Evolutionary Bioinformatics

ORIGINAL RESEARCH

Phylogenomic Study of Lipid Genes Involved in Microalgal Biofuel Production—Candidate Gene Mining and Metabolic Pathway Analyses

Namrata Misra, Prasanna Kumar Panda, Bikram Kumar Parida and Barada Kanta Mishra

Bioresources Engineering Department, CSIR-Institute of Minerals and Materials Technology (Formerly Regional Research Laboratory), Bhubaneswar, Odisha, India. Corresponding author emails: pkpanda@immt.res.in; pkrpan@gmail.com

Abstract: Optimizing microalgal biofuel production using metabolic engineering tools requires an in-depth understanding of the structure-function relationship of genes involved in lipid biosynthetic pathway. In the present study, genome-wide identification and characterization of 398 putative genes involved in lipid biosynthesis in *Arabidopsis thaliana Chlamydomonas reinhardtii, Volvox carteri, Ostreococcus lucimarinus, Ostreococcus tauri* and *Cyanidioschyzon merolae* was undertaken on the basis of their conserved motif/domain organization and phylogenetic profile. The results indicated that the core lipid metabolic pathways in all the species are carried out by a comparable number of orthologous proteins. Although the fundamental gene organizations were observed to be invariantly conserved between microalgae and *Arabidopsis* genome, with increased order of genome complexity there seems to be an association with more number of genes involved in triacylglycerol (TAG) biosynthesis and catabolism. Further, phylogenomic analysis of the genes provided insights into the molecular evolution of lipid biosynthetic pathway in microalgae and confirm the close evolutionary proximity between the Streptophyte and Chlorophyte lineages. Together, these studies will improve our understanding of the global lipid metabolic pathway and contribute to the engineering of regulatory networks of algal strains for higher accumulation of oil.

Keywords: microalgae, biofuel, lipid biosynthetic genes, phylogenomics, bioinformatics

Evolutionary Bioinformatics 2012:8 545–564

doi: 10.4137/EBO.S10159

This article is available from http://www.la-press.com.

Introduction

Growing levels of atmospheric pollution, mounting energy demand, and the incessant rise in crude oil prices are some of the issues which have in recent times driven global efforts in biofuel research. Currently, commercial-scale biofuels are sourced primarily from a variety bioenergy crops that include sugarcane (*Saccharum officinarum*), sugar beet (*Beta vulgaris*), switch grass (*Panicum virgatum*), soybean (*Glycine max*), canola (*Brassica napus*) and sunflower (*Helianthus annus*).[1] Although the environmental benefits of biofuels as compared to fossil fuels are well established, concerns are being raised about their long-term sustainability, especially against the backdrop of diversion of arable land for biofuel-based cropping systems and their corresponding adverse impact on the global food supply chain.[2] In consequence, algae-based biofuels are increasingly gaining the attention of researchers due to their rapid growth rate coupled with high carbon dioxide uptake, high lipid content and comparatively low, marginal land usage rates.[3]

Notwithstanding the many advantages of biofuels and their technical feasibility, the commercial viability of the algal biofuel process is still an area of concern requiring better strain development and improved post-harvest process engineering.[4] The major challenge is to achieve accumulation of improved lipid profiles with concomitant reduction in energy inputs in order to minimize the cost of production.[2] The enhancement of lipid production in microalgal cells under controlled stress conditions and engineering metabolic pathways are promising strategies to obtain large amounts of standard biofuel for industry. Despite positive experimental reports on enhanced microalgal lipid accumulation under physiological or nutritional stress regimes, many contrasting studies have indicated a concomitant reduction in overall biomass yield under such conditions.[5] In this context, harnessing the potential of genome-scale metabolic engineering has been suggested as a promising area of research to boost oil production in microalgal strains, including modification of algal lipid profile for improved biofuel properties.[6,7]

Over the past few years various studies have been carried out concerning alteration of fatty acid composition in plants through genetic engineering approaches, along with the development and deployment of a number of plant lipid-related genomics databases.[8–11] Comparative genomics analyses using bioinformatics tools have also been performed recently to identify genes involved in lipid biosynthesis in various oleaginous plants. For example, a total of 1003 maize lipid-related genes were cloned and annotated by Lin et al,[12] while Sharma and Chauhan[13] identified a total of 261 lipid genes from the genome of *Arabidopsis*, *Brassica,* soybean and castor. Complete or near complete genome sequences have been reported for several algae.[6] Yet, lack of adequate knowledge regarding the structure-function of lipid biogenesis genes in an evolutionary context is a major impediment in engineering metabolic pathways of algae for over-production of fuel precursors.[14] Various experimental techniques like insertional mutagenesis and targeted gene disruption have been employed to analyze gene function in a few algae. However, many of these approaches are tedious, time-consuming, fiscally prohibitive and limited by a number of biological constraints.[15] As an alternative, phylogenomics is now increasingly used to gain insights into metabolic pathways at the molecular level by comparative genomics and co-evolutionary analyses of related gene.[16] Therefore the present work was designed to identify the genes involved in lipid metabolic pathway from the genomes of microalgae (including *Chlamydomonas reinhardtii, Volvox carteri, Ostreococcus lucimarinus, Ostreococcus tauri* and *Cyanidioschyzon merolae)* using sequence similarity search with *Arabidopsis thaliana* homologs. In addition phylogenomics protocols have been employed to study the structure-function relationship of the encoded proteins and to gain much needed insights into their phylogenetic evolution. We hope that the present study contributes to the biochemical and molecular information needed for augmentation of lipid synthesis in microalgae.

Materials and Methods

Gene retrieval and annotation

An initial set of lipid genes was obtained from the *Arabidopsis thaliana* lipid gene database (http://www.plantbiology.msu.edu/lipids/genesurvey/index.html) to construct a query protein set. The *Arabidopsis* lipid gene database is a convenient and reliable source of genes covering all the major biochemical events responsible for biosynthesis and catabolism of

plant lipids.[17] Subsequently, each protein in the query dataset was used to identify homologs in microalgae by subjecting it to BLASTp[18] search with e-value inclusion threshold set to 0.001 against microalgal genome databases provided by Joint Genome Institute. These include *Cyanidioschyzon merolae* http://merolae.biol.s.u-tokyo.ac.jp/), *Chlamydomonas reinhardtii* (http://genome.jgi-psf.org/chlamy/chlamy.info.html), *Volvox carteri* (http://www.phytozome.net/volvox.php), *Ostreococcus lucimarinus* (http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.home.html), *Ostreococcus tauri* (http://genome.jgi-psf.org/Ostta4/Ostta4.home.html). Based on multiple alignments and/or the presence of conserved motif patterns, some initial sequences "hits" were then discarded. Functional descriptions of genes or gene products were performed by annotation of Cluster of Orthologous groups (COGs) using KOGnitor program,[19] the latter being a widely used tool in the field of computational genomics for detecting candidate set of orthologs in prokaryotes and eukaryotes.[19] In addition, assignment of Gene Ontology (GO) terms describing biological processes and molecular function was annotated by the GO browser and annotation tool AmiGO.[20] The Gene Ontology is currently the pre-eminent approach for functional annotation of homologous genes and protein sequences in multiple organisms.[20]

Metabolic pathway study

Metabolic pathways were subsequently analyzed using the KEGG pathway database,[21] an extensively employed biochemical pathway database to analyze lipid pathways in diverse organisms.[22] To enrich the pathway annotation, sequences were submitted to the KEGG Automatic Annotation Server (KAAS) to identify the orthologous gene groups.[23] KAAS annotates every submitted sequence with a KEGG ortholog (KO) identifier that allows identification of orthologous and paralogous relationships between the genes of interest. Further, a set of six reference pathway maps, namely fatty acid biosynthesis, fatty acid metabolism, fatty acid elongation, glycerolipid metabolism, glycerophospholipid metabolism and pathway map for biosynthesis of unsaturated fatty acids, were downloaded from the KEGG database. This dataset contains a complete biochemical description of the pathways related to the lipid metabolism observed in different organisms. They were used as templates for comprehensive examination of the lipid biosynthetic genomic repertoire of microalgae by correlating genes in the genome with gene products (enzymes), in accordance with their respective Enzyme Commission (EC) number.

Prediction of subcellular localization

Three different protein targeting prediction programs were used to determine the putative subcellular localization of the candidate proteins: TargetP,[24] ChloroP[25] and WolfPsort.[26] Each program is based on different terminology and predictions. The location assignment of TargetP is based on the presence of any of the N-terminal presequences: chloroplast transit peptides (cTP), mitochondrial targeting peptide (mTP) or secretory pathway signal peptide (SP). The ChloroP server predicts the presence of chloroplast transit peptides (cTP) in protein sequences and the location of potential cTP cleavage sites. WolfPsort is an extension of the PSORT II program for protein subcellular localization prediction. It classifies protein into more than 10 location sites, including dual localization such as proteins which shuttle between the cytosol and nucleus. The sensitivity and specificity of this program has been experimentally verified to be 70%.

Physico-chemical characterization and secondary structure prediction

Physico-chemical properties like length, molecular weight, isoelectric point (pI), total number of positive and negative residues, Instability Index,[27] Aliphatic Index[28] and Grand Average hydropathy (GRAVY)[29] were computed using the Expasy's ProtParam server.[30] GOR IV server[31] was employed for the prediction of secondary structural features like alpha helices, extended strands and random coils in terms of percentage in the protein sequences.

Calculation of the GC content

The GC content of the predicted genes was determined using Genscan web server.[32]

Motif identification

Protein sequence motifs for each gene family were identified using the MEME program.[33] The analyses parameters were set as follows: number of repetitions-zero or one per sequence; maximum number of motifs—1; minimum and maximum width—6 and 50, respectively.

The motif profile for each gene family is presented schematically. Domain arrangements along sequences were predicted using InterProscan[34] to determine protein homolog relationships among species.

