forests

MDPI

Article

Transcriptome Analysis Reveals Putative Genes Involved in the Lipid Metabolism of Chaulmoogra Oil Biosynthesis in *Carpotroche brasiliensis* (Raddi) A.Gray, a Tropical Tree Species

Letícia Maróstica de Vasconcelos [1,*], Flora Bittencourt [2], Ramon Oliveira Vidal [3], Edson Mario de Andrade Silva [4], Eduardo Almeida Costa [5], Fabienne Micheli [6], Matias Kirst [7], Carlos Priminho Pirovani [8] and Fernanda Amato Gaiotto [8]

1 Programa de Pós-Graduação em Genética e Biologia Molecular, Centro de Biotecnologia e Genética, Universidade Estadual de Santa Cruz (UESC), Ilhéus 45662-901, Brazil
2 Programa de Pós-Graduação de Estudos em Cultura e Território, Universidade Federal do Tocantins (UFT), Araguaína 77824-838, Brazil
3 Max-Delbrück-Center for Molecular Medicine (MDC), Berlin Institute for Medical Systems Biology, 13125 Berlin, Germany
4 Programa Interunidades de Pós-Graduação em Bioinformática, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte 31270-901, Brazil
5 Centro de Computação Avançada e Multidisciplinar (CCAM), Universidade Estadual de Santa Cruz (UESC), Ilhéus 45662-901, Brazil
6 UMR AGAP, CIRAD, 34398 Montpellier, France
7 School of Forest Resources & Conservation, University of Florida, Gainesville, FL 32603, USA
8 Centro de Biotecnologia e Genética, Universidade Estadual de Santa Cruz (UESC), Ilhéus 45662-901, Brazil
* Correspondence: lmvasconcelos@uesc.br or leticia.ufrb@gmail.com

Citation: Vasconcelos, L.M.d.; Bittencourt, F.; Vidal, R.O.; Silva, E.M.d.A; Costa, E.A.; Micheli, F.; Kirst, M.; Pirovani, C.P.; Gaiotto, F.A. Transcriptome Analysis Reveals Putative Genes Involved in the Lipid Metabolism of Chaulmoogra Oil Biosynthesis in *Carpotroche brasiliensis*, a Tropical Tree Species. *Forests* **2022**, *13*, 1806. https://doi.org/10.3390/f13111806

Academic Editor: Evi Alizoti

Received: 18 September 2022
Accepted: 24 October 2022
Published: 29 October 2022

Abstract: Chaulmoogra oil is found in the seeds of *Carpotroche brasiliensis* (Raddi) Endl. (syn. Mayna brasiliensis Raddi), an oil tree of the Achariaceae family and native to Brazil's Atlantic Forest biome, which is considered the fifth most important biodiversity hotspot in the world. Its main constituents are cyclopentenic fatty acids. Chaulmoogra oil has economic potential because of its use in the cosmetics industry and as a drug with anti-tumor activity. The mechanisms related to the regulation of oil biosynthesis in *C. brasiliensis* seeds are not fully understood, especially from a tissue-specific perspective. In this study, we applied a de novo transcriptomic approach to investigate the transcripts involved in the lipid pathways of *C. brasiliensis* and to identify genes involved in lipid biosynthesis. Comparative analysis of gene orthology, expression analysis and visualization of metabolic lipid networks were performed, using data obtained from high-throughput sequencing (RNAseq) of 24 libraries of vegetative and reproductive tissues of *C. brasiliensis*. Approximately 10.4 million paired-end reads (Phred (Q) > 20) were generated and re-assembled into 107,744 unigenes, with an average length of 340 base pairs (bp). The analysis of transcripts from different tissues identified 1131 proteins involved in lipid metabolism and transport and 13 pathways involved in lipid biosynthesis, degradation, transport, lipid bodies, and lipid constituents of membranes. This is the first transcriptome study of *C. brasiliensis*, providing basic information for biotechnological applications of great use for the species, which will help understand chaulmoogra oil biosynthesis.

Keywords: oil biosynthesis; gene expression; gene orthology; unsaturated fatty acids; fatty acid desaturase; tropical rainforest

1. Introduction

Carpotroche brasiliensis (Raddi) Endl. (syn. Mayna brasiliensis Raddi) is a tree of the Achariaceae family; the seeds are composed of 70% oil [1], including mainly linear cyclic fatty acids and triacylglycerols [2]. Chaulmoogra oil is composed mainly of cyclopentenyl fatty acids (AGCs) [3], and to a lesser extent, fatty acids such as palmitic, palmitoleic, stearic, oleic, and linoleic acids [4]. The knowledge about the biosynthetic pathway of fatty acids present in the seeds of *C. brasiliensis* is still scarce. Existing information is restricted to the chemical reactions and components associated with its production [4].

Cyclopentenyl fatty acids present in chaulmoogra oil form a well-defined chemical structure comprising a 5-membered unsaturated ring attached to a linear side chain terminated by a carboxylic acid [5]. Evidence suggests that AGCs arise from the elongation of the aleprolic acid chain, possibly obtained from cyclopentenylglycine. The conversion of cyclopentenylglycine into aleprolic acid may occur by transamination and oxidative decarboxylation. Thereby, activated aleprolic acid is stretched to cyclopentenyl fatty acids [6]. These findings indicate that (i) cyclopentenylglycine is the precursor of cyclopentenyl fatty acids, and (ii) unsaturated linear chain acids undergo cyclization. Studies of the biosynthesis pathway of cyclopentenic fatty acids in the leaves and chloroplasts of *Culoncoha echinata* (Oncobeae) indicate that cyclopentenic fatty acids are synthesized from aspartate and pyruvate or glutamate and acetate [2]. Therefore, the chaulmoogra oil biosynthesis pathways suggested in previous studies are quite controversial.

Given the scarcity of information about the components of the chaulmoogra oil biosynthetic pathway, identifying genes/transcripts encoding enzymes involved in synthesizing these fatty acids may provide valuable information for future studies. The discovery of genes encoding enzymes involved in plants specialized metabolism (such as the synthesis of chaulmoogra oil) represents a unique challenge. Complex enzymatic networks can produce several byproducts instead of a single compound [7]. An additional complication is that different plant species vary considerably in their oil content and fatty acid composition [8]. This is particularly true of the pathway of cyclopentenic acids, where there are few studies. Additionally, chaulmoogra oil-producing species do not have genomic information available in public databases, making studies related to omics difficult. Such studies are relevant for accessing and understanding biosynthetic pathways, identifying new genes and understanding the mechanisms of oil accumulation.

Our study aimed to uncover the mechanisms involved in chaulmoogra oil biosynthesis, answering the following questions: (i) Which lipid metabolism genes are present in different tissues of *C. brasiliensis*?and (ii) What is the level of these genes expressed in different tissues?

Here, we report the first transcriptome of the tropical tree *C. brasiliensis* and seek new insights into the molecular bases of lipid biosynthesis. Our study provides rich genetic information that will be useful for understanding the synthesis of chaulmoogra oil.

2. Materials and Methods

2.1. RNA-Sequencing and Analysis

2.1.1. Plant Material

C. brasiliensis trees were randomly chosen to compose the sample group, the source of the collections was carried out in cocoa farms with authorization from the cooperative of seed collectors of the agroforestry system Cabruca in the municipalities of Camamu and Maraú, Bahia, Brazil (Supplementary Table S1). Voucher specimens have been stored at the Santa Cruz State University herbarium (RH-Uesc 20315–RH-Uesc 20316), the species was identified by José Lima da Paixão. Based on the quality of the extracted RNA, 24 samples were selected for library development. The material was immediately frozen in liquid nitrogen and stored in an ultrafreezer at −80 °C before RNA extraction. Thus, each library was developed from the extraction of 1 tissue from 1 tree. The samples were

collected in triplicate to compose the 24 developed libraries, using 12 buds and 12 other plant tissues, namely: 2 leaves, 2 roots, 2 flowers, 3 seeds, pulp, and the skin pool of 3 fruits.

2.1.2. RNA Extraction and cDNA Library Preparation

The RNAqueous® Total RNA Isolation Kit-AM1912 was used to perform the total RNA extraction, following the manufacturer's recommendations. The integrity and amount of RNA sample was evaluated in a NanoDrop spectrophotometer (Nano-Drop, Wilmington, DE, USA), with 2% agarose gel and an Agilent TapeStation 2200 instrument, considering an RNA Integrity Number value > 5.

The cDNA libraries were built from 10 ng total RNA NEBNext Ultra RNA Library Prep Kit for Illumina (E7530S) and NEBNext Multiplex (E7335) Oligos for Illumina to develop cDNA libraries, following the manufacturer's protocols (Illumina, CA, USA). The libraries were quantified with the KAPA Library quantification kit with an Agilent 2100 Bioanalyzer. A total of 24 cDNA libraries were sequenced on Illumina MiSeq 2 × 250 bp paired-end. Raw sequencing data were deposited at NCBI Sequence Read Archive (SRA) under project number PRJNA858666.

2.1.3. Quality Control and Assembly

The Trimmomatic v.0.32 software (https://github.com/timflutre/trimmomatic, accessed on 18 October 2022) was used to verify the quality of the reads. We used a Phred score of Q > 20 as a threshold to trim low-quality bases from the ends of the reads. High-quality reads were used to assemble a de novo transcriptome using the Trinity v.2.1.1 software (https://github.com/trinityrnaseq/trinityrnaseq, accessed on 18 October 2022) following an analysis pipeline (Supplementary Figure S1).

