

RESEARCH ARTICLE

Exploring Triacylglycerol Biosynthetic Pathway in Developing Seeds of Chia (*Salvia hispanica* L.): A Transcriptomic Approach

Sreedhar R. V.¹, Priya Kumari^{1,3}, Sunny D. Rupwate², Ram Rajasekharan², Malathi Srinivasan^{1,3*}

1 CSIR-Lipidomic Centre (CSIR-LIPIC), CSIR-Central Food Technological Research Institute (CSIR-CFTRI) Resource Centre, Allalandra, GKVK Post, Bangalore—560 065, Karnataka, India, **2** CSIR-Lipidomic Centre (CSIR-LIPIC), CSIR-Central Food Technological Research Institute (CSIR-CFTRI), Mysore—570 020, Karnataka, India, **3** Academy of Scientific and Innovative Research, CSIR, New Delhi—110 025, India

☯ These authors contributed equally to this work.

* m.srinivasan@cftri.res.in



OPEN ACCESS

Citation: R. V. S, Kumari P, Rupwate SD, Rajasekharan R, Srinivasan M (2015) Exploring Triacylglycerol Biosynthetic Pathway in Developing Seeds of Chia (*Salvia hispanica* L.): A Transcriptomic Approach. PLoS ONE 10(4): e0123580. doi:10.1371/journal.pone.0123580

Academic Editor: Amit Dhingra, Washington State University, UNITED STATES

Received: October 7, 2014

Accepted: March 4, 2015

Published: April 13, 2015

Copyright: © 2015 R. V. et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All raw paired end sequence data files are available from the NCBI database (accession number PRJNA196477).

Funding: This work was financially supported by Council of Scientific and Industrial Research (CSIR), INDIA through 12th Five year plan project LIPIC (BSC0401). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Chia (*Salvia hispanica* L.), a member of the mint family (Lamiaceae), is a rediscovered crop with great importance in health and nutrition and is also the highest known terrestrial plant source of heart-healthy omega-3 fatty acid, alpha linolenic acid (ALA). At present, there is no public genomic information or database available for this crop, hindering research on its genetic improvement through genomics-assisted breeding programs. The first comprehensive analysis of the global transcriptome profile of developing *Salvia hispanica* L. seeds, with special reference to lipid biosynthesis is presented in this study. RNA from five different stages of seed development was extracted and sequenced separately using the Illumina GAIIx platform. *De novo* assembly of processed reads in the pooled transcriptome using Trinity yielded 76,014 transcripts. The total transcript length was 66,944,462 bases (66.9 Mb), with an average length of approximately 880 bases. In the molecular functions category of Gene Ontology (GO) terms, ATP binding and nucleotide binding were found to be the most abundant and in the biological processes category, the metabolic process and the regulation of transcription-DNA-dependent and oxidation-reduction process were abundant. From the EuKaryotic Orthologous Groups of proteins (KOG) classification, the major category was “Metabolism” (31.97%), of which the most prominent class was ‘carbohydrate metabolism and transport’ (5.81% of total KOG classifications) followed by ‘secondary metabolite biosynthesis transport and catabolism’ (5.34%) and ‘lipid metabolism’ (4.57%). A majority of the candidate genes involved in lipid biosynthesis and oil accumulation were identified. Furthermore, 5596 simple sequence repeats (SSRs) were identified. The transcriptome data was further validated through confirmative PCR and qRT-PCR for select lipid genes. Our study provides insight into the complex transcriptome and will contribute to further genome-wide research and understanding of chia. The identified novel UniGenes will facilitate gene discovery and creation of genomic resource for this crop.

Introduction

A low ratio of ω -6/ ω -3 fatty acids (5 or lower) in the diet is considered healthy for combating cardiovascular diseases, cancer, inflammation and autoimmune diseases. However, this ratio averages 16.74 in the US, 15 in the UK and northern Europe and, alarmingly, as high as 38–50 in urban India [1]. Flax seed, marine algae and fish are potent sources of ω -3 fatty acids, but their use is limited as a result of certain anti-nutritional factors, fishy flavor, low availability, contamination with mercury compounds and non-vegetarian nature [2]. Chia (*Salvia hispanica* L.), a member of the mint family (Lamiaceae), is native to Mexico and parts of South America. The seeds contain approximately 26–35% oil by weight and have the highest known content of the omega-3 fatty acid, α -linolenic acid (approximately 60% of the total fatty acids) [3]. Chia oil, with very high omega-3 fatty acid (α -linolenic acid/ ALA) content, represents an excellent alternative source of this nutritive compound. Chia oil has a ω -6/ ω -3 fatty acid ratio of <1 , and regular use can assist in obtaining a healthy overall dietary ratio. α -Linolenic acid is a known precursor for the synthesis of very long-chain polyunsaturated fatty acids such as eicosapentaenoic acid and docosahexaenoic acid, which have an important role in brain development, cognition, behavior and eyesight [4] and reduce the risk of coronary heart diseases [1]. Chia seeds are also a rich source of protein (up to 26%) [3]; total dietary fiber (up to 40%) [5]; antioxidants; essential minerals such as calcium, magnesium, phosphorus, iron, zinc and copper; and vitamins such as thiamine, niacin, riboflavin, vitamin C and vitamin E [6,7], which potentially supports their reputation as a ‘superfood.’

Chia was an important staple food crop of pre-Columbian Central America along with maize, amaranth and beans. This crop was re-discovered recently due to its superior nutritional qualities. Chia, which is considered a dietary nutritional supplement, is safe and is exempted from regulation by the US Food and Drug Administration. The European Commission authorizes the sale of chia seeds in the Union as a novel food ingredient with a daily intake of up to 15 grams, and chia seeds can be used in baked products, breakfast cereals, and fruit, nut and seed mixes in levels of up to 10% [8]. The seeds can be eaten raw as whole or ground seeds or as sprouts. The seed flavor is very mild, and can be considered almost flavorless. The seeds can be added to almost anything, including cereals, salads, soups, yogurt, sweets, baby foods, health bars, breads, cakes and other foods. Commercially available products made with chia include sports nutrition bars, breakfast bars, chocolate bars, oil soft-gel capsules, seed flour and various types of cakes, breads and other bakery products. Presently, chia is cultivated commercially in Argentina, Australia, Bolivia, Ecuador, Guatemala and Mexico [9]. We have for the first time, successfully evaluated its adaptability and suitability for commercial cultivation under tropical South Indian conditions. We are currently in the process of genetically improving this crop to develop lines with a high yield, higher oil content and modified fatty acid profiles to be suitable for cooking and to meet other nutritional requirements.

Modern molecular breeding techniques with biotechnological intervention have revolutionized the process of genetic improvement in plants. Various genomic tools have greatly facilitated the development of improved genotypes or varieties in several crops. Whole-genome crop sequences are very useful in crop improvement programs. With high resource and time demands, the generation of these sequences has been restricted to the most important crops. Transcriptome sequencing by next-generation sequencing (NGS) technologies and analysis with powerful bioinformatic tools has become a fast and inexpensive alternative, especially for non-model or orphan crop species in which reference genomes or sequence data are unavailable. The generation of large-scale expressed sequence tags (ESTs) or transcript data and molecular markers have accelerated breeding research in these crops. The availability of a large number of genetic markers developed through NGS technologies is facilitating trait mapping and linkage mapping, making marker-assisted breeding more feasible [10].

In addition to aiding in genomics-assisted breeding programs, transcriptomic analysis through NGS can facilitate the genetic modification of plants. Genetic engineering of plants to modify their nutritional compositions has been successful, particularly in oil-seed crops such as canola, brown mustard, soybean, safflower and linseed [11]. Transgenic safflower with seed oil containing >70% gamma linolenic acid (GLA) [12] has been recently commercialized in the USA, and transgenic soybeans producing 15–30% stearidonic acid (SDA) and 5–8% GLA developed by Monsanto and Solae LLC (Limited Liability Company) are close to commercialization after achieving successful safety assessments and GRAS (generally recognized as safe) notice [13]. Oil deposition has been also enhanced by the over-expression of diacylglycerol acyltransferase (AtDGAT) cDNA in wild-type *Arabidopsis thaliana* [14] and by the up-regulation of two soybean Dof-type transcription factor (GmDof) genes associated with fatty acid biosynthesis [15].

