein	4
ACX	Acyl-CoA oxidase	5
ACDM	Acyl-CoA dehydrogenase	1
KAT	3-ketoacyl-CoA thiolase	4

Table 6. Summary of putative regulators of FA and TAG biosynthesis found in Sacha inchi's annotation.

TF oil metabolism related genes		
Enzyme abreviation	Long name	Number of putative genes
14-3-3	14-3-3	2
bZIP67	BASIC LEUCINE ZIPPER 67	1
ABI3	ABSCISIC ACID INSENSITIVE 3	1
BPM1	E3 ligase adaptor BTB/POZMATH 1	3
DREB2C	Dehydration-responsive element-binding 2C	2

FUS3	FUSCA3	1
GL2	GLABRA 2	1
KIN10	SNF1 KINASE HOMOLOG 10	2
LEC1	LEAFY COTYLEDON 1	2
MED15	MEDIATOR 15	1
SHN1	SHINE 1	2
SPT	SPATULA	1
TT2	TRANSPARENT TESTA 2	1
TT8	TRANSPARENT TESTA 8	1
TTG1	TRANSPARENT TESTA GLABRA 1	1
WRI1	WRINKLED1	1
WRKY6	WRKY6	4

3.3 Transcriptome analysis

3.4 Cross library comparison

When comparing the similarities between proprietary and publicly available RNA-seq data from *P. volubilis*, in a PCA (Figure 7), we found different "Transcriptional states". These states correlate with the tissues and the time the samples were collected. Our principal motivation with this analysis was to check which of our samples was more or less correlated with interesting libraries (regarding ALA biosynthesis in developing seeds) from other transcriptome studies, that in some cases, had more reliable metadata. From this analysis we were able to identify 3 transcriptional states (Figure 7): "Other-tissues", "Early-seed" and "Late-seed". Tissues different from seed clustered together, including our leaf and flower tissue samples, as expected (Figure 7). This sample cluster is classified as "Other-tissues", as they weren't of particular interest.

In the "Early-seed" cluster, libraries of seed tissue at stages SI-1 to SI-4 from [4]. Seeds from the SI-5 stage recorded the highest level of ALA and other fatty acids from that study [4]. These libraries clustered closely (almost all of the red points in the far left section in Figure 7), indicating low transcriptional variability in the 0-10 DAP (Days after pollination) to 80-110 DAP that they span. All of our E2 libraries clustered together close to those points, indicating that our E2 libraries are probably from an "Early-seed" state before the seed starts accumulating high amounts of ALA and other FAs.

In the "Late-seed" cluster, we find that most of our seed tissue libraries (E3 to E5) clustered together with high similarity (far right triangle points in the green cluster Figure 7). Our E3 to E5 libraries are closest to late-seed tissue from publicly available data, in particular to the late seed tissue (60 DAP) from [6] and the SI-5 stage from [4], which presented the highest level of ALA and other fatty acids from their study. This suggests that our E3, E4, and E5 stages probably correspond to a late seed tissue where the highest oil accumulation occurs. This cluster has a higher variation that may be due to differences in *P. volubilis* strains between the samples, or environmental factors from the location the samples were collected (Colombia versus China [4, 6]). Hence, the most informative comparison to make for differentially expressed genes (DEGs) for our samples is E2 against any of the E3, E4, or E5 stages.

When looking at the first principal component of the PCA (Figure 7) we find that the "Early-seed" and "Other-tissues" cluster in the left side of the PC1 component. Although with higher variability, the "Late-seed" cluster tends to the right side of PC1. This suggests that PC1 captures the variability in the transcriptional state transition from primary metabolism to a transcriptional state where the high production of ALA and other FAs occurs. "Early-seed" states are more similar to leaves, stem, flower, and other tissues than "Late-seed" states.

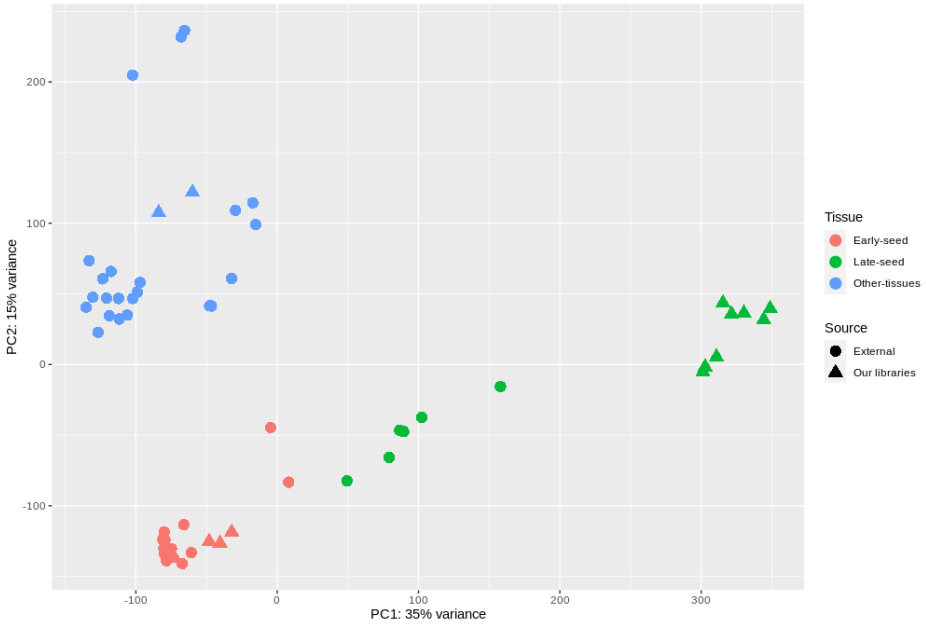

Figure 7. Principal Component Analysis (PCA) of the log2 DESeq2 normalized counts of the 55 libraries used for transcriptome analysis in this work. Circle points correspond to external RNA-seq data, mainly from China. Triangles denote our RNA-seq libraries. In Red and Green libraries are represented the "Early-seed" and "Late-seed" tissue/state, respectively. In blue are represented tissues other than seed. A detailed description of the libraries is given in the Table S6.

3.5 Differential expression results

When analyzing the results of the differential expression analysis, we found considerable variability in the expression changes across the copies of the same genes (*e.g* one PK copy is upregulated while another copy is downregulated), which made the analysis more complicated. It is also important to highlight that we ran a differential expression analysis with *Liu et. al.* data (in this section "*Liu et. al.*" refers to this study: [4]) with a different method than the one they used. They used RPKM as a normalizing count method while we used DESeq2, which uses the median of ratios method [66], a more robust metric when comparing different RNA-seq libraries from several studies.

Confirming the results of the PCA analysis (Figure 7), we found that most of the differentially expressed genes were detected when comparing the E2 seed stage with any of the other seed stages (E3, E4, E5). See Table S2.

3.5.1 Acetyl-CoA production

The LPD subunit of PDHC was upregulated in the late-seed stage while the other subunits presented variable changes or downregulation. PK was mostly downregulated across late-seed stages. PEPC and PPC show variable changes in our libraries. In contrast, PPC shows consistent upregulation in the SI-5 stage in the *Liu et. al.* data , suggesting some upregulation in the protein biosynthesis pathway. Overall, in our libraries, the production of acetyl-CoA is reduced in the late-seed stage, which may indicate a limitation in the carbon flux toward FA biosynthesis in mature seeds.

3.5.2 ALA biosynthesis

ACC subunit copies presented variable expression changes. Some subunits reported significant up-regulation between early and late-seed stages while others the contrary. MCMT, HAD, KASII, and

FATA showed significant upregulation in the late-seed stage in our data. Which is coherent with the previous report [4]. KAR and EAR presented variable fold changes, with some genes being upregulated while others being downregulated. FATB did not show any changes in our libraries, but one copy was downregulated in early stages in *Liu et. al.* data (between the SI-2 and SI-3). KASI only presented upregulation in *Liu et. al.* data . SAD-ACP showed upregulation in the E2 versus E3 stages, although it was downregulated in the E4 to E5 stages. In *Liu et. al.* data SAC-ACP also presented upregulation in the SI-5 stage, consistent with their results.

Most of the FAD2/6 copies were significantly downregulated (one of those by more than 200 fold, in our data and in *Liu et. al.* data), directly inconsistent with the previously reported result [4], were FAD2 was upregulated in the SI-5 stage. In an older study, FAD2 also presented upregulation [3]. This inconsistencies may be due to the lack of replicates (only in [3]) or by the use of RPKM normalizing method. Other factor behind this inconsistency may be the use of a genome-guided transcriptome analysis. Further analysis of the expression this gene presents trough seed development are recommended.