2.1.4. Quality of Transcripts, Complete Transcripts, and Super Transcripts

The transcripts' quality and completeness were assessed using the Benchmarking Universal Single-Copy Orthologs (BUSCO 3.1.0) (https://github.com/openpaul/busco, accessed on 18 October 2022), based on ortholog groups from *Arabidopsis thaliana* and Solanaceae. The Busco software was chosen because it allowed us to detect the natural variation of conserved genes within clades, based on evolutionarily informed expectations of gene content of near-universal single-copy orthologs for a variety of eukaryotic clades. The indexing and alignment of transcripts was carried out using Bowtie2 2.2.5 (https://github.com/BenLangmead/bowtie, accessed on 18 October 2022). The trimmed transcripts of each library were aligned against the indexed reference and subsequently analyzed with RSEM (http://deweylab.github.io/RSEM/, accessed on 18 October 2022) to estimate the expression values normalized by transcripts per million (TPM) in each RNA-Seq library.

The TransDecoder software 5.5.0 (https://github.com/TransDecoder/TransDecoder/, accessed on 18 October 2022) was used to identify candidate coding regions for starting and ending transcript sequences. For this purpose, the Basic Local Alignment Search Tool (BLAST) was used versus the Uniref and Pfam 31.0 databases (http://pfam.xfam.org, accessed on 18 October 2022). This allowed predicting complete sequences of isoforms selected by TransDecoder, as well as protein sequences. Pfam is an extensive collection of protein families, represented by multi-sequence alignments, using hidden Markov chains (HMMs). In this sense, only complete sequences (i) with start and end sequence transcription, (ii) showing high similarity with cured sequences from the Uniref database, and (iii) with functional domains based on the statistical alignment of Pfam were considered.

The Trinity software (https://github.com/trinityrnaseq/trinityrnaseq, accessed on 18 October 2022) was used to reconstruct the primary transcript sequence and obtain super transcripts. This analysis recognizes unique and common sequence regions between isoforms and merges these isoforms into a single linear sequence. The super transcripts

are useful in the context of de novo free genomes, since they yield a gene sequence similar to what can be obtained when sequencing the genome.

2.1.5. Functional Annotation

The functional annotation of the *C. brasiliensis* transcript was generated with the EggNOG mapper database using HMM to obtain the EggNOG mapper 4.5 orthology data (Supplementary Figure S1). For all transcripts, the clusters of orthologous groups (COGs) and archaeal clusters of orthologous genes (arCOG) were assigned [9]. The sequences of *C. brasiliensis* proteins were also annotated based on the InterProScan databases (http://www.ebi.ac.uk/InterProScan/ accessed on 18 October 2022) and the Pannzer 2 tool (protein annotation using Z-score) [10]. These analyses recovered the EC protein codes, which were used in the KEGG mapper [11] for the identification of lipid pathways. Annotations with ARGOT_PPV greater than 0.35 were selected from the results obtained with Pannzer 2. The gene ontology (GO) categories of fruit and seed libraries were retrieved and formatted according to the native format of the WEGO platform (Web Gene Ontology Annotation Plot) (http://wego.genomics.org.cn/cgi-bin/wego/index.pl, accessed on 18 October 2022) to obtain a bar plot for categorization into cellular components, molecular functions, and biological processes.

2.2. Phylogenetic Analysis

The 18S ribosomal sequence of *C. brasiliensis* was identified by searching for putative homologs of the 18S *A. thaliana* sequence. The 18S transcription of *C. brasiliensis* was performed against the Non-Redundant Database (NR) of the National Center for Biotechnology Information (NCBI) to obtain the 18S data of species used in the phylogeny construction. Redundant sequences were excluded and at least one copy of each genus was maintained. Multiple nucleotide alignment (ClustalW) of 79 sequences was performed using the Mega X 10.1 software [12]. The model test was performed using MEGA X 10.1. Genetic relationships between accessions were calculated using a distance matrix obtained by the Tamura-Nei nucleotide substitution model [13]. The identified taxa were grouped based on the maximum likelihood method. The consistency of the clustering patterns was assessed by 1000 bootstrap replications. The generated dendrogram was edited with the Figtree v.1.4.4 software (http://tree.bio.ed.ac.uk/ accessed on 18 October 2022).

2.3. Analysis of Orthologous Gene Families

The ARALIP website platform was used to select gene families related to lipids in *A. thaliana* (http://aralip.plantbiology.msu.edu/pathways/pathways accessed on 18 October 2022). Searches for the term "fatty acids" were carried out in the Phytozome database v.12 (https://phytozome.jgi.doe.gov/ accessed on 18 October 2022), including searches for gene sequences related to lipid metabolism in *A. thaliana*, *Glycine max*, and *Populus trichocarpa*.

BLASTp was used to compare the lipid families of the model species against the *C. brasiliensis* protein database. The BLASTp result was used to further analyze gene orthology in OrthoMCL v.1.4 [14] with the standard inflation parameter I = 1.5. OrthoMCL was used to obtain the best hits and explore similarity measures to ortholog and paralog groups. The Markov clustering algorithm was applied, and only families with one species were excluded, but those with at least one *C. brasiliensis* gene were kept.

All *C. brasiliensis* genes belonging to at least one of the families (identified by OrthoMCL) were associated with lipid synthesis in the species. The lipid genes obtained in OrthoMCL were used to perform the analysis in different tissues.

2.4. Heatmap

A heatmap was constructed with all the lipids identified in all libraries using the OrthoMCL, Pannzer 2, EggNOG mapper, KEGG Mapper, and InterProScan. A heatmap was constructed to visualize the expression of the *C. brasiliensis* lipid transcripts identified in

the different tissues with the aid of the "ComplexHeatmap" package (https://github.com/jokergoo/ComplexHeatmap accessed on 18 October 2022) in the R version 4.0.1 environment (R Core Team). Thereafter, the values of TPM (transcripts per million) were transformed into z-scores with the function "scale". Euclidean distance and grouping with complete linkage were used to add the dendrogram to the heatmap. From the functional annotation analyses, an ontology gene enrichment (GO) analysis was performed for each heatmap cluster, using the BINGO [15].

2.5. Construction of the Lipid Metabolism Model

The following tools were used to build the lipid metabolism model: (i) the OrthoMCL software to identify the orthologous lipid gene families; (ii) the Mercator4 v5.0 software (https://www.plabipd.de/portal/mercator4, accessed on 18 October 2022) to map the unigene file previously identified, obtaining a text file with one or more BINs per protein; (iii) the MapMan program to visualize the expression and meta-analysis data, to annotate the plant omics data [16], to correlate each unigene to its expression level, and to identify the lipid metabolism pathways (X4.2 Lipid metabolism R2.0); (iv) Pannzer 2 [10] and lnterProScan (http://www.ebi.ac.uk/InterProScan/ accessed on 18 October 2022) for protein annotation (focusing on keywords such as fatty, lipid, and desat to contemplate the largest gene number from the lipid pathway; (v) the Kegg Mapper platform [11] to rescue the protein EC (Enzyme Commission Numbers), to automatically assign KO identifiers (KEGG Orthology; https://www.kegg.jp/ or https://www.genome.jp/kegg/ accessed on 18 October 2022) and to map the described lipid metabolism pathways (biosynthesis of fatty acids and unsaturated fatty acids, and elongation of fatty acids); and (vi) the "ComplexHeatmap" package (https://github.com/jokergoo/ComplexHeatmap, accessed on 18 October 2022) in *R* version 4.0.1 to generate a heatmap from the rescued EC. For heatmap generation, the TPM values were transformed to z-scores with the "scale" function, while for clustering, the Euclidean distance and clustering with complete linkage function were used. The lipid pathways and the expression data were manually associated in a general scheme.

3. Results

3.1. RNA Sequencing: The First Transcriptomic Information of C. brasiliensis

The Illumina MiSeq sequencing (2 × 250 bp) of *C. brasiliensis* generated about 10.4 million paired-end reads (Phred > 20) for 24 distinct cDNA tissue libraries. The assembly of reads from the 24 libraries of vegetative and reproductive tissues generated a total of 263,562 unigenes, with an average length of 340 base pairs (bp). The average length of the transcripts varied by 554.34 bp, with N50 length of 637 bp and the GC content of 42.68% (Supplementary Table S2). The GC content was 41.37% and N50 totalized 696,928 bp.

We used TransDecoder to reduce the number of total transcripts and the number of unigenes, but to increase the average N50 length (Supplementary Table S2). The total number of initial unigenes was 263,562 and after treatment with the TransDecoder, the total number of unigenes was 12,908.

The quality of assembly allowed us to detect 73.5% (BUSCO, Table S3) of the genes found in *A. thaliana*. This result can be considered satisfactory since the species *C. brasiliensis* is a non-model organism, and studies with genomes of non-model species generally report BUSCO scores ranging from 50% to 95%, depending on aspects of the species (genome size, number of repetitive elements), and their taxonomic position [17].

A total of 38,841 proteins were detected. Of them, 21,393 (55.07%) proteins were annotated with GO terms, based on functional annotation of the transcripts. In the three main GO categories (biological processes, cellular components, and molecular functions), proteins assigned to the activity subcategories "metabolic process", "cell" and "cell process" were found in higher percentages (Supplementary Figure S2). Similar results were observed in RNA-Seq data for *Brassica napus* L. during seed maturation [18]. We managed

to detect new proteins for *C. brasiliensis*, as well as finding similarities in functional categories of proteins in comparison to model species.