Exon-intron structure and phylogenetic analyses

The exon-intron structural patterns of the lipid biosynthetic genes were analyzed using the gene prediction algorithm of Genscan.[32] To construct the phylogenetic tree, amino acid sequences were aligned using the ClustalX program implemented in BioEdit[35] (v 7.1.3) with default settings and then manually refined by trimming of poorly conserved N and C termini. ClustalX[36] has been demonstrated to be a user-friendly tool for providing good, biologically accurate alignments within a reasonable time limit. Many options are provided such as the realignment of selected sequences or blocks of conserved residues and the possibility of building up difficult alignments, making ClustalX an ideal tool for working interactively on alignments.[36] Subsequently, sequence alignment of genes predicted to be in similar families were used as an input file for the MEGA 4 software.[37] Phylogenetic tree was built via the neighbor-joining (NJ) method with evaluation of 1000 rounds of bootstrapping test, followed by identification of sub-tree.

Results and Discussion

Comparative genomic analyses of lipid genes in microalgal species

Interest in microalgae as a potential feedstock for biofuel production and other valuable biomaterials is rooted in the ability of microalgae to rapidly accumulate significant amounts of neutral lipids.[38] Under optimal conditions, microalgae synthesize fatty acids used primarily for esterification into polar glycerol-based membrane lipids like glycosylglycerides and phosphoglycerides, whereas under stress conditions, many microalgae tend to accumulate storage lipids called triacylglycerol (TAGs).[16] Although global fatty acid biosynthetic mechanisms are known in higher plants,[39] pathways responsible for lipid accumulation in microalgae are not well studied. Hence, in order to bridge our existing knowledge gap regarding algal lipid metabolism, comparative metabolic pathway analyses have been performed across five microalgal genomes, using homologous plant genes as reference with an objective of functional characterization of predicted genes. EC numbers, Cluster of Orthologous Groups (COGs), protein domain family and GO terms were determined for the respective candidate genes. The above in silico approach has been reviewed recently to be reliable enough for accurate function prediction of uncharacterized proteins encoded by genes in a genome.[40]

In the present study, using the *Arabidopsis* annotation data as the BLAST input query set, a total of 398 orthologous genes present in *A. thaliana, C. reinhardtii, V. carteri*, *O. lucimarinus*, *O. tauri* and *C. merolae* genomes were identified. The above approach to identify candidate genes involved in biosynthesis and accumulation of storage oil has been successfully demonstrated in plants by Sharma and Chauhan.[13] These 398 genes clustered into 40 gene families and includes 142, 56, 59, 47, 41 and 53 genes from *A. thaliana, C. reinhardtii*, *V. carteri*, *O. lucimarinus*, *O. tauri* and *C. merolae* genomes, respectively (Table 1). The identified genes are involved in the synthesis of phospholipids, glycerolipid and storage lipids like TAG. We further divided the predicted genes into categories like gene-coding enzymes involved in biosynthesis and catabolism of fatty acid, TAG and membrane lipid. The comprehensive list of candidate genes along with experimental evidence of the respective enzyme action influencing lipid accumulation is presented in Table 1.[41–74] Approximately 47% of the predicted gene products found in the present study were previously annotated as 'predicted', 'probable', 'putative uncharacterized' and 'similar' or 'hypothetical' proteins (Table 1). The annotation of these sequences has been improved and a role in lipid biosynthetic process was assigned to each of them by similarity search with homologous plant genes, annotation of Gene Ontology, and through identification of conserved domains or motifs. Furthermore, on comparison to the previous report on lipid gene identification in *C. merolae* genome by Sato and Moriyama,[75] the present study has identified 20 additional genes involved in lipid biosynthesis.

To investigate metabolic processes responsible for the synthesis of microalgal biofuel precursors, KO identifiers were assigned to the predicted 398 genes

representing 36 unique EC numbers, which were subsequently used to study metabolic pathway maps available in KEGG pathway database. KEGG is considered one of the most important bioinformatics resources for understanding higher-order functional meaning and the utilities of the organism from its genome information. It hosts information on the majority of well-known metabolic pathways, including lipid pathways for several organisms such as higher plants, bacteria and algae. Recently, it has been used successfully by Rismani-Yazdi et al[14] to identify pathways and the underlying gene responsible for production of biofuel precursors in *Dunaliella tertiolecta*, a potential microalgal biofuel feedstock. Using the above approach, a total of 79 lipid genes including 22 from *A. thaliana*, 21 from *C. merolae*, 10 from *C. reinhardtii*, 10 from *O. lucimarinus* and 8 each from *V. carteri* and *O. tauri* were recognized that were not earlier indexed in KEGG metabolic pathway database (Table 1).

The global synthesis pathway of TAG begins with the basic fatty acid precursors, acetyl-CoA, and continues through fatty acid biosynthesis, complex lipid assembly and saturated fatty acid modification until TAG bodies are finally formed.[76] A simplified overview of TAG biosynthetic pathway in microalgae is shown as Figure 1. Comparative analyses with the genomes of *C. reinhardtii*, *V. carteri*, *O. lucimarinus*, *O. tauri, C. merolae* and *A. thaliana* indicates that the majority of genes involved in lipid production are orthologous among these species. Additionally, the extensive amino acid sequence conservation (more than 60% pair-wise sequence identity) among the genes involved in lipid biosynthesis provides indications of functional equivalence between *Arabidopsis* and microalgal genes. Thus, the present results demonstrate that the underlying fatty acid and TAG biosynthesis process are directly analogous to those reported in higher plants.[16] It may further be noted that although algae predominantly share similar lipid biosynthetic pathways with higher plants, the present in silico analyses revealed that the sizes of the gene families responsible for lipid biosynthesis in microalgae are smaller than *Arabidopsis*. Certain specific pathways were also observed to be absent in microalgae, including the fatty acid biosynthesis termination mechanism by FAT homologs in *C. merolae.* The above computational analyses find support from the previous experimental reports on the algal lipid metabolism.[75]

Furthermore, our results conclusively indicate that enzymes that are responsible for higher lipid accumulation in plants and other eukaryotes, either through over-expression or gene knock-out strategies, are present not only in oleaginous algal species (*C. reinhardtii*) but also in other algal species, notably *O. tauri* and *C. merolae* (Fig. 2). Comparison of the number of genes in each step of lipid metabolic pathway suggests that the green algae *C. reinhardtii* and *V. carteri* have an expanded array of genes involved in TAG biosynthesis and catabolism, including fatty acid thioesterase, long chain acyl-CoA synthase, acyl-CoA oxidase, desaturase, glycerol-3-phosphate acyltransferase, and diacylglycerol acyltransferase. Additionally, the proportion of these gene copy numbers appear to be correlated with the genome complexity of the organisms under study (Fig. 2).

Prediction of subcellular location

The prediction of subcellular localization of proteins is essential to elucidate the spatial organization of proteins according to their function and to refine our knowledge of cellular metabolism.[77] Thus, prediction of subcellular location provides valuable information about the function of proteins as well as the interconnectivity of biological processes.[78] In the present study, subcellular location of lipid biosynthetic proteins by tools such as TargetP, ChloroP and WolfPsort showed different locations using several unique algorithms. The objective of using more than one analytical tool was to improve the specificity of the prediction, as various studies have shown that combined results from several prediction programs are advantageous to rule out false positives and false negatives.[78] The available localization prediction tools show different strengths and no tool is clearly and globally optimal.[77] Moreover, it is known that some localizations are badly predicted by all the algorithms, especially in the case of proteins exhibiting dual targeting to plastids and mitochondria, which could be a phenomenon more common than previously thought.[79] This analyses showed that majority of the predicted proteins are located in four compartments: plastids (31%), mitochondria (26%), cytoplasmic (28%) and nucleus

Table I. Candidate genes involved in lipid biosynthetic pathway of *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Volvox carteri, Ostreococcus lucimarinus, Ostreococcus tauri* and *Cyanidioschyzon merolae* genome.

Gene/symbol	EC no.	KOG no.	KEGG ID	Gene ontology	Corresponding homologous enzymes in algal species (SwissProt accession ID)					JGI protein ID	Ref**
					A. thaliana	*C. reinhardtii*	*V. carteri*	*O. lucimarinus*	*O. tauri*	*C. merolae*	
Fatty acid biosynthesis											
Homomeric acetyl-CoA carboxylase (ACC)	6.4.1.2	KOG0368	K11262	GO:0004075	Q9C8G0, Q38970		D8UA31*	A4RRC3, A4S479¶	Q01GA9, Q00ZG8¶	CMM188C	41–46
Heteromeric ACC biotin carboxylase subunit (BCC)	6.4.1.2/ 6.3.4.14	KOG0238	K01961	GO:0004075 GO:0003989	O04983, F4JYE1, F4JYE0	A8JGF4, A8JEW0	D8UF54	A4S140¶,†	Q013U7¶	CMS299C	
ACC carboxyl-transferase α-subunit (ACC CT α)	6.4.1.2	KOG0238	K01962	GO:0003989	Q9LD43	A8J646	D8TNY0			CMV056C¶	
ACC CT β subunit (ACCCT β)	6.4.1.2	KOG0540	K01963	GO:0003989	P56765	A8JHU1	D8U455¶,*			CMV207C¶	
ACC biotin carboxyl carrier protein (ACC-BCCP)	6.4.1.2	KOG0540	K02160	GO:0003989	Q42533, F4KE21, Q9LLC1	A8JDA7	D8U256*			CMV134C¶	
Malonyl-CoA-ACP transacylase (MCT)	2.3.1.39	KOG2926	K00645	GO:0004314	Q8RU07, Q8L5U2*,¶, F4IMR0	A8HP61	D8TTQ7¶	A4S2U9†, A4SAC5†	Q011G6•, Q00S12	CMT420C	
β-ketoacyl-ACP synthase I (KAS I)	2.3.1.41	KOG1394	K00647	GO:0004315	P52410, F4KHF4	A8JEF7	D8UDW0, D8TXC7	A4RSM2, A4S713†	Q01EI4	CMM286C	47,48
β-ketoacyl-ACP synthase II (KAS II)	2.3.1.179	KOG1394	K09458	GO:0033817	Q9C9P4, Q8L3X9	A8JCK1, A8IG50	D8TJC9	A4S7B9, A4RTJ7†	Q00V56, Q01DP0*	CML329C	
β-ketoacyl-ACP synthase III (KAS III)	2.3.1.180	KOG1394	K00648	GO:0033818	P49243, B9DHF9¶	A8JHL7†	D8TXF1*	A4S7P4†	Q00V15	CMD118C	
β-ketoacyl-ACP reductase (KAR)	1.1.1.100	KOG1200	K00059	GO:0004316	P33207, Q9SQR4, Q9SQR2	A8JBX4, Q84X75	D8TK78*, D8TV99*	A4RQY6†	Q01GL3	CMS393C¶	
3-hydroxyacyl-ACP dehydrase (HAD)	4.2.1.-	–	K02372	GO:0008659	Q9LX13, Q9SIE3 Q8LBU6*,¶	A8IX17†	D8TV61*	A4RUS8†		CMI240C¶	

