

RESEARCH ARTICLE

Open Access

Transcriptome sequencing and annotation of the microalgae *Dunaliella tertiolecta*: Pathway description and gene discovery for production of next-generation biofuels

Hamid Rismani-Yazdi^{1,2}, Berat Z Haznedaroglu¹, Kyle Bibby¹, Jordan Peccia^{1*}

Abstract

Background: Biodiesel or ethanol derived from lipids or starch produced by microalgae may overcome many of the sustainability challenges previously ascribed to petroleum-based fuels and first generation plant-based biofuels. The paucity of microalgae genome sequences, however, limits gene-based biofuel feedstock optimization studies. Here we describe the sequencing and *de novo* transcriptome assembly for the non-model microalgae species, *Dunaliella tertiolecta*, and identify pathways and genes of importance related to biofuel production.

Results: Next generation DNA pyrosequencing technology applied to *D. tertiolecta* transcripts produced 1,363,336 high quality reads with an average length of 400 bases. Following quality and size trimming, ~ 45% of the high quality reads were assembled into 33,307 isotigs with a 31-fold coverage and 376,482 singletons. Assembled sequences and singletons were subjected to BLAST similarity searches and annotated with Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) identifiers. These analyses identified the majority of lipid and starch biosynthesis and catabolism pathways in *D. tertiolecta*.

Conclusions: The construction of metabolic pathways involved in the biosynthesis and catabolism of fatty acids, triacylglycerols, and starch in *D. tertiolecta* as well as the assembled transcriptome provide a foundation for the molecular genetics and functional genomics required to direct metabolic engineering efforts that seek to enhance the quantity and character of microalgae-based biofuel feedstock.

Background

Global demand for petroleum as a transportation and heating fuel is predicted to increase 40% by 2025 [1]. Liquid biofuels from plants and microalgae feedstock represent a renewable sustainable alternative to petroleum energy when greenhouse gases released during the combustion of these biofuels are partially neutralized by the carbon dioxide required for their growth. The greatly minimized acreage estimates, high lipid or starch content, and biomass production rates that surpass those of terrestrial plants suggest that biodiesel or ethanol derived from lipids or starch produced by microalgae may circumvent many of the limitations ascribed to

petroleum fuel and first generation plant-based biofuels [2-7]. The most commonly stated paradigm for producing biodiesel from microalgae is to grow these microorganisms in open pond or closed reactor systems, extract the lipids or starch, and transform them into biodiesel by transesterification or ethanol by fermentation, respectively.

Unlike ethanol or other plant biofuels, technology to economically grow microalgae with high lipid or starch content is in the early stages of development [2]. Economic viability and environmental sustainability require the optimization of characteristic microalgae strains and ecologies in order to increase the per cell enrichment of lipids or starch and to improve fuel production and performance properties [6]. An in-depth knowledge of microalgae genomics precludes these necessary increases in biological efficiency. Numerous studies concerning

* Correspondence: jordan.peccia@yale.edu

¹Department of Chemical and Environmental Engineering, Yale University, New Haven, CT 06511, USA

Full list of author information is available at the end of the article

the effects of stress conditions on lipid and starch contents of microalgae have been documented in the literature [8-14]. However, an understanding of how microalgae respond to physiological stress at molecular level is largely limited to model organisms [15,16], and the relevant pathways in microalgae have not been fully documented [17]. Although transcribed gene and pathway information is requisite for planning and introducing stable and successful genetic manipulations in these microalgae, these efforts have been hampered by the lack of sequenced genomes of biofuel relevant microalgae. Due to the large efforts that are required to sequence these medium size (~100 mb) eukaryotic genomes, only seven microalgae genomes have been completed as of 2010 [7]. Alternatively, transcriptome sequencing can be a more efficient approach for obtaining microalgae functional genomics information. Transcriptome sequencing targets only coding DNA and this reduced sequencing requirement coupled with the rapidly evolving next-generation sequencing methods can result in high transcriptome coverage depth and facilitates the *de novo* assembly of transcriptomes from species where full genomes do not exist [18-21]. The more rapid and economic creation of these transcriptomes enables researchers to focus on organisms of direct biofuels interest and reduce the reliance on model organisms [19].

The objectives of this study are to discover genes that encode enzymes involved in the biosynthesis of biofuel precursors in the microalgae *Dunaliella tertiolecta* and to describe the relevant metabolic pathways. *D. tertiolecta* is a flagellated unicellular marine microalga belonging to the Chlorophyta phylum. The rationale for selecting *D. tertiolecta* as a non-model organism in this study lies in its ability to produce large quantities of lipids and starch (up to 67% and 27% of organism dry weight, respectively), rapid growth rate in hyper saline environments which aids in maintenance of pure cultures, and lack of a rigid cell wall which eases product extraction and genetic manipulation [22-27]. These unique physiological and structural traits gives *D. tertiolecta* considerable advantages over model organisms such *Chlamydomonas reinhardtii* as feedstock for biofuel production.

D. tertiolecta was cultured under nitrogen- and osmotic-inducing stress conditions and total RNA was extracted from cells during log and stationary growth phases. Libraries of cDNA constructed from total RNA were normalized and sequenced using the 454 GS FLX platform with Titanium chemistry. The transcriptome was assembled using the pool of sequencing data obtained from all cDNA libraries, and resulting individual transcripts (isotigs) and singletons were annotated. Sequences were screened to identify enzymes-encoding genes present, and relevant lipid and starch pathways

were reconstructed. Results demonstrate the capability of using transcriptome data from next-generation sequencing to identify pathways of interest and potential targets for metabolic engineering in microalgae, and enable functional genomics studies on a non-model species relevant for the production of next-generation biofuel.