Search for proteins based on sequence identity found 38,841 sequences, which were categorized into 25 functional groups. Proteins with "unknown function" represented the largest group (8416, or 21.6%), followed by "post-translational modifications" proteins (2685, or 6.9%), proteins related to "signal transduction mechanisms" (2669, or 6.8%). Proteins associated with "cellular motility" (41, or 0.10%), and "extracellular structures" (64, or 0.16%) were the smallest groups. No proteins were associated with the category "general function prediction" (Figure 1). The shorter sequences may lack a characterized protein domain or may be too short to show sequence matches, resulting in false-negative results. Because genomic and transcriptomic information is currently lacking for *C. brasiliensis* in databases, these cases of no hits can be considered putative novel protein sequences.

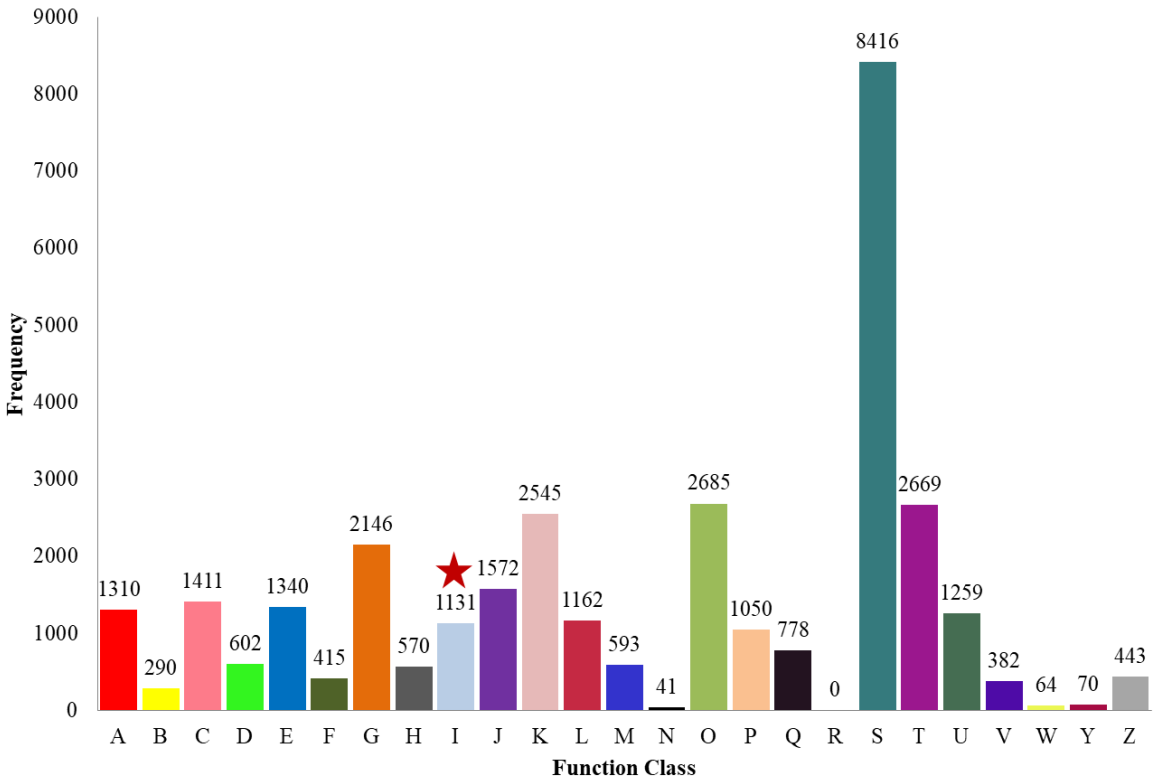

Figure 1. Histogram of clusters of orthologous groups (COG) classification according to the number of protein found in *C. brasiliensis* transcriptome. (A) RNA processing and modification; (B) Chromatin structure and dynamics; (C) Energy production and conversion; (D) Cell cycle control, cell division, chromosome partitioning; (E) Amino acid transport and metabolism; (F) Nucleotide transport and metabolism; (G) Carbohydrate transport and metabolism; (H) Coenzyme transport and metabolism; (I) Lipid transport and metabolism; (J) Translation, ribosomal structure and biogenesis; (K) Transcription; (L) Replication, recombination and repair; (M) Cell wall/ membrane/envelope biogenesis; (N) Cell motility; (O) Posttranslational modification, protein turnover, chaperones; (P) Inorganic ion transport and metabolism; (Q) Secondary metabolites biosynthesis, transport and catabolism; (R) General function prediction only; (S) Function unknown; (T) Signal transduction mechanisms; (U) Intracellular trafficking, secretion, and vesicular transport; (V) Defense mechanisms; (W) Extracellular structures; (Y) Nuclear structure; (Z) Cytoskeleton. The star highlights the category (I) Lipid transport and metabolism.

The metabolic network analysis of *C. brasiliensis* revealed 148 pathways, but the combined global map showed 742 metabolic pathways (ko01100) and 334 secondary metabolite biosynthesis enzymes (ko01110). This result reveals that *C. brasiliensis* can be

considered a plant rich in secondary metabolites. Due to the commercial importance of the oil in *C. brasiliensis*, our study focused on "lipid metabolism", for which 13 pathways were identified (Figure 2).

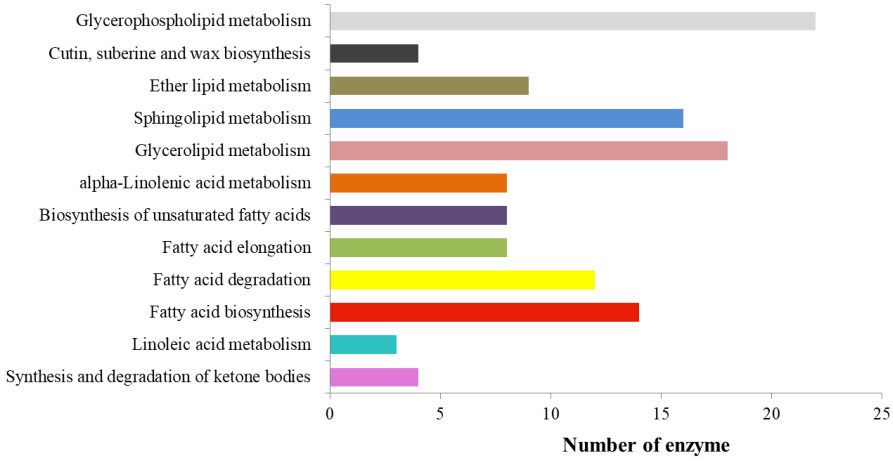

Figure 2. Pathways involved in *Carpotroche brasiliensis* lipid metabolism.

A large number of sequences involved in lipid synthesis and metabolism was predicted, as well as the biochemical pathways involved in the synthesis of chaulmoogra oil. We found 1131 unigenes that possibly encode proteins associated with metabolism and lipids, representing 2.9% of total proteins, in all libraries of *C. brasiliensis* (Figure 1).

3.2. Phylogenetic Analyses with 18S: A Search for Related Species Producing Lipids of Economic Importance

Searches for transcripts including the ribosomal 18S from *C. brasiliensis* revealed 78 similar sequences, which matched lipid producing tree species or other plant species that produce secondary metabolites, such as *Cannabis sativa*, *Triadica sebifera*, and *Citrus clementina.* When comparing these RNAs with *A. thaliana* and other oilseed models, such as Glycine *max*, *Populus trichocarpa*, and species of the families Achariaceae and defunct Flacourtiaceae, the phylogenetic tree indicated a division of all species into three major groups (Supplementary Figure S3). *C. brasiliensis* showed high proximity with *Camptostylus manii*, *Dasylepis brevipedicellata*, and *Erythrospermum phytolaccoides,* which belong to the Achariaceae family (Supplementary Figure S3).

It was interesting to find that the model plant species *Arabidopsis thaliana* is relatively closely related to *C. brasiliensis* (Supplementary Figure S3) according to the 18S gene sequence analysis, as well as in terms of gene completeness (Supplementary Table S3). Thus, based on the phylogeny results obtained with 18S sequences in the present study, we chose *A. thaliana* as a biological model in the subsequent analyses of groups of orthologous genes.

3.3. Orthologous Groups Involved in Lipid Metabolism

The main family of lipid genes found in our results was the chaperones (Figure 3). A previous study with the Chinese tallow tree (*Triadiaca sebifera*) evaluated proteins in lipid droplets of the fruit mesocarp, and numerous proteins were found related to signal transduction and activity similar to molecular chaperones [19]. This study identified proteins with similar functions to chaperones, such as annexin D8, aspartic proteinases, bcl-2-associated athanogene (BAG), and aquaporin. Therefore, we believe that chaperones are important for the synthesis of lipids.