A prerequisite for introducing successful and stable genetic manipulation is to understand the transcribed gene and pathway information. Insight into the spatial and temporal control of gene expression and an integrated model involving biochemical pathways and differentially expressed genes (DEGs) during lipid biosynthesis can reveal the gene expression profile, important genes and their levels of expression. Such insight helps to identify, characterize and modify the potential transcripts of interest.

With no public genomic information (except for a partial nucleotide sequence of the chloroplast *rbcl* gene for rubisco; GenBank: Z37442.1) or database entries available for chia, our major aim was to establish a public information platform for further functional genomic studies. We also aimed at identifying the transcripts involved in lipid biosynthesis. In the present study, we present the first comprehensive analysis of a global transcriptome profile of developing chia seeds. We used Illumina GAIIX sequencing platform to sequence the transcriptome and to investigate gene expression in developing seeds of chia. To provide the most comprehensive representation of the developing seed transcriptome, we extracted and separately sequenced RNA from five different stages of seed development (3, 7, 14, 21 and 28 days after flower opening), assuming differential gene expression levels through the developmental stages of the seeds.

Results and Discussion

With the aim of characterizing the transcriptome of developing seeds of *S. hispanica*, we sequenced five cDNA samples from five different stages of seed development using the Illumina GAIIX platform. Each run yielded approximately 35 million raw reads, ultimately yielding 3.5GB data. The raw paired-end sequence data in the FASTQ format was deposited in the National Center for Biotechnology Information's (NCBI) BioProject database (as a Short Read Archive; SRA) under accession number PRJNA196477. A pooled data set was created by combining the reads from five different sequenced libraries and subjected to *de novo* assembly. *De novo* assembly of processed reads using Trinity yielded 76,014 transcripts after filtering out those shorter than 200 bases. The details of the pooled transcriptome are provided in [Table 1](#).

The total transcript length was 66,944,462 bases (66.9 Mb), with an average length of approximately 880 bases. The transcripts were 54.19% AT, which was marginally higher than the GC content of 45.81%. There were 24,754 transcripts with a length of more than 1000 bp. In another species of *Salvia* (*Salvia splendens*) [16], *de novo* assembly and annotation of transcriptome yielded 83,093 and 81,127 transcripts in two different strains (strain 35 and Cailinghong) with average length of 905 and 919 bases, respectively. A total of 28,197 and 27,992 transcripts, respectively, were more than 1000 bp. There was also difference in the GC content which was approximately 50% in *S. splendens* transcriptome compared to 45.81% in *S. hispanica*. The N50

Table 1. Transcriptome summary statistics.

Total number of transcripts	76,014
Minimum transcript length (bp)	201
Maximum transcript length (bp)	11,579
Average transcript length (bp)	880.69
Total transcript length (bp)	66,944,462
Transcripts >500 bp	43,435
Transcripts >1 kbp	24,754
Transcripts >5 kbp	129
N50 contig size (bp)	1338
GC %	45.81

doi:10.1371/journal.pone.0123580.t001

value of our assembly (1338 bases) was almost similar to that of *S. splendens* (1346 bases) [16]. However, N50 contig size of chia assembly was 1338 bases which is higher than most of the recently published plant transcriptome assemblies like celery (1204 and 1193 bases) [17], *Houttuynia cordata* Thumb. (1051 bases) [18], Chinese hazelnut (799 bases) [19] and Geraniaceae (478 and 1319 bases) [20], indicating good transcriptome assembly. In medicinal plant *Salvia miltiorrhiza*, the transcriptome sequencing of leaf and root tissues using 454 GS-FLX system and its assembly yielded 64,139 unigenes with average length of 413 bases [21]. The N50 value of isotigs was 750 bases. Sequencing of transcriptome of *Salvia sclarea* L calyx using 454 GS-FLX system resulted in identification of 45,822 unique sequences with average read size of 214 bp [22]. The average GC content was 45% which is similar to that of our present study in *S. hispanica*.

Functional annotation and candidate genes encoding enzymes involved in lipid biosynthesis

Functional annotation of the pooled transcripts was conducted using BLASTX (E value ≤ 0.001) against the UniProt Plant Protein database, which resulted in 83% of the transcripts being annotated. A total of 27,122 unique genes were mapped. The top five homologous plant genomes contributing to predicted transcript annotation were *Vitis vinifera* (44%), *Populus balsamifera* (22%), *Ricinus communis* (19%), *A. thaliana* (3%) and *Glycine max* (3%) (S1 Table). A similar trend was also observed in *S. splendens* [16]. The top proteins contributing were pentatricopeptide repeat-containing protein-putative (18%), DNA binding protein-putative (13%), an uncharacterized protein (12%), protein-binding protein-putative (11%), transcription factor-putative (9%), zinc finger protein-putative (8%) and Cytochrome P450 (8%). There were additional homology searches made using BLASTX to further annotate genes in transcriptome by using known sequence of lipid related genes like peanut soluble MGAT [23].

Pathways contributing to lipid biosynthesis and modification have been well documented in major oil seed crops. Through our study, most of the known enzymes involved in lipid biosynthesis were successfully identified in chia based on the annotation of the transcripts obtained. The percentage distribution of lipid biosynthesis-related transcripts identified in chia is provided in Fig 1. Glycerophospholipid metabolism, inositol phosphate metabolism, the phosphatidylinositol signaling system, glycosylphosphatidylinositol-anchor biosynthesis and glycerolipid metabolism-related transcripts were the highest in number, and transcripts related to α -linolenic acid metabolism, linoleic acid metabolism and arachidonic acid metabolism were also prominent.

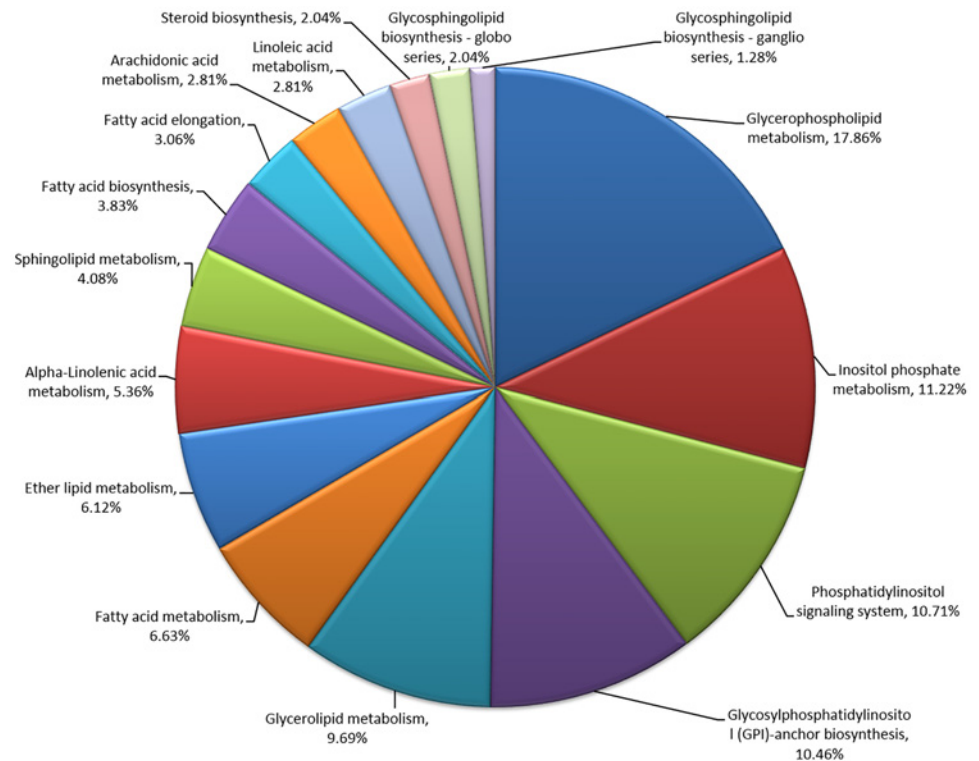


Fig 1. Percentage distribution of lipid synthesis-related transcripts.

doi:10.1371/journal.pone.0123580.g001

Metabolic pathway related to oil accumulation

Triacylglycerol (TAG) is biosynthesized *de novo* by a series of enzymatic acylation reactions of the glycolytic intermediate glycerol-3-phosphate (G3P) that are acyl-CoA dependent. Certain dephosphorylation steps are also required for TAG synthesis.