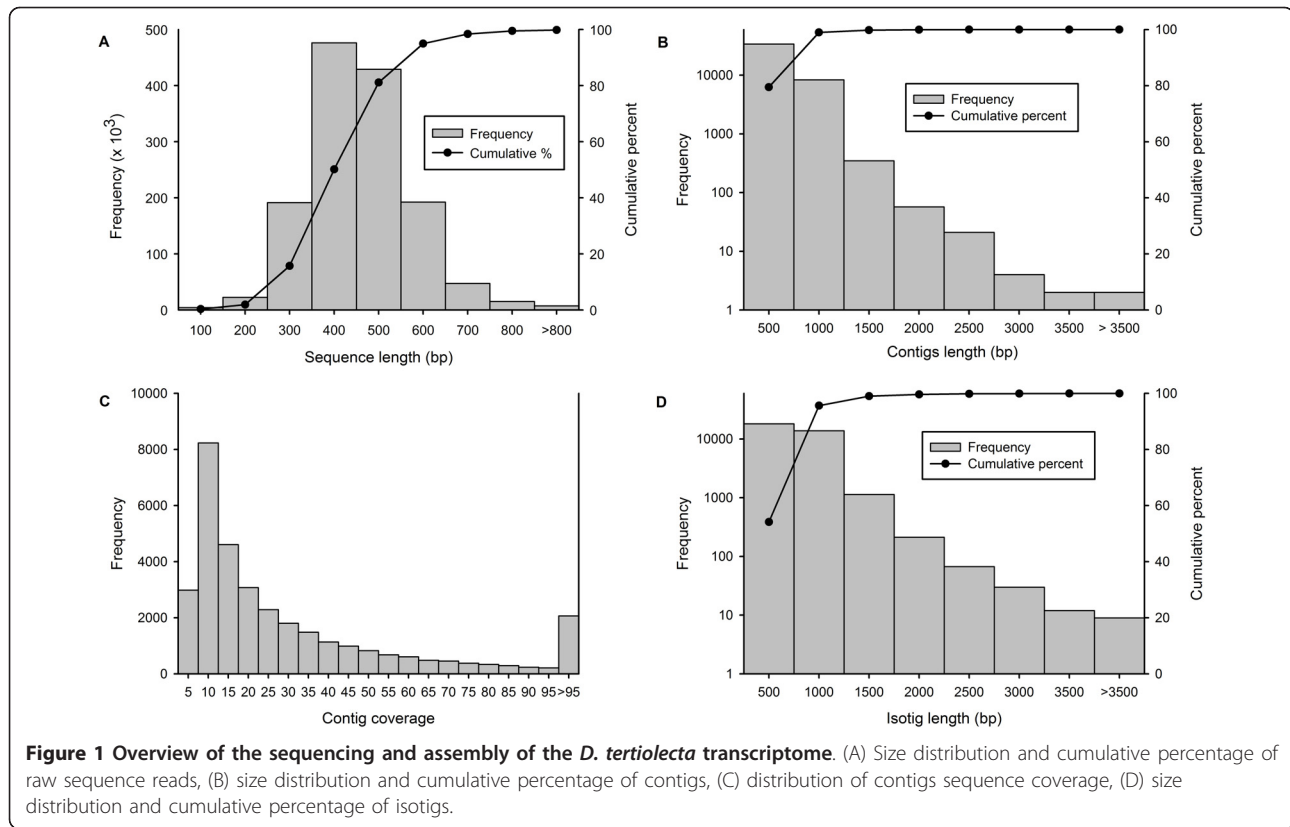
Results and Discussion

Sequencing and *de novo* assembly of the transcriptome

To identify genes and reconstruct the metabolic pathways involved in the production of biofuel precursors in *D. tertiolecta*, pure cultures were grown under nitrogen rich and nitrogen depleted conditions, and high salt concentrations. Cells were harvested in the log and stationary growth phases. These conditions are known to influence the production and accumulation of lipids and starch in microalgae [23,28-30], and were therefore used to increase the expression and maximize the diversity of genes related to these processes. Responses for nitrogen deprivation resulted in starch concentration doubling to over 25% of the cell dry weight with no increase in lipid content. The elevated salt concentration did not affect the starch content of nitrogen sufficient cells harvested during the stationary phase, but resulted in 22% increase in the total lipid content of the cells. Harvesting of microalgae in the exponential growth phase resulted in a near doubling of the lipid content to greater than 35% of the cell dry weight versus the stationary phase. The normalized cDNA libraries of cells grown under the above conditions were pooled and sequenced using the 454 GS FLX Titanium, and the *D. tertiolecta* transcriptome was assembled from the resulting sequencing reads.

Sequencing of cDNA libraries generated a total 1,385,389 raw reads, with an average length of 410 bp. The size distribution of raw reads is shown in Figure 1A, and a summary of sequencing and assembly results are presented in Table 1 and Additional file 1. After trimming for the adaptors and primer sequences, 20,036 sequences were removed due to their short length, low complexity, and overall low quality scores. This pre-assembly cleaning and trimming step resulted in 1,365,353 high quality (HQ) reads, with an average length of 400 bp, corresponding to 98.5% of the original raw sequences.

Trimmed and cleaned sequences were assembled using the cDNA assembly feature of Newbler software v.2.3. (Roche, IN, USA). A total of 609,149 HQ reads were assembled into 34,301 contiguous sequences (contigs), and 376,482 reads were identified as singletons (i.e., reads not assembled into contigs). The size of contigs ranged from 86 to 4,258 bp, with an average length of 377 ± 227 bp. The sequencing coverage ranged from



1 to 653 with an average of 31. The distribution of contigs size and coverage are shown in Figures 1B and 1C, respectively. Contiguous sequences were further assembled into 33,307 isotigs. Isotigs are the putative transcripts constructed using the overlapping contig reads provided as input to the Newbler cDNA assembler. The size distribution of isotigs which ranged from 101 to over 4,941 bp, with an average length of 532 ± 263 bp, are shown in Figure 1D. More than 95% of the assembled bases were incorporated into isotigs greater than 552 bp ($N_{50} = 552$ bp). The coverage depth for isotigs ranged from 1 to 14, with an average of 2.1 contigs assembled into each isotig. The isotigs and singletons together resulted in 409,789 unique sequences.

Functional annotation

All unique sequences were aligned against the sequences in the National Center for Biotechnology Information (NCBI) non-redundant (nr) protein database using the BLASTx algorithm. Using an E-value threshold of 10^{-6} , a total of 8,466 isotigs (25% of total isotigs), and 15,888 singleton sequences (4% of total singletons) had significant BLAST matches (Table 2). The frequency of annotated isotigs is consistent with the 20 to 40% values

previously reported for *de novo* transcriptome assemblies of eukaryotes [20,21,31,32]. Sequences that did not have BLASTx matches but met quality control may still have biological significance and may be important in future, directed studies on *D. tertiolecta* metabolism. In total, 24,354 unique sequences were identified through BLAST searches. Analysis of BLAST matches demonstrated a distinct microalgae character of the transcribed genes. The top-hit species distribution of BLAST matches is shown in Figure 2. Approximately 60% of the sequences had significant matches with closely related microalgae species, predominantly *Volvox carteri* and *Chlamydomonas reinhardtii*. A phylogenetic tree inferring the evolutionary relationship between *D. tertiolecta* and these organisms is presented in the Additional file 2. *C. reinhardtii* is a model microalga with a sequenced genome and has been the focus of most physiological, molecular, and genetic studies in microalgae. *V. carteri* is a multicellular alga and a draft genome sequence has been completed for this organism. BLAST search also identified 13% of the sequences as being similar with the marine algicidal bacterium *Kordia algicida*, and the remaining 27% were related to plant species and other marine microalgae and bacteria. The similarity of BLAST results with *Kordia algicida* indicates that this microorganism might share some genetic information

Table 1 D. tertiolecta transcriptome sequencing and assembly summary

	Sequences	Bases (Mbp)
Sequencing		
Raw sequencing reads	1,385,389	567.6
Average read length	410 bp	
Pre-assembly		
Trimmed	352,422	
Trashed	20,036	
Reads used in assembly	1,365,353	546.6
Average read length	400 bp	
¹ Assembly		
Contigs		
Reads assembled as contigs	609,149	258
Number of contigs	34,301	
Average length of contigs	377 bp	
Range of contigs length	86-4,258 bp	
Depth on contigs	31	
Isotigs		
Number of isotigs	33,307	17.8
Average length of isotigs	535 bp	
Range of isotigs length	101-4,941 bp	
Depth on isotigs	2.1	
Singletons		
Unique sequence	376,482	70.1

¹While this manuscript was in the review process, Newbler v2.5 became publicly available. We reassembled the *D. tertiolecta* transcriptome using Newbler v2.5 and the outcomes of assembly in comparison with those obtain from v2.3 are presented as the Additional file 1. The differences observed between the outputs of two versions were not significant to impact the major or minor conclusions of this study, and are consistent with reports from Ewen-Campen et al. [77].

with *D. tertiolecta*, or be present in our samples. The latter, however, is very unlikely since we used pure cultures of *D. tertiolecta*, and constructed the cDNA libraries using poly-T primers.