Just one of the three FAD3 copies was significantly upregulated, in *Liu et. al.* data and our libraries (up to 32 fold upregulation). FAD3 was also found upregulated in late-seed stages [3, 4], Consistent with the high content of ALA in Sacha inchi's seeds. LACS reported variable fold changes, with most of the genes significantly upregulated, which suggests high activity in the regeneration of the acyl-CoA pool, partially consistent with [4] report.

3.5.3 TAG biosynthesis

GPDH shows significant upregulation both in the SI-5 stage and in the late-seed stages of our data. GPAT copies show general downregulation in the late-seed stages. In *Liu et. al.* data LPAAT copies show no change or upregulation in the SI-5 stage, while variable fold changes across copies were detected in our data. Most of the PP copies were downregulated in the late-seed stages in our data. Slight upregulation in the SI-5 stage was detected in the *Liu et. al.* data . DGAT copies presented variable fold changes in the late seed stages both in our and *Liu et. al.* data . PLC, CK, and CCT were mostly downregulated in the late-seed stages in both our and *Liu et. al.* data . Slight upregulation was detected in the SI-3 and SI-4 stages. OLE presented significant upregulation in practically all of his copies in the late-seed stage, consistent with their role of packaging TAGs and with the results of previous reports [4].

Overall, the transcriptional changes of this pathway present high variability. This may suggest that other regulatory mechanisms that act outside the transcriptome (and hence beyond the dynamics this study could reveal) may be in place to stimulate the production of TAGs. We know the production of TAGs is active, given the high oil content of late-stage seeds and the significant upregulation of OLEs.

3.5.4 FA catabolism

MFP2 show significant upregulation in the SI-5 stage of the *Liu et. al.* data (up to ca. 2 fold upregulation), while in our data, just one of the copies presents downregulation. All of the ACX copies show significant upregulation in the late-seed stages (up to ca. 10 fold upregulation), this was observed both in our data and in the *Liu et. al.* data . KAT also presented upregulation in our data and in *Liu et. al.* libraries (up to ca. 10 upregulation). Overall this suggests that regulation of FAs is active in the high oil content stage of the seed. This is apparently contradictory: the catabolism of FAs being particularly active despite having the highest amount of FAs at this stage. Perhaps other factors come into play to explain why this activity is not phenotypically translated. Other explanation is an balanced catabolism which is activated but the anabolism is so much higher that it cannot compete. However this is less likely since usually when catabolism is on, anabolism is turned off and vice versa.

3.5.5 Regulators of FA/TAG biosynthesis

Some of the expression changes detected for ortholog regulators will be covered in this section. The rest will be covered in the "Co-expression network analysis" as synergistic effects detected in the co-expression modules made more sense to describe them there.

ABI3 was the regulator with the highest upregulation fold-change in the late-seed stage of all the regulators analyzed in this work. It was also significantly upregulated in the SI-4 stage of the *Liu et. al.* data [4] which may suggest a role before the late seed stage. LEC1 can act as a positive regulator upstream of ABI3 and FUS3 in control of seed maturation and possibly in the regulation of FA metabolism [101]. This is consistent with LEC1 and FUS3 being also upregulated in the late-seed stage.

TT8 and TTG1 were downregulated in the late-seed stages, which is consistent with the upregulation of LEC1 and FUS3 (targets of TT8 and TTG1 transcriptional downregulation [86, 88, 91]) in the same stage. One of the 14-3-3 protein copies was downregulated in the late seed stage. 14-3-3 proteins have roles stabilizing WRI1 protein structure [94]. This might suggest that, in the late stage of the seed, degradation of WRI1 is promoted. MED15 might be interacting with WRI1 as both are upregulated in the late-seed stage.

SPT is upregulated in the late-seed stages and the SI-4 stage of the *Liu et. al.* libraries [4]. Given its role of inhibition of seed storage protein(SSPs) genes [98], it is suggested that SPT is redirecting the carbon flux in the seed development towards FA biosynthesis which is consistent with the stage were most of the FA and TAG accumulation occrus in *P. volubilis*

3.6 Co-expression network analysis

A total of 18769 genes that passed the filters of minimum counts and variance were used to construct the weighted gene co-expression network resulting in the detection of 25 co-expression modules (Table S7). To select which modules were of special interest the following criteria were used: 1. The number of FA and TAG related genes in the modules and 2. Enriched GO terms related to FA and TAG production, in particular **lipid biosynthetic process** (GO:0008610), **fatty acid biosynthetic process** (GO:0006633) and **chloroplast** (GO:0009507). The criteria helped to identify three modules of special interest: Cyan, Blue, and Turquoise each one containing 843, 4373, and 2834 total genes respectively, FA and TAG related genes were 21, 32, and 30 respectively. The Cyan module was enriched for GO including **lipid biosynthetic process**, **lipid transport** and **carbohydrate metabolic process** (Table S3). The Blue module was enriched for **chloroplast**, **lipid storage** and and **unsaturated fatty acid biosynthetic process** (Table S4). The Turquoise module was enriched for **chloroplast**, **lipid biosynthetic process** (Table S5).

From each module of the network, the top Hub genes, which represent the most connected genes in each module, were analyzed to search for potential regulators related to FA and TAG production in each module.

3.6.1 Cyan module

The genes found in the Cyan module are summarized in Table S8. One of the most interesting findings in this module is the grouping of 4 important regulators of oil biosynthesis in *A. thaliana*: bZIP67, LEC1, FUS3, and WRI1, finding them in a single module indicates they are transcriptionally related. All of these genes, excepting LEC1, presented significant upregulation in late-seed stages. This is consistent with the reported interaction of these regulators in *A. thaliana* [86, 90]. Of special interest is WRI1, since it has been modified to improve oil production in *Camelina sativa* [102], *Nicotiana benthamiana* [103] , *Zea mays* [104], *A. thaliana* [105], *Lepidium campestre* [106] and *Jatropha curcas* [107]. The effects of WRI1 overexpression on fatty acid composition are heterogeneous and possibly

species-dependent [106]. In some of these modifications, increases in the production of C18:3 fatty acids have been reported [102, 103] while in others the composition is not affected [104]. In the *Jatropha curcas* study (also from the *Euphorbiaceae* family), no changes in the composition of C18:3 in seeds were analyzed [107]. Although they do report an increase in C18:1 and a reduction in C18:2, which may also suggest reductions in the C18:3 ratio. Overall this suggests that overexpression of WRI1 in *P. volubilis* would increase total seed oil content, but the composition of FAs may be altered. A FAD3/7/8 copy was also found in the cyan module and it was significantly upregulated in the late-seed stages. This is in line with the model where LEC1 induces FUS3 which then, either independently or cooperatively with LEC1, triggers the induction of bZIP67 which promotes the expression of FAD3, regulating the content of ALA in the developing seed [83].

GL2 was also found in the cyan module and significantly down-regulated in the late-seed stage. This is in line with previous reports were downregulation of GL2 via TTG1 was related to changes in seed oil accumulation by reducing the carbon flux to seed coat mucilage and flavonoid pigments, which compete with oil for photosynthates [86, 91]. TTG1 was also downregulated, which suggests that other mechanisms could be behind the downregulation of GL2.

No genes in the FA catabolism pathway were found in this module, just acetyl-CoA production, FA biosynthesis, and TAG biosynthesis. All of the genes presented either no change or upregulation in the late-seed stages. This is coherent with the upregulation of the putative ortholog regulators in this same module.

When looking at the Hub genes of the cyan module, 3 out of the top 10 hub genes of the module were of special interest given their supported relationship with FAs or TAG biosynthesis in other studies. One of the genes was the FAD3/7/8 copy (gene id: PVL00010686), which suggests that this gene is of great importance in the module and that it undergoes considerable regulation. The second gene (gene id: PVL00044451) is functionally annotated as a "brodomain containing protein 9", and it presented significant upregulation in the late-seed stages of the *Liu et. al.* data. This gene is homologous to a TF identified in the unicellular red alga *Cyanidioschyzon merolae* BRD1 ("Brodomain-containing protein" id: CMK212C), which was found to be important in the TAG accumulation in *Cyanidioschyzon merolae* [108]. The proposed pathway for this regulation is via upregulation of LPAT1 (the equivalent to LPAAT in our case) [108]. No LPAAT copies were found in the cyan module, which suggests that PVL00044451 may have a different mechanism in *P. volubilis* than the reported in *C. merolae*. The third gene (gene id: PVL00034290) was also upregulated in the late-seed stages of both the *Liu et. al.* data and our libraries. PVL00034290 is an *A. thaliana* ortholog of "Triacylglycerol lipase 2" (LIP2). This ortholog may be related to TAG catabolism and mobilizes stored lipids (possibly TAG) into free fatty acids that are necessary for growth and development [109]. Lipases can have important roles in determining the TAG content and fatty acid composition during seed development [110].