Enoyl-ACP reductase (EAR)	1.3.1.9	KOG0725	K00208	GO:0004318	Q9SLA8, Q9M672, O04942, Q9FEF2	A8JFI7†	D8UC03*	A4S0L7†	Q014N2	CMT381C	
Acyl-ACP thioesterase/ Fatty acid thioesterase (FAT)	3.1.2.14	–	K10782	GO:0000036	Q42561, Q9SV64, Q9SJE2, Q42562, Q42558, Q41917	A8HY17*	D8TJT0*,¶	A4RS92†	Q01FC4•		49–53
Fatty acid elongation											
3-hydroxyacyl-CoA dehydrogenase (CHAD)	1.1.1.35	KOG2304	K00074	GO:0008691 GO:000385	Q9LDF5¶	A8IVP3†,¶	D8UMK6*,¶	A4RUY4†	Q01C53*,¶	CMC137C¶	
Enoyl-CoA hydratase (ECH)	4.2.1.17	KOG1680 KOG1679	K01692	GO:0004300	Q6NL24, O23468¶, Q0WRQ2¶, Q9T0K7	A8I9B0†,¶	D8TRG5*	A4SBD9†	Q010Z7	CMK139C CMT074C¶	
Enoyl-CoA reductase (TER)	1.3.1.38	KOG1639	K10258	GO:0019166	Q8LCU7•, F4J6R6*, Q9M2U2	A8HM32†, A8JAQ9†	D8THB1*, D8U5N0*,¶	A4RUU7†, A4RU17†	Q01D21	CMD146C△	
Fatty acid catabolism											
Long chain acyl-CoA synthase (LACS)	6.2.1.3	KOG1256	K01897	GO:0004467	Q9T0A0, Q9T009, Q8LPS1, Q8LKS5, Q9SJD4, Q9CAP8 Q9C7W4, Q9XIA9 O22898	A8JH58†, A8HRV2†,¶	D8TMY5*, D8TKU*, D8TP15*, D8TNJ2*, D8TS64*	A4RWX1†, A4S5G5†	Q00Y52*, Q00UP7	CML197C•, CME186C	54
Acyl-CoA oxidase (AOX)	1.3.3.6	KOG0135	K00232	GO:0003997	O65201, F4KG18, O65202, F4JMK8, Q96329, Q9ZQP2, Q9LMI7, P0CZ23	A8ISE5†, A8JGC8†, A8JB97†	D8U3F9*, D8TVM2*, D8U064*, D8U3J5*	A4RR33†	Q01GH2	CMK115C	55
Acyl-CoA dehydrogenase (ACADM)	1.3.99.3	KOG0139	K00249	GO:0003995	Q8RWZ3, Q0WM98¶, Q67ZU5¶, Q9M7Y7¶	A8J3M3¶	D8U2A4*	A4RQF1†	Q01H50*,¶	CML080C	

(*Continued*)

Table I. (*Continued*)

Gene/symbol	EC no.	KOG no.	KEGG ID	Gene ontology	Corresponding homologous enzymes in algal species (SwissProt accession ID)					JGI protein ID	Ref**
					A. thaliana	*C. reinhardtii*	*V. carteri*	*O. lucimarinus*	*O. tauri*	*C. merolae*	
Enoyl-CoA hydratase (ECH)	4.2.1.17	KOG1680 KOG1679	K01692	GO:0004300	Q6NL24, O23468¶, Q0WRQ2¶, Q9T0K7	A8I9B0†,¶	D8TRG5*	A4SBD9†, A4S307†	Q010Z7	CMK139C CMT074C¶	
3-hydroxyacyl-CoA dehydrogenase (CHAD)	1.1.1.35	KOG2304	K00074	GO:0008691 GO:0003857	Q9LDF5¶, Q9ZPI5, Q9ZPI6	A8IVP3†,¶	D8UMK6*,¶	A4RUY4†	Q01C53*,¶	CMC137C¶	
Acetyl-CoA acetyl-transferase (THIL)	2.3.1.9	KOG1390	K00626	GO:0003985	Q8S4Y1, Q9FIK7, F4JYM8*, B9DGQ1, Q3E8F0	A8J0X4†	D8UKX0*, D8TZN7*			CMA042C CME087C·	
Fatty acid desaturation											
Δ^9 acyl-aCP desaturase (Δ^9D)	1.14.19.1	KOG1600	K00507	GO:0004768	Q9SID2, O65797, Q9FPD5, Q9LM13, Q9LM14, Q9LND8, Q9LND9, Q949X0¶, Q9LVZ3	A8J015, A8JEN2, C6ZE81¶, A8IQB8·	D8U961, D8TRE9*,¶	A4S9D8†	Q00T63	CMJ201C CMM045C	56–59
Δ^{12} acyl-aCP desaturase (Δ^{12}D)	1.14.19.-	KOG:TW OG0155	K10256 K10255	GO:0045485	P46313, P46312, Q8LFZ8¶, Q19MZ0	A8IR24, O48663	D8UB74, D8TTW0	A4RWB5†	Q01DF5	CMK291C¶	
Triacylglycerol (TAG) biosynthesis and catabolism											
Glycerol kinase (GK)	2.7.1.30	KOG2517	K00864	GO:0004370	F4HS76, Q9M8L4, A0JPS9¶, C0Z2P8	A8IT31†	D8TXT9*	A4RTW5†	Q01D72*,¶	CMJ173C	
Glycerol-3-phosphate dehydrogenase (G3PDH)	1.1.5.3	KOG0042	K00111	GO:0004368	Q9SS48	A8HTE5†	D8TSE3*	A4RU40	Q01CZ8	CML209C	60

Glycerol-3-phosphate acyltransferase (GPAT)	2.3.1.15	KOG2898	K00631 K00630	GO:0004366	Q43307, Q9LHS7, Q8GWG0, Q9SYJ2, Q9LMM0, Q9FZ22, Q9SHJ5, Q0WPD4, O80437, Q9CAY3	A8J0R2, A8HVM5†	D8TVT7*, D8TIB3*	A4RT23•, A4S945†	Q01F77	CMK217C$^{\Delta,\P}$ CMJ027C CMA017C•	61,62
1-acylglycerol-3-phosphate acyltransferase/ Lysophosphatidi Acid acyl-transferase (AGPAT/LPAT)	2.3.1.51	KOG1505	K00655 K13519	GO:0003841	Q8GXU8, Q8LG50, Q9SYC8, Q8L4Y2•, Q9LHN4•	A8J0J0	D8U1V6*, D8TWQ3*	A4S0H0†	Q014T8*, Q00SS2†	CME109C$^{\bullet,\P}$ CMF185C CMJ021C$^{\bullet,\P}$	63,64
Phosphatidate phosphatase (PP)	3.1.3.4	KOG3030	K01080	GO:0008195	Q9ZU49¶, Q3EC91¶ Q8LFD1, A8MR10*, F4IX65¶, Q9XI60¶, Q9LJQ8¶	A8JGB5†,¶	D8U3B0*	A4RU93†,¶	Q01CT9*,¶	CMR054C$^{\Delta,\P}$ CMR488C$^{\Delta,\P}$	
Diacylglycerol Acyltransferase (DGAT)	2.3.1.20	KOG0831 KOG0380	K00635 K11155	GO:0004144	Q9SLD2, Q9ASU1¶, Q93ZR6¶	A8IXB2¶	D8UGA9*, D8UHL*	A4S872†¶	Q00UG1*,¶	CMJ162C$^{\Delta,\P}$ CMQ199C¶ CME100C	65–74
Triacylglycerol lipase (TAGL)	3.1.1.3	KOG4569	K01046	GO:0004806	Q9LZA6, Q9M1I6, F4JY30¶	D5LAZ6¶, D5LAW3¶, A8HYG2¶,†	D8TT81*, D8U4S5*,¶	A4RQN3†,¶, A4S9E4†, A4RZ46†	Q00T58†, Q016Q6†	CMS254C¶ CMT151C$^{\Delta,\P}$	
Membrane lipid biosynthesis											
Ethanolamine phospho-transferase (EPT1)	2.7.8.1	KOG2877	K00993	GO:0004307	O82567, F4HQU9	Q6U9W9	D8TWP7*	A4S097†,¶	Q01BV3¶	CMF133C¶	
CDP-Diacylglycerol synthase (CDS1)	2.7.7.41	KOG1440	K00981	GO:0004605	Q1PE48, F4JL60, O49639, F4JL62, O04928	A8ILG5, A8IRM0 A8IRL9	D8TPH2, D8TK01	A4RZR8, A4RWB0	Q01AN2, Q015S5	CMM311C¶ CMN215C CMS056C	
Phosphatidyl glycerol lphosphate synthase (PGP3)	2.7.8.5	KOG1617	K00995	GO:0008444	O80952, Q67ZP8*,¶ Q9M2W3	A8JEJ8	D8U650*, D8UDS7*,¶	A4S5X3†	Q00W48	CMN196C CMJ134C¶	

(*Continued*)

Table I. (*Continued*)