Using the Blast2Go platform [33], sequences with a BLAST match were further annotated with Gene Ontology (GO) terms and enzyme commission (EC) numbers.

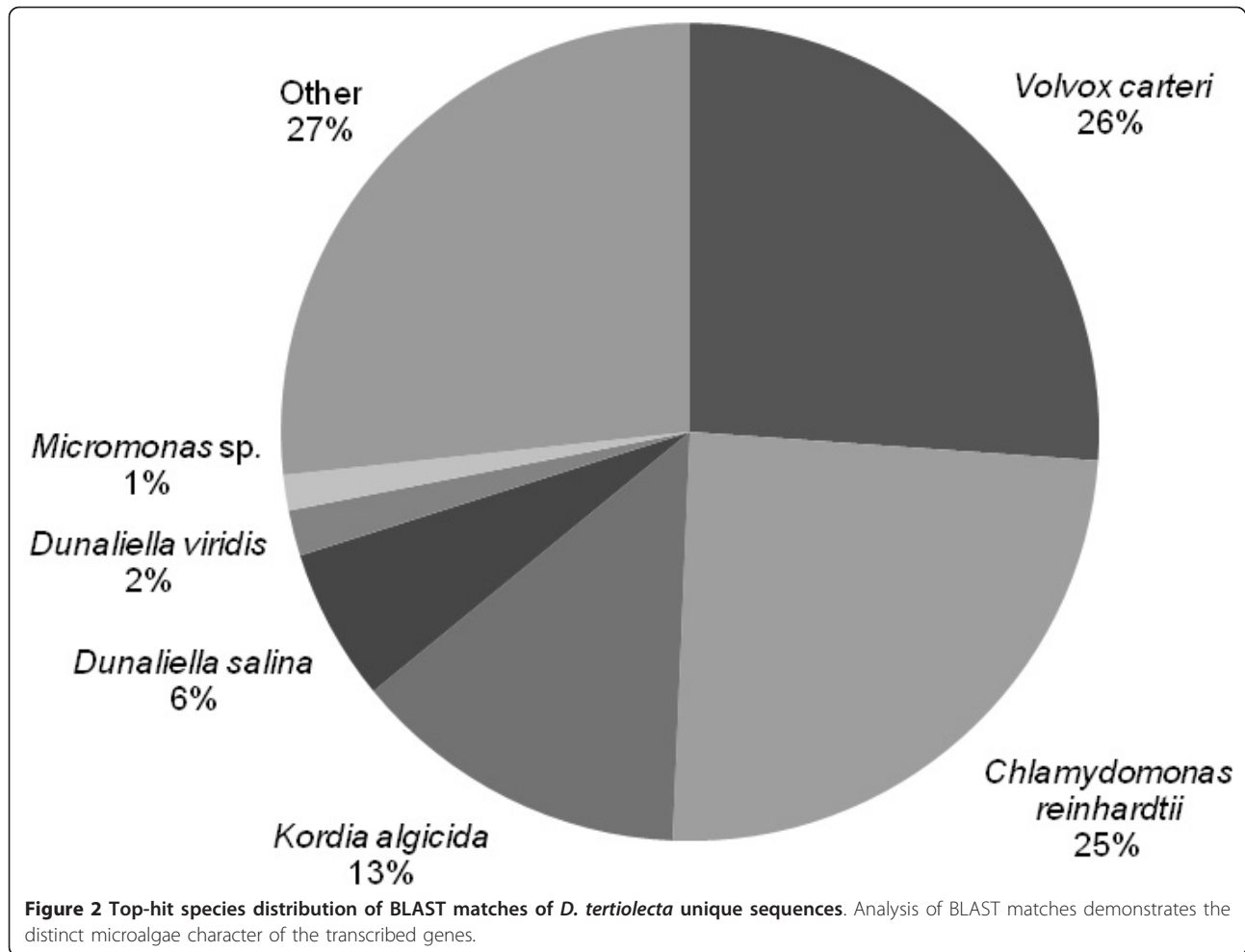
GO terms were assigned to a majority of isotigs (63%) and singletons (65%) (Table 2). The distribution of most abundant GO terms for biological processes, molecular functions, and cellular components is presented in Figure 3. Of the 15,686 sequences annotated with GO terms, 7,077 sequences were assigned with EC numbers (2,289 of isotigs, and 4,788 of singletons) (Table 2). To further enrich the annotation of our transcriptome library, unique sequences were assigned with KEGG orthology (KO) identifiers using KEGG Automatic Annotation Server (KAAS) [34], and subsequently mapped to BRITE functional hierarchies. The BRITE functional hierarchies linked many of the annotated sequences with biological systems such as genetic and environmental information processing, and cellular processes in *D. tertiolecta* [Additional file 3].

Pathway classification by KEGG

To reconstruct the metabolic pathways involved in *D. tertiolecta*, annotated sequences were mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) using the Blast2Go platform [35]. We identified transcripts coding for all enzymes related to the major metabolic pathways in *D. tertiolecta* (Table 3). The completeness of these reconstructed pathways indicates that the gene function assignments were biologically meaningful and that the EC numbers have been correctly assigned to annotated sequences. Comparative analysis of enzyme-coding sequences between *D. tertiolecta* and model organisms, *Volvox carteri*, and *Chlamydomonas reinhardtii*, using BLASTx analysis revealed relatively low homology between *D. tertiolecta* and these organisms for the enzymes described in this study (Table 4, 5, 6). These differences indicate that functional genomics and metabolic engineering of *D. tertiolecta* cannot be fully based on the sequence information obtained from model organisms, and further demonstrates the importance of annotated *D. tertiolecta* transcriptome as a genetic framework required for carrying out such studies. Because of their relevance to production of precursors for biofuel production, the metabolic pathways associated with biosynthesis and catabolism of lipids and starch were given further treatment below. A more fundamental understanding of these pathways in

Table 2 D. tertiolecta transcriptome annotation summary

	Number of sequences		
	Isotigs	Singletons	Total unique sequences
Total number of sequences	33,307	376,482	409,789
Sequences with BLAST matches	8,466	15,888	24,354
Sequences annotated with Gene Ontology (GO) terms	5,354	10,332	15,686
Sequences assigned with Enzyme Commission (EC) numbers	2,289	4,788	7,077