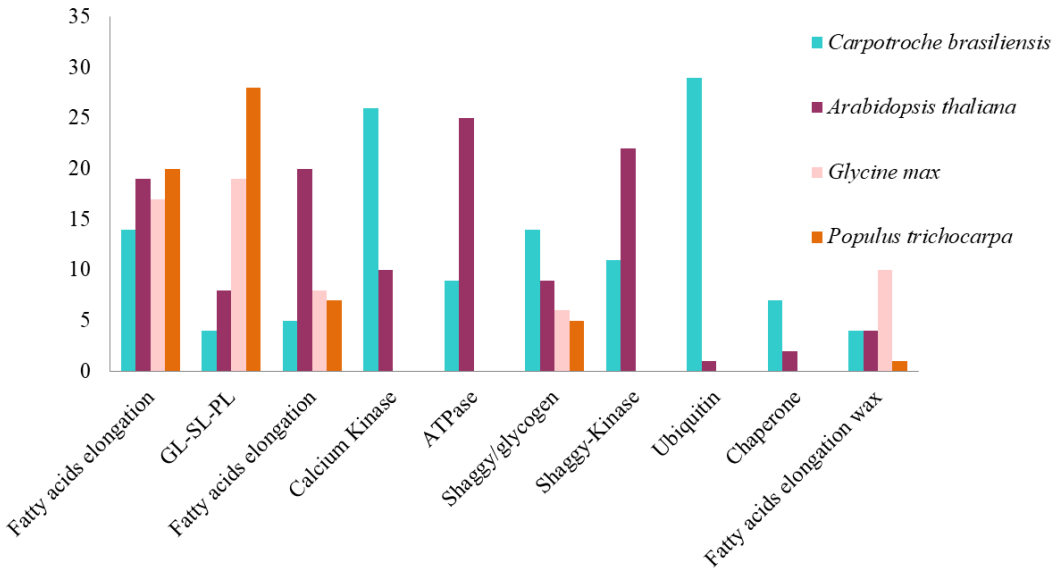

Figure 3. Bar plot graph of the number of the most representative lipid families generated in the OrthoMCL of the species *Carpotroche brasiliensis*, *Arabidopsis thaliana*, *Glycine max* and *Populus trichocarpa*. GL: galactolipid, SL: sulfolipid, PL: phospholipid.

The second most numerous gene family among species was that of lipids involved with cell membranes, such as galactolipids, sulfolipids, and phospholipids (Figure 3).

Our results suggest that seven putative enzymes identified in the fast oil accumulation stage may be involved in the synthesis of triacylglycerols (1-acyl-sn-glycerol-3-phosphate acyltransferase; phosphatidate phosphatase; diacylglycerol O-acyltransferase; phospholipid:diacylglycerol acyltransferase; phosphatidylcholine:diacylglycerol cholinephosphotransferase; lysophospholipid acyltransferase; and phospholipase) (Supplementary Table S4), suggesting their involvement in oil accumulation in *C. brasiliensis* seeds.

3.4. Expression of Genes Involved in Lipid Biosynthesis

3.4.1. Special Characteristics of Lipids in Fruits and Seeds

In *C. brasiliensis*, lipids are primarily stored in seeds. We identified a high level of gene expression for oil synthesis in the seeds and immature fruits of *C. brasiliensis.* Five clusters represent the biological processes involved in lipid metabolism: Cluster 1 (acetyl-CoA metabolism and lipid A metabolism); Cluster 2 (lipid A metabolism and seed oilbody biogenesis); Cluster 3 (unsaturated fatty acid metabolism and acylglycerol metabolism); Cluster 4 (fatty acid oxidation and glycerollipid biosynthesis) (Figure 4).

In *C. brasiliensis* seeds, we identified acyl CoA proteins (Cluster 1) (Figure 4), 32 acetyl-CoA carboxylase enzyme (ACCase) subunits (8 for biotin carboxylase and 24 for biotin carboxyl transport proteins, respectively), 5 malonyl-CoA ACP transacylase (MAT) enzymes, 10 3-ketoacyl-ACP synthase (KAS) enzymes (5 for KAS I, 2 for KAS II, and 3 for KAS III, respectively), and 27 enzymes involved in other components of synthase fatty acids (10 for hydroxyacyl-ACP dehydrase HAD, and 17 for enoyl-ACP reductase EAR) (Supplementary Table S4). We also identified 7 acyl-ACP thioesterase FAT enzymes, which catalyze reactions that produce free fatty acids (Cluster 4) (Figure 4).

We identified 8 diacylglycerol acyltransferase enzymes (EC 2.3.1.20), 11 glycerol-3-phosphate acyltransferase enzymes (EC 2.3.1.51), and 2 lysophospholipid acyltransferase enzymes (EC 2.3.1.23) in the triacylglycerol assembly pathway (Kennedy pathway) (Supplementary Table S4). We also observed 106 fatty acid enzymes (LACS, HCD, SAD, FAD5, oleosin, caleosin) Cluster 2 (Figure 4 and Supplementary Table S4). Additionally, we

identified four oleosins and three caleosins that may play a role in the accumulation and maintenance of chaulmoogra oil.

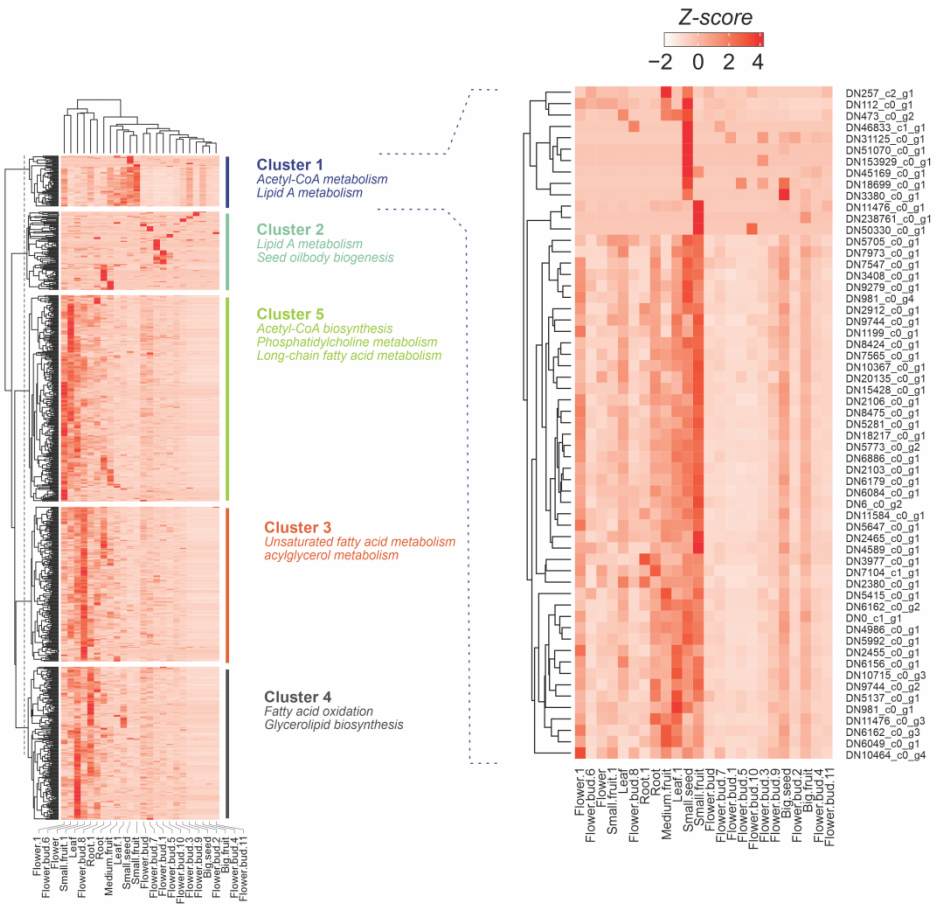

Figure 4. Heatmap of the analysis of the RNA-Seq transcriptome of 24 tissue libraries of *Carpotroche brasiliensis*. The heatmap of gene expression (Z-score) for 24 libraries, indicated by scale marks in the figure (columns) and lipids (lines). The color corresponds to the z-score per gene that is calculated from TPM (Transcripts Per Million). The redder the higher the expression values. The dendrogram shows the gene cluster according to the Euclidean distance. Five clusters represent the biological processes involved in lipid metabolism. Cluster 1, highlighted, represents the biological processes: lipid and acetyl-CoA metabolism and the genes that expressed *Carpotroche brasiliensis* libraries.

For the formation of unsaturated fatty acids (Cluster 3) (Figure 4), such as chaulmoogra oil, six proteins that encode fatty acid desaturase (FAD) have been identified, including six types of FAD (SAD, FAD5) and stearoyl-ACP desaturase (SAD), which removes two hydrogen atoms from stearic acid (18C: 0) to form oleic acid (18C: 1) (Supplementary Table S4).

One unigene (phospholipase A) encoding lipases (Supplementary Table S4) was identified from our libraries. Lipases present in developing castor bean may be involved in the remodeling of TGs after synthesis [20]. The function of these lipases in developing *C. brasiliensis* seeds remains unclear.

3.4.2. Special Features of Lipids in Flower Tissue

The genes involved in lipid metabolism were analyzed and identified on a scale (Z-score) with relative levels between transcripts and tissues, with a specific focus on the synthesis of fatty acids and triacylglycerol accumulation/storage pathways (Figure 4). Most of the genes had a strong contrast between seed and non-seed lipid genes, showing

substantially different levels of expression among tissues. The libraries with the highest levels of expression were the flower, floral bud, immature fruit, and leaf libraries (Figure 4). Although seeds are by far the largest commercial sources of oils from *C. brasiliensis*, oil is also abundant in many other tissues.

3.4.3. Special Features of Root and Leaf Lipids

We detected transcripts related to the synthesis of lipids in the root library of *C. brasiliensis* (Cluster 5) (Figure 4). Such results were expected, since suberin layers serve as an infection barrier. The wax surface influences plant–insect interactions, and helps to prevent germination of pathogenic microbes [21].