Gene/symbol	EC no.	KOG no.	KEGG ID	Gene ontology	Corresponding homologous enzymes in algal species (SwissProt accession ID)					JGI protein ID	Ref**
					A. thaliana	*C. reinhardtii*	*V. carteri*	*O. lucimarinus*	*O. tauri*	*C. merolae*	
Ethanolamine kinase (EKT1)	2.7.1.82	KOG4720	K00894	GO:0004305	O81024*, Q8LAQ2*	A8J2J5	D8TJH5	A4S0V5†	Q014D1*,¶	CMR011C•,¶	
CTP: phospho-ethanolamine cytidyl transferase (ECT)	2.7.7.14	KOG2803	K00967	GO:0004306	Q9ZVI9	Q84JV7	D8TWX6*	A4S2P2†	Q011M7	CMS052C	
UDP-sulfoquinovose synthase (SQD)	3.13.1.1 2.4.1.-	KOG1371 KOG1111	K06118 K06119	GO:0046507 GO:0046510	O48917, Q8S4F6	Q763T6, A8JB95 A8HMC2	D8U760*, D8U5J8*	A4S476, A4S792†	Q00ZH1, Q00V96	CMR012C CMR015C	
Monogalactosyl diacylglycerol synthase (MGDGS)	2.4.1.46		K03715	GO:0046509	Q9SI93, O81770	A8HUF1†	D8TQW6*	A4RT08		CMI271C	
Digalactosyl diacylglycerol synthase (DGDGS)	2.4.1.241		K09480	GO:0035250	Q9S7D1, Q8W1S1	A8HU66	D8TQZ2*,¶	A4S4N5†,¶, A4S0F1¶,†	Q00Z06, Q014V9¶,†		
Inositol phospho-transferase (PIS)	2.7.8.11	KOG3240	K00999	GO:0003881	Q8LBA6, Q8GUK6, F4JTR2*	A8ICX2	D8TPK4	A4SAF2†	Q00RY0	CMM125C	

Notes: *Putative uncharacterized proteins; †predicted proteins; •probable proteins; ᐞsimilar protein; ¶absent in KEGG pathway database; **relevant references on experimental evidences of the respective enzyme action influencing lipid accumulation.

(6%) (Fig. S1 and Table S1). The above results are consistent with the experimental observations that de novo synthesis of fatty acids occurs primarily in the plastid and/or mitochondria.[5] About 19% of the proteins revealed the presence of both the mitochondrial target peptide and chloroplast transit peptide in the sequences. Recent reports have shown an unexpectedly high frequency of dual targeting of proteins to both the mitochondria and chloroplast, hence making it difficult to predict the correct location of these proteins within a cell.[80,81] Furthermore, approximately 3% of the predicted proteins were located in more than one compartment ie, nucleus and cytoplasm, which were the same highly paired compartments as identified in Arabidopsis[82] and sugarcane[83] proteome, suggesting that there is a significant amount of interactions between these two organelles.

Hyunjong et al[84] have reported that targeting a particular enzyme to several compartments simultaneously in the same plant will augment its production when compared to its individual compartments in the same plant. Hence the predicted localization information would certainly aid in targeting the lipid biosynthetic enzymes to enhance oil accumulation in microalgae.

Physico-chemical characterization and secondary structure prediction

Various physico-chemical parameters were computed using Expasy's ProtParam tool (Fig. 3 and Table S2). Molecular weight was observed between the ranges of 1116.818–299171.0 for all lipid biosynthetic proteins in microalgae. The majority of the predicted proteins were found to have a pI greater than 7, indicating that proteins involved in lipid biosynthesis are generally

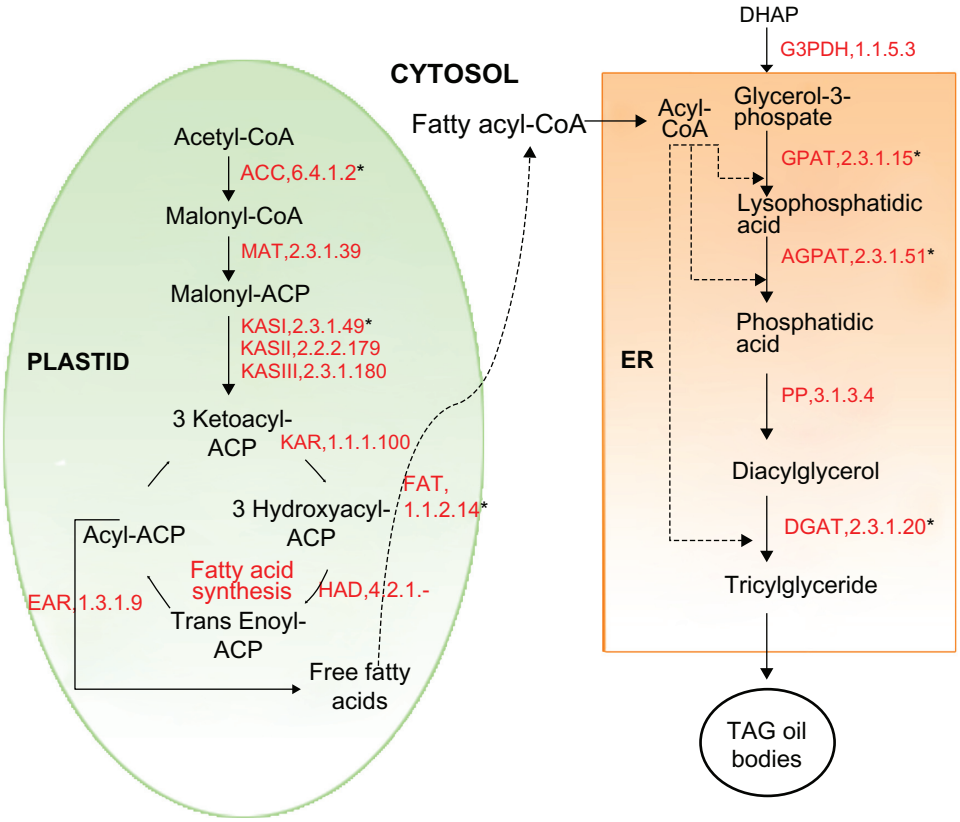

Figure 1. Schematic overview of Triacylglyceride (TAG) biosynthetic pathway in microalgae.
Notes: Free fatty acids and TAG are synthesised in the chloroplast and endoplasmic reticulum respectively. The vital enzymes reported by various experimental studies to be involved in accelerated lipid accumulation are marked with an *asterisk*.
Abbreviations: ACC, Acetyl-CoA carboxylase; MAT, Malonyl-CoA-ACP transacylase; KAS, 3-ketoacyl-ACP synthase; KAR, 3-ketoacyl-ACP reductase; HAD, 3-hydroxyacyl-ACP dehydratases; EAR, Enoyl-ACP reductase; FAT, Fatty acid thioesterase; G3PDH, Glycerol-3-phosphate dehydrogenase; GPAT, Glycerol-3-phosphate acyltransferase; AGPAT, 1-acylglycerol-3-phosphate acyltransferase also known as LPAT, lysophosphatidic acid acyl transferase; PP, Phosphatidate phosphatase; DGAT, Diacylglycerol acyltransferase.

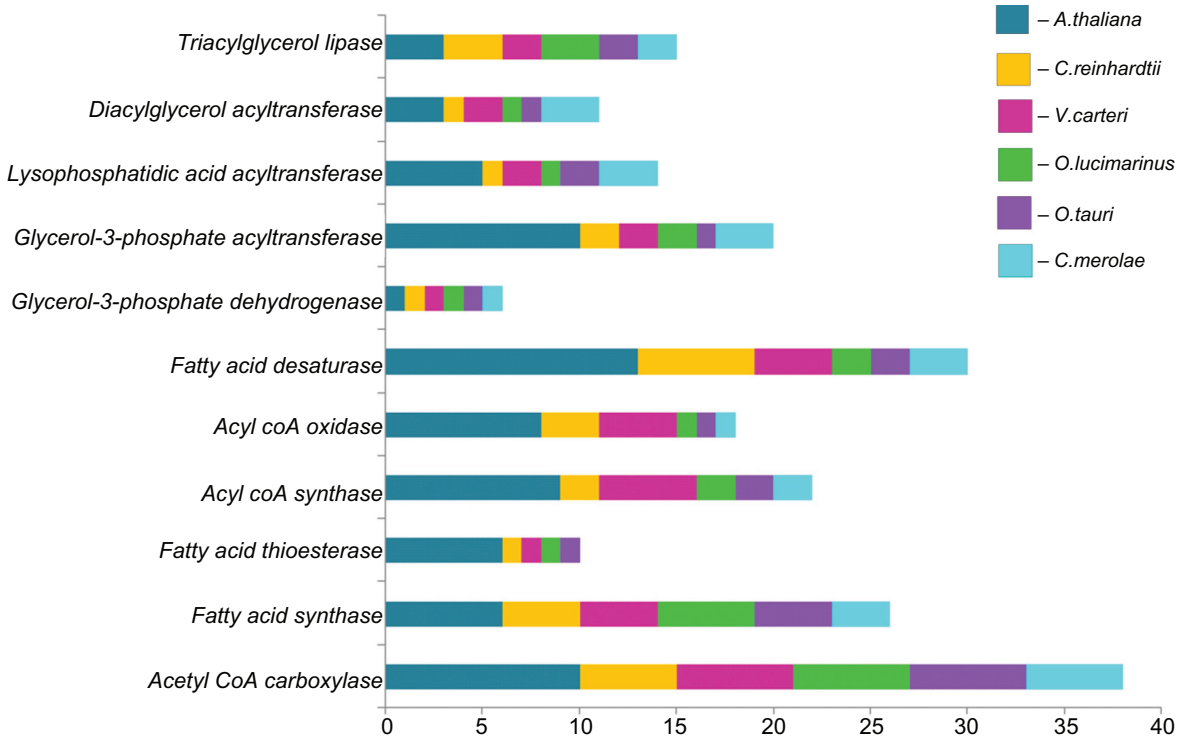

Figure 2. Number of gene homologues in the TAG biosynthetic pathway in *A. thaliana*, *C. reinhardtii*, *V. carteri*, *O. lucimarinus*, *O. tauri* and *C. merolae*.
Notes: For each reaction, coloured squares denotes the number of homologous genes in *A. thaliana* (blue), *C. reinhardtii* (yellow), *V. carteri* (pink), *O. lucimarinus* (green), *O. tauri* (purple) and *C. merolae* (light blue).