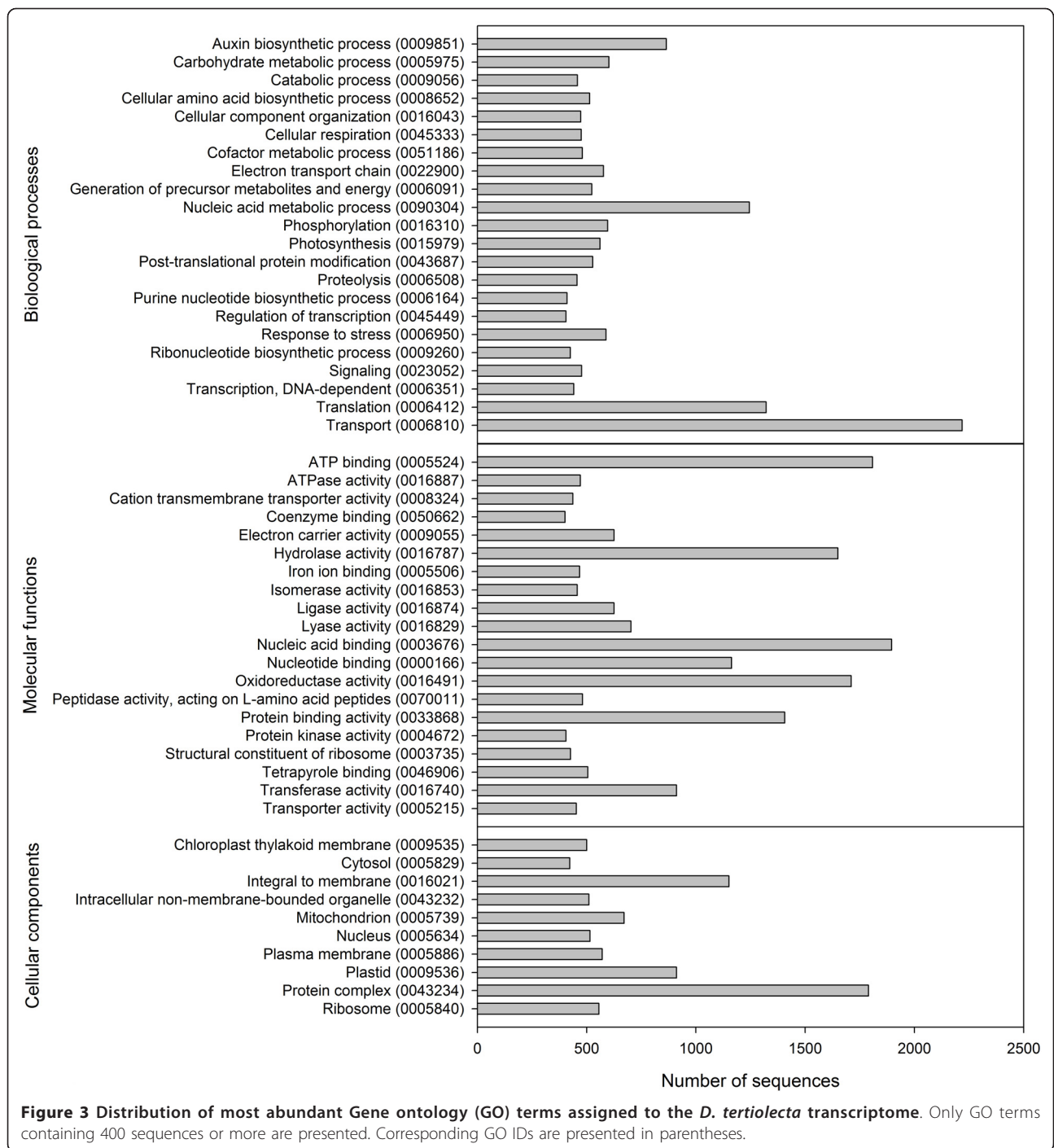
microalgae is required to direct efforts to enhance the microalgae-biofuel production and produce the fuel characteristics needed for commercialization.

Fatty acid biosynthesis

Interest in microalgae as a potential feedstock for the production of biofuels and other valuable biomaterials is rooted in their ability to rapidly accumulate significant amounts of neutral lipids [36]. Under optimal conditions, microalgae synthesize fatty acids primarily for esterification into polar glycerol-based membrane lipids that consist of glycosylglycerides and phosphoglycerides, whereas under environmental stress conditions, many microalgae accumulate neutral triacylglycerols (TAGs) [17]. Although global fatty acid biosynthesis pathways are known in eukaryotes [37], biosynthesis and regulation of fatty acids in microalgae are not well studied.

Based on the functional annotation of the transcriptome, we have successfully identified the genes encoding for key enzymes involved in the biosynthesis and catabolism of fatty acids in *D. tertiolecta*. Table 4 lists the

enzymes involved in fatty acid biosynthesis in *D. tertiolecta*. On average, four transcripts were identified per enzyme. The reconstructed pathway based on these identified enzymes is depicted in Figure 4. The fatty acid biosynthesis pathway in microalgae primarily occurs in the chloroplast, and produces saturated C16:0 (palmitic acid) and C18:0-ACP (acyl carrier protein). These fatty acids can then be used as precursors for the synthesis of neutral lipids (mainly TAGs). Fatty acid biosynthesis in *D. tertiolecta* starts with acetyl-CoA carboxylase (ACC, EC: 6.4.1.2), which catalyzes the formation of malonyl-CoA from acetyl-CoA and bicarbonate. Next, malonyl-CoA ACP transacylase (MAT, EC: 2.3.1.39) catalyzes the transfer of malonyl-CoA to malonyl-ACP, the carbon donor for subsequent elongation reactions. During the elongation, the malonyl group of malonyl-ACP participates in a series of condensation reactions with acyl ACP (or acetyl-CoA) acceptors that are catalyzed by the multiple isoforms of the condensing enzyme, ketoacyl-ACP synthase (KAS) (Table 4). The first condensation reaction is catalyzed by beta-ketoacyl-ACP



synthase III (KAS III, EC: 2.3.1.180) and results in a four-carbon product. The other condensing enzyme, beta-ketoacyl-ACP synthase I (KAS I, EC: 2.3.1.41), produces varying chain lengths (6 to 16 carbons). Finally, beta-ketoacyl-ACP synthase II (KASII, EC: 2.3.1.179) is responsible for the elongation of C16:0-ACP to C18:0-ACP. The initial product of each condensation reaction is a beta-ketoacyl-ACP. Three additional reactions

(reduction, dehydration, and another reduction) occur after each condensation. To form a saturated fatty acid, the beta-ketoacyl ACP product is reduced by the enzyme beta-ketoacyl-ACP reductase (KAR, EC: 1.1.1.100), dehydrated by 3R-hydroxyacyl-ACP dehydrase (HAD, EC: 4.2.1.-), and then reduced by enoyl-ACP reductase (EAR, EC: 1.3.1.9). These four reactions lead to a lengthening of the precursor fatty acid by two

Table 3 Essential metabolic pathways annotated in the *D. tertiolecta* transcriptome

Pathway	Enzymes found	Known enzymes
Photosynthetic carbon fixation (Calvin cycle)	12	13
Glycolysis/Gluconeogenesis	10	10
Pentose phosphate	5	5
Citrate cycle	10	10
Fatty acid biosynthesis	6	6
TAG biosynthesis	3	4
Starch biosynthesis	4	4

carbons for every round of cycle. To produce 16 or 18 carbon fatty acids, this cycle is repeated for 6 or 7 times respectively (Figure 4).