Coding DNA sequences (CDS) encoding enzymes in the TAG biosynthetic pathway were identified in the transcriptome, including those in both the Kennedy pathway [24, 25] and the soluble TAG biosynthesis pathway [23] (Fig 2). The biosynthetic machinery responsible for the formation of TAG must compete with phospholipid biosynthesis for a common intermediate precursor, i.e., phosphatidic acid (PA). In yeast, carbon flux in these pathways was recently shown to be regulated by a phospholipase C [26]. The exploration of such regulation in plant systems like chia, using its transcriptome data, could be interesting. The first step in the biosynthesis of glycerolipid/glycerophospholipid is the acylation of glycerol-3-phosphate (G3P) at its two free hydroxyl positions to produce lysophosphatidic acid (LPA) and then PA by the enzymes G3P acyltransferase (GPAT) and LPA acyltransferase (LPAT), respectively. Both of these products can be dephosphorylated to monoacylglycerol (MAG) and diacylglycerol (DAG); PA is dephosphorylated by phosphatidic acid phosphatase (PAP). DAG is then acylated to produce TAG. The PA-independent pathway leading to the formation of TAG from DAG via MAG involves two acylation steps after the action of lipid phosphatase on LPA [27]. The acylation of monoacylglycerol by MAG acyltransferase (MGAT) has been described in plants, and it is interesting to note that the acylation is carried out by two types of enzymes, a soluble MGAT [27] and the oil body-associated protein, oleosin. Previously, oleosins were thought to be structural proteins, but recently, they have been shown to exhibit MGAT and

GLYCEROLIPID METABOLISM

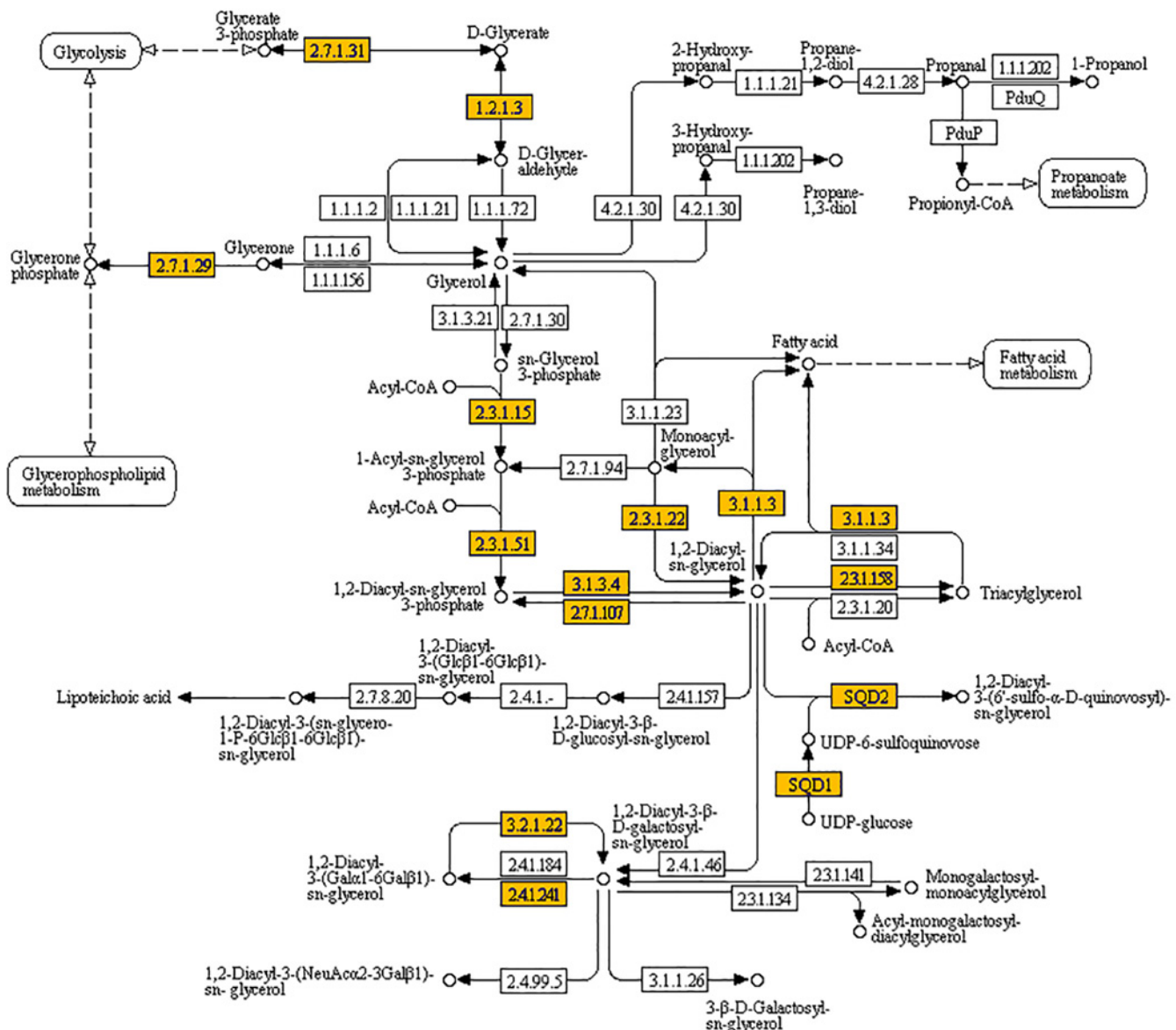


Fig 2. Glycerolipid metabolism pathway. Genes highlighted are those identified in chia in the present study.

doi:10.1371/journal.pone.0123580.g002

phospholipase activities [28]. In our analysis of the chia transcriptome, four CDS corresponding to oleosin and two CDS corresponding to soluble MGAT were identified.

In oilseeds like chia, TAG is sequestered in the oil bodies, which are made up of a hydrophobic core mainly comprising TAG and sterol ester that is surrounded by a phospholipid monolayer containing proteins such as oleosin, steroleosin and caleosin, which are vital for the appropriate accumulation and storage of TAG. Many other proteins play crucial roles in supporting TAG biosynthesis, including the oil body-associated protein discussed above, acyl carrier protein and other seed-specific proteins.

elongated by a sequence of condensation reactions with either acyl-ACP or acetyl-CoA by multiple isoforms of β -ketoacyl-ACP synthase. Furthermore, CDS encoding fatty acid desaturase enzymes were also identified. The other family of enzymes that can dictate the fatty acid profile of an organism is acyl-ACP thioesterases, and the corresponding reads were identified in the chia transcriptome.

Gene Ontology (GO) annotation

GO is used widely for the standard representation of genes across species and to provide a controlled vocabulary of terms for describing gene products [31]. The contigs from the pooled transcriptome were assigned GO terms based on BLASTP searches against sequences with previously identified products (S2 Table). The three categories of GO terms (biological processes, cellular components and molecular functions) were represented by 870, 236 and 1203 terms, respectively (Fig 4). A similar pattern was observed in *S. miltiorrhiza* transcriptome [21]. The top ten GO terms in each component are also shown. In the molecular functions category, ATP binding and nucleotide binding were found to be the most abundant terms, whereas in the biological processes category, the metabolic process and the regulation of transcription-DNA-dependent and oxidation-reduction process were abundant. Predominance of metabolic process terms indicates metabolism as major process in developing seeds. In *S. splendens*, biological process ontology distribution pattern indicated multiple developmental processes [16] which may be due to the transcriptome analysis of tissues from leaf, stem, shoot and root tissues. Membrane, integral to membrane and nucleus classes were prominent GO terms in the cellular components category.