basic in nature. However, the deduced sequences for genes such as acetyl-CoA carboxylase, acetyl-CoA acetyltransferase, glycerol kinase, ethanolamine kinase and phosphoethanolamine cytidyl transferase were determined to be acidic. These values of isoelectric point (overall charge) will be useful for developing a buffer system for purification of the enzymes by an isoelectric focusing method. Instability Index analyses reveals the presence of certain dipeptides occurring at significantly different frequencies between stable and unstable proteins. Proteins with an instability index less than 40 are predicted to be stable while those with a value greater than 40 are assumed to be unstable. In the present study the high occurrence frequency of unstable proteins may be explained in the context of the recent work of Cao,[85] who observed such a phenomenon in many plants and microorganisms due to the possible inherent feedback mechanism that regulates the optimal level of accumulation of cellular metabolites. The aliphatic index refers to the relative volume of a protein that is occupied by aliphatic side chains (eg, alanine, isoleucine, leucine and valine) and contributes to the increased thermal stability observed for globular proteins. Aliphatic index for the screened proteins ranged from 70.24 to 119.16. The very high aliphatic index for all sequences indicated that their structures are more stable over a wide range of temperature. The GRAVY index indicates the solubility of the protein. The lipid biosynthetic proteins which showed large negative values indicated that these proteins are relatively more hydrophobic when compared to proteins with less negative values.

The secondary structure of the microalgal proteins involved in lipid metabolism were analyzed by submitting the amino acid sequence to the GOR IV program, which has been experimentally cross validated to have a mean accuracy of 64.4% for the three state prediction.[32] The secondary structure indicates whether a given amino acid lies in a helix, strand or a coil. Secondary structure features of the proteins are represented in Table S3. The results revealed that random coil to be predominant followed by alpha helices and extended strands in the majority of sequences.

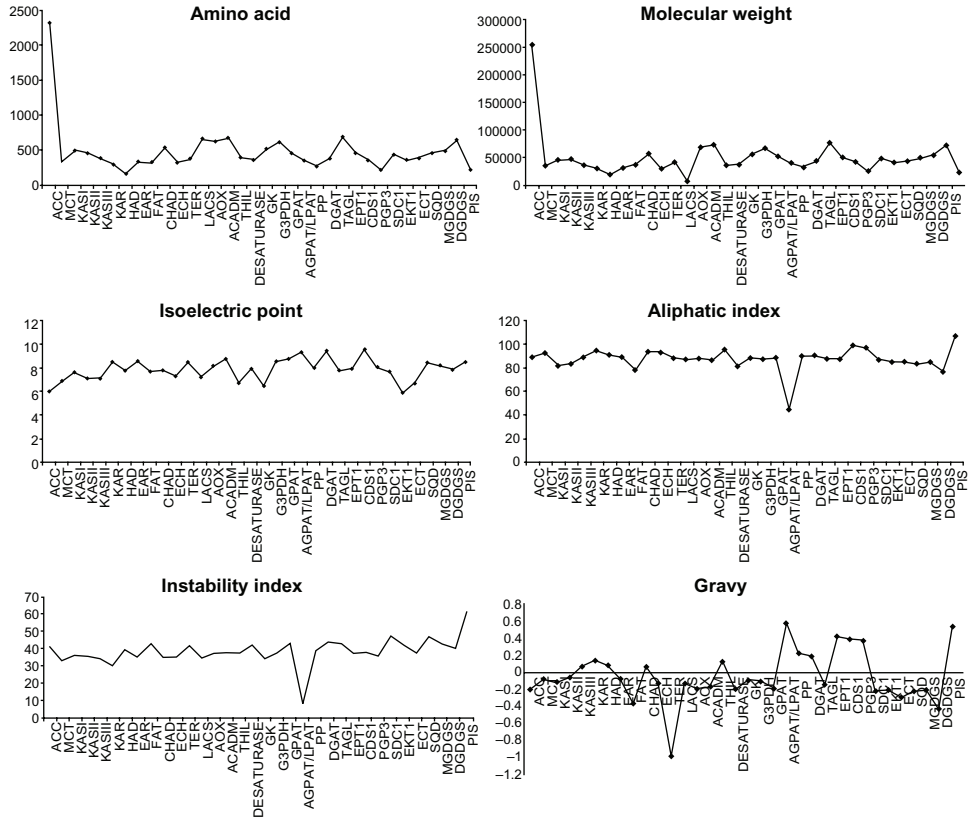

Figure 3. Distribution of various physico-chemical characteristics of putative proteins encoded by lipid genes in *A. thaliana*, *C. reinhardtii*, *V. carteri*, *O. lucimarinus*, *O. tauri* and *C. merolae*.
Note: The individual physico-chemical values for each protein as calculated by ProtParam server is provided in Supplementary Table 2.

GC-content analyses

The variations in the guanine (G) and cytosine (C) content observed between species is one of the central issues in evolutionary bioinformatics. The average GC-content of the lipid biosynthetic genes, as calculated by the Genscan server, was 39.89%, 63.35%, 56.92%, 59.88%, 59.04% and 55.57% for *A. thaliana*, *C. reinhardtii*, *V. carteri*, *O. lucimarinus O. tauri* and *C. merolae* respectively. The GC values lie close to the calculated GC-content of the whole genome of the respective organisms under study.[86–89] However, a slightly higher GC-content for the gene sequences was observed in contrast to the background GC-content for the entire genome of all the studied species. Among the microalgae, the highest GC-content was observed in *C. reinhardtii*. The GC-content of *C. reinhardtii* is also experimentally reported to be higher than that of the multicellular organisms.[90] Comparative analyses of the GC-content of the individual genes revealed minor variations among the microalgal genomes (Fig. 4 and Table S4). The above finding is in congruence with the earlier report stating that eukaryotic genomes vary less in their GC content.[91] Furthermore, GC-content analyses indicated that the genes with high GC-content were also identified to be stable by ProtParam server as compared to genes having low GC-content. This may apparently be due to the fact that GC pair is bound by 3 hydrogen bonds (H-bonds), compared to 2 H-bonds in AT, thus contributing to the greater stability of the gene products. In addition, analyses of individual predicted genes in *O. lucimarinus* and *O. tauri* revealed more or less similar GC-content in both the subspecies.

Motif and domain architecture

A motif is a sequence pattern found conserved in a group of related protein or gene sequences.[34] An exhaustive search of the protein motifs using the MEME program identified 36 core conserved sequences in the lipid biosynthetic genes of microalgae predicted

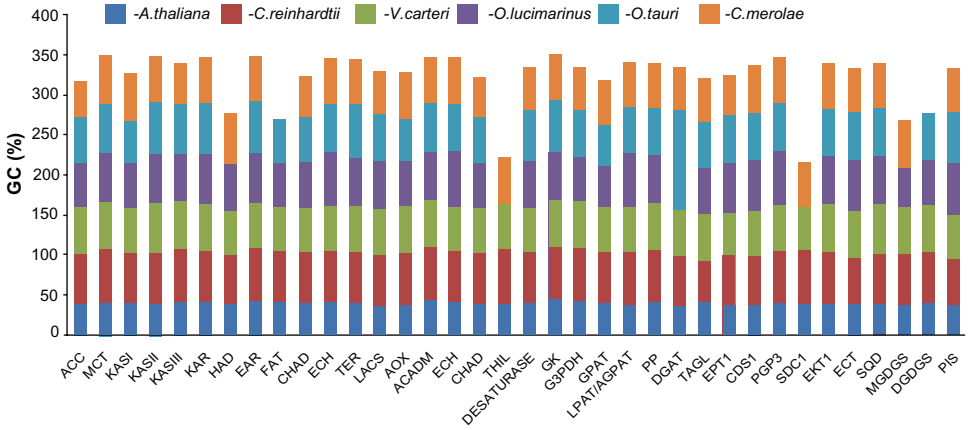

Figure 4. Comparison of the GC-content of lipid biosynthetic genes among five unicellular algae and the vascular plant, *A. thaliana*.
Notes: Columns represent the average GC content of the genes (in percentage) of each organism: *A. thaliana* (blue), *C. reinhardtii* (red), *V. carteri* (green), *O. lucimarinus* (purple), *O. tauri* (blue) and *C. merolae* (orange) in a down to up order. The individual GC-content values of each gene as calculated by Genscan web server are given in Supplementary Table 4.

in the present study (Fig. 5). The overall height of each stack indicates the sequence conservation at that position, whereas the height of symbols within each stack reflects the relative frequency of the corresponding amino acid (Fig. 5). The sequence logos showed that majority of the predicted motifs are basically composed of hydrophobic and polar uncharged residues. It is likely that these conserved residues are critical for the catalytic activity of the enzymes and may be involved in substrate binding, direct catalysis, and maintenance of the protein structure. In addition to motif analyses, a detailed comparison of the domain architectures of the gene products at the whole genome level is given in Figure 5. Results indicate that the majority of domains observed in genes involved in lipid biosynthesis are present in all microalgal species under study. Therefore, the critical amino acid residues present in the conserved motif and domain of the lipid genes will certainly act as a framework for better understanding their structure-function relationship.