For the synthesis of unsaturated fatty acids in plastid, a double bond is introduced to the acyl group esterified to ACP via the enzyme acyl-ACP desaturase (AAD, EC: 1.14.19.2). The elongation of fatty acids in the chloroplast is terminated when the acyl group is removed from ACP by acyl-ACP thioesterase enzymes, oleoyl-ACP hydrolase (OAT, EC: 3.1.2.14), or when acyl-ACP thioesterase A (FatA) hydrolyze the acyl-ACP and releases the free fatty acid, or when acyl transferases in the chloroplast transfers the fatty acid directly from ACP to glycerol-3-phosphate or monoacylglycerol-3-phosphate. The final fatty acid composition is determined by the activities of enzymes that use these acyl-ACPs at the termination phase of fatty acid synthesis. We have also identified desaturation enzymes $\Delta^{12}(\omega^6)$ -desaturase ($\Delta^{12}D$, EC: 1.4.19.6), which desaturates oleic acid (18:1n-9) to form linoleic acid (18:2n-6), and $\Delta^{15}(\omega^3)$ -desaturase ($\Delta^{15}D$, EC: 1.4.19.-), which further desaturates linoleic acid to form α -linolenic acid (18:3n-3). The annotation of *D. tertiolecta* transcriptome did not identify any genes encoding enzymes involved in further desaturation and elongation of linoleic and linolenic acids that could result in production of longer chain polyunsaturated fatty acids. The lack of identification of these enzymes is consistent with the profile of fatty acids produced by *D. tertiolecta* [38,39].

In addition to synthesis, all the enzymes involved in fatty acid catabolism (β -oxidation pathway) of *D. tertiolecta* were successfully identified and are presented in Table 4 (coded by six transcripts on average). The fatty acid catabolism pathway is provided in Additional file 4. The β -oxidation pathway in microalgae involves four enzymes: acyl-coA oxidase (AOx, EC: 1.3.3.6); enoyl-CoA hydratase (ECH, EC: 4.2.1.17); 3-hydroxyacyl-CoA dehydrogenase (CHAD, EC: 1.1.1.35); and acetyl-CoA acyltransferase (ACAT, EC: 2.3.1.16), which collectively catalyze the cleavage of two carbons from the acyl chain

during each cycle. The resulting acetyl-CoA is then used to produce energy for the cell via the citrate cycle.

The *D. tertiolecta* transcriptome presented here contains most of the enzymes required for the biosynthesis, elongation, and metabolism of fatty acids (Table 4), and the subsequent reconstructed pathways are consistent with those proposed for model microalgae *C. reinhardtii* [17], and plants in general [37,40-42]. These findings contribute to the biochemical and molecular information needed for metabolic engineering of fatty acid synthesis in microalgae. The most commonly stated strategy is the overexpression of ACC, the rate-limiting step in fatty-acid biosynthesis [43]. The condensing enzymes that are identified in this study are also potential targets for improving fatty acid biosynthesis. For example, Verwoert et al. has shown that the overexpression of KAS III in *Brassica* seeds alter the composition of fatty acids but does not change the per cell quantity [44]. A final example approach for per cell fatty acid enrichment in microalgae is to block lipid catabolism [45], which could then result in increased lipid storage.

Triacylglycerols (TAG) biosynthesis and catabolism

Some species of microalgae are capable of producing and accumulating high amounts of neutral storage lipids, mainly TAGs, under stress-inducing conditions. TAGs can serve as precursors for production of biodiesel and other bio-based products such as plastics, cosmetics, and surfactants [36]. Although the global pathways for TAG biosynthesis are known, the existing knowledge on the pathways and enzymes involved in TAG synthesis in microalgae is limited [46,47]. Based on the KEGG pathway assignment of the functionally annotated sequences, transcripts coding for most of the enzymes involved in TAG biosynthesis were found in *D. tertiolecta*. These enzymes are presented in Table 5, and the suggested pathway for TAG synthesis in *D. tertiolecta* is shown in Figure 5. The precursors for TAG biosynthesis are glycerol-3-phosphate (G-3-P), and acyl-CoA. G-3-P is produced by the catabolism of glucose (glycolysis) or to a lesser extent by the action of the enzyme glycerol kinase (GK, EC: 2.7.1.30) on free glycerol. We identified 16 transcripts in *D. tertiolecta* transcriptome library coding for this enzyme. The acyl-CoA, on the other hand, is generated via esterification of fatty acids produced in the chloroplast to coenzyme A. The first two steps of TAG biosynthesis involve sequential esterification of acyl chains from acyl-CoA to positions 1 and 2 of G-3-P, resulting in the formation of phosphatidic acid (PA), which is a key intermediate in the biosynthesis of all glycerolipids. These steps are catalyzed by glycerol-3-phosphate O-acyltransferase (GPAT, EC: 2.3.1.15) and 1-acyl-sn-glycerol-3-phosphate

Table 4 Enzymes involved in fatty acid biosynthesis and metabolism identified by annotation of the *D. tertiolecta* transcriptome