3.6.2 Blue module

The genes found in the Blue module are described in Table S9. In the blue module, four ortholog regulators were found: BPM1, KIN10, WRKY6, and DREB2C. All of these presented upregulation in the late-seed stages either in our data or the *Liu et. al.* data. BPM1 and KIN10 act as suppressors of WRI1 activity by promoting its degradation [92, 93, 96]. This suggests that BPM1 and KIN10 regulate WRI1 activity in the late-seed stages of *P. volubilis*. Similar observations may arise from WRKY6, which in *A. thaliana* regulates the expression of genes such as FAD2, FAD3, FUS3, and ABI3 [90]. DREB2C upregulation could be related to the regulation of FAD3/7/8.

It is worth noting that most of the OLE genes and many TAG biosynthesis genes were found in this module. Genes such as PDAT, DGAT, and LPAAT (particularly important for TAG biosynthesis) were mostly upregulated in the late-seed stages in this module.

One of the hub genes in this module (gene id: PVL00027076) was annotated as a "Heat shock transcription factor". In *C. merolae* a Heat shock TF (CML277C) was found to be important in the regulation of TAG biosynthesis via LPAT1 (LPAAT in this work) [108]. LPAAT is found upregulated

in this module which may be related to PVL00027076. Heat shock TF regulate TAG biosynthesis as the fatty acid composition of the cell membranes is an important factor for plants to adapt to heat stress [108].

3.6.3 Turquoise module

The genes found in the Turquoise module are described in Table S10. Two hub genes in this module were found to be related with FA catabolism. The first one (gene id: PVL00027001) is annotated as a "Enoyl-CoA hydratase/isomerase", which is related to fatty acid catabolism [111]. PVL00027001 is downregulated in late-seed stages. This pattern may be important for the accumulation of TAG during seed development in *P. volubilis*. The second gene (gene id: PVL00000344) is annotated as a "monoacylglycerol lipase", which is related to TAG degradation [112]. This gene was upregulated in the late-seed stage. This module is the one with the most FA catabolism genes, which is also consistent with two of their Hub genes being related to FA or TAG catabolism.

4 Conclusions

In this work, we continued and expanded previous efforts to understand *Plukenetia volubilis* genome and its oil biosynthesis.

The first high-quality gene annotation for the species is reported. This includes a total of 47531 functionally annotated genes with high completeness (93.12 %). Annotations are important for understanding the role of particular genes in a non-model species, the evolutionary history of the species, and its particular features. It also opens the door for devising ways to edit the species genome, either for research or commercial applications.

Although the annotation is complete for *P. volubilis*, we found a high number of genes compared to other species in the family. A contributing factor to this is the high heterozygosity this species has, which was not entirely resolved during the assembly of the genome.

The reported gene numbers related to ACC and PEPC, and the oil/protein composition of *P. volubilis* seeds are in line with the hypothesis that the ACC/PEPC gene number ratio may influence the total oil proportion in plant seeds [81].

Including published RNA-seq data for *P. volubilis* seeds across many development stages and other tissues, made it possible for us to differentiate differentiate transcriptional states on a temporal and a per-tissue basis. We also were able to correlate the transcriptional states of our seed libraries with the reported seed stages of other studies given their well-curated metadata regarding oil accumulation in the seed. This allowed us to contrast and strengthen our results, especially regarding the differential expression analyses.

The overall results from the differential expression analysis are considerably variable, even within the same gene families. Some of the clearest findings are the significant upregulation of almost all the OLE genes in the late-seed stage, which are related to TAG accumulation. Another finding is the general upregulation of the FA catabolism pathway in the late-seed stages, suggesting that FA catabolism is active even though the accumulation of FAs and TAGs is in process at the same stage. Some detected inconsistencies with previous transcriptome studies may arise from the different approaches to make the analysis presented in this work.

Building a co-expression network, we found three co-expression modules (Cyan, Blue, and Turquoise) of particular interest regarding FA and TAG biosynthesis, which group highly connected and correlated genes. The Cyan module grouped important ortholog regulators of FA and TAG biosyntheses like WRI1, FUS3, and LEC1. WRI1 is a common target gene for increasing oil production in plants. The blue module grouped suppressors of WRI1 activity, and it was mainly related to TAG biosynthesis.

The turquoise module had two highly connected genes that were related to FA and TAG catabolism, and it was the module with the most FA catabolism-related genes. These regulators may serve as targets for genetic modification in *P. volubilis* as they can help to understand further the dynamics behind the high PUFA content in *P. volubilis*, or as candidates for genetic enhancement of this crop.

We present a high-quality genome annotation for Sacha Inchi's genome, which provides a foundation for future studies on the pathways behind features of interest this crop (be it FA and TAG accumulation or others) and the evolutionary history of *P. volubilis*, and related species. Further research is needed to investigate the regulator genes identified in this work and the *in vivo* mechanisms of their influence in FA and TAG pathways. As this regulators are potential targets for genetic improvement of *P. volubilis*. The insights learned here can also help in the metabolic engineering of desirable traits in other plants.

5 Acknowledgements

We would like to thank EAFIT University and the Apolo supercomputing center team for giving us the computational resources required to make possible most of this work. Also thanks to Luis A. Arteaga-Figueroa for giving us valuable insights throughout this work.

References

[1] Nete Kodahl. "Sacha inchi (Plukenetia volubilis L.)—from lost crop of the Incas to part of the solution to global challenges?" In: *Planta* 251 (4 2020), pp. 1–22. ISSN: 14322048. DOI: 10.1007/s00425-020-03377-3. URL: https://doi.org/10.1007/s00425-020-03377-3.

[2] Luis A. Follegatti-Romero et al. "Supercritical CO2 extraction of omega-3 rich oil from Sacha inchi (Plukenetia volubilis L.) seeds". In: *Journal of Supercritical Fluids* 49 (3 July 2009), pp. 323–329. ISSN: 08968446. DOI: 10.1016/j.supflu.2009.03.010.

[3] Xiaojuan Wang et al. "Transcriptome analysis of Sacha Inchi (Plukenetia volubilis L.) seeds at two developmental stages". In: *BMC Genomics* 13 (1 2012). ISSN: 14712164. DOI: 10.1186/1471-2164-13-716.

[4] Guo Liu et al. "Transcriptome analyses reveals the dynamic nature of oil accumulation during seed development of Plukenetia volubilis L." In: *Scientific Reports* 10 (1 2020), pp. 1–17. ISSN: 20452322. DOI: 10.1038/s41598-020-77177-w. URL: https://doi.org/10.1038/s41598-020-77177-w.

[5] Rosana Chirinos et al. "Sacha inchi (Plukenetia volubilis): A seed source of polyunsaturated fatty acids, tocopherols, phytosterols, phenolic compounds and antioxidant capacity". In: *Food Chemistry* 141 (3 Dec. 2013), pp. 1732–1739. ISSN: 18737072. DOI: 10.1016/j.foodchem.2013.04.078.

[6] Xiao Di Hu et al. "De novo transcriptome assembly of the eight major organs of Sacha Inchi (Plukenetia volubilis) and the identification of genes involved in -linolenic acid metabolism". In: *BMC Genomics* 19 (1 2018), pp. 1–14. ISSN: 14712164. DOI: 10.1186/s12864-018-4774-y.

[7] Artemis P. Simopoulos. *The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases*. June 2008. DOI: 10.3181/0711-MR-311. URL: http://journals.sagepub.com/doi/10.3181/0711-MR-311.

[8] Qiantang Fu et al. "De novo transcriptome assembly and comparative analysis between male and benzyladenine-induced female inflorescence buds of Plukenetia volubilis". In: *Journal of Plant Physiology* 221 (December 2017 2018), pp. 107–118. ISSN: 01761617. DOI: 10.1016/j.jplph.2017.12.006. URL: https://doi.org/10.1016/j.jplph.2017.12.006.