KOG annotation

KOG terms derived from the annotation of 8,288 transcripts using KOG proteins from eukaryotic clusters are provided in S3 Table. The KOG classifications with multiple assignments were individually assessed and assigned to transcripts (individual KOG assignment; the redundant

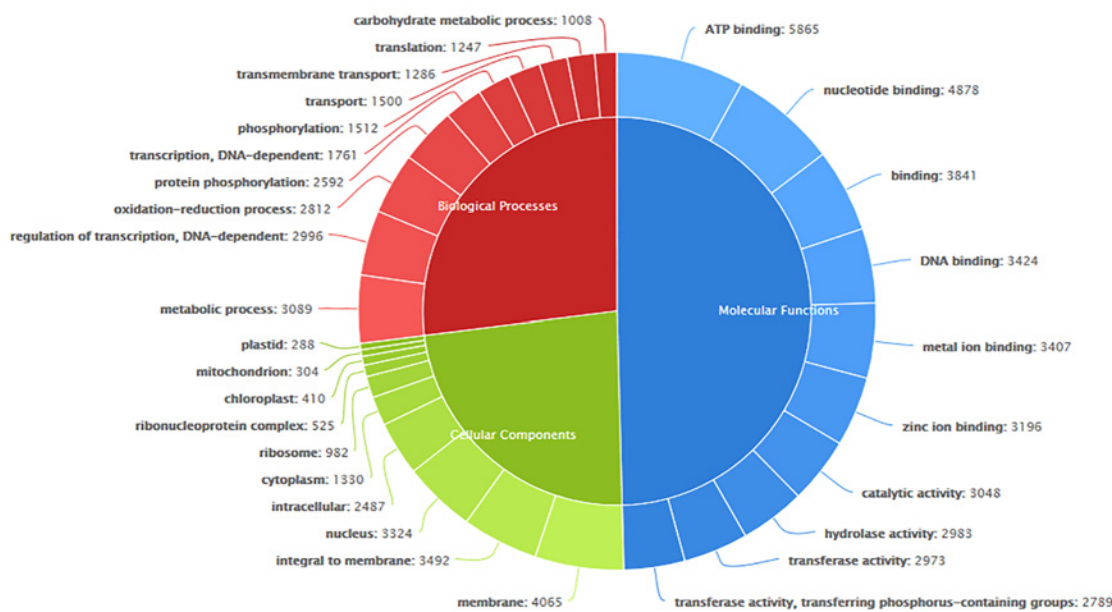


Fig 4. GO classification. The top 10 GO terms in Cellular components, Molecular function and Biological processes are displayed.

doi:10.1371/journal.pone.0123580.g004

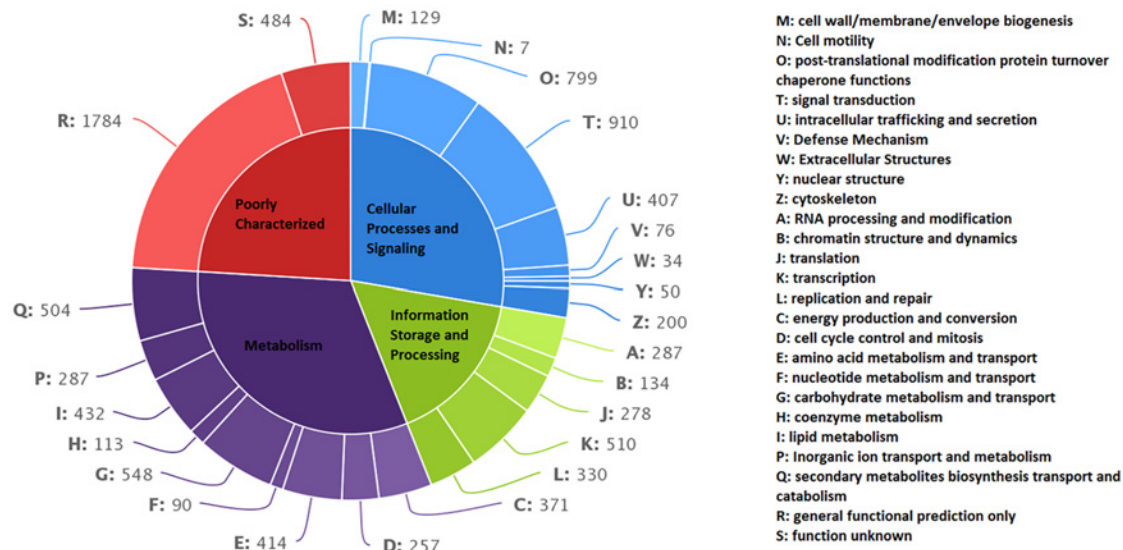


Fig 5. KOG classification. KOG functional categories of the transcripts after annotation against the KOG proteins.

doi:10.1371/journal.pone.0123580.g005

transcript list is shown in Fig 5). From the KOG classification, the major category was “Metabolism” (31.97%), of which the most prominent class was ‘carbohydrate metabolism and transport’ (5.81% of total KOG classifications) followed by ‘secondary metabolite biosynthesis transport and catabolism’ (5.34%) and ‘lipid metabolism’ (4.57%). It is noteworthy that in the “metabolism” category, the ‘carbohydrate metabolism and transport’ and ‘lipid metabolism’ classes were the best represented, highlighting the data integrity at the sequencing and analysis levels. Developing chia seeds accumulate carbohydrates and lipids at high levels. This phenomenon is reflected in proteins related to metabolism being prominently represented in the KOG classification. The second major category was “Cellular processes and Signaling” (27.68%), in which the most frequently observed classes were ‘signal transduction’ (9.64%), ‘post-translational modification protein turnover’ (8.47%) and ‘intracellular trafficking and secretion’ (4.31%). In the “Information storage and processing category”, the highly occurring classes were ‘transcription’ (5.41%), ‘replication and repair’ (3.5%) and ‘RNA processing and modification’ (3.04%). Of the “Poorly characterized” category annotations, ‘general functional prediction only’ represented 18.91% and ‘function unknown’ represented 5.1% of the transcripts. The prominent presence of the “Poorly characterized” category may have occurred because chia is distantly related to the organisms comprising the eukaryotic KOG database.

Validation of Transcriptome

Since transcript assembly needs validation and as our interest is in chia lipid metabolism, we chose to validate eight lipid genes. The transcripts of all the 8 lipid genes, viz., *MGAT*, *OLE1*, *DGAT1*, *DGAT2*, *DGAT3*, *PDAT*, *Thiolase* and *Desaturase* were confirmed by PCR, as was observed in an agarose gel. The amplicon sizes matched with the expected size of the gene based on assembled transcripts (Fig 6). *MGAT*, *OLE1*, *DGAT2*, *DGAT3*, *Delta-15-desaturase* and *omega-3 desaturase* PCR products were further sequenced for confirmation, giving ~100% identity with the assembled transcriptome sequences (S1 File). This is the first time that these genes have been identified in *S. hispanica*. The sequences have been deposited at NCBI as mRNA sequences with the accession numbers, GenBank: KM016233 and GenBank: KM016234 for *OLE1* and *MGAT*, respectively. Further, the protein sequences of *MGAT*, *OLE1*,

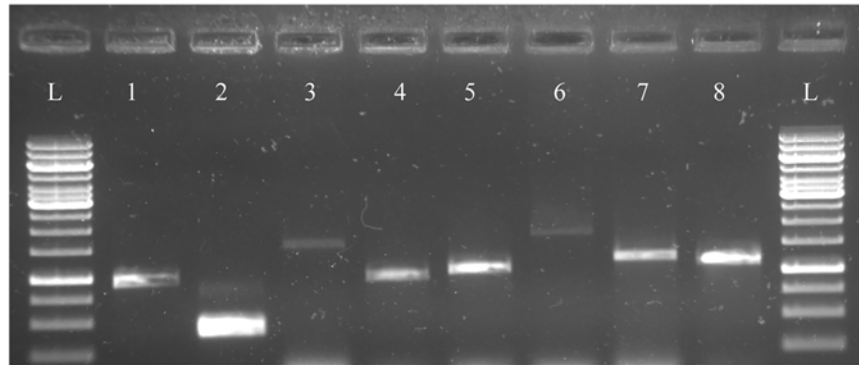


Fig 6. Detection and validation of different lipid genes by PCR. cDNA was used as template to amplify lipid genes. The amplified fragments were analyzed by 0.8% agarose gel electrophoresis; 1 kb DNA ladder containing 14 bands (from down 250, 500, 750, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 5000, 6000, 8000, 10000 bp; GeneRuler 1 Kb DNA ladder from Thermo Scientific, cat no. SM0311) was used. Lane L: 1 kb DNA ladder, Lane1: *MGAT* (963 bp), Lane2: *OLE1* (429 bp), Lane3: *DGAT1* (1620 bp), Lane4: *DGAT2* (1014 bp), Lane5: *DGAT3* (1098 bp), Lane6: *PDAT* (1881 bp), Lane7: *Desaturase* (1300 bp), Lane8: *Thiolase* (1179 bp)

doi:10.1371/journal.pone.0123580.g006

DGAT2, *DGAT3*, *Delta-15-desaturase* and *omega-3 desaturase* of chia were aligned with known sequences of other oilseed plants for homology ([S2 File](#)).