Enzyme	Symbol	EC Number	Number of transcripts	¹ %Sequence alignment with corresponding enzymes in model organisms (Accession #)	
				<i>C. reinhardtii</i>	<i>V. carteri</i>
Fatty acid biosynthesis					
Biotin carboxylase	BC	6.3.4.14	3	² NM	46 (EFJ41621.1)
Acetyl-CoA carboxylase	ACC	6.4.1.2	7	28 (XP_001700442.1)	28 (EFJ46223.1)
malonyl-CoA-ACP transacylase	MAT	2.3.1.39	4	NM	62 (EFJ49352.1)
Beta-ketoacyl-ACP synthase I	KAS I	2.3.1.41	7	NM	NM
Beta-ketoacyl-ACP synthase II	KAS II	2.3.1.179	12	NM	NM
Beta-ketoacyl-ACP synthase III	KAS III	2.3.1.180	1	NM	NM
Beta-ketoacyl-ACP reductase	KAR	1.1.1.100	7	63 (XP_001703473.1)	61 (EFJ52014.1)
3R-hydroxyacyl-ACP dehydrase	HAD	4.2.1.-	2	NM	NM
Enoyl-ACP reductase (NADH)	EAR	1.3.1.9	5	NM	NM
Oleoyl-ACP thioesterase	OAT	3.1.2.14	1	NM	NM
Acyl-ACP thioesterase A	FatA	3.1.2.14, 3.1.2.-	1	NM	NM
Acyl-ACP thioesterase B	FatB	3.1.2.14, 3.1.2.-	0	NM	NM
Fatty acid desaturation					
Δ^9 Acyl-ACP desaturase	AAD	1.14.19.2	5	76 (XP_001691597.1)	76 (EFJ49192.1)
$\Delta^{12}(\omega^6)$ -Desaturase	$\Delta^{12}D$	1.14.19.6	3	54 (XP_001691669.1), 47 XP_001693068.1)	54 (XP_002955859.1), 47 (XP_002949932.1)
$\Delta^{15}(\omega^3)$ -Desaturase	$\Delta^{15}D$	1.14.19.-	2	NM	NM
Fatty acid elongation					
3-Hydroxyacyl-CoA dehydrogenase	CHAD	1.1.1.35	9	NM	NM
Long-chain-3-hydroxyacyl-CoA dehydrogenase	LCHAD	1.1.1.211	2	NM	NM
Enoyl-CoA hydratase	ECH	4.2.1.17	16	NM	70 (EFJ49010.1)
Trans-2-enoyl-CoA reductase (NADPH)	TER	1.3.1.38	3	90 (XP_001690095.1)	NM
Palmitoyl-CoA hydrolase	PCH	3.1.2.22	1	19 (XP_001698944.1)	19 (EFJ40725.1)
Fatty acid catabolism					
Long-chain acyl-CoA synthetase	ACSL	6.2.1.3	3	77 (XP_001693692.1)	77 (EFJ51208.1)
Acyl-CoA oxidase	AOx	1.3.3.6	2	39 (XP_001699193.1)	NM
Acyl-CoA dehydrogenase	ACADM	1.3.99.3	16	45 (XP_001699193.1), 84 (XP_001693484.1), 73 (XP_001695945.1)	NM
Enoyl-CoA hydratase	ECH	4.2.1.17	16	NM	70 (EFJ49010.1)
Long-chain 3-hydroxyacyl-CoA dehydrogenase	LCHAD	1.1.1.211	2	NM	NM
Acetyl-CoA acyltransferase	ACAT	2.3.1.16	1	NM	NM
3-hydroxyacyl-CoA dehydrogenase	CHAD	1.1.1.35	9	NM	NM
Acetyl-CoA C-acetyltransferase	thiL	2.3.1.9	6	39 (XP_001694888.1)	34 (EFJ47048.1), 34 (EFJ39622.1), 34 (EFJ46961.1)
3-hydroxybutyryl-CoA epimerase		5.1.2.3	8	NM	NM
Enoyl-CoA isomerase	Eci	5.3.3.8	8	NM	NM
Alcohol dehydrogenase	ADH	1.1.1.1	3	80 (XP_001693934.1)	53 (EFJ45245.1)
Aldehyde dehydrogenase (NAD ⁺)	ALDH	1.2.1.3	3	24 (XP_001696928.1)	21 (EFJ40853.1)
Ferredoxin-NAD ⁺ reductase	FNR	1.18.1.3	1	NM	NM

¹In cases where multiple transcripts have been aligned with the associated enzymes in the model organisms, average similarity is reported.

²NM denotes that the annotated transcripts did not match the sequence of corresponding enzyme in model organisms.

Table 5 Enzymes involved in TAG biosynthesis and catabolism identified by annotation of the *D. tertiolecta* transcriptome

Enzyme	Symbol	EC Number	Number of transcripts	¹ %Sequence alignment with corresponding enzymes in model organisms (Accession #)	
				<i>C. reinhardtii</i>	<i>V. carteri</i>
TAG biosynthesis					
Glycerol kinase	GK	2.7.1.30	16	42 (A8IYH1)	NM
Glycerol-3-phosphate O-acyltransferase	GPAT	2.3.1.15	0	² NM	NM
1-Acyl-sn-glycerol-3-phosphate O-acyltransferase	AGPAT	2.3.1.51	2	NM	NM
Phosphatidate phosphatase	PP	3.1.3.4	1	NM	NM
Diacylglycerol O-acyltransferase	DGAT	2.3.1.20	2	NM	NM
Phospholipid:diacylglycerol acyltransferase	PDAT	2.3.1.158	1	NM	NM
TAG catabolism					
Triacylglycerol lipase	TAGL	3.1.1.3	2	NM	NM

¹In cases where multiple transcripts have been aligned with the associated enzymes in the model organisms, average similarity is reported.

²NM denotes that the annotated transcripts did not match the sequence of corresponding enzyme in model organisms.

O-acyltransferase (AGPAT, EC: 2.3.1.51), respectively. Genes encoding for AGPAT were identified in the *D. tertiolecta* transcriptome library, however, genes encoding for GPAT were not identified. The next reaction in the pathway is catalyzed by the enzyme phosphatidate phosphatase (PP, EC: 3.1.3.4), which removes the phosphate group from PA and produces the diacylglycerol (DAG). One transcript was annotated as coding for this enzyme in the *D. tertiolecta* transcriptome library. The DAG is an essential intermediate in the biosynthesis of phosphatidylcholine (PC) and phosphatidylethanolamine (PE). Finally, the resultant DAG is acylated, at the position 3 using an acyl donor to form the TAG. We identified transcripts coding for two distinct enzymes that catalyze this reaction. These enzymes differ in their source of acyl-donor. The first enzyme is diacylglycerol O-acyltransferase (DGAT, EC: 2.3.1.20), which uses acyl-

CoA as an acyl-donor and is believed to be the main pathway for biosynthesis of TAG [17,48]. The second enzyme is phospholipid:diacylglycerol acyltransferase (PDAT, EC: 2.3.1.158), and uses phospholipids as acyl donors. This acyl-CoA independent route for TAG biosynthesis had been previously reported in some plants and yeast [49], and the gene encoding for PDAT has been identified in the sequenced genome of *C. reinhardtii* [46]. Identification of PDAT in *D. tertiolecta* provides further evidence that some microalgae might have the potential to channel the fatty acids incorporated in membrane lipids (e.g. PC), into the TAG synthesis. The identification of this alternative route for the flux of fatty acids into and out of TAG synthesis can also provide insight into the connection between rapid degradation of membrane lipids with concurrent accumulation of TAGs when microalgae are exposed to stress

Table 6 Enzymes involved in starch biosynthesis and metabolism identified by annotation of the *D. tertiolecta* transcriptome

Enzyme	Symbol	EC Number	Number of transcripts	¹ %Sequence alignment with corresponding enzymes in model organisms (Accession #)	
				<i>C. reinhardtii</i>	<i>V. carteri</i>
Starch biosynthesis					
Phosphoglucomutase	PGM	5.4.2.2	10	² NM	NM
ADP-glucose pyrophosphorylase	AGPase	2.7.7.27	7	49 (EDP04344.1), 84 (EDP08701.1)	NM
Starch synthase	SS	2.4.1.21	17	92 (EDP00372.1)	NM
1,4- α -Glucan branching enzyme	BE	2.4.1.18	6	87 (EDP05581.1)	82 (EFJ46288.1)
Sucrose phosphorylase	SuPase	2.4.7.1	3	NM	NM
Hexokinase	HXX	2.7.1.1	6	NM	NM
Starch catabolism					
α -Amylase	(α -AMY)	3.2.1.1	9	98 (EDP00963.1)	NM
oligo-1,6-Glucosidase	(O1, 6G)	3.2.1.10	1	NM	NM
Starch phosphorylase	(SPase)	2.4.1.1	10	43 (EDP02951.1), 78 (EDO98385.1)	NM

¹In cases where multiple transcripts have been aligned with the associated enzymes in the model organisms, average similarity is reported.