3.4.4. Metabolic Pathways Related to Oil Biosynthesis

The organelles involved in the synthesis and elongation of fatty acids are chloroplasts (de novo synthesis of fatty acids) and rough endoplasmic reticulum (Figure 5).

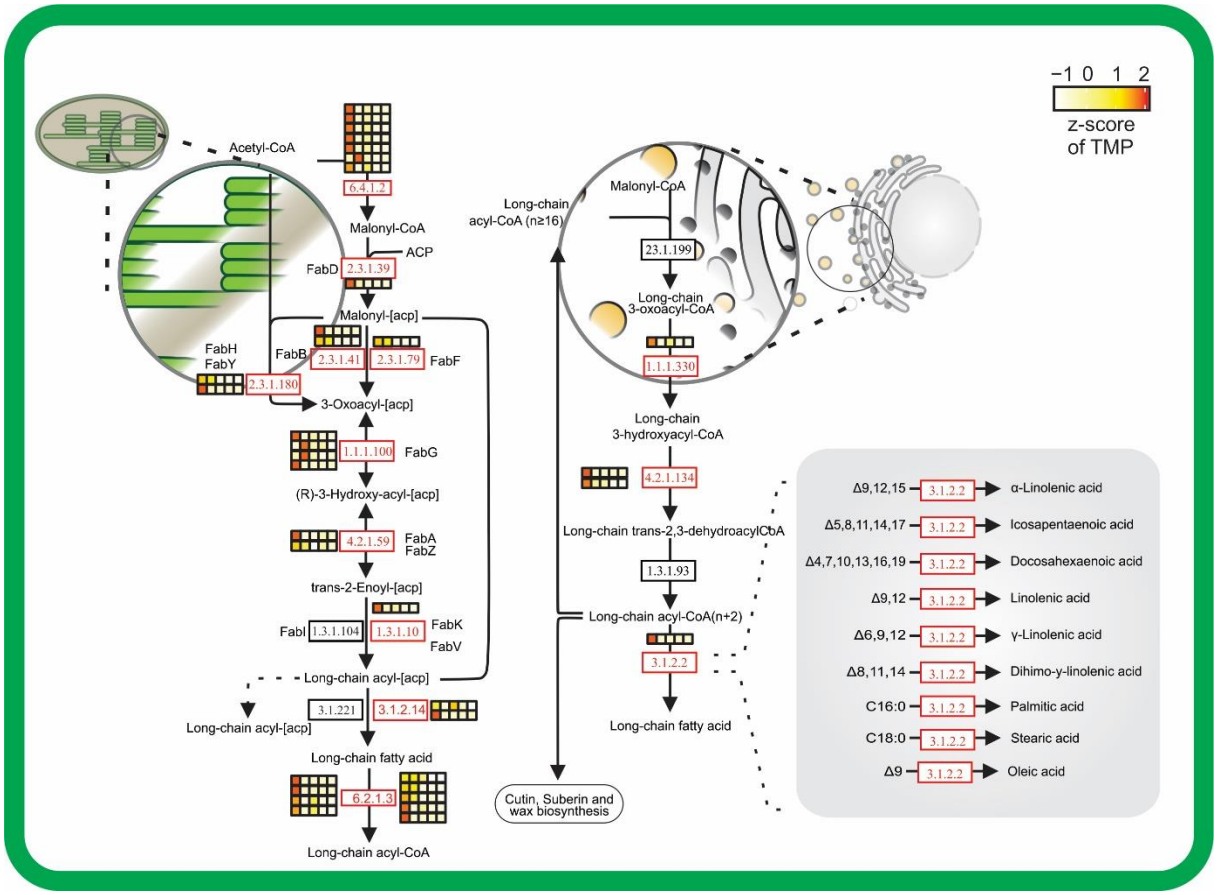

Figure 5. Lipid biosynthesis pathways and transcript expression patterns in immature and mature *Carpotroche brasiliensis* seeds and fruits. The transcript patterns for enzymes involved in glycolysis reactions in mitochondria. Fatty acid synthesis occurs in plastids, and in the endoplasmic reticulum, triacylglycerol synthesis occurs—the elongation and establishment of fatty acids. The gene expression heatmap (Z-score), where an average of the expression of the libraries of immature and ripe fruits was performed. The color corresponds to the Z-score per gene that is calculated from TPM. The color intensity indicates the relative level of expression. The first column represents the library of immature seed, immature fruit, immature fruit, ripe seed, and ripe fruit, respectively. The enzymes found on this route are marked in red. The gray box represents unsaturated fatty acids. Abbreviation: FabA, 3-hydroxyacyl-ACP dehydratase II; FabB, β-ketoacyl-ACP synthase I; FabD, ACP-S-malonyltransferase; FabF, 3-oxoacyl-ACP synthase II; FabG, 3-oxoacyl-ACP reductase; FabH, 3-oxoacyl-ACP synthase III; FabY, Acetoacetyl-ACP synthase; FabV, Enoyl-ACP reductase; FabK, Enoyl-ACP reductase II; FabZ, 3-hydroxyacyl-ACP dehydratase; ACP, acyl-carrier protein.

In terms of the synthesis of fatty acids in plastids, genes were regulated in libraries of immature and ripe fruits and seeds. For example, the genes for malonyl-CoA:ACP-S-malonyltransferase (FabD), Enoyl-ACP reductase (FabV), and Enoyl-ACP reductase II (FabK) have higher levels of expression in the tissues of immature seeds and fruits (Figure 5). On the other hand, in mature seeds and fruits, these genes are down-regulated.

The table in Figure 5 shows that the lipids from *C. brasiliensis* seeds are composed mainly of saturated fatty acids, such as palmitic acid (C16: 0) and stearic acid (C18: 0), and unsaturated fatty acids, such as oleic acid (C18: 1), linoleic acid (C18: 2), alpha-linolenic acid (C18: 3), and palmitoleic acid (C16: 1). In *C. brasiliensis*, we observed the presence of 3-oxoacyl-[acyl-carrier-protein] synthase II (EC: 2.3.1.179), the main enzyme involved in determining the fatty acid chain length (that is, the ratio of fatty acids from 16C to 18C). We also observed many saturated fatty acids. Thus, we believe that as already reported in another study of *Carya illinoinensis* [22], these lipid genes are important in the developing embryo.

4. Discussion

This is the first transcriptome and the first characterization of gene expressions in leaves, flowers, floral buds, roots, fruits, and seeds of *C. brasiliensis*. Hence, it represents a unique transcriptomic resource available for this chaulmoogra oil-producing species, endemic to the Atlantic Forest biome.

Observing the assembling results, we detected that *C. brasiliensis* transcriptome is close to those previously identified in members of the defunct Flacourtiaceae family, such as *Idesia polycarpa* [23]. These results suggest that our data are robust and of similar quality to the transcriptomes from other oilseed species. The 12,908 unigenes are considered a good estimate of the total genes for a tree species, since plant genomes are expected to have between 12,000 and 45,000 genes [24], suggesting that we detected more complete transcripts.

Additionally, we included *Arabidopsis thaliana* as a possible model species because it is a relatively close group (Supplementary Figure S3), as well as its gene completeness (Supplementary Table S3). Thus, from the phylogeny results obtained with 18S sequences in the present study, we confirmed *A. thaliana* as the biological model in subsequent analyses of orthologous gene groups. We highlight that the 18S phylogeny data were critical in improving the understanding of the evolutionary history of *C. brasiliensis*, as well as providing additional information on the family Achariaceae. By analyzing orthologous groups involved in lipid metabolism, we identified clusters of lipid gene families, including orthologous genes among *C. brasiliensis*, *A. thaliana*, *G. max* and *P. trichocarpa*. Consequently, it was possible to identify orthologous genes with common ancestry among different species, since these genes tend to preserve the function of their ancestor. Therefore, our results indicate that the *C. brasiliensis* transcriptome can be used as reference for studies of other phylogenetically close oil tree species. When other models were used for comparative analysis of the transcriptome assembly of *C. brasiliensis* and *Solanum lycopersicum*, the result found was 49.6% in the groups researched using BUSCO (Supplementary Table S3). Such findings may be related to the evolutionary distance of the botanical families used to compare gene completeness. In this regard, the species *A. thaliana* would be more adequate to measure the gene completeness and quality verification of the assembly of the *C. brasiliensis* transcriptome. Studies using RNA-Seq data from 24 species of vascular plants were previously reported with BUSCO scores between 60% and 85% [25].

In most plants, the main unsaturated fatty acids (UFAs) are three: oleic (18:1), linoleic (18:2), and α-linolenic (18:3). The biosynthetic pathway in *Arabidopsis* is taken as an example. Briefly, in plastids, fatty acids are synthesized again from acetyl-coenzyme A (CoA), due to the joint action of acetyl-CoA carboxylase (ACC) and fatty acid synthase (FAS). Once produced, 18: 0, conjugated to the acyl carrier protein (ACP), enters mainly the unsaturation pathway administered by a series of desaturases, and 18: 1-ACP is rapidly created by stearoyl-ACP desaturase (SAD). However, the biosynthesis of polyunsaturated

fatty acids is coupled with that of membrane glycerolipids, which is conducted in two parallel pathways—the 'prokaryotic' in the plastids and the 'eukaryotic' in the endoplasmic reticulum [26]. These identified lipid unigenes provided critical clues to clone and identify key functional genes involved in unsaturated fatty acids and triacylglycerol biosynthesis in *C. brasiliensis* seeds.