Exon-intron structure and phylogenetic analyses

In order to gain insights into the evolution of the lipid biosynthetic genes, we analyzed exon-intron structure patterns of the predicted gene homologs (Table S5). The results revealed that the exon-intron spilt pattern of *C. reinhardtii* and *V. carteri* genes were homologous to that of *Arabidopsis*, although insertion, deletion and intron-size variations were common. Likewise, conservation with respect to exon-intron number and size were observed between *O. lucimarinus* and *O. tauri*. The *C. merolae* genome is remarkable for its paucity of introns[88] and in our study we also could not detect its presence in any of the predicted genes. *O. lucimarinus* and *O. tauri* genes contained fewer introns as compared to *C. reinhardtii*, *V. carteri* and *A. thaliana* and our present results confirms the previous report that *C. reinhardtii* lipid biosynthetic genes contain a higher number of introns.[92] A phylogenetic tree was constructed to evaluate the evolutionary relationship among the predicted genes (Fig. 6). The phylogenetic tree showed that in the majority of predicted genes with similar functions and sharing similar intron-exon structure, conserved motif patterns were clustered together in the tree because of their common ancestry and in accordance with our expectations. In most of the gene families, it was observed that the protein sequence of the two sub-species *O. lucimarinus* and *O. tauri* (Prasinophytes) were present as sister clades and that it falls within the green algal cluster comprising of *C. reinhardtii*, *V. Carteri* (Chlorophytes) and *A. thaliana* (Streptophytes). The Chlorophytes and Streptophytes lineages are a part of the green plant lineage (Viridiplantae).[93] Further, the phylogenetic analyses suggest that protein homologs of *C. merolae* (Rhodophytes) seem to diverge from the root of the green lineage. Overall, we found that components

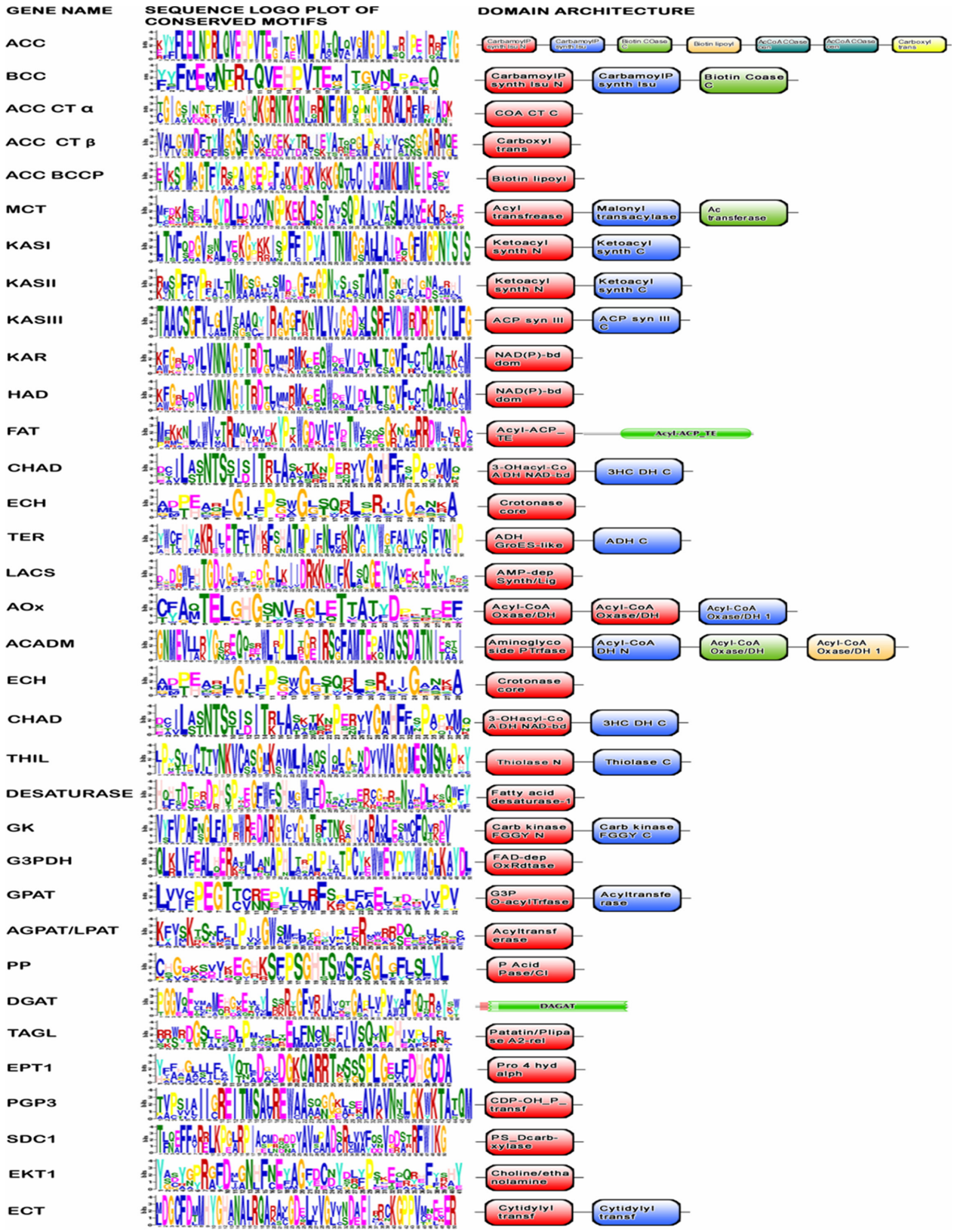

Figure 5. Conserved domain architectures and sequence logo plots of lipid biosynthetic genes using InterProscan and MEME programs, respectively.
Notes: The overall height of each stack indicated the sequence conservation at that position, whereas the height of symbols within each stack reflects the relative frequency of the corresponding amino acid. The amino acids are colour coded as: A, C, F, I, L, V, W and M (Blue-Most hydrophobic); N, Q, S and T (Green-Polar, non-charged and non-aliphatic residues); D and E (Magenta-Acidic); K and R (Red-Positively charge).

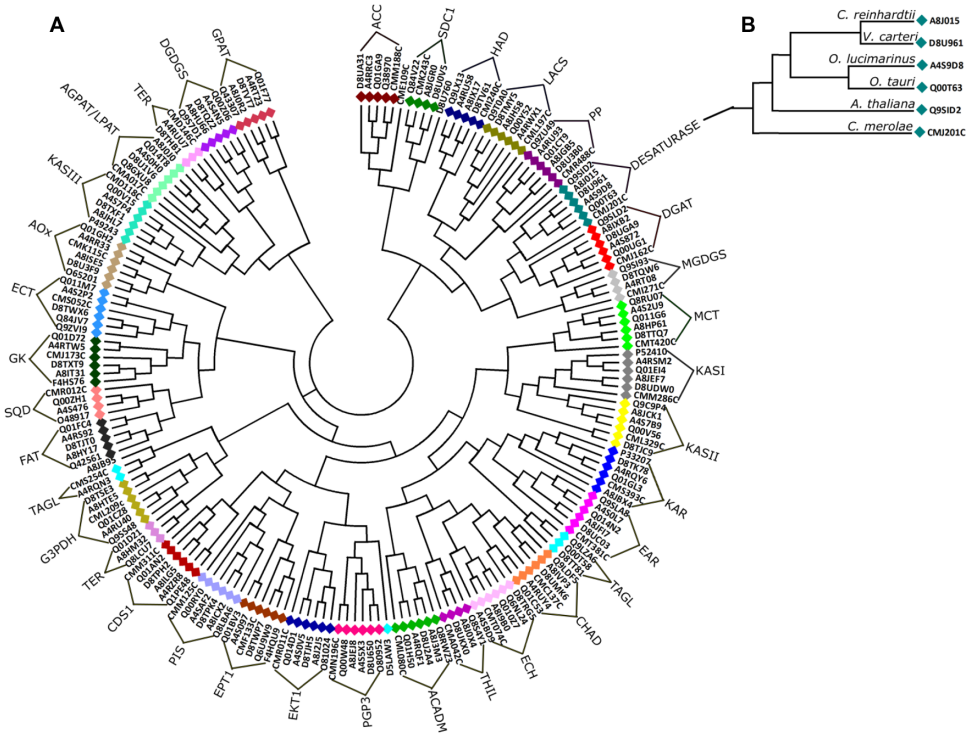

Figure 6. (**A**) Phylogenetic tree inferred from the amino acid sequences of lipid genes in *A. thaliana*, *C. reinhardtii*, *V. carteri*, *O. lucimarinus*, *O. tauri* and *C. merolae*. Proteins with identical functional characterization are represented by similar colour coded diamond shapes. Protein accession numbers are represented while organism names to which proteins belong are given in Table 1. Some homologous proteins were omitted to increase clarity of the remaining groups. The tree indicates that proteins with similar functions were clustered together and further, in most of the gene families for instance in desaturase (**B**), the protein sequence of the two sub-species *O. lucimarinus* and *O. tauri* were present as sister clades and falls within the green algal cluster comprising of *C. reinhardtii*, *V. Carteri* and *A. thaliana*, while the protein homologs of *C. merolae* seem to diverge from the root of the green lineage.

of lipid biosynthetic pathway are remarkably well conserved, particularly among the Viridiplantae lineage.

Conclusion

Identification of genes responsible for oil accumulation is a pre-requisite to targeting microalgae for enhanced yields of biofuel precursors using metabolic engineering. A comprehensive computational analyses of the predicted genes of microalgae against *Arabidopsis* was performed through gene annotation, subcellular localization, physico-chemical characterization, exon-intron pattern, motif/domain organization and phylogenomics studies. The results revealed that although each of the algal species maintains the basic genomic repertoire required for lipid biosynthesis, they possess additional lineage-specific gene groups. Additionally, the extensive sequence and structure conservation of the putative genes indicates functional equivalence between microalgae and *Arabidopsis*. Phylogenetic analyses demonstrated that genes of lipid biosynthetic pathway from Prasinophytes, Chlorophytes, Streptophytes and Rhodophytes were clustered according to their conserved motif pattern, exon-intron structure and functional equivalence. The in-depth broad investigation of each individual gene and their encoded products across the microalgal genome will certainly facilitate metabolic engineering of microalga for biofuel production.

Acknowledgement

The research fellowship granted to NM by the Government of India's Department of Biotechnology (under a grant-in-aid project) and subsequently a Senior Research Fellowship award by Council of Scientific & Industrial Research, India is gratefully acknowledged.