²NM denotes that the annotated transcripts did not match the sequence of corresponding enzyme in model organisms.

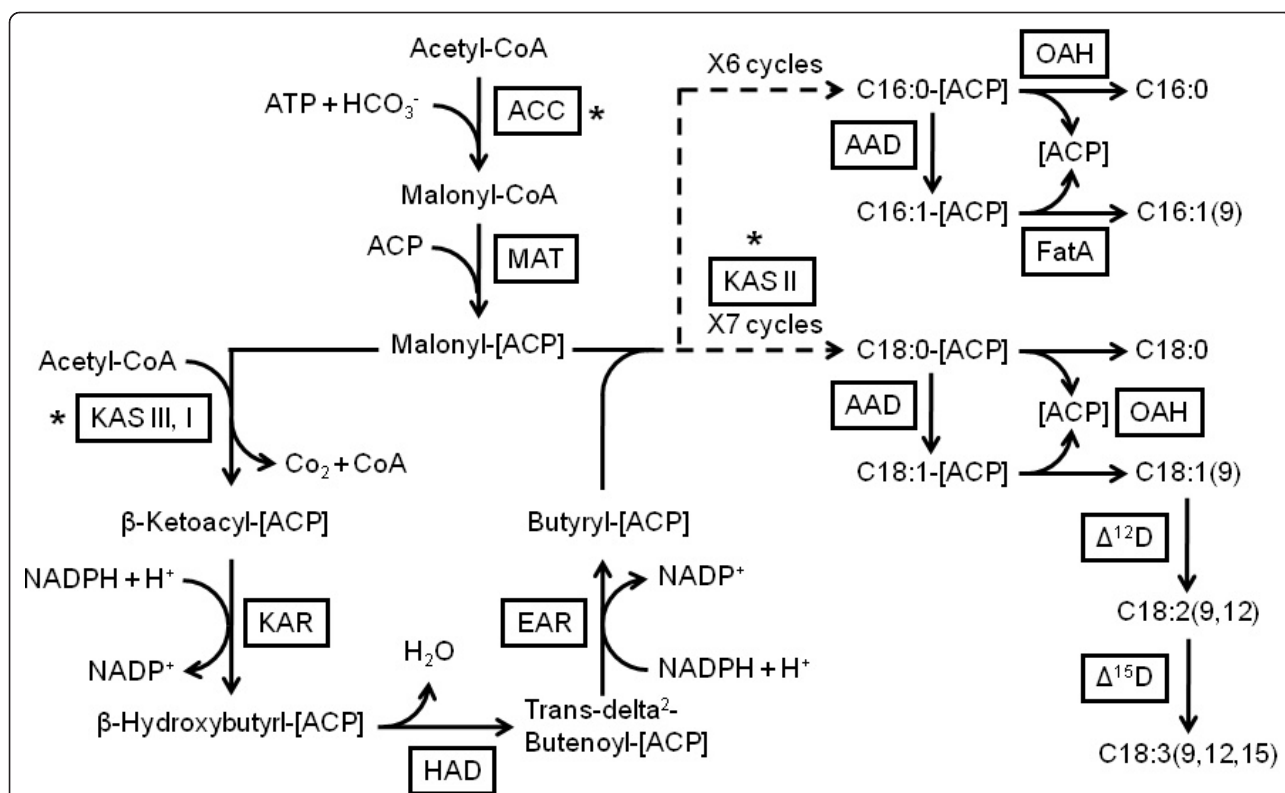


Figure 4 Fatty acid biosynthesis pathway reconstructed based on the *de novo* assembly and annotation of *D. tertiolecta* transcriptome. Identified enzymes are shown in boxes and include: ACC, acetyl-CoA carboxylase (EC: 6.4.1.2); MAT, malonyl-CoA ACP transacylase (EC: 2.3.1.39); KAS, beta-ketoacyl-ACP synthase (KAS I, EC: 2.3.1.41; KASII, EC: 2.3.1.179; KAS III, EC: 2.3.1.180); KAR, beta-ketoacyl-ACP reductase (EC: 1.1.1.100); HAD, beta-hydroxyacyl-ACP dehydrase (EC: 4.2.1.-); EAR, enoyl-ACP reductase (EC: 1.3.1.9); AAD, acyl-ACP desaturase (EC: 1.14.19.2); OAH, oleoyl-ACP hydrolase (EC: 3.1.2.14); FatA, Acyl-ACP thioesterase A (EC: 3.1.2.-); $\Delta^{12}D$, $\Delta^{12}(\omega^6)$ -desaturase (EC: 1.4.19.6); $\Delta^{15}D$, and $\Delta^{15}(\omega^3)$ -desaturase (EC: 1.4.19.-). The fatty acid biosynthesis pathway in *D. tertiolecta* produces saturated, C16:0 (palmitic acid) and C18:0 (stearic acid), and unsaturated fatty acids C18:1 (9) (oleic acid), C18:2 (9,12) (linoleic acid), and C18:3(9,12,15) (α -linolenic acid). Key enzymes are shown with an asterisk next to the boxes.