In comparison with other studies related to oil production in plants, our results indicate better coverage. In a study carried out with pecan (*Carya illinoinensis*), 153 unigenes associated with lipid biosynthesis were identified, including 107 unigenes for fatty acid biosynthesis, 34 for triacylglycerol biosynthesis, 7 for oily bodies, and 5 for transcription factors [22]. In another work using peanut (*Arachis hypogaea*), the authors identified 654 unigenes involved in lipid metabolism and transport, which represented 4.5% of the total number of unigenes [27]. Furthermore, the total number of enzymes involved in the metabolism of glycerophospholipids in *C. brasiliensis* was 22, which is similar to that observed in studies with seeds of *Camellia meiocarpa* and *Camellia oleifera*, where 10 unigenes related to the metabolism of glycerophospholipids were found [28]. The similar results between these two genera (*Carpotroche* and *Camellia*) indicate that oil-producing species have similar lipid metabolism enzymes, even though they produce oils with different properties. The high number of libraries used in the present study may have contributed to obtaining higher and more robust values of unigenes related to the transport and metabolism of lipids in *C. brasiliensis*. Our metabolic network results show that the sequences related to the biosynthesis of secondary metabolites will be a good resource for future research into the biosynthesis mechanism of flavonoids, phytosterols, saponins, and other medicinal compounds of the species.

Our study also provides important information on the biological sequences of species belonging to the Achariaceae family and can contribute to elucidate the evolutionary history of this group. The phylogenetic relationships with similar species that also produce lipids (important compounds for the production of biodiesel) and essential oils of economic importance can help to better understand the mechanisms of oil production in plants. The knowledge of the evolutionary relationships of *C. brasiliensis* also improves the understanding of the homology of unknown genes, from genes previously described in closely related species. This knowledge can help in the discovery of new genes involved in several metabolic pathways.

Previously, *C. brasiliensis* was classified as belonging to the Flacourtiaceae family. However, most taxonomic groups producing cyanogenic cyclopentene glycosides and flowers with unequal numbers of sepals and petals have already been reclassified [29]. This was the case of our target species and others that produce (AGC). Therefore, they were classified in the family Achariaceae [29], with acceptance in the last classification of Angiosperms (APG IV 2016) [30]. Consequently, Achariaceae needs broader phylogenetic and basic studies, which could identify uses of plants of the family beyond potential drug production. We consider that our 18S phylogeny data were fundamental to reveal the level of evolutionary relationships with *A. thaliana*, improving the understanding of the evolutionary history of *C. brasiliensis*, as well as providing additional information from the family Achariaceae.

Furthermore, we identified clusters of lipid gene families, including orthologous genes among *C. brasiliensis*, *A. thaliana*, *G. max*, and *P. trichocarpa*. Consequently, it was possible to identify orthologous genes with common ancestry among different species, since these genes tend to preserve the function of their ancestor. One interesting example of orthologous groups involved in lipid metabolism was the family of the chaperone lipid genes (Figure 3). A previous study with the Chinese tallow tree (*Triadiaca sebifera*) evaluated proteins in lipid droplets of the fruit mesocarp, and numerous proteins were found related to signal transduction and activity similar to molecular chaperones [19]. Our results identified proteins with similar functions to chaperones, such as annexin D8, aspartic proteinases, bcl-2-associated athanogene (BAG), and aquaporin. Therefore, we believe

that chaperones are important for the synthesis of lipids such as chaulmoogra oil in *C. brasiliensis*.

Other important gene families for plant lipid synthesis are related with cell membranes. The assembly of triacylglycerols (TG) in the endoplasmic reticulum is integrated with the assembly of membrane glycerol lipids, which has two possible routes leading to the formation of triacylglycerol [31]. In the Kennedy pathway, glycerol-3-phosphate acyltransferase (EC: 2.3.1.15), 1-Acyl-sn-glycerol-3-phosphate acyltransferase (EC: 2.3.1.51), phosphatidic acid phosphatase (EC: 3.1.3.4), and diacylglycerol acyltransferase (EC: 2.3.1.20) are the sequential enzymes involved in the synthesis of triacylglycerols [31]. Recently, some additional independent acyl-CoA reactions have been identified in the development of seeds and other plant tissues that contribute to the synthesis of TGs, although the relative contributions of these alternative pathways to the accumulation of TG varies depending on the tissue and/or species [31].

Regarding oilseed and medicinal species, it was expected that the most synthesis of oil genes was found in seeds and immature fruits. These findings corroborate those of similar studies carried out with the species *Hydnocavpus anthelminthica,* belonging to the same family as *C. brasiliensis.* A study evaluating the synthesis (AGCs) showed a decline in the synthesis of cyclopentenyl fatty acids as the seed maturity progressed, indicating that the activity ceases almost completely at full maturity [6].

The results of gene expression involved in lipid biosynthesis imply that the Acyl-CoA dependent TG biosynthesis pathway might be an active pathway in TG biosynthesis in *C. brasiliensis* seeds. Similar results were found for the oil species Sacha Inchi (*Plukenetia volubilis* L.) [32] and *Brassica napus* L. [18]. These identified proteins are putative enzymes in the synthesis of lipids and are also found in other oleaginous species. Furthermore, it is important to highlight that the expression of oleosin genes is generally closely associated with oil accumulation in developing seeds [33]. A similar pattern of lipid gene expression has been shown in studies with fruits of *Vitellaria paradoxa* [34] and with *Idesia polycarpa* [23], although there are no studies identifying the production of chaulmoogra oil by these species.

As we know, oleic and linoleic acids are constituents of chaulmoogra oil in *C. brasiliensis* seeds. Based on this knowledge, it is interesting to note that the functions of the six enzymes identified can be the molecular basis for the formation of polyunsaturated fatty acids in the seeds. In addition, the Acyl-CoA-independent TG assembly pathway, including acyl editing and phosphatidylcholine/1,2-sn-diacylglycerol interconversion, is believed to facilitate the incorporation of polyunsaturated fatty acids into TG in some plant species [35]. Other works with oilseed plants have also identified classes of desaturases, such as Sacha Inchi (*Plukenetia volubilis* L.) [32] and two species of *Camellia* [28].

Additionally, the results based on flower tissue reveal some very specific changes in expression patterns (Figure 4). High levels of expression in non-seed tissues of *C. brasiliensis* for transcripts associated with lipid synthesis can indicate not only distinct roles during oil accumulation and seed development but also tissue-specific differences in their functions, such as in flower buds and flowers. The flowers were found to contain several secondary metabolites, pigments, and complex molecules involved in cellular processes, which probably occur throughout development. Therefore, these genes are necessary for the proper regulation of cell proliferation and expansion, the development of reproductive tissues, and the sculpting of the final shape of the different organs [36]. Our results draw attention to the importance of lipids such as phosphatidylcholine, a phospholipid found in *C. brasiliensis*, which is usually found in biological membranes (Figure 3). This class of lipids exhibits circadian oscillation and is also correlated with flowering, pointing to the important role of lipids in lowers in general [37]. On the other hand, the vital importance of plant surface wax in protecting tissue from environmental stresses is reflected in the huge commitment of epidermal cells to cuticle formation [38]. Another common characteristic of all terrestrial plants, identified in the annotation in this study, was the cutin (Figure 2). Cutin is a hydrophobic substance that covers surfaces exposed to air to

prevent the non-stomatal loss of water and protect against pathogens. Cutin and wax contain derivatives of very-long-chain fatty acids, such as alkanes and alcohols, with chain lengths > 20 carbons. The composition of the wax varies according to the species, organ, and developmental stage [38].

Similar to other studies carried out to investigate the species of the Achariaceae family (e.g., *Hydnocarpus anthelminthica* and *Culoncoha echinata*) [6], our results indicate that cyclopentenyl fatty acids are synthesized not just in seeds, but in other tissues as well, including leaves, as is the case of many other families of oil plants.

Our analysis of orthology, metabolic pathways of lipids, and functional annotation were used to annotate the transcripts of *C. brasiliensis*-encoding orthologs of putative plant enzymes involved in the biosynthesis of saturated and unsaturated fatty acids, besides fatty acid elongation. These data were integrated and compiled to propose metabolic routes that lead to the accumulation of lipids in *C. brasiliensis* seeds (Figure 5). The de novo biosynthesis of fatty acids in plants occurs in plastids, performed by a dissociable complex of monofunctional fatty acid synthase enzymes [22]. Briefly, the pyruvate dehydrogenase (PDH) complex generates acetyl-CoA, the component used in the production of fatty acids. The first step in fatty acid biosynthesis is the conversion of acetyl-CoA to malonyl-CoA by acetyl-CoA carboxylase (ACC). The malonyl group is then transferred from CoA to the acyl carrier protein (ACP), and the condensation between malonyl-ACP and acetyl-CoA is catalyzed by the fatty acid synthase complex. This is the first in a series of sequential reactions of condensation, reduction, and dehydration, adding two units of carbon to the lengthening acyl chain. The acyl chains are finally hydrolyzed by the acyl-ACP thioesterases, which release fatty acids [39]. Fatty acid synthases rely on a small protein, the acyl transporter protein (ACP), to carry the fatty acid load from enzyme to enzyme, for the elongation and synthesis of fatty acids.