[9] Xiaojuan Wang and Aizhong Liu. "Expression of genes controlling unsaturated fatty acids biosynthesis and oil deposition in developing seeds of sacha inchi (Plukenetia volubilis L.)" In: *Lipids* 49 (10 2014), pp. 1019–1031. ISSN: 15589307. DOI: 10.1007/s11745-014-3938-z.

[10] Jeffrey A Martin and Zhong Wang. "Next-generation transcriptome assembly". In: *Nature Reviews Genetics* 12.10 (2011), pp. 671–682.

[11] Bingxin Lu, Zhenbing Zeng, and Tieliu Shi. "Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq". In: *Science China Life Sciences* 56.2 (2013), pp. 143–155.

[12] A Marchant et al. "Comparing de novo and reference-based transcriptome assembly strategies by applying them to the blood-sucking bug Rhodnius prolixus". In: *Insect biochemistry and molecular biology* 69 (2016), pp. 25–33.

[13] Simón Villanueva Corrales, Javier Correa Álvarez, et al. "Genome profiling and assembly of the non-model oilseed crop Plukenetia volubilis L." PhD thesis. Universidad EAFIT, 2020.

[14] David Ellinghaus, Stefan Kurtz, and Ute Willhoeft. "LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons". In: *BMC bioinformatics* 9.1 (2008), p. 18.

[15] Zhao Xu and Hao Wang. "LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons". In: *Nucleic acids research* 35.suppl_2 (2007), W265–W268.

[16] Shujun Ou and Ning Jiang. "LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons". In: *Mobile DNA* 10.1 (2019), pp. 1–3.

[17] Shujun Ou and Ning Jiang. "LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons". In: *Plant physiology* 176.2 (2018), pp. 1410–1422.

[18] Weijia Su, Xun Gu, and Thomas Peterson. "TIR-Learner, a new ensemble method for TIR transposable element annotation, provides evidence for abundant new transposable elements in the maize genome". In: *Molecular plant* 12.3 (2019), pp. 447–460.

[19] Jieming Shi and Chun Liang. "Generic Repeat Finder: a high-sensitivity tool for genome-wide de novo repeat detection". In: *Plant physiology* 180.4 (2019), pp. 1803–1815.

[20] Wenwei Xiong et al. "HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes". In: *Proceedings of the National Academy of Sciences* 111.28 (2014), pp. 10263–10268.

[21] Ren-Gang Zhang et al. "TEsorter: lineage-level classification of transposable elements using conserved protein domains". In: *bioRxiv* (2019), p. 800177.

[22] Maximo Rivarola et al. "Castor Bean Organelle genome sequencing and worldwide genetic diversity analysis". In: *PLoS ONE* 6 (7 2011). ISSN: 19326203. DOI: 10.1371/journal.pone.0021743. URL: https://pubmed.ncbi.nlm.nih.gov/21750729/.

[23] Lin Zhang et al. "Global analysis of gene expression profiles in physic nut (Jatropha curcas L.) seedlings exposed to salt stress". In: *PLoS ONE* 9 (5 May 2014). ISSN: 19326203. DOI: 10.1371/journal.pone.0097878. URL: https://pubmed.ncbi.nlm.nih.gov/24837971/.

[24] Henry Daniell et al. "The complete nucleotide sequence of the cassava (Manihot esculenta) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron". In: *Theoretical and Applied Genetics* 116 (5 Mar. 2008), pp. 723–737. ISSN: 00405752. DOI: 10.1007/s00122-007-0706-y. URL: https://pubmed.ncbi.nlm.nih.gov/18214421/.

[25] Chaorong Tang et al. "The rubber tree genome reveals new insights into rubber production and species adaptation". In: *Nature Plants* 2 (6 June 2016). ISSN: 20550278. DOI: 10.1038/NPLANTS.2016.73. URL: https://pubmed.ncbi.nlm.nih.gov/27255837/.

[26] Grzegorz M. Boratyn et al. "Magic-BLAST, an accurate RNA-seq aligner for long and short reads". In: *BMC Bioinformatics* 20 (1 July 2019), p. 405. ISSN: 14712105. DOI: 10.1186/s12859-019-2996-x. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2996-x.

[27] Sam Kovaka et al. "Transcriptome assembly from long-read RNA-seq alignments with StringTie2". In: *Genome Biology* 20 (1 Dec. 2019), p. 278. ISSN: 1474760X. DOI: 10.1186/s13059-019-1910-1. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1910-1.

[28] Tomáš Brůna et al. "BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database". In: *NAR Genomics and Bioinformatics* 3 (1 Jan. 2021). ISSN: 2631-9268. DOI: 10.1093/nargab/lqaa108. URL: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa108/6066535.

[29] Katharina J Hoff et al. "BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS". In: *Bioinformatics* 32.5 (2016), pp. 767–769.

[30] Katharina Hoff et al. "Whole-genome annotation with BRAKER". In: *Methods in molecular biology (Clifton, NJ)* 1962 (2019), p. 65.

[31] Tomas Bruna, Alexandre Lomsadze, and Mark Borodovsky. "GeneMark-EP and-EP+: automatic eukaryotic gene prediction supported by spliced aligned proteins". In: *bioRxiv* (2020), pp. 2019–12.

[32] Benjamin Buchfink, Chao Xie, and Daniel H Huson. "Fast and sensitive protein alignment using DIAMOND". In: *Nature methods* 12.1 (2015), pp. 59–60.

[33] Osamu Gotoh. "A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence". In: *Nucleic acids research* 36.8 (2008), pp. 2630–2638.

[34] Hiroaki Iwata and Osamu Gotoh. "Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features". In: *Nucleic acids research* 40.20 (2012), e161–e161.

[35] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[36] Derek W Barnett et al. "BamTools: a C++ API and toolkit for analyzing and managing BAM files". In: *Bioinformatics* 27.12 (2011), pp. 1691–1692.

[37] Alexandre Lomsadze, Paul D Burns, and Mark Borodovsky. "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm". In: *Nucleic acids research* 42.15 (2014), e119–e119.

[38] Mario Stanke et al. "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources". In: *BMC bioinformatics* 7.1 (2006), pp. 1–11.

[39] Mario Stanke et al. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding". In: *Bioinformatics* 24 (5 Mar. 2008), pp. 637–644. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btn013. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btn013.

[40] Alexandre Lomsadze et al. "Gene identification in novel eukaryotic genomes by self-training algorithm". In: *Nucleic Acids Research* 33 (20 Nov. 2005), pp. 6494–6506. ISSN: 03051048. DOI: 10.1093/nar/gki937. URL: https://academic.oup.com/nar/article/33/20/6494/1082033.

[41] Carson Holt and Mark Yandell. "MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects". In: *BMC Bioinformatics* 12 (1 Dec. 2011), p. 491. ISSN: 14712105. DOI: 10.1186/1471-2105-12-491. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-491.

[42] Ian Korf. "Gene finding in novel genomes". In: *BMC Bioinformatics* 5 (1 May 2004), p. 59. ISSN: 14712105. DOI: 10.1186/1471-2105-5-59. URL: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-5-59.

[43] Alaina Shumate and Steven L Salzberg. "Liftoff: accurate mapping of gene annotations". In: *Bioinformatics* (Dec. 2020). Ed. by Alfonso Valencia. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa1016. URL: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btaa1016/6035128.

[44] Mao Sheng Chen et al. "De novo genome assembly and Hi-C analysis reveal an association between chromatin architecture alterations and sex differentiation in the woody plant Jatropha curcas". In: *GigaScience* 9 (2 Feb. 2020), pp. 1–12. ISSN: 2047217X. DOI: 10.1093/GIGASCIENCE/GIAA009. URL: http://orcid.org/0000-0002-8339-1857http://orcid.org/0000-0001-6045-5865.

[45] Pedro Queirós et al. "Mantis: flexible and consensus-driven genome annotation". In: *GigaScience* 10.6 (2021), giab042.

[46] Jaime Huerta-Cepas et al. "EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D309–D314. ISSN: 13624962. DOI: 10.1093/nar/gky1085. URL: https://academic.oup.com/nar/article/47/D1/D309/5173662.

[47] D. H. Haft et al. "TIGRFAMs: A protein family resource for the functional identification of proteins". In: *Nucleic Acids Research* 29 (1 Jan. 2001), pp. 41–43. ISSN: 03051048. DOI: 10.1093/nar/29.1.41. URL: www.tigr.org/CMR2/db_assignmentextver2.html.