Determination of mRNA Copy Number

Quantitative Reverse transcriptase polymerase chain reaction based method was used for absolute quantification of copy number of mRNA of the genes *MGAT*, *OLE1*, *DGAT2*, *DGAT3*, *Delta-15-desaturase* and *omega-3 desaturase*. A standard curve constructed by using linear DNA as template, was highly linear for all genes (R^2 value was ≥ 0.98). PCR amplification efficiency (E) was more than 85% and the slopes of the standard curves were between 3.2 and 3.7. The expression levels of the six genes across the various developmental stages as determined by qRT-PCR were generally in agreement with the levels observed using NGS sequencing ([Fig 7](#), [S4 Table](#)).

SSR mining

Microsatellite or SSR markers are the repeats of 2–6 nucleotide bases. These markers are the most important molecular markers, with broad applications in many genetic and genomic studies. The development of EST-SSR markers through transcriptome sequencing has greater advantages than genomic SSRs due to their direct association with transcribed genes and inexpensive development. EST-SSRs are more likely to be tightly linked to a specific gene function and may play direct roles in controlling important agronomic characteristics [32]. EST-SSRs are expected to be more transferable than random genomic SSRs across lines, populations and species [33]. In the present study, 5596 SSRs were identified in 5415 transcripts ([S5 Table](#)), out of which 454 sequences contained more than one SSR. The most abundant SSRs were trinucleotide type motifs (2818, 50.4%) followed by dinucleotide type motifs (1980, 35.4%), hexanucleotide type motifs (360, 6.4%), tetranucleotide type motifs (237, 4.2%) and pentanucleotide type motifs (201, 3.6%) ([Fig 8](#)). On the contrary, out of the 2,453 SSRs discovered, dinucleotide repeat motifs (39.9%) were the most abundant followed by trinucleotide (29.3%), tetranucleotide (0.8%), pentanucleotide (0.7%) and hexanucleotide (0.7%) repeat motifs in *S. splendens* [16]. This may be mainly because of species variation or difference in the tissue used for transcriptome sequencing. The most prominent repeat motif type was (AG/CT) followed by

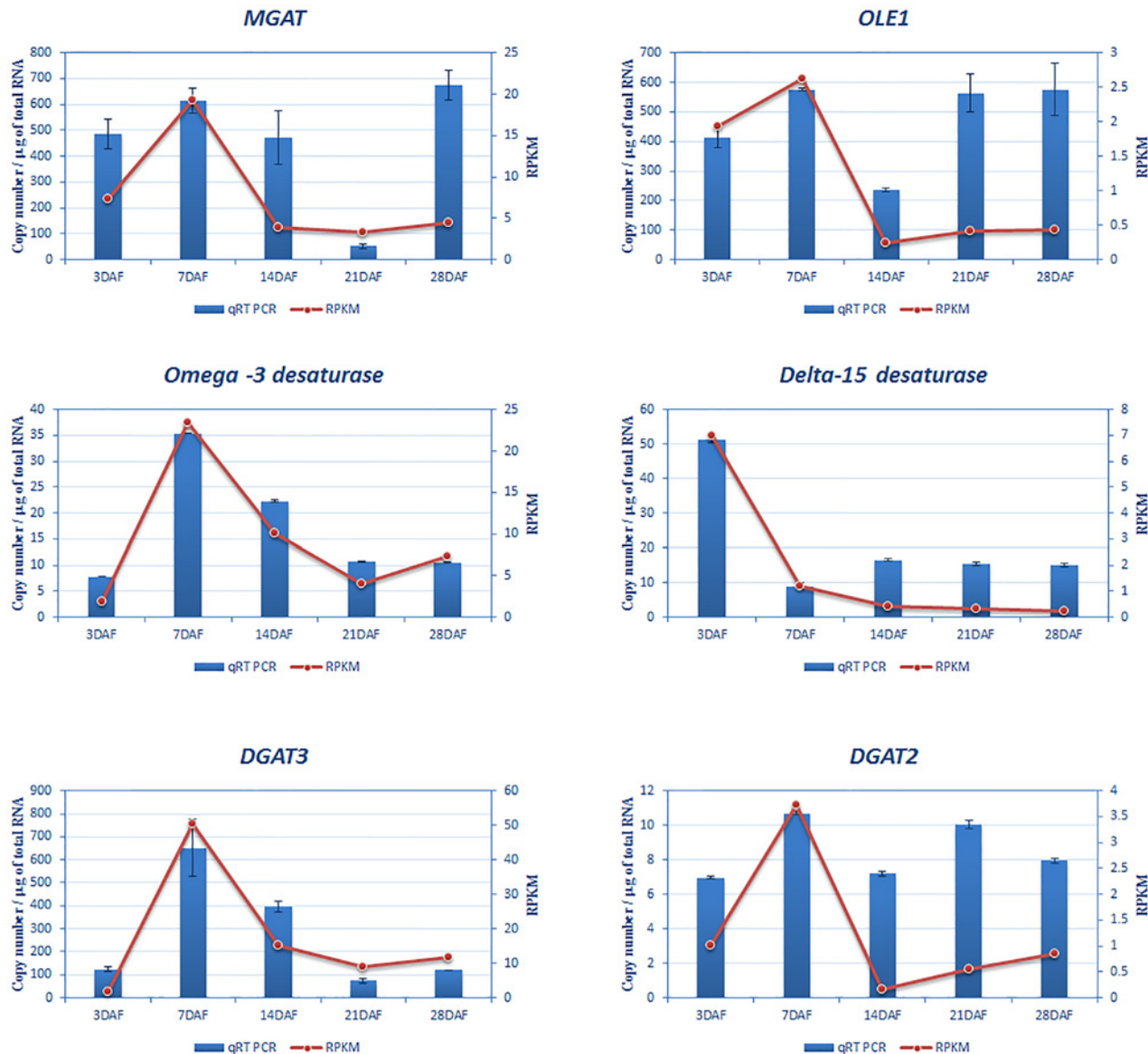


Fig 7. Real Time PCR analysis to validate the transcriptome data. Validation of expression pattern of transcriptome of *OLE1*, *MGAT*, *DGAT2*, *DGAT3*, *Delta 15 desaturase* and *Omega-3 desaturase* by quantitative qRT-PCR was carried out using total RNA isolated from different stages of Chia seeds (3, 7, 14, 21 and 28 DAF). Absolute copy number of transcripts was calculated. Error bars represent standard deviation between three replicates. Expression value RPKM obtained by NGS for same stages is also represented.

doi:10.1371/journal.pone.0123580.g007

(GA/TC), (CCG/CGG), (CA/TG), (GAA/TTC), (GGA/TCC), (CTC/GAG), (CAG/CTG), (AGA/TCT) and (CGC/GCG) (Fig 9). Chia, being an oil seed plant, we also identified the transcripts related to TAG biosynthesis that have SSRs, most of which were highly expressed throughout the various developmental stages (S6 Table). In a non-model plant such as *S. hispanica* that has no genomic sequences available; SSRs such as those mined in our study can be invaluable resource for genetic improvement and the construction of genetic map.

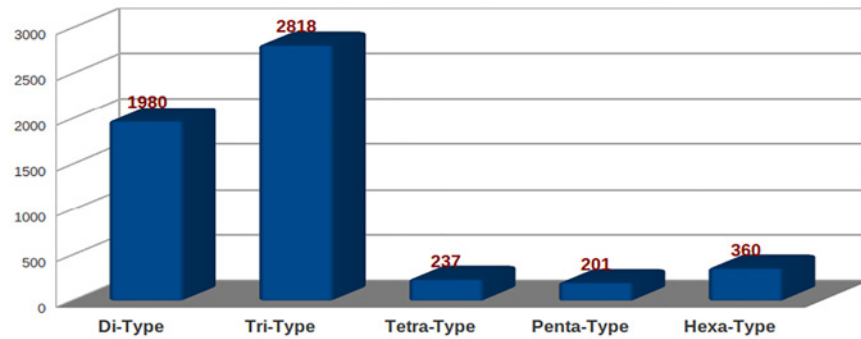


Fig 8. Number of each type of SSR detected in chia transcriptome.

doi:10.1371/journal.pone.0123580.g008

Developmental stage based differences in gene expression levels

Although we have provided a comprehensive account of the chia transcriptome pooled from 5 different seed developmental stages, detailed analyses of the assembled contigs in each developmental stage is underway. For instance, we have analyzed the differential gene expression levels, especially of genes involved in lipid metabolism, across the five stages. Heat maps representing the gene expression levels of the TAG biosynthetic genes like LPAT, GPAT, MGAT, DGAT etc. and of the fatty acid biosynthetic genes like desaturases and elongases across the five stages of chia seed development have been provided ([S7](#) and [S8](#) Tables).