Our results suggest that fatty acid biosynthesis, fatty acid elongation, and the tricarboxylic acid (TCA) cycle are all activated in the seed and fruit development process. Many of the enzymes involved in the metabolism of fatty acids have been increased or underregulated, and specific enzymes are critical in the biosynthesis of cyclopentenic fatty acids.

5. Conclusions

We conclude from the results generated and the complexity of the analyses carried out that *C. brasiliensis* can significantly broaden the understanding of the metabolism and lipid synthesis since we developed a very comprehensive lipid unigene resource. It was possible to identify 13 lipid pathways and 1131 proteins involved with lipid metabolism and transport, as well as to identify cyclopentenic acids, such as oleic acid, linoleic acid, palmitic acid, and stearic acid, in chaulmoogra oil. From our results, we also conclude that the synthesis of chaulmoogra oil starts in immature seeds, where the highest number and expression values of lipid transcripts were found. However, we also found that in all tissues, transcribed putative genes were involved in the processes of synthesis, metabolism, transport, and degradation of lipids.

This is the first transcriptome study of *C. brasiliensis*, providing basic information for biotechnological applications of great use for the species, which will help understand chaulmoogra oil biosynthesis. The dataset developed in this study expands the database of *C. brasiliensis*. These resources can contribute to the discovery of new genes in developing seeds, which reinforces the importance of the present study. The basic information produced can be applied in future biotechnological approaches providing a booter to obtain improved varieties via genetic engineering, as well as for the development of molecular markers. Finally, the set of identified unigenes can also contribute to the future annotation and assembly of the complete genome sequence of *C. brasiliensis*. To our knowledge, this is one of the most complete and extensive efforts to annotate lipid genes from tropical non-timber trees.

6. Declarations

Ethics Approval and Consent to Participate

Ethical approval for the botanical material sampling of *Carpotroche brasiliensis* for this study was granted by the Brazillian "Biodiversity Authorization and Information System" (SISBIO-42397-1). Additionally, we obtained the genetical accessing authorization from the "National System for the Management of Genetic Heritage and Associated Traditional Knowledge (SisGen-A99B6B8), for this research work, in compliance with relevant institutional, national, and international guidelines and legislation. The authors declare to be in accordance with the IUCN Policy on Research Involving Endangered Species and the Convention on Trade in Endangered Species of Wild Fauna and Flora. However, the species *Carpotroche brasiliensis* is not in any list of species at risk or threat of extinction.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/f13111806/s1, Figure S1: Schematic overview of the structure of *Carpotroche brasiliensis* RNA-Seq data analysis; Figure S2: Gene Ontology (GO) analysis of the categories for immature and mature fruits and seeds from *Carpotroche brasiliensis*; Figure S3: Phylogeny based on the Maximum likelihood method with 1000 bootstrap replicates and 18S gene sequences; Table S1: Data from samples and sequencing of RNA extraction from tissues of *Carpotroche brasiliensis*.; Table S2: Assembly statistics of *C. brasiliensis* transcriptome by RNA-Seq and TransDecoder; Table S3: Assessment of transcriptome quality by BUSCO; Table S4: Enzymes involved in fatty acid biosynthesis and catabolism identified by the annotation of *Carpotroche brasiliensis* transcriptome.

Author Contributions: Conceived and designed the experiments: F.B., C.P.P., F.M., M.K. and F.A.G.; Performed the experiments: F.B. and F.A.G.; Analyzed the data: L.M.d.V., F.B., R.O.V., E.M.d.A.S. and E.A.C.; Wrote the manuscript: L.M.d.V. All authors have read and agreed to the published version of the manuscript.

Funding: We would like to acknowledge the "Fundação de Amparo à Pesquisa do Estado da Bahia" (FAPESB # TSC0017/2014) for the financial support. We also thank the following "Conselho Nacional de Desenvolvimento Científico e Tecnológico" (CNPq), "Coordenação de Aperfeiçoamento de Pessoal de Nível Superior" (CAPES) and Fapesb agencies for the scholarships (LMV, FB), and research fellowships for FAG, CCP (FAG, CNPq #306160/2017-0; CPP, CNPq #303765/2019-4).

Data Availability Statement: The datasets generated and/or analyzed during the current study are available in the National Center for Biotechnology repository, raw sequencing data were deposited at NCBI Sequence Read Archive (SRA) under project number PRJNA858666 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA858666/).

Acknowledgments: The authors thank Fernando Santana for technical support and authorization of plant material collections in cocoa farms in the Camamu region, Bahia, Brazil, and to members of the BIOPREA group (Caio Argolo, Ceslaine Barbosa, Dalma Brito, Edson Mário, Jonathan Mucherino, Patrícia Gonzalez, Raner Santana and Valter Magalhães) for their scientific discussions. The authors thank Annette Fahrenkrog, Bárbara Müller, Leandro Neves and Christopher Dervinis, who contributed to the transcriptome.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

References

1. Pinto, L.C.; de Souza, M.P.C.; Lopes, M.V.; Figueiredo, C.A.V. Teor de Fenólicos Totais e Atividade Antioxidante Das Sementes Da Carpotroche Brasiliensis (Raddi). *Rev. Ciências Médicas Biológicas* **2012**, *11*, 170. https://doi.org/10.9771/cmbio.v11i2.6680.
2. Rehfeldt, A.G.; Schulte, E.; Spener, F. Occurrence and Biosynthesis of Cyclopentenyl Fatty Acids in Leaves and Chloroplasts of Flacourtiaceae. *Phytochemistry* **1980**, *19*, 1685–1689. https://doi.org/10.1016/S0031-9422(00)83795-4.
3. Oliveira, A.S.; Lima, J.A.; Rezende, C.M.; Pinto, A.C. Cyclopentenyl Acids from Sapucainha Oil (Carpotroche Brasiliensis Endl, Flacourtiaceae): The First Antileprotic Used in Brazil. *Quim. Nova* **2009**, *32*, 139–145. https://doi.org/10.1590/S0100-40422009000100027.
4. Waktola, H.D.; Kulsing, C.; Nolvachai, Y.; Rezende, C.M.; Bizzo, H.R.; Marriott, P.J. Gas Chromatography–Mass Spectrometry of Sapucainha Oil (Carpotroche Brasiliensis) Triacylglycerols Comprising Straight Chain and Cyclic Fatty Acids. *Anal. Bioanal. Chem.* **2019**, *411*, 1479–1489. https://doi.org/10.1007/s00216-019-01579-7.