[48] Takuya Aramaki et al. "KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold". In: *Bioinformatics* 36 (7 Apr. 2020). Ed. by Alfonso Valencia, pp. 2251–2252. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btz859. URL: https://academic.oup.com/bioinformatics/article/36/7/2251/5631907.

[49] Sara El-Gebali et al. "The Pfam protein families database in 2019". In: *Nucleic Acids Research* 47 (D1 Jan. 2019), pp. D427–D432. ISSN: 13624962. DOI: 10.1093/nar/gky995. URL: https://pfam.xfam.org.

[50] Fernando Mora-Márquez et al. "TOA: A software package for automated functional annotation in non-model plant species". In: *Molecular Ecology Resources* 21 (2 Feb. 2021), pp. 621–636. ISSN: 1755-098X. DOI: 10.1111/1755-0998.13285. URL: https://onlinelibrary.wiley.com/doi/10.1111/1755-0998.13285.

[51] Michiel Van Bel et al. "PLAZA 4.0: An integrative resource for functional, evolutionary and comparative plant genomics". In: *Nucleic Acids Research* 46 (D1 Jan. 2018), pp. D1190–D1196. ISSN: 13624962. DOI: 10.1093/nar/gkx1002. URL: https://academic.oup.com/nar/article/46/D1/D1190/4561641.

[52] Yuki Moriya et al. "KAAS: An automatic genome annotation and pathway reconstruction server". In: *Nucleic Acids Research* 35 (SUPPL.2 July 2007), W182–W185. ISSN: 03051048. DOI: 10.1093/nar/gkm321. URL: http://www.genome.jp/kegg/kaas/.

[53] David M Emms and Steven Kelly. "OrthoFinder: phylogenetic orthology inference for comparative genomics". In: *Genome biology* 20.1 (2019), pp. 1–14.

[54] David M Emms. "OrthoFinder best practices". In: *OrthoFinder Tutorials Phylogenetic orthology inference for comparative genomics* (2019), p. 1. URL: https://davidemms.github.io/orthofinder_tutorials/orthofinder-best-practices.html (visited on 06/22/2021).

[55] Tanya Z Berardini et al. "The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome". In: *genesis* 53.8 (2015), pp. 474–485.

[56] Yan-Jing Liu, Xiao-Ru Wang, and Qing-Yin Zeng. "De novo assembly of white poplar genome and genetic diversity of white poplar population in Irtysh River basin in China". In: *Science China Life Sciences* 62.5 (2019), pp. 609–618.

[57] Suyun Wei, Yonghua Yang, and Tongming Yin. "The chromosome-scale assembly of the willow genome provides insight into Salicaceae genome evolution". In: *Horticulture research* 7.1 (2020), pp. 1–12.

[58] Paris Veltsos et al. "Early sex-chromosome evolution in the diploid dioecious plant Mercurialis annua". In: *Genetics* 212.3 (2019), pp. 815–835.

[59] Jin Liu et al. "The chromosome-based rubber tree genome provides new insights into spurge genome evolution and rubber biosynthesis". In: *Molecular plant* 13.2 (2020), pp. 336–350.

[60] Saakshi Jalali et al. "Exploitation of Hi-C sequencing for improvement of genome assembly and in-vitro validation of differentially expressing genes in Jatropha curcas L." In: *3 Biotech* 10.3 (2020), pp. 1–9.

[61] Maximo Rivarola et al. "Castor bean organelle genome sequencing and worldwide genetic diversity analysis". In: *PloS one* 6.7 (2011), e21743.

[62] Diego Villanueva-Mejía et al. "Transcriptome of Plukenetia volubilis reveals important genes in the Polyunsaturated Fatty Acids pathways during seed maduration". In: (Sept. 2017).

[63] Eliah G Overbey et al. "NASA GeneLab RNA-seq consensus pipeline: standardized processing of short-read RNA-seq data". In: *Iscience* 24.4 (2021), p. 102361.

[64] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29 (1 Jan. 2013), pp. 15–21. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/bts635. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts635.

[65] Bo Li and Colin N. Dewey. "RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome". In: *BMC Bioinformatics* 12 (1 Aug. 2011), p. 323. ISSN: 14712105. DOI: 10.1186/1471-2105-12-323. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-323.

[66] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15 (12 Dec. 2014), p. 550. ISSN: 1474760X. DOI: 10.1186/s13059-014-0550-8. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

[67] Kayla A Johnson and Arjun Krishnan. "Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data". In: *bioRxiv* (2020).

[68] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (2010), pp. 139–140.

[69] Siliang Chen et al. "Identification of crucial genes in abdominal aortic aneurysm by WGCNA". In: *PeerJ* 7 (2019), e7873.

[70] Peter Langfelder and Steve Horvath. "WGCNA: an R package for weighted correlation network analysis". In: *BMC bioinformatics* 9.1 (2008), pp. 1–13.

[71] Peter Langfelder and Steve Horvath. "2.a Automatic network construction and module detection". In: *Tutorial for the WGCNA package for R:I. Network analysis of liver expression data in female mice* (2014). URL: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/ (visited on 06/22/2021).

[72] Bin Zhang and Steve Horvath. "A general framework for weighted gene co-expression network analysis". In: *Statistical applications in genetics and molecular biology* 4.1 (2005).

[73] Adrian Alexa and Jörg Rahnenführer. "Gene set enrichment analysis with topGO". In: *Bioconductor Improv* 27 (2009), pp. 1–26.

[74] Patricia P. Chan and Todd M. Lowe. *tRNAscan-SE: Searching for tRNA genes in genomic sequences*. 2019. DOI: 10.1007/978-1-4939-9173-0_1. URL: https://link.springer.com/protocol/10.1007/978-1-4939-9173-0_1.

[75] Ioanna Kalvari et al. "Rfam 14: Expanded coverage of metagenomic, viral and microRNA families". In: *Nucleic Acids Research* 49 (D1 Jan. 2021), pp. D192–D200. ISSN: 13624962. DOI: 10.1093/nar/gkaa1047. URL: https://rfam.org/covid-19..

[76] E. P. Nawrocki and S. R. Eddy. "Infernal 1.1: 100-fold faster RNA homology searches". In: *Bioinformatics* 29 (22 Nov. 2013), pp. 2933–2935. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btt509. URL: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt509.

[77] Karin Lagesen et al. "RNAmmer: consistent and rapid annotation of ribosomal RNA genes". In: *Nucleic acids research* 35.9 (2007), pp. 3100–3108.

[78] Mingcheng Wang et al. "High-quality genome assembly of an important biodiesel plant, Euphorbia lathyris L". In: *DNA Research* (2021).

[79] Wenquan Wang et al. "Cassava genome from a wild ancestor to cultivated varieties". In: *Nature communications* 5.1 (2014), pp. 1–9.

[80] Jianjun Lu et al. "A Chromosome-level Assembly of a Wild Castor Genome Provides New Insights into the Adaptive Evolution in a Tropical Desert". In: *bioRxiv* (2021).

[81] Lin Zhang et al. "Tung tree (Vernicia fordii) genome provides a resource for understanding genome evolution and improved oil production". In: *Genomics, proteomics & bioinformatics* 17.6 (2019), pp. 558–575.

[82] Jungmin Ha et al. "Genome sequence of Jatropha curcas L., a non-edible biodiesel plant, provides a resource to improve seed-related traits". In: *Plant biotechnology journal* 17.2 (2019), pp. 517–530.

[83] Ana Mendes et al. "bZIP67 regulates the omega-3 fatty acid content of Arabidopsis seed oil by activating fatty acid desaturase3". In: *The Plant Cell* 25.8 (2013), pp. 3104–3116.

[84] Fiona M Bryant et al. "Basic LEUCINE ZIPPER TRANSCRIPTION FACTOR67 transactivates DELAY OF GERMINATION1 to establish primary seed dormancy in Arabidopsis". In: *The Plant Cell* 31.6 (2019), pp. 1276–1288.

[85] Que Kong and Wei Ma. "WRINKLED1 as a novel 14-3-3 client: function of 14-3-3 proteins in plant lipid metabolism". In: *Plant signaling & behavior* 13.8 (2018), e1482176.

[86] Mei He et al. "Plant unsaturated fatty acids: Biosynthesis and regulation". In: *Frontiers in Plant Science* 11 (2020), p. 390.

[87] Jinye Mu et al. "LEAFY COTYLEDON1 is a key regulator of fatty acid biosynthesis in Arabidopsis". In: *Plant physiology* 148.2 (2008), pp. 1042–1054.