Conclusions

For the first time, we have established a public genomic information platform for chia (*S. hispanica* L.), a rediscovered crop of immense health and nutritional importance that is also the best known plant source of omega-3 fatty acids. This comprehensive transcriptome analysis data can form the basis for accelerated genomics-assisted genetic improvement programs and facilitate a better understanding and more effective manipulation of biochemical pathways, including lipid biosynthesis. The data across five different seed developmental stages, with special reference to genes involved in lipid metabolism can serve as very valuable information, especially in plant lipid research focusing on improving the quality and quantity of chia seed oil and in promoting it as commercial edible oil. This knowledge can subsequently enable more focused and intensive research on this underutilized crop, leading to the development of better

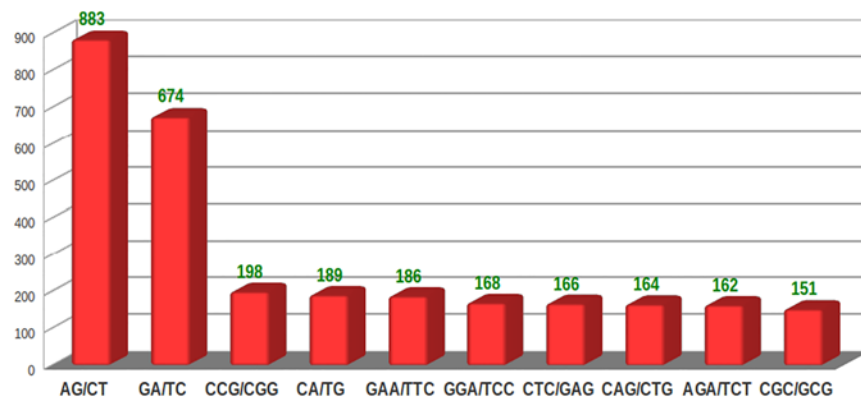


Fig 9. Frequency of classified repeat types.

doi:10.1371/journal.pone.0123580.g009

cultivars or lines promoting its large-scale commercial cultivation and utilization, especially in developing countries with rampant malnutrition.

Materials and Methods

Ethics statement

No specific permissions were required for these locations or activities and the study did not involve any endangered or protected species, since plants were grown in our own private field at CSIR-Central Food Technological Research Institute Resource Centre (CSIR-Central Institute of Medicinal and Aromatic Plants Research Farm) in Bangalore located at 12°58' N latitude and 77°35' E longitude with an altitude of 930 msl and annual mean rainfall of 870 mm in red loamy soil. As a result of our genetic improvement program involving chia, a promising, high-yielding line producing blue flowers and white seeds (CHIampion B1) has been developed by us and was used in this study. It is not an endangered species.

Total RNA extraction, library construction and deep sequencing

Developing seeds at 3, 7, 14, 21 and 28 days after flower opening (DAF) were collected from five different plants and the samples were pooled. The 28 days after flower opening seeds represent completely matured embryo or seed. RNA from each pool was isolated using a Qiagen RNeasy Mini Kit following the manufacturer's protocol. The yield and quality were determined using an Agilent 2100 Bioanalyzer and a NanoDrop spectrophotometer. For RNA library construction and deep sequencing, equal quantity of RNA from each pool was used. Approximately 6 µg of RNA representing each group was sequenced using the Illumina GAIIx platform. The workflow of the present study is represented in [S1 Fig](#).

Raw data preprocessing

Raw reads were subjected to quality check using SeqQC. High-quality bases ($Q \geq 20$) comprised more than 85% of the bases of both the forward and reverse paired-end reads. The raw data was filtered to remove adaptor reads and low-quality (Phred Score < 20) reads using the NGS QC Toolkit [34], yielding a dataset consisting of clean reads.

De novo transcriptome assembly

High-quality, filtered paired and orphan sequence reads (Phred Score ≥ 20) obtained from all 5 samples were pooled to serve as a representative transcriptome and were provided as an input for transcriptome assembly. Trinity Assembler [35] was used to generate unique transcripts (UTs). Trinity assembly is based on the de Bruijn graph [36]. The assembled UTs with sequence lengths longer than 200 bp were considered for downstream transcript annotation and quantification.

Annotation and quantification of the transcriptome

Annotation for all the UTs (≥ 200 bp) was performed using a BLAST [37] homology search against the UniProt Plant Protein Database [38]. Blast hits with an E-value of $\leq 1e-3$ were considered annotated homologous proteins. GO and KEGG pathway information for the annotated transcripts was assigned based on the homologous proteins. The expression levels of all the unique transcripts (UTs) in individual libraries (3, 7, 14, 21 and 28 DAF) were measured by mapping the HQ filtered reads using the BOWTIE2 tool [39] with final assembled UTs as reference and calculating the RPKM (reads per kilobase per million reads) [40] values based on the mapped reads and the coverage of individual UTs. The RPKM method corrects for biases

in total gene exon size and normalizes to the total number of short-read sequences obtained in each library. The fold expression of the UTs was further obtained by calculating the Log base 2 of RPKM values.

Validation of Transcriptome

To assess the precision and quality of the assembly and annotation data that were derived from the Illumina sequencing, PCR amplification with gene specific primers for eight lipid genes (*MGAT*, *OLE1*, *DGAT1*, *DGAT2*, *DGAT3*, *PDAT*, *Thiolase* and *Desaturase*) that are known to be involved in plant lipid metabolism, using cDNA as template was carried out. Total RNA was extracted from immature chia seeds using TRI Reagent (Sigma). The cDNA was synthesized from 1µg of total RNA using SuperScript III First-Strand Synthesis System (Invitrogen) with oligo dT primers. A set of gene specific primers for the eight genes to be assessed were designed based on the transcriptome sequence and were used in a polymerase chain reaction with the above cDNA as the template, in order to validate the assembled sequence. The primer sequences are provided in Table 2. PCR condition used is as follows: 95°C for 1 min. followed by 35 cycles of 95°C for 30 sec, annealing temperature for 30 sec, extension temperature 72°C for 2 min., followed by final extension 72°C for 5 min. The PCR products were electrophoresed using a 0.8% agarose gel. Further, the PCR amplicons of *MGAT*, *OLE1*, *DGAT2*, *DGAT3*, *Delta-15-desaturase* and *omega-3 desaturase* were Sanger sequenced and compared with the assembled sequences to confirm their identities.

Real time Quantitative Reverse Transcriptase PCR

We confirmed the mRNA expression levels of six lipid genes, viz., *MGAT*, *OLE*, *DGAT2*, *DGAT3*, *Delta-15-desaturase* and *ω-3 desaturase* in seeds collected from five different stages by qRT-PCR. The mRNA samples were prepared from seeds and were used to synthesize the cDNA. 5µg total RNA was used with Super Script First-Strand (Invitrogen; cat.no. #11904–018) with oligo dT primer. A set of gene specific RT primers for all six genes were designed to quantify the copy number. 20µl qRT-PCR reaction mixture containing 1µl of template cDNA,

Table 2. List of PCR primers used for validation.