Author Contributions

Conceived and designed the experiments: NM, PKP. Analysed the data: NM, PKP, BKM. Wrote the first

draft of the manuscript: NM, PKP, BKP. Contributed to the writing of the manuscript: NM, BKP, PKP. Agree with manuscript results and conclusions: NM, PKP, BKP, BKM. Jointly developed the structure and arguments for the paper: PKP, BKM. Made critical revisions and approved final version: NM, PKP, BKP, BKM. All authors reviewed and approved of the final manuscript.

Funding

Author(s) disclose no funding sources.

Competing Interests

Author(s) disclose no potential conflicts of interest.

Disclosures and Ethics

As a requirement of publication author(s) have providedto the publisher signed confirmation of compliance with legal and ethical obligations including but not limited to the following: authorship and contributorship, conflicts of interest, privacy and confidentiality and (where applicable) protection of human and animal research subjects. The authors have read and confirmed their agreement with the ICMJE authorship and conflict of interest criteria. The authors have also confirmed that this article is unique and not under consideration or published in any other publication, and that they have permission from rights holders to reproduce any copyrighted material. Any disclosures are made in this section. The external blind peer reviewers report no conflicts of interest.

References

1. Yuan JS, Tiller KH, Al-Ahmad H, Stewart NR, Stewart CN. Plants to power: bioenergy to fuel the future. *Trends Plant Sci*. 2008;13:421–7.
2. Wijffels RH, Barbosa MJ. An outlook on Microalgal Biofuels. *Science*. 2010;329:796–9.
3. Chisti Y. Biodiesel from microalgae. *Biotechnol Adv*. 2007;25:294–306.
4. Pienkos PT, Darzins A. The promise and challenges of microalgal-derived biofuels. *Biofuels Biprod Bioref*. 2009;3:431–40.
5. Courchesne NMD, Parisien A, Wang B, Lan CQ. Enhancement of lipid production using biochemical, genetic and transcription factor engineering approaches. *J Biotechnol*. 2009;141:31–41.
6. Goldberg IK, Cohen Z. Unravelling algal lipid metabolism, recent advances in gene identification. *Biochimie*. 2011;93:91–100.
7. Radakovits R, Jinkerson RE, Darzins A, Posewitz MC. Genetic Engineering of Algae for Enhanced Biofuel Production. *Eukaryot Cell*. 2010;9:486–501.
8. Rogalski M, Carrer H. Engineering plastid fatty acid biosynthesis to improve food quality and biofuel production in higher plants. *Plant Biotechnol J*. 2011;9:554–64.
9. Napier JA. The production of unusual fatty acids in transgenic plants. *Annu Rev Plant Biol*. 2007;58:295–319.
10. Topfer R, Martini N, Schell J. Modification of Plant Lipid synthesis. *Science*. 1995;268:681–6.
11. Mao F, Yin Y, Zhou FF, et al. pDAWG: an integrated database for plant cell-wall genes. *Bioenerg Res*. 2009;2:209–16.
12. Lin L, Hui L, Ying LJ, et al. Genome-wide survey of maize lipid-related genes: candidate genes mining, digital expression profiling and co-location with QTL for maize kernel oil. *Sci China Life Sci*. 2010;53:690–700.
13. Sharma A, Chauhan R. In-silico Identification and Comparative Genomics of Candidate Genes Involved in Biosynthesis and Accumulation of Seed Oil in Plants. *Comp Funct Genom*. 2012. Article ID914843:1–14. doi:10.1155/2012/914843.
14. Rismani-Yazdi H, Haznedaroglu BZ, Bibby K, Peccia J. Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: Pathway description and gene discovery for production of next-generation biofuels. *BMC Genomics*. 2011;12:148–65.
15. Cerutti H, Ma X, Msanne J, Repas T. RNA-Mediated silencing in algae: biological roles and tools for gene function. *Eukaryot Cell*. 2011;10: 1164–72.
16. Reyes-Prieto A, Yoon HS, Bhattacharya D. Phylogenomics: its growing impact on Algal Phylogeny and evolution. *Algae*. 2006;21:1–10.
17. Beisson F, Koo AJK, Ruuska S, et al. Arabidopsis genes involved in acyl lipid metabolism: a 2003 census of the candidates, a study of the distribution of expressed Sequence Tags in Organs, and a web based database. *Plant Physiol*. 2003;132:682–97.
18. Altschul SF, Gish W, Miller W, Myer EW, Lipman DJ. Basic Local Alignment Search Tool. *J Mol Biol*. 1990;215:403–10.
19. Tatusov RL, Federova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. *Bioinformatics*.2003;4:41–55.
20. Carbon S, Ireland A, Mungali CJ, Shu SQ, Marshal B, Lewis S. AmiGo-online access to ontology and annotation data. *Bioinformatics*. 2009;25:288–9.
21. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 2004;32:277–80.
22. Wheelock CE, Goto S, Yetukuri L, et al. Bioinformatics strategies for the analyses of lipids. *Methods Mol Biol*. 2009;580:339–68.
23. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res*. 2000;35:W182–5.
24. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequences. *J Mol Biol*. 2000;300:1005–16.
25. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. ChloroP: a neural network based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci*. 1999;8:978–84.
26. Horton P, Park KJ, Obayashi T, et al. WolfPsort-protein localization predictor. *Nucleic Acids Res*. 2007;35:W585–7.
27. Guruprasad K, Reddy BVP, Pandit MV. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Prot Eng*. 1990;4: 155–64.
28. Ikai AJ. Thermo stability and aliphatic index of globular proteins. *J Biochem*. 1980;88:1895–1898.
29. Doolittle RFK. A simple method for displaying the hydropathic character of a protein. *J Mol Biol*. 1982;157:105–32.
30. Gasteiger E, Hoogland C, Gattiker A, et al. Protein identification and analyses tools on the Expasy's server. *Methods Mol Biol*. 1999;112:531–52.
31. Garnier J, Gibrat JF, Robson B. GOR IV secondary structure prediction method version IV. *Method Enzymol*. 1996;266:540–53.
32. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 1997;268:78–94.
33. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res*. 2006;34: W369–73.
34. Quevillon E, Silventoinen V, Pillai S, et al. InterProscan: protein domains identifier. *Nucleic Acids Res*. 2005;33: W1216–20.
35. Hall TA. BioEdit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acid Symp Ser*. 1999;41: 95–8.