[88] Mingxun Chen et al. "TRANSPARENT TESTA8 inhibits seed fatty acid accumulation by targeting several seed development regulators in Arabidopsis". In: *Plant Physiology* 165.2 (2014), pp. 905–916.

[89] Zhong Wang et al. "TRANSPARENT TESTA 2 regulates embryonic fatty acid biosynthesis by targeting FUSCA 3 during the early developmental stage of A rabidopsis seeds". In: *The Plant Journal* 77.5 (2014), pp. 757–769.

[90] Ge Song et al. "The WRKY6 transcription factor affects seed oil accumulation and alters fatty acid compositions in Arabidopsis thaliana". In: *Physiologia plantarum* 169.4 (2020), pp. 612–624.

[91] Chengxiang Li et al. "Site-specific phosphorylation of TRANSPARENT TESTA GLABRA1 mediates carbon partitioning in Arabidopsis seeds". In: *Nature communications* 9.1 (2018), pp. 1–13.

[92] Que Kong et al. "Molecular basis of plant oil biosynthesis: Insights gained from studying the WRINKLED1 transcription factor". In: *Frontiers in plant science* 11 (2020), p. 24.

[93] Liyuan Chen et al. "Arabidopsis BPM proteins function as substrate adaptors to a cullin3-based E3 ligase to affect fatty acid metabolism in plants". In: *The Plant Cell* 25.6 (2013), pp. 2253–2264.

[94] Wei Ma et al. "14-3-3 protein mediates plant seed oil biosynthesis through interaction with AtWRI1". In: *The Plant Journal* 88.2 (2016), pp. 228–235.

[95] Mi Jung Kim, In-Cheol Jang, and Nam-Hai Chua. "The mediator complex MED15 subunit mediates activation of downstream lipid-related genes by the WRINKLED1 transcription factor". In: *Plant physiology* 171.3 (2016), pp. 1951–1964.

[96] Zhiyang Zhai, Hui Liu, and John Shanklin. "Phosphorylation of WRINKLED1 by KIN10 results in its proteasomal degradation, providing a link between energy homeostasis and lipid biosynthesis". In: *The Plant Cell* 29.4 (2017), pp. 871–889.

[97] Zhiyang Zhai et al. "Trehalose 6-phosphate positively regulates fatty acid synthesis by stabilizing WRINKLED1". In: *The Plant Cell* 30.10 (2018), pp. 2616–2627.

[98] Bohan Liu et al. "The SPATULA transcription factor regulates seed oil content by controlling seed specific genes in Arabidopsis thaliana". In: *Plant Growth Regulation* 82.1 (2017), pp. 111–121.

[99] Ning Liu et al. "Overexpression of WAX INDUCER1/SHINE1 gene enhances wax accumulation under osmotic stress and oil synthesis in Brassica napus". In: *International journal of molecular sciences* 20.18 (2019), p. 4435.

[100] Yumei Yin et al. "AmDREB2C, from Ammopiptanthus mongolicus, enhances abiotic stress tolerance and regulates fatty acid composition in transgenic Arabidopsis". In: *Plant Physiology and Biochemistry* 130 (2018), pp. 517–528.

[101] Xiaodong Wang et al. “New insights into the genetic networks affecting seed fatty acid concentrations in Brassica napus”. In: *BMC plant biology* 15.1 (2015), pp. 1–18.

[102] Dahee An and Mi Chung Suh. “Overexpression of Arabidopsis WRI1 enhanced seed mass and storage oil content in Camelina sativa”. In: *Plant Biotechnology Reports* 9.3 (2015), pp. 137–148.

[103] Thomas Vanhercke et al. “Synergistic effect of WRI1 and DGAT1 coexpression on triacylglycerol biosynthesis in plants”. In: *Febs Letters* 587.4 (2013), pp. 364–369.

[104] Bo Shen et al. “Expression of ZmLEC1 and ZmWRI1 increases seed oil production in maize”. In: *Plant physiology* 153.3 (2010), pp. 980–987.

[105] Harrie Van Erp et al. “Multigene engineering of triacylglycerol metabolism boosts seed oil content in Arabidopsis”. In: *Plant Physiology* 165.1 (2014), pp. 30–36.

[106] Emelie Ivarson et al. “Effects of overexpression of WRI1 and hemoglobin genes on the seed oil content of Lepidium campestre”. In: *Frontiers in plant science* 7 (2017), p. 2032.

[107] Jian Ye et al. “Overexpression of a transcription factor increases lipid content in a woody perennial Jatropha curcas”. In: *Frontiers in plant science* 9 (2018), p. 1479.

[108] Sota Takahashi et al. “Identification of Transcription Factors and the Regulatory Genes Involved in Triacylglycerol Accumulation in the Unicellular Red Alga Cyanidioschyzon merolae”. In: *Plants* 10.5 (2021), p. 971.

[109] AK Padham et al. “Functional Analysis of Arabidopsis Lipase Genes”. In: *Advanced Research on Plant Lipids*. Springer, 2003, pp. 263–266.

[110] Masatake Kanai et al. “Soybean (Glycine max L.) triacylglycerol lipase GmSDP1 regulates the quality and quantity of seed oil”. In: *Scientific reports* 9.1 (2019), pp. 1–10.

[111] Aner Gurvitz et al. “Peroxisomal Δ3-cis-Δ2-trans-enoyl-CoA isomerase encoded by ECI1 is required for growth of the yeast Saccharomyces cerevisiae on unsaturated fatty acids”. In: *Journal of Biological Chemistry* 273.47 (1998), pp. 31366–31374.

[112] Qikui Wu et al. “Transcriptome analysis of metabolic pathways associated with oil accumulation in developing seed kernels of Styrax tonkinensis, a woody biodiesel species”. In: *BMC plant biology* 20.1 (2020), pp. 1–17.

1 Methodology related

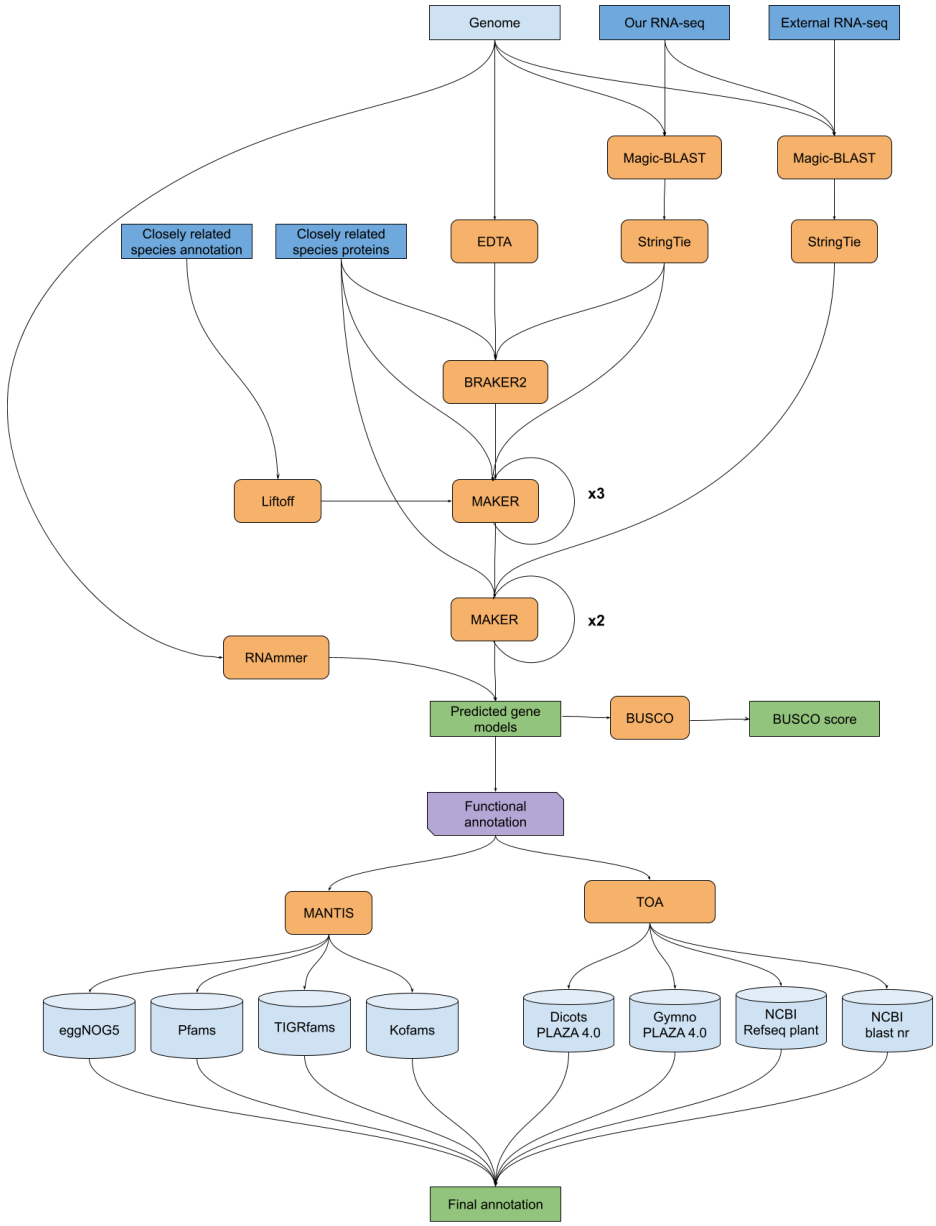

Figure 1. Gene prediction and Genome annotation schematic of pipeline. In orange boxes software is represented, in dark blue boxes is the base data used to make homology searches. In green boxes are the results of the pipeline.