Gene name	Primer sequence	Expected size (bp)	Location of amplicon	T _a (°C)
<i>MGAT</i>	FW: 5' ATGTCGCCGGAAAATCC 3'	963	Full length	58
	RV: 5' ATTCTTCTTACCATATCTCTCAACT 3'			
<i>OLE1</i>	FW: 5' ATGGCTGATCAACACTACGG 3'	429	Full length	58
	RV: 5' AGAGCTTTGCGCAACGG 3'			
<i>DGAT1</i>	FW: 5' ATGGCGATCACGGACTC 3'	1620	Full length	62
	RV: 5' CCTTGTGCCAGCTTTTCG 3'			
<i>DGAT2</i>	FW: 5' ATGTCGTCTGAATCCAACGG 3'	1014	Full length	50
	RV: 5' TAGAATCCTAAGCTGCAAGTCTG 3'			
<i>DGAT3</i>	FW: 5' ATGGACGCCGCTGCCATG 3'	1098	Full length	65
	RV: 5' AGATGCAGCAGCAAACCAA 3'			
<i>PDAT</i>	FW: 5' AAGAAGCAGCGAAATGGT 3'	1881	91–1871	55.9
	RV: 5' GATCCTTTCGGACCACTTGA 3'			
<i>Thiolase</i>	FW: 5' TGGAAAAACACCAGGAAAGG 3'	1179	63–1241	50
	RV: 5' CCAAAGCTTCCGATCTTGTC 3'			
<i>Desaturase</i>	FW: 5' TAGACTCGCACACTCTGGCTTT 3'	1300	9–1308	65
	RV: 5' AAGGAATGTGATGGGCTGAG 3'			

doi:10.1371/journal.pone.0123580.t002

Table 3. Details of qRT PCR primers.

Gene name	Primer sequence for qRT PCR	Ta (°C)	Amplicon length (bp)	Location of amplicon (bp)
<i>MGAT</i>	FW: 5' CAAGACGCTGAAGCTGTACG3'	58.3	100	828–927
	RW: 5' AATCCAAGCCCTCATGTGACG3'			
<i>OLE1</i>	FW: 5' AGTACATGACCGGGAAGCAC3'	59.4	98	302–399
	RV: 5' GTCCACCCTGTCCCTCATGT3'			
<i>DGAT2</i>	FW: 5' ATGCCCGGTTTGTGGAAG3'	61	83	907–989
	RV: 5' AATCCTAAGCTGCAAGTCTGG3'			
<i>DGAT3</i>	FW: 5' GCACCTCAAGCAGAGTGCAG3'	61	83	779–861
	RV: 5' CTTGCCTCCCATAACACACT3'			
<i>Delta 15 desaturase</i>	FW: 5' CATTAGTGGAGGCGACGAG3'	50	82	955–1036
	RV: 5' ACGGAACGGCTCCAGATT3'			
<i>ω-3 desaturase</i>	FW: 5' CATGACTGTGGCCATGGAAG3'	50	97	91–187
	RV: 5' ATCCATGGTAGGGCACCAA3'			

doi:10.1371/journal.pone.0123580.t003

0.8μl (200nM) each of forward and reverse primers and 10μl of SsoFast Evagreen Super mix (Bio-Rad; cat.no#172-5201AP) was used. The primers used for qRT-PCR are shown in [Table 3](#) with their respective Ta and amplicon sizes. The Tm was determined by running a gradient PCR from 55–65°C. The reaction was carried out under following conditions: 95°C for 30 sec followed by 40 cycles of 95°C for 1 sec, respective annealing temperature for 5 sec + plate read, go to 2 for 40 times, melt curve, 65°C–95°C (in 0.5°C increments), plate read. The intensity of fluorescence was captured at each cycle using a Real Time System Machine (Bio-Rad, CFX96). A threshold for the increase in fluorescence was arbitrarily set in the exponential phase of the amplification plots, and a standard curve was drawn to show the copy number of the standard DNA (linear) vs the threshold cycle.

SSR identification

SSRs were identified using MISA (Micro Satellite identification tool: <http://pgrc.ipk-gatersleben.de/misa/>), wherein dinucleotide and trinucleotide repeats were given a minimum threshold of 6 and 4 repeats, respectively. A minimum threshold of 3 repeats was given for pentanucleotide and hexanucleotide repeats. The maximum distance between two SSRs was specified as 100 bases.

Supporting Information

S1 Fig. Workflow.

(TIF)

S1 File. Sequence alignment between transcriptome assembly and Sanger sequencing.

(DOCX)

S2 File. Sequence alignment of some Chia lipid genes with known oil seed plants.

(DOCX)

S1 Table. Summary of annotated transcripts of unique genes mapped.

(XLSX)

S2 Table. Details of Gene Ontology assigned.

(XLSX)

S3 Table. Summary of KOG annotated transcripts.

(XLSX)

S4 Table. Expression of all transcripts over different stages of seed development.

(XLSX)

S5 Table. Details of SSR sequences identified.

(XLSX)

S6 Table. Transcripts related to TAG biosynthesis having SSRs.

(XLSX)

S7 Table. Expression levels of some of the genes involved in TAG biosynthesis across the 5 developmental stages.

(XLSX)

S8 Table. Expression levels of some important genes involved in fatty acid synthesis across the 5 developmental stages.

(XLSX)

Acknowledgments

Authors are thankful to Bionivid Technology Pvt. Ltd., Bangalore for sequencing and analyses, and to the Council of Scientific and Industrial Research (CSIR), India, for financial support.

Author Contributions

Conceived and designed the experiments: RR MS. Performed the experiments: SRV PK SDR. Analyzed the data: SRV SDR. Contributed reagents/materials/analysis tools: SRV MS. Wrote the paper: SRV RR MS.

References

1. Simopoulos AP (2008) The importance of the omega-6/omega-3 fatty acid ratio in cardiovascular disease and other chronic diseases. *Exp Biol Med* 233:674–688. doi: [10.3181/0711-MR-311](https://doi.org/10.3181/0711-MR-311) PMID: [18408140](https://pubmed.ncbi.nlm.nih.gov/18408140/)
2. Ayerza R, Coates WE (2008) Chia seeds and the Columbus concept. In: Meester, editor. *Wild-type food in health promotion and disease prevention*. Humana Press Inc, Totowa, NJ. pp 377–392.
3. Ayerza R, Coates W (2011) Protein content, oil content and fatty acid profiles as potential criteria to determine the origin of commercially grown chia (*Salvia hispanica* L.). *Ind Crop Prod* 34(2):1366–1371.
4. Bourre JM (2004) Roles of unsaturated fatty acids (especially omega-3 fatty acids) in the brain at various ages and during ageing. *J Nutr Health Aging* 8(3):163–174. PMID: [15129302](https://pubmed.ncbi.nlm.nih.gov/15129302/)
5. Reyes-Caudillo E, Tecante A, Valdivia-Lopez MA (2008) Dietary fiber content and antioxidant activity of phenolic compounds present in Mexican chia (*Salvia hispanica* L.) seeds. *Food Chem* 107:656–663.
6. USDA: National Nutrient Database for Standard Reference, Release 24 (2011) Nutrient Data Laboratory Home Page, Edited by U.S. Department of Agriculture, Agriculture Research Service.
7. Munoz LA, Cobos A, Diaz O, Aguilera JM (2013) Chia seed (*Salvia hispanica*): An ancient grain and a new functional food. *Food Rev Int* 29(4):394–408.
8. Commission implementing decision of 22 January 2013, Official Journal of the European Union, L 21/34, 24.1.2013
9. Small E (2011) Blossoming treasures of biodiversity. *Biodiversity* 12(1):49–56.
10. Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* 27(9):522–530. doi: [10.1016/j.tibtech.2009.05.006](https://doi.org/10.1016/j.tibtech.2009.05.006) PMID: [19679362](https://pubmed.ncbi.nlm.nih.gov/19679362/)
11. Haslam RP, Ruiz-Lopez N, Eastmond P, Moloney M, Sayanova O, Napier JA, (2013) The modification of plant oil composition via metabolic engineering—better nutrition by design. *Plant Biotechnol J* 11:157–168. doi: [10.1111/pbi.12012](https://doi.org/10.1111/pbi.12012) PMID: [23066823](https://pubmed.ncbi.nlm.nih.gov/23066823/)