36. Chenna R, Sugawara H, Koike T, et al. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*. 2003;31:3497–500.
37. Tamura K, Dudley J, Nei M, Kumar S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol*. 2007;24: 1596–9.
38. Scott SA, Davey MP, Dennis JS, et al. Biodiesel from algae: challenges and prospects. *Curr Opin Biotechnol*. 2010;21:277–86.
39. Coleman RA, Lee DP. Enzymes of triacylglycerol synthesis and their regulation. *Prog Lipid Res*. 2004;43:134–76.
40. Janga, SC, Diaz-Meijia JJ, Moreno-Hagelsieb G. Network-based function prediction and interactomics: The case for metabolic enzymes. *Metab Eng*. 2011;13:1–10.
41. Davies MS, Solbiati J, Cronan JE Jr. Overproduction of acetyl-CoA carboxylase activity increases the rate of fatty acid biosynthesis in *Escherichia coli*. *J Biol Chem*. 2000;275:28593–8.
42. Dunahay TG, Jarvis EE, Roessler PG. Genetic transformation of the diatoms *Cyclotella cryptica* and *Navicula saprophila*. *J Phycol*. 1995;31: 1004–12.
43. Klaus D, Ohlrogge JB, Neuhaus HE, Dorman P. Increased fatty acid production in potato by engineering of acetyl-CoA carboxylase. *Planta*. 2004;219:389–96.
44. Ratledge C. Fatty acid biosynthesis in microorganisms being used for single cell oil production. *Biochimie*. 2004;86:807–15.
45. Roesler K, Shintani D, Savage L, Bodderpalli S, Ohlrogge J. Targeting of the Arabidopsis homomeric acetyl-coenzyme A carboxylase to plastids of rapeseeds. *Plant Physiol*. 1997;13:75–81.
46. Thelen JJ, Ohlrogge JB. Metabolic engineering of fatty acid biosynthesis in plants. *Metab Eng*. Jan 2002;4(1):12–21.
47. Dehesh K, Edward P, Fillatti J, Slabaugh M, Byrne J. KAS IV: a 3-ketoacyl-ACP synthase from Cuphea sp is a medium chain specific condensing enzyme. *Plant J*. 1998;15:383–90.
48. Dehesh K, Tai H, Edwards P, Byrne J, Jaworski JG. Overexpresssion of 3-ketoacyl-acyl-carrier protein synthase IIIs in plants reduces the rate of lipid synthesis. *Plant Physiol*. 2001;125:1103–14.
49. Jiang P, Cronan JE. Inhibition of fatty acid synthesis in *Escherichia coli* in the absence of phospholipid synthesis and release of inhibition by thioesterase action. *J Bacteriol*. 1994;176:2814–21.
50. Voelker TA, Davies HM. Alteration of the specificity and regulation of fatty acid synthesis of *Escherichia coli* by expression of a plant medium-chain acyl-acyl carrier protein thioesterase. *J Bacteriol*. 1994;176: 7320–7.
51. Voelker TA, Worrell AC, Anderson L, et al. Fatty acid biosynthesis redirected to medium chains in transgenic oil seed plants. *Science.* 1992; 257:72–4.
52. Dehesh K, Jones A, Knutzon DS, Voelker TA. Production of high levels of 8:0 and 10:0 fatty acids in transgenic canola by overexpresssion of *Ch FatB*2, a thioesterase cDNA from *Cuphea hookeriana*. *Plant J*. 1996;9: 167–72.
53. Eccleston VS, Ohlrogge JB. Expression of lauroyl-acyl carrier protein thioesterase in Brassica napus seeds induced pathways for both fatty acid oxidation and biosynthesis implies a set point for TAG accumulation. *Plant Cell*. 1998;10:613–22.
54. Fulda M, Schnurr J, Abbadi A, Heinz E. Peroxisomal Acyl-CoA synthetase activity is essential for seedling development in *Arabidopsis thaliana*. *Plant Cell*. 2004;16:394–405.
55. Rylott EL, Rogers CA, Gilday AD, Edgell T, Larson TR, Graham IA. Arabidopsis mutants in short and medium chain acyl-CoA oxidase activities accumulate acyl-CoAs and reveal that fatty acid beta oxidation is essential for embryo development. *J Biol Chem*. 2003;278:21370–7.
56. Bondaruk M, Johnson S, Degafu A, et al. Expression of a cDNA encoding palmitoyl-acyl carrier protein desaturase from cats claw (*Doxantha unguis-cati* L.) in *Arabidopsis thaliana* and *Brassica napus* leads to accumulation of unusual unsaturated fatty acids and increased stearic acid content in the seed oil. *Plant Breeding*. 2007;126:186–94.
57. Cahoon EB, Lindqvist Y, Schneider G, Shanklin J. Redesign of soluble fatty acid desaturases from plants for altered substrate specificity and double bond position. *P Natl Acad Sci U S A*. 1997;94:4872–7.
58. Cahoon EB, Shah S, Shanklin J, Browse J. A determinant of substrate specificity predicted from the acyl-acyl carrier protein desaturase of developing cat's claw seed. *Plant Physiol*. 1998;117:593–8.
59. Cahoon EB, Shanklin J. Substrate-dependent mutant complementation to select fatty acid desaturase variants for metabolic engineering of plant seed oils. *P Natl Acad Sci U S A*. 2000;97:12350–5.
60. Vigeolas H, Waldeck P, Zank T, Geigenberger P. Increasing seed oil content in oilseed rape *Brassica napus* (L) by overexpresssion of a yeast glycerol-3-phosphate dehydrogenase under the control of a seed specific promoter. *Plant Biotechnol J*. 2007;5:431–41.
61. Jain RK, Coffey M, Lai K, Kumar A, Mackenzie SL. Enhancement of seed oil content by expression of glycerol-3-phosphate acyltransferase genes. *Biochem Soc Trans*. 2000;28:959–60.
62. Vigeolas H, Geigenberger P. Increased levels of glycerol-3-phosphate lead to a stimulation of flux into triacylglycerol synthesis after supplying glycerol to developing seeds of *Brassica napus L. Planta*. 2004;219:827–35.
63. Taylor DC, Katavic V, Zou JT, et al. Field testing of transgenic rapeseed cv. Hero transformed with a yeast sn-2 acyltransferase results in increased oil content, erucic acid content and seed yield. *Mol Breed*. 2002;8:317–22.
64. Zou JT, Katavic V, Giblin EM, et al. Modification of seed oil content and acyl composition in the Brassicaceae by expression of a yeast sn-2 acyltransferase gene. *Plant Cell*. 1997;9:902–23.
65. Jako C, Kumar A, Wei Y, et al. Seed specific overexpresssion of an Arabidopsis cDNA encoding a diacylglycerol acyltransferase enhances seed oil content and seed weight. *Plant Physiol*. 2001;126:861–74.
66. Lardizabal K, Effertz R, Levering C, et al. Expression of *Umbelopsis ramaniana* DGAT2A in seed increases in soybean. *Plant Physiol*. 2008;148: 89–96.
67. Zheng P, Allen WB, Roesler K, et al. A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nat Genet*. 2008;40: 367–71.
68. Bouvier-Nave P, Benveniste P, Oelkers P, Sturley SL, Schaller H. Expression in yeast and tobacco of plant cDNA encoding acyl-CoA: diacylglycerol acyltransferase. *Eur J Biochem*. 2000;267:85–96.
69. Routaboul JM, Benning C, Caboche M, Lepininc L. The TAG1 locus of Arabidopsis encodes for a diacylglycerol acyltransferase. *Plant Physiol Biochem*. 1999;37:831–40.
70. Katavic V, Reed DW, Taylor DC, et al. Alteration of seed fatty-acid composition by an ethyl methane sulfonate- induced mutation in *Arabidopsis thaliana* affecting diacylglycerol acyltransferase activity. *Plant Physiol*. 1995;108:399–409.
71. Taylor DC, Katavic V, Zou JT, et al. Field testing of transgenic rapeseed cv. Hero transformed with a yeast sn-2 acyltransferase results in increased oil content, erucic acid content and seed yield. *Mol Breed*. 2002;8:317–22.
72. Taylor DC, Zhang Y, Kumar A, et al. Molecular modification of triacylglycerol accumulation by over-expression of *DGAT*1 to produce canola with increased seed oil content under field conditions. *Botany*. 2009;87:533–43.
73. Burgal J, Shockey J, Lu C, et al. Metabolic engineering of hydroxyl fatty acid production in plants. RcDGAt2 drives dramatic increases in ricinoleate levels in seed oil. *Plant Biotechnol J*. 2008;8:819–31.
74. Graham IA. Seed storage oil mobilization. *Annu Rev Plant Biol*. 2008;59: 115–42.
75. Sato N, Moriyama T. Genomic and Biochemical Analysis of Lipid Biosynthesis in the Unicellular Rhodophyte *Cyanidioschyzon merolae*: Lack of a Plastidic Desaturation Pathway Results in the Coupled Pathway of Galactolipid Synthesis. *Eukaryot Cell*. 2007;6:1006–17.
76. Ohlrogge J, Browse J. Lipid biosynthesis. *Plant Cell*. 1995;7:957–70.
77. Liley KS, Dupree P. Plant organelle proteomics. *Curr Opin Plant Biol*. 2007;10:594–9.
78. Shen YO, Burger G. 'Unite and conquer': enhanced prediction of protein subcellular localization by integrating multiple specialized tools. *BMC Bioinformatics*. 2007;8:420–31.
79. Heazlewood JL, Verboom Re, Tonti-Filippini J, Small I, Millar AH. SUBA: the Arabidopsis subcellular database. *Nucleic Acids Res*. 2007;35: D213–8.
80. Peeters N, Small J. Dual targeting to mitochondria and chloroplast. *Biochim Biophys Acta*. 2001;1541:54–63.

81. Millar AH, Whelan J, Small I. Recent surprises in protein targeting to mitochondria and plastids. *Curr Opin Plant Biol*. 2006;9:610–5.
82. Li S, Ehrhardt DW, Rhee SY. Systematic analyses of Arabidopsis organelles and a protein localization database for facilitating fluorescent tagging of full-length Arabidopsis proteins. *Plant Physiol*. 2006;141:527–39.
83. Vicentini R, Menossi M. The predicted subcellular localisation of the sugarcane proteome. *Funct Plant Biol*. 2009;36:242–50.
84. Hyunjong B, Lee DS, Hwang I. Dual targeting of xylanase to chloroplasts and Peroxisomes as a means to increase protein accumulation in plant cells. *J Exp Bot*. 2006;57:161–9.
85. Cao H. Structure-Function analyses of Diacyl-glycerol Acyltransferase sequences from 70 organisms. *BMC Research Notes*. 2011;4:249–72.
86. Misumi O, Yoshida Y, Nishida K, et al. Genome analyses and its significance in four unicellular algae, *Cyanidioschyzon merolae*, *Ostreococcus tauri*, *Chlamydomonas reinhardtii* and *Thalassiosira pseudonana*. *J Plant Res*. 2008;121:3–17.
87. Palenik B, Grimwood J, Aerts A, et al. The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *P Natl Acad Sci Biol*. 2007;104:7705–10.
88. Matsuzaki M, Misumi O, Shin IT, et al. Genome sequence of the ultrasmall unicellular red algae *Cyanidioschyzon merolae* 10D. *Nature*. 2004;428: 653–7.
89. Prochnik SE, Umen J, Nedelcu AM, et al. Genomic Analysis of Organismal Complexity in the Multicellular Green Alga *Volvox carteri*. *Science*. 2010; 329:223–6.
90. Labadorf A, Link A, Rogers MF, et al. Genome-wide analyses of alternative splicing in Chlamydomonas reinhardtii. *BMC Genomics*. 2010;11: 114–24.
91. Ferguson AA, Jiang N. Pack-MULES: recycling and reshaping genes through GC-biased acquisition. *Mobile Genetic Elements*. 2011;1:135–8.
92. Merchant SS, Prochnik SE, Vallon O, et al. The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*. 2007;318: 245–50.
93. Yoon HS, Hackett JD, Cinglia C, Pinto G, Bhattacharya D. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol*. 2004;21:809–18.

Supplementary Data

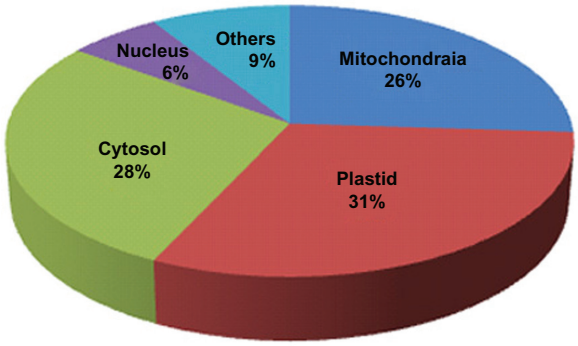

Figure S1. Classification of microalgal lipid biosynthetic proteins on the basis of subcellular localization using TargetP, ChloroP and WolfPsort prediction tools.

Table S1. Subcellular localisation prediction of proteins encoded by lipid biosynthetic genes in *A. thaliana, C. reinhardtii, V. carteri, O.lucimarinus, O. tauri* and *C. merolae*, using TargetP, ChloroP and WolfPsort programs.

Table S2. Various physico-chemical characters exhibited by putative proteins encoded by genes involved in lipid metabolism in *A. thaliana, C. reinhardtii, V. carteri, O. lucimarinus, O. tauri* and *C. merolae*, as calculated by ProtParam server.

Table S3. The calculated secondary structures of the proteins encoded by lipid biosynthetic genes, using GOR IV program.

Table S4. GC-content values of lipid biosynthetic genes as calculated by Genscan web server.

Table S5. Exon-intron coordinates of lipid biosynthetic genes in *A .thaliana, C. reinhardtii, V. carteri, O. lucimarinus, O. tauri* and *C. merolae*.