2 Annotation related

Table 1. Cross species comparison of ACC and PEPC genes with seed oil and protein content. Gene numbers were taken from [1]. Seed oil and protein content were taken from the respective references. *P. volubilis* data comes form this work.

Species	ACC genes	PEPC genes	Oil content	Protein content	Reference
Vernicia fordii	9	3	60%	5%	[1]
Jatropha curcas	6	4	27%	32%	[2]
Ricinus communis	6	7	36%	25%	[3]
Arabidopsis thaliana	4	3	30%	30%	[4]
Sesamum indicum	6	4	49%	23%	[5]
Glycine max	10	16	20%	40%	[1]
Plukenetia volubilis	8	3	50%	27%	-

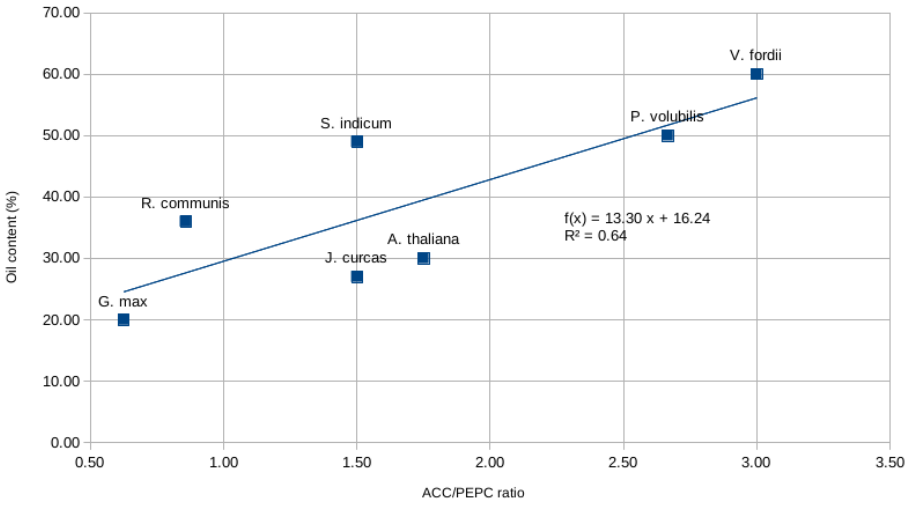

Figure 2. ACC/PEPC ratio vs seed Oil content (%)

3 Differential expression analysis

Table 2. Number of differentially expressed genes between seed stages. Most of the change is detected between E2 and E3 or E4 or E5. .

	E2	E3	E4	E5
E2	-	10463	11091	12581
E3	10463	-	245	3200
E4	11091	245	-	3144
E5	12581	3200	3144	-

4 Co-expression network analysis

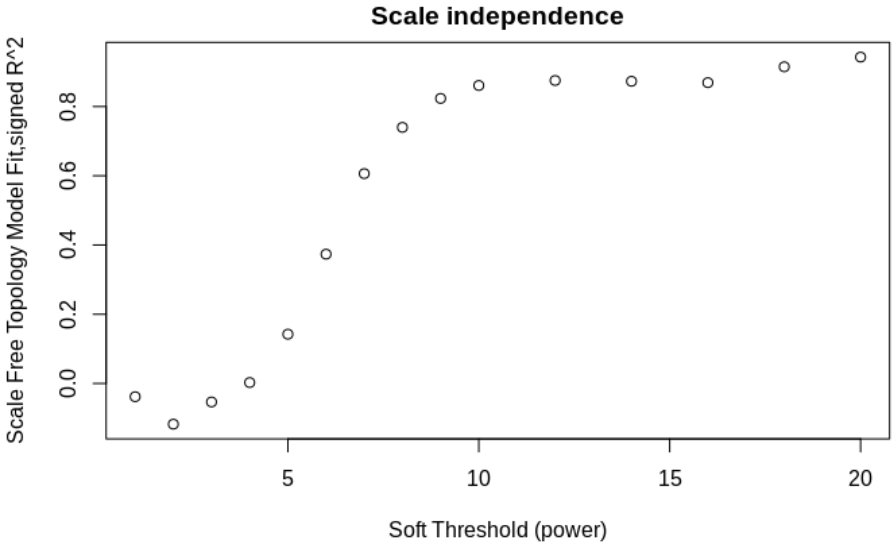

Figure 3. Scale-free fitindex (y-axis) as a function of the soft-thresholding power (x-axis). In this work a power 12, which is c.a the lowest power for which the scale-free topology fitindex curve flattens out upon reaching a high value. This was done following recomendations from [6]

Table 3. Cyan module GO enrichment analysis result from topGO, table was filtered to show interest GO terms.

GO.ID	Term	Annotated	Significant	Expected	elimFisher	GO_type
GO:0034219	carbohydrate transmembrane transport	31	6	0.48	6.8E-06	BP
GO:0006629	lipid metabolic process	1148	58	17.63	2.4E-05	BP
GO:0002933	lipid hydroxylation	6	3	0.09	7E-05	BP
GO:0008610	lipid biosynthetic process	574	25	8.81	0.00022	BP
GO:0005975	carbohydrate metabolic process	1083	41	16.63	0.00041	BP
GO:0006869	lipid transport	146	9	2.24	0.00045	BP
GO:0008289	lipid binding	361	19	5.67	0.00025	MF

Table 4. Blue module GO enrichment analysis result from topGO, table was filtered to show interest GO terms.

GO.ID	Term	Annotated	Significant	Expected	elimFisher	GO_type
GO:0009507	chloroplast	985	203	137.87	3.9E-07	CC
GO:0009706	chloroplast inner membrane	39	17	5.46	6.8E-06	CC
GO:0005758	mitochondrial intermembrane space	12	8	1.68	4.3E-05	CC
GO:0019915	lipid storage	18	8	1.50	4.7e-05	BP
GO:0006636	UFA biosynthetic proccess	10	4	0.83	0.00674	BP

Table 5. Turquoise module GO enrichment analysis result from topGO, table was filtered to show interest GO terms.

GO.ID	Term	Annotated	Significant	Expected	elimFisher	GO_type
GO:0005975	carbohydrate metabolic process	1083	153	59.46	1.7E-10	BP
GO:0008610	lipid biosynthetic process	574	97	31.52	8.8E-08	BP
GO:0009507	chloroplast	985	126	81.99	0.00051	CC
GO:0009941	chloroplast envelope	192	27	15.98	0.00496	CC
GO:0009570	chloroplast stroma	185	26	15.4	0.0058	CC

Table 6. Tissue/stage of all the RNA-seq libraries in this study. For our libraries and the libraries from [7] a DAP(Days after pollination) temporal assignation for the samples was not found. The assigned PCA cluster is specified. It is worth to mention that the 60 DAP seed libraries form [8] and the >120 DAP seeds from [9] were quite similar(the green circles in the close to the midle part of PC1 in Figure) even though they differ by a factor of two in DAP, this suggests that the seeds from [9] had a longer seed development than the seeds from [8], this may be due to strain differences.