12. Nykiforuk CL, Shewmaker C, Harry I, Yurchenko OP, Zhang M, Reed C, et al. (2012) High level accumulation of gamma linolenic acid (C18:3Δ6,9,12 *cis*) in transgenic safflower (*Carthamus tinctorius*) seeds. *Transgenic Res* 21(2):367–381. doi: [10.1007/s11248-011-9543-5](https://doi.org/10.1007/s11248-011-9543-5) PMID: [21853296](https://pubmed.ncbi.nlm.nih.gov/21853296/)
13. Monsanto Company (2009) Stearidonic (SDA) Omega-3 Soybean Oil GRAS Notice. Available: http://www.accessdata.fda.gov/scripts/fcn/gras_notices/gm000283.pdf.
14. Jako C, Kumar A, Wei Y, Zou J, Barton DL, Giblin EM, et al. (2001) Seed-Specific Over-Expression of an Arabidopsis cDNA encoding a diacylglycerol acyltransferase enhances seed oil content and seed weight. *Plant Physiol* 126(2):861–874. PMID: [11402213](https://pubmed.ncbi.nlm.nih.gov/11402213/)
15. Wang H-W, Zhang B, Hao Y-J, Huang J, Tian A-G, Liao Y, et al. (2007) The soybean Dof-type transcription factor genes, *GmDof4* and *GmDof11*, enhance lipid content in the seeds of transgenic Arabidopsis plants. *Plant J* 2(4):716–729.
16. Ge X, Chen H, Wang H, Shi A, Liu K (2014) *De novo* assembly and annotation of *Salvia splendens* transcriptome using the Illumina platform. *PLoS ONE* 9(3): e87693. doi: [10.1371/journal.pone.0087693](https://doi.org/10.1371/journal.pone.0087693) PMID: [24622329](https://pubmed.ncbi.nlm.nih.gov/24622329/)
17. Li MY, Wang F, Jiang Q, Ma J, Xiong AS (2014) Identification of SSRs and differentially expressed genes in two cultivars of celery (*Apium graveolens* L.) by deep transcriptome sequencing. *Hort Res* 1:10.
18. Wei L, Li S, Liu S, He A, Wang D, Wang J, et al. (2014) Transcriptome analysis of *Houttuynia cordata* Thumb. by Illumina paired-end RNA sequencing and SSR marker discovery. *PLoS ONE* 9(1): e84105. doi: [10.1371/journal.pone.0084105](https://doi.org/10.1371/journal.pone.0084105) PMID: [24392108](https://pubmed.ncbi.nlm.nih.gov/24392108/)
19. Huima, Lu Z, Liu B, Qiu Q, Liu J (2013) Transcriptome analysis of a Chinese hazelnut species *Corylus mandshurica*. *BMC Plant Biol* 13: 152. doi: [10.1186/1471-2229-13-152](https://doi.org/10.1186/1471-2229-13-152) PMID: [24093758](https://pubmed.ncbi.nlm.nih.gov/24093758/)
20. Zhang J, Ruhlman TA, Mower JP, Jansen RK (2013) Comparative analyses of two Geraniaceae transcriptomes using next-generation sequencing. *BMC Plant Biol* 13: 228. doi: [10.1186/1471-2229-13-228](https://doi.org/10.1186/1471-2229-13-228) PMID: [24373163](https://pubmed.ncbi.nlm.nih.gov/24373163/)
21. Yang L, Ding G, Lin H, Cheng H, Kong Y, Wei Y, et al. (2013) Transcriptome analysis of medicinal plant *Salvia miltiorrhiza* and identification of genes related to tanshinone biosynthesis. *PLoS ONE* 8(11): e80464. doi: [10.1371/journal.pone.0080464](https://doi.org/10.1371/journal.pone.0080464) PMID: [24260395](https://pubmed.ncbi.nlm.nih.gov/24260395/)
22. Legrand S, Valot N, Nicole F, Moja S, Baudino S, Jullien F, et al. (2010) One-step identification of conserved miRNAs, their targets, potential transcription factors and effector genes of complete secondary metabolism pathways after 454 pyrosequencing of calyx cDNAs from Labiate *Salvia sclarea* L. *Gene* 450: 55–62. doi: [10.1016/j.gene.2009.10.004](https://doi.org/10.1016/j.gene.2009.10.004) PMID: [19840835](https://pubmed.ncbi.nlm.nih.gov/19840835/)
23. Saha S, Enugutti B, Rajakumari S, Rajasekharan R (2006) Cytosolic triacylglycerol biosynthetic pathway in oilseeds. Molecular cloning and expression of peanut cytosolic diacylglycerol acyltransferase. *Plant Physiol* 141(4):1533–1543. PMID: [16798944](https://pubmed.ncbi.nlm.nih.gov/16798944/)
24. Kennedy EP (1961) Biosynthesis of complex lipids. *Fed Proc* 20: 934–940. PMID: [14455159](https://pubmed.ncbi.nlm.nih.gov/14455159/)
25. Ohlrogge J B, Somerville CR (1991) The genetics of plant lipids. *Biochim Biophys Acta* 1082(1):1–26. PMID: [1901223](https://pubmed.ncbi.nlm.nih.gov/1901223/)
26. Rupwate SD, Rupwate PS, Rajasekharan R (2012) Regulation of lipid biosynthesis by phosphatidylinositol-specific phospholipase C through the transcriptional repression of upstream activating sequence inositol containing genes. *FEBS Lett* 586(10):1555–1560. doi: [10.1016/j.febslet.2012.04.022](https://doi.org/10.1016/j.febslet.2012.04.022) PMID: [22673525](https://pubmed.ncbi.nlm.nih.gov/22673525/)
27. Vijayaraj P, Jashal CB, Vijayakumar A, Rani SH, Rao DKV, Rajasekharan R, (2012) A bifunctional enzyme that has both monoacylglycerol acyltransferase and acyl hydrolase activities. *Plant Physiol* 160(2):667–683. doi: [10.1104/pp.112.202135](https://doi.org/10.1104/pp.112.202135) PMID: [22915575](https://pubmed.ncbi.nlm.nih.gov/22915575/)
28. Parthibane V, Rajakumari S, Venkateshwari V, Iyappan R, Rajasekharan R (2012) Oleosin is bifunctional enzyme that has both monoacylglycerol acyltransferase and phospholipase activities. *J Biol Chem* 287(3):1946–1954. doi: [10.1074/jbc.M111.309955](https://doi.org/10.1074/jbc.M111.309955) PMID: [22128159](https://pubmed.ncbi.nlm.nih.gov/22128159/)
29. Bates PD, Fatihi A, Snapp AR, Carlsson AS, Browse J, Lu C (2012) Acyl editing and head group exchange are the major mechanisms that direct polyunsaturated fatty acid flux into triacylglycerols. *Plant Physiol* 160:1530–1539. doi: [10.1104/pp.112.204438](https://doi.org/10.1104/pp.112.204438) PMID: [22932756](https://pubmed.ncbi.nlm.nih.gov/22932756/)
30. Wang L, Shen W, Kazachkov M, Chen G, Chen Q, Carlsson AS, et al. (2012) Metabolic interactions between the Lands cycle and the Kennedy pathway of glycerolipid synthesis in Arabidopsis developing seeds. *Plant Cell* 24:4652–4669. doi: [10.1105/tpc.112.104604](https://doi.org/10.1105/tpc.112.104604) PMID: [23150634](https://pubmed.ncbi.nlm.nih.gov/23150634/)
31. Blake JA, Dolan M, Drabkin H, Hill DP, Ni L, Sitnikov D, et al. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res* 40:D559–D564. doi: [10.1093/nar/gkr1028](https://doi.org/10.1093/nar/gkr1028) PMID: [22102568](https://pubmed.ncbi.nlm.nih.gov/22102568/)
32. Bozhko M, Riegel R, Schubert R, Muller-Starck G (2003) A cyclophilin gene marker confirming geographical differentiation of Norway Spruce populations and indicating viability response on excess soil-borne salinity. *Mol Ecol* 12:3147–3155. PMID: [14629393](https://pubmed.ncbi.nlm.nih.gov/14629393/)

33. Park YH, Alabady MS, Ulloa M, Sickler B, Wilkins TA, Yu J, et al (2005) Genetic mapping of new cotton fiber loci using EST-derived microsatellites in an interspecific recombinant inbred line cotton population. *Mol Genet Genomics* 274:428–441. PMID: [16187061](#)
34. Patel RK, Jain M (2012) NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE* 7(2): e30619. doi: [10.1371/journal.pone.0030619](#) PMID: [22312429](#)
35. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29: 644–652. doi: [10.1038/nbt.1883](#) PMID: [21572440](#)
36. Compeau PEC, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol* 29:987–991. doi: [10.1038/nbt.2023](#) PMID: [22068540](#)
37. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *J Mol Biol* 215: 403–410. PMID: [2231712](#)
38. UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucleic Acids Res* 41:D43–D47. doi: [10.1093/nar/gks1068](#) PMID: [23161681](#)
39. Langmead B, Salzberg S (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. doi: [10.1038/nmeth.1923](#) PMID: [22388286](#)
40. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628. doi: [10.1038/nmeth.1226](#) PMID: [18516045](#)