5. Gunstone, F.D. *The Chemistry of Oils and Fats: Sources, Composition, Properties, and Uses-Blackwell*; John Wiley & Sons: Hoboken, NJ, USA, 2009.
6. Cramer, U.; Spener, F. Biosynthesis of Cyclopentenyl Fatty Acids. *Biochim. Biophys. Acta Lipids Lipid Metab.* **1976**, *450*, 261–265. https://doi.org/10.1016/0005-2760(76)90098-9.
7. Hall, D.E.; Zerbe, P.; Jancsik, S.; Quesada, A.L.; Dullat, H.; Madilao, L.L.; Yuen, M.; Bohlmann, J. Evolution of Conifer Diterpene Synthases: Diterpene Resin Acid Biosynthesis in Lodgepole Pine and Jack Pine Involves Monofunctional and Bifunctional Diterpene Synthases. *Plant Physiol.* **2013**, *161*, 600–616. https://doi.org/10.1104/pp.112.208546.
8. Xiao, M.; Zhang, Y.; Chen, X.; Lee, E.J.; Barber, C.J.S.; Chakrabarty, R.; Desgagné-Penix, I.; Haslam, T.M.; Kim, Y.B.; Liu, E.; et al. Transcriptome Analysis Based on Next-Generation Sequencing of Non-Model Plants Producing Specialized Metabolites of Biotechnological Interest. *J. Biotechnol.* **2013**, *166*, 122–134. https://doi.org/10.1016/j.jbiotec.2013.04.004.
9. Huerta-cepas, J.; Forslund, K.; Coelho, L.P.; Szklarczyk, D.; Jensen, L.J.; Mering, C. Von; Bork, P.; Delbru, M. Fast Genome-Wide Functional Annotation through Orthology Assignment by EggNOG-Mapper. *Mol. Biol. Evol.* **2017**, *34*, 2115–2122. https://doi.org/10.1093/molbev/msx148.
10. Koskinen, P.; Törönen, P.; Nokso-Koivisto, J.; Holm, L. PANNZER: High-Throughput Functional Annotation of Uncharacterized Proteins in an Error-Prone Environment. *Bioinformatics* **2015**, *31*, 1544–1552. https://doi.org/10.1093/bioinformatics/btu851.
11. Kanehisa, M.; Sato, Y. KEGG Mapper for Inferring Cellular Functions from Protein Sequences. *Protein Sci.* **2020**, *29*, 28–35. https://doi.org/10.1002/pro.3711.
12. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* **2018**, *35*, 1547–1549. https://doi.org/10.1093/molbev/msy096.
13. Tamura, K.; Nei, M. Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans And Chimpanzees. *Mol. Biol. Evol.* **1993**, *10*, 512–526.
14. Chen, F.; Mackey, A.J.; Jr, C.J.S.; Roos, D.S. OrthoMCL-DB : Querying a Comprehensive Multi-Species Collection of Ortholog Groups. *Nucleic Acids Res.* **2006**, *34*, 363–368. https://doi.org/10.1093/nar/gkj123.
15. Maere, S.; Heymans, K.; Kuiper, M. BiNGO: A Cytoscape Plugin to Assess Overrepresentation of Gene Ontology Categories in Biological Networks. *Bioinformatics* **2005**, *21*, 3448–3449. https://doi.org/10.1093/bioinformatics/bti551.
16. Thimm, O.; Bla, O.; Gibon, Y.; Nagel, A.; Meyer, S.; Kru, P.; Selbig, J.; Mu, L.A.; Rhee, S.Y.; Stitt, M. MAPMAN: A User-Driven Tool to Display Genomics Data Sets onto Diagrams of Metabolic Pathways and Other Biological Processes. *Plant J.* **2004**, *37*, 914–939. https://doi.org/10.1111/j.1365-313X.2004.02016.x.
17. Yu, X.H.; Cahoon, R.E.; Horn, P.J.; Shi, H.; Prakash, R.R.; Cai, Y.; Hearney, M.; Chapman, K.D.; Cahoon, E.B.; Schwender, J.; et al. Identification of Bottlenecks in the Accumulation of Cyclic Fatty Acids in Camelina Seed Oil. *Plant Biotechnol. J.* **2018**, *16*, 926–938. https://doi.org/10.1111/pbi.12839.
18. Liao, P.; Woodfield, H.K.; Harwood, J.L.; Chye, M.L.; Scofield, S. Comparative Transcriptomics Analysis of Brassica Napus L. During Seed Maturation Reveals Dynamic Changes in Gene Expression between Embryos and Seed Coats and Distinct Expression Profiles of Acyl-CoA-Binding Proteins for Lipid Accumulation. *Plant Cell Physiol.* **2019**, *60*, 2812–2825. https://doi.org/10.1093/pcp/pcz169.
19. Zhi, Y.; Taylor, M.C.; Campbell, P.M.; Warden, A.C.; Shrestha, P.; El Tahchy, A.; Rolland, V.; Vanhercke, T.; Petrie, J.R.; White, R.G.; et al. Comparative Lipidomics and Proteomics of Lipid Droplets in the Mesocarp and Seed Tissues of Chinese Tallow (Triadica Sebifera). *Front. Plant Sci.* **2017**, *8*, 1–20. https://doi.org/10.3389/fpls.2017.01339.
20. Brown, A.P.; Kroon, J.T.M.; Swarbreck, D.; Febrer, M.; Larson, T.R.; Graham, I.A.; Caccamo, M.; Slabas, A.R. Tissue-Specific Whole Transcriptome Sequencing in Castor, Directed at Understanding Triacylglycerol Lipid Biosynthetic Pathways. *PLoS ONE* **2012**, *7*, e30100. https://doi.org/10.1371/journal.pone.0030100.
21. Delude, C.; Fouillen, L.; Bhar, P.; Cardinal, M.J.; Pascal, S.; Santos, P.; Kosma, D.K.; Joubès, J.; Rowland, O.; Domergue, F. Primary Fatty Alcohols Are Major Components of Suberized Root Tissues of Arabidopsis in the Form of Alkyl Hydroxycinnamates. *Plant Physiol.* **2016**, *171*, 1934–1950. https://doi.org/10.1104/pp.16.00834.
22. Huang, R.; Huang, Y.; Sun, Z.; Huang, J.; Wang, Z. Transcriptome Analysis of Genes Involved in Lipid Biosynthesis in the Developing Embryo of Pecan (Carya Illinoinensis). *J. Agric. Food Chem.* **2017**, *65*, 4223–4236. https://doi.org/10.1021/acs.jafc.7b00922.
23. Li, R.J.; Gao, X.; Li, L.M.; Liu, X.L.; Wang, Z.Y.; Lü, S.Y. De Novo Assembly and Characterization of the Fruit Transcriptome of Idesia Polycarpa Reveals Candidate Genes for Lipid Biosynthesis. *Front. Plant Sci.* **2016**, *7*, 1–18. https://doi.org/10.3389/fpls.2016.00801.
24. Sterck, L.; Rombauts, S.; Vandepoele, K.; Peer, Y. Van De; Rouze, P. How Many Genes Are There in Plants (… and Why Are They There)? *Curr. Opin. Plant Biol.* **2007**, *10*, 199–203. https://doi.org/10.1016/j.pbi.2007.01.004.
25. Marx, H.; Jorgensen, S.A.; Wisely, E.; Li, Z.; Katrina, M.D.; Barker, M.S. Progress Towards Plant Community Transcriptomics: Pilot RNA-Seq Data from 24 Species of Vascular Plants at Harvard Forest. *SELL J.* **2020**, *5*, 55. https://doi.org/10.1101/2020.03.31.018945.
26. He, M.; Qin, C.X.; Wang, X.; Ding, N.Z. Plant Unsaturated Fatty Acids: Biosynthesis and Regulation. *Front. Plant Sci.* **2020**, *11*, 1–13. https://doi.org/10.3389/fpls.2020.00390.

27. Yin, D.; Wang, Y.; Zhang, X.; Li, H.; Lu, X.; Zhang, J.; Zhang, W.; Chen, S. De Novo Assembly of the Peanut (Arachis Hypogaea L.) Seed Transcriptome Revealed Candidate Unigenes for Oil Accumulation Pathways. *PLoS ONE* **2013**, *8*, e73767. https://doi.org/10.1371/journal.pone.0073767.
28. Feng, J.L.; Yang, Z.J.; Bai, W.W.; Chen, S.P.; Xu, W.Q.; El-Kassaby, Y.A.; Chen, H. Transcriptome Comparative Analysis of Two Camellia Species Reveals Lipid Metabolism during Mature Seed Natural Drying. *Trees Struct. Funct.* **2017**, *31*, 1827–1848. https://doi.org/10.1007/s00468-017-1588-5.
29. Chase, M.W.; Zmarzty, S.; Lledó, M.D.; Wurdack, K.J.; Swensen, S.M.; Fay, M.F. When in Doubt, Put It in Flacourtiaceae : A Molecular Phylogenetic Analysis Based on Plastid RbcL DNA Sequences. *Kew Bull.* **2002**, *57*, 141–181. https://doi.org/10.2307/4110825
30. Chase, M.W.; Christenhusz, M.J.M.; Fay, M.F.; Byng, J.W.; Judd, W.S.; Soltis, D.E.; Mabberley, D.J.; Sennikov, A.N.; Soltis, P.S.; Stevens, P.F.; et al. An Update of the Angiosperm Phylogeny Group Classification for the Orders and Families of Flowering Plants: APG IV. *Bot. J. Linn. Soc.* **2016**, *181*, 1–20. https://doi.org/10.1111/boj.12385.
31. Chapman, K.D.; Dyer, J.M.; Mullen, R.T. Biogenesis and Functions of Lipid Droplets in Plants: Thematic Review Series: Lipid Droplet Synthesis and Metabolism: From Yeast to Man. *J. Lipid Res.* **2012**, *53*, 215–226. https://doi.org/10.1194/jlr.R021436.
32. Wang, X.; Xu, R.; Wang, R.; Liu, A. Transcriptome Analysis of Sacha Inchi (Plukenetia Volubilis L.) Seeds at Two Developmental Stages. *BMC Genom.* **2012**, *13*, 716. https://doi.org/10.1186/1471-2164-13-716.
33. Voelker, T.; Kinney, A.J. Variations in the Biosynthesis of Seed-Storage Lipids. *Lipids* **2001**, *42*, 358–365.
34. Wei, Y.; Ji, B.; Siewers, V.; Xu, D.; Halkier, B.A.; Nielsen, J. Identification of Genes Involved in Shea Butter Biosynthesis from Vitellaria Paradoxa Fruits through Transcriptomics and Functional Heterologous Expression. *Appl. Microbiol. Biotechnol.* **2019**, *103*, 3727–3736. https://doi.org/10.1007/s00253-019-09720-3.
35. Bates, P.D.; Durrett, T.P.; Ohlrogge, J.B.; Pollard, M. Analysis of Acyl Fluxes through Multiple Pathways of Triacylglycerol Synthesis in Developing Soybean Embryos. *Plant Physiol.* **2009**, *150*, 55–72. https://doi.org/10.1104/pp.109.137737.
36. Irish, V.F. The Flowering of Arabidopsis Flower Development. *Plant J.* **2010**, *61*, 1014–1028. https://doi.org/10.1111/j.1365-313X.2009.04065.x.
37. Nakamura, Y.; Teo, N.Z.W.; Shui, G.; Chua, C.H.L.; Cheong, W.F.; Parameswaran, S.; Koizumi, R.; Ohta, H.; Wenk, M.R.; Ito, T. Transcriptomic and Lipidomic Profiles of Glycerolipids during Arabidopsis Flower Development. *New Phytol.* **2014**, *203*, 310–322. https://doi.org/10.1111/nph.12774.
38. Samuels, L.; Kunst, L.; Jetter, R. Sealing Plant Surfaces: Cuticular Wax Formation by Epidermal Cells. *Annu. Rev. Plant Biol.* **2008**, *59*, 683–707. https://doi.org/10.1146/annurev.arplant.59.103006.093219.
39. Li-beisson, Y.; Shorrosh, B.; Beisson, F.; Andersson, M.X.; Arondel, V.; Bates, P.D.; Bird, D.; Debono, A.; Durrett, T.P.; Franke, R.B.; et al. Acyl-Lipid Metabolism. *Arab. Book* **2013**, *11*, e0161. https://doi.org/10.1199/tab.0161.