Tissue/Stage	Stage	PCA cluster	Reference
Seed	5-10 DAP	Red	[10]
Seed(SI-1)	0-10 DAP	Red	[9]
Seed(SI-1)	0-10 DAP	Red	[9]
Seed(SI-1)	0-10 DAP	Red	[9]
Seed(SI-2)	20-40 DAP	Red	[9]
Seed(SI-2)	20-40 DAP	Red	[9]
Seed(SI-2)	20-40 DAP	Red	[9]
Seed(SI-3)	50-70 DAP	Red	[9]
Seed(SI-3)	50-70 DAP	Red	[9]
Seed(SI-3)	50-70 DAP	Red	[9]
Seed(SI-4)	80-110 DAP	Red	[9]
Seed(SI-4)	80-110 DAP	Red	[9]
Seed(SI-4)	80-110 DAP	Red	[9]
Seed(E2)	-	Red	Our data
Seed(E2)	-	Red	Our data
Seed(E2)	-	Red	Our data
Seed	60 DAP	Green	[8]
Seed	60 DAP	Green	[8]
Seed	60 DAP	Green	[8]
Seed(SI-5)	>120 DAP	Green	[9]
Seed(SI-5)	>120 DAP	Green	[9]
Seed(SI-5)	>120 DAP	Green	[9]
Seed(E3)	-	Green	Our data
Seed(E3)	-	Green	Our data
Seed(E4)	-	Green	Our data
Seed(E4)	-	Green	Our data
Seed(E4)	-	Green	Our data
Seed(E5)	-	Green	Our data
Seed(E5)	-	Green	Our data
Seed(E5)	-	Green	Our data
Female Flower	60 DAP	Blue	[8]
Female Flower	60 DAP	Blue	[8]
Female Flower	60 DAP	Blue	[8]
Female inflorescence bud	-	Blue	[7]
Fruit	60 DAP	Blue	[8]
Fruit	60 DAP	Blue	[8]
Fruit	60 DAP	Blue	[8]
Leaves	60 DAP	Blue	[8]
Leaves	60 DAP	Blue	[8]
Leaves	60 DAP	Blue	[8]
Male Flower	60 DAP	Blue	[8]
Male Flower	60 DAP	Blue	[8]
Male Flower	60 DAP	Blue	[8]
Male inflorescence bud	-	Blue	[7]
Root	60 DAP	Blue	[8]
Root	60 DAP	Blue	[8]
Root	60 DAP	Blue	[8]
Shoot apex	60 DAP	Blue	[8]
Shoot apex	60 DAP	Blue	[8]
Shoot apex	60 DAP	Blue	[8]
Stem	60 DAP	Blue	[8]

Stem	60 DAP	Blue	[8]
Stem	60 DAP	Blue	[8]
Flowers(PF)	-	Blue	Our data
Leaves(PH)	-	Blue	Our data

Table 7. Total genes and interest genes per module. "FA genes" refers to FA and TAG biosynthesis related genes. Regulator genes correspond to ortholog regulators found in the modules.

Module	Genes	FA genes	Regulator genes
blue	4373	32	6
cyan	843	21	5
turquoise	2834	30	2
darkgreen	233	1	2
darkolivegreen	327	0	2
greenyellow	2626	22	1
mediumpurple3	504	2	1
darkgrey	401	0	1
darkred	508	0	1
yellowgreen	1114	10	0
lightgreen	286	5	0
pink	1192	5	0
red	799	4	0
green	914	3	0
lightyellow	280	3	0
white	176	3	0
darkorange	252	2	0
skyblue	158	2	0
violet	382	2	0
paleturquoise	133	1	0
sienna3	108	1	0
darkmagenta	113	0	0
lightcyan1	38	0	0
lightsteelblue1	40	0	0
steelblue	135	0	0

Table 8. Summary of genes found in the "cyan" coexpression module.

Gene	Type	Counts
KASII	FA gene	2
LACS	FA gene	2
PK	FA gene	2
KASI	FA gene	1
DGAT	FA gene	1
DHLAT	FA gene	1
FATB	FA gene	1
GDPH	FA gene	1
KAR	FA gene	1
LPCAT	FA gene	1
PDAT	FA gene	1
ACC	FA gene	1
b-PDHC	FA gene	1
DAG-CPT	FA gene	1
FAD3/7/8	FA gene	1
FATA	FA gene	1
LPD	FA gene	1
PP	FA gene	1

bZIP67	Regulator	1
FUS3	Regulator	1
GL2	Regulator	1
LEC1	Regulator	1
WRI1	Regulator	1

Table 9. Summary of genes found in the "blue" coexpression module.

Gene	Type	Counts
OLE	FA gene	6
LPAAT	FA gene	4
LACS	FA gene	3
DGAT	FA gene	2
FATB	FA gene	2
KAR	FA gene	2
PLA2	FA gene	2
PPC	FA gene	2
DHLAT	FA gene	1
GDPH	FA gene	1
LPCAT	FA gene	1
PDAT	FA gene	1
CCT	FA gene	1
KAT	FA gene	1
PEPC	FA gene	1
PLC	FA gene	1
DREB2C	Regulator	2
WRKY6	Regulator	2
BPM1	Regulator	1
KIN10	Regulator	1

Table 10. Summary of genes found in the "turquoise" coexpression module.

Gene	Type	Counts
PK	FA gene	3
LACS	FA_gene	2
DGAT	FA_gene	2
FATB	FA_gene	2
SAD-ACP	FA_gene	2
GPAT	FA_gene	2
MFP2	FA_gene	2
FAD2/6	FA_gene	2
a-PDHC	FA_gene	2
LPAAT	FA_gene	1
KAR	FA_gene	1
DHLAT	FA_gene	1
GDPH	FA_gene	1
LPCAT	FA_gene	1
KAT	FA_gene	1
KASII	FA_gene	1
FAD3/7/8	FA_gene	1
ACX	FA_gene	1
ACDM	FA_gene	1
CK	FA_gene	1
WRKY6	Regulator	1
SHN1	Regulator	1

References

[1] Lin Zhang et al. "Tung tree (Vernicia fordii) genome provides a resource for understanding genome evolution and improved oil production". In: *Genomics, proteomics & bioinformatics* 17.6 (2019), pp. 558–575.

[2] Azza A Abou-Arab and Ferial M Abu-Salem. "Nutritional quality of Jatropha curcas seeds and effect of some physical and chemical treatments on their antinutritional factors". In: *African Journal of Food Science* 4.3 (2010), pp. 93–103.

[3] TO Akande AA Odunsi, OS Olabode, and TK Ojediran. "Physical and nutrient characterisation of raw and processed castor (Ricinus communis L.) seeds in Nigeria". In: *World Journal of Agricultural Sciences* 8.1 (2012), pp. 89–95.

[4] Masatake Kanai et al. "Extension of oil biosynthesis during the mid-phase of seed development enhances oil content in Arabidopsis seeds". In: *Plant biotechnology journal* 14.5 (2016), pp. 1241–1250.

[5] AH Bahkali, MA Hussain, and AY Basahy. "Protein and oil composition of sesame seeds (Sesamum indicum, L.) grown in the Gizan area of Saudi Arabia". In: *International journal of food sciences and nutrition* 49.6 (1998), pp. 409–414.

[6] Peter Langfelder and Steve Horvath. "2.a Automatic network construction and module detection". In: *Tutorial for the WGCNA package for R:I. Network analysis of liver expression data in female mice* (2014). URL: https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/ (visited on 06/22/2021).

[7] Qiantang Fu et al. "De novo transcriptome assembly and comparative analysis between male and benzyladenine-induced female inflorescence buds of Plukenetia volubilis". In: *Journal of Plant Physiology* 221 (December 2017 2018), pp. 107–118. ISSN: 01761617. DOI: 10.1016/j.jplph.2017.12.006. URL: https://doi.org/10.1016/j.jplph.2017.12.006.

[8] Xiao Di Hu et al. "De novo transcriptome assembly of the eight major organs of Sacha Inchi (Plukenetia volubilis) and the identification of genes involved in -linolenic acid metabolism". In: *BMC Genomics* 19 (1 2018), pp. 1–14. ISSN: 14712164. DOI: 10.1186/s12864-018-4774-y.

[9] Guo Liu et al. "Transcriptome analyses reveals the dynamic nature of oil accumulation during seed development of Plukenetia volubilis L." In: *Scientific Reports* 10 (1 2020), pp. 1–17. ISSN: 20452322. DOI: 10.1038/s41598-020-77177-w. URL: https://doi.org/10.1038/s41598-020-77177-w.

[10] Xiaojuan Wang et al. "Transcriptome analysis of Sacha Inchi (Plukenetia volubilis L.) seeds at two developmental stages". In: *BMC Genomics* 13 (1 2012). ISSN: 14712164. DOI: 10.1186/1471-2164-13-716.