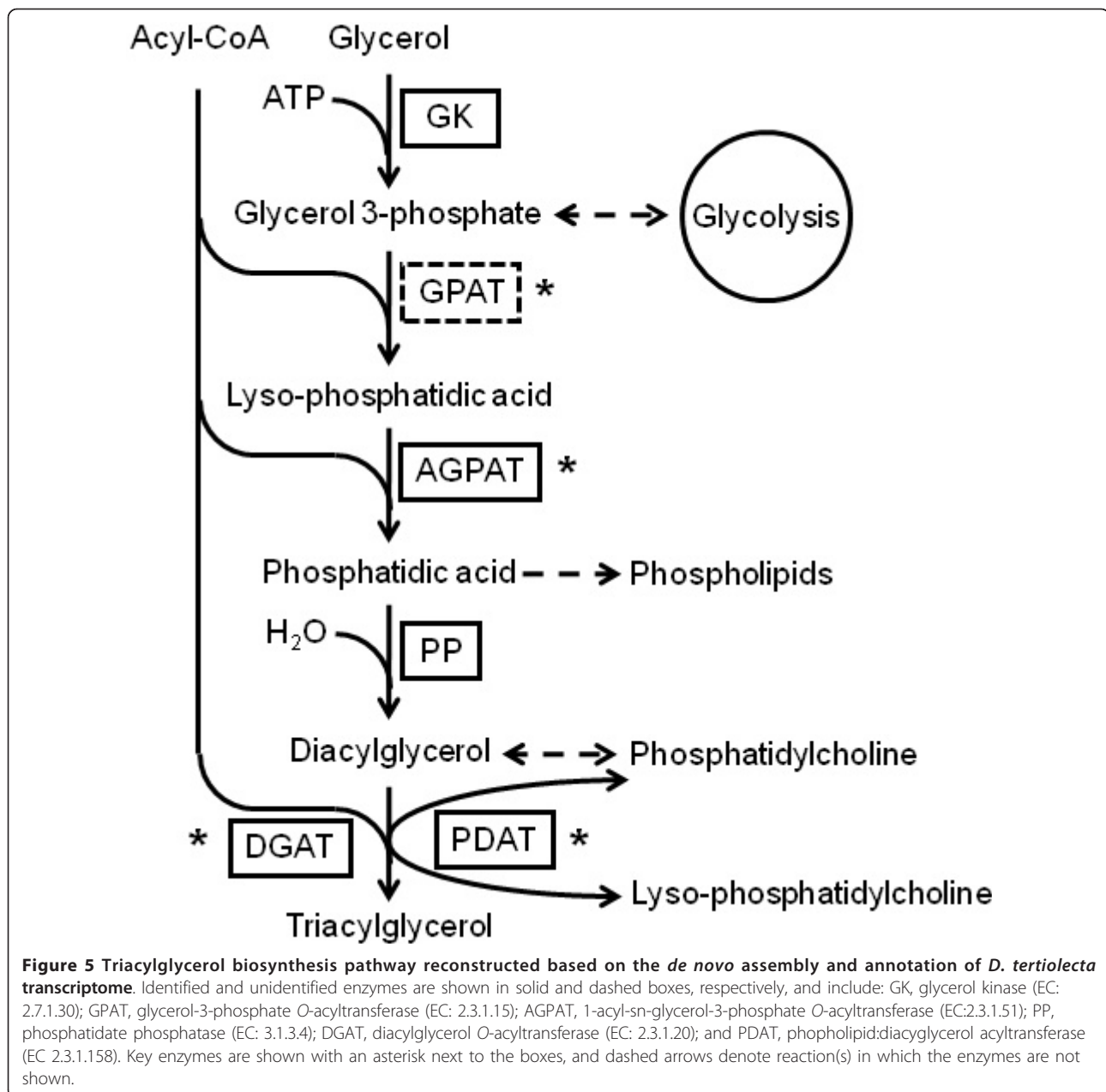
conditions [17]. The functional importance of PDAT in TAG biosynthesis, however, remains to be determined via gene-knockout experiments and analysis of lipid profiles.

Additionally, we identified transcripts coding for enzymes related to catabolism of TAG. The complete breakdown of TAG takes place in two stages. First, hydrolysis of ester bonds that link fatty acyl chains to the glycerol backbone is catalyzed by lipases. We found two transcripts in the *D. tertiolecta* transcriptome coding for triacylglycerol lipase (TAGL, EC: 3.1.1.3), which releases fatty acids from DAG and TAG. In the second stage, the fatty acids that are liberated may be further broken down by oxidation or follow other metabolic pathways including re-esterification with glycerol to form new acylglycerols [42]. Suppression of enzymes involved in TAG degradation, such as TAGL, could potentially increase the TAG content. Though this approach has previously resulted in elevated levels of TAG in transgenic plants, it severely limits plant growth [50]. Another potential

approach includes manipulation of acyltransferases enzymes, as they are the key determinant of content and acyl composition of glycerolipids [51-54]. Identification of transcripts coding for these enzymes in *D. tertiolecta* provide the first step for attempts to genetically engineer this organism to increase the production and modify the composition of the lipids [45,46].

Starch biosynthesis and catabolism

As the main assimilatory product of photosynthesis, some species of microalgae synthesize a significant amount of starch as storage materials in their plastids [29,55]. The accumulated starch is an attractive substrate for the production of a variety of biofuels, including ethanol, butanol, and hydrogen [56-59]. Production of biofuel from microalgae-based starch potentially overcomes the sustainability and pretreatment process disadvantages ascribed to using plant-based starch and lignocellulosic materials as ethanol feedstock [57,60]. Though the pathways associated with biosynthesis and



degradation of starch are well studied in plants and the model microalgae *C. reinhardtii*, such knowledge is scarce in non-model microalgae with direct biofuel potentials.

Based on the KEGG pathway assignments, we identified numerous transcripts coding for enzymes involved in the biosynthesis and catabolism of starch in *D. tertiolecta* (Table 6). The pathway of starch synthesis in *D. tertiolecta* involves the enzymes phosphoglucomutase (PGM, EC: 5.4.2.2), which generates glucose-1-phosphate (Glc-1-P) from glucose-6-phosphate (Glc-6-P), and ADP-glucose pyrophosphorylase (AGPase, EC:

2.7.7.27), which uses Glc-1-P and ATP to generate ADPGlc and inorganic pyrophosphate. ADPGlc is the substrate for starch synthase (SS, EC: 2.4.1.21), which generates ADP and an amylose, a linear chain of glucose residues connected by α -1,4-glycosidic bonds. The final step in the pathway involves α -1,4-glucan branching enzyme (BE, EC: 2.4.1.18), which catalyzes the formation of α -1,6-glycosidic bonds in the elongated glucans that comprise starch. Based on the identified enzymes from the retrieved transcripts, biosynthesis of starch from Glc-6-P in *D. tertiolecta* resembles the classical pathway of starch synthesis proposed in plants and *C. reinhardtii*

[61]. We also identified transcripts that code for sucrose phosphorylase (SuPase, EC: 2.4.7.1), which generates G-1-P from sucrose as an alternative source of substrate for starch synthesis. This alternative pathway is consistent with the newly proposed model for starch synthesis in *Arabidopsis* [62]. Experimental research (e.g. gene knock-out analysis), however, is required to support the role of SuPase in starch biosynthesis in *D. tertiolecta*. Figure 6 shows the reconstructed pathway for biosynthesis of starch in *D. tertiolecta*. Although our results do not allow speculation on the cellular location of starch biosynthesis in *D. tertiolecta*, previous studies on *D. marina* starch pathway suggest that it is exclusively intraplasmidic [63], which is consistent with starch synthesis and storage in the *Cryptophyceae* class of microalgae [64].

The starch biosynthesis pathway was well represented in our library as indicated by the number of transcripts assigned to each enzyme (on average 9 transcripts per enzyme). All of known enzymes in the starch synthesis pathway presented were identified (Table 6). Genetic

manipulation of key enzymes, mainly AGPase and to less extent SS, involved in the process has been tried to increase starch contents in crop plants. Much of the efforts have been focused on AGPase as this enzyme catalyses a rate-limiting step in the biosynthesis of starch, and thus increase in its activity could lead to increased rate of starch synthesis [65].

The starch accumulated in green microalgae is considered to be mostly utilized in the respiration. We identified two distinct pathways, namely hydrolytic and phosphorolytic, involved in starch catabolism in *D. tertiolecta* (Figure 6). Twenty transcripts in our library were annotated as coding for enzymes involved in these pathways (Table 6). The enzymes related to the hydrolytic pathway include α -amylase (α -AMY, EC: 3.2.1.1), and oligo-1,6-glucosidase (O1, 6G, EC: 3.2.1.10). These two enzymes catalyze the hydrolysis of starch to oligosaccharides (i.e. dextrin) and further to α -D-glucose, respectively. The released α -D-glucose may be further degraded through glycolysis or be phosphorylated by hexokinase (HXK, 2.7.1.1), for reentry into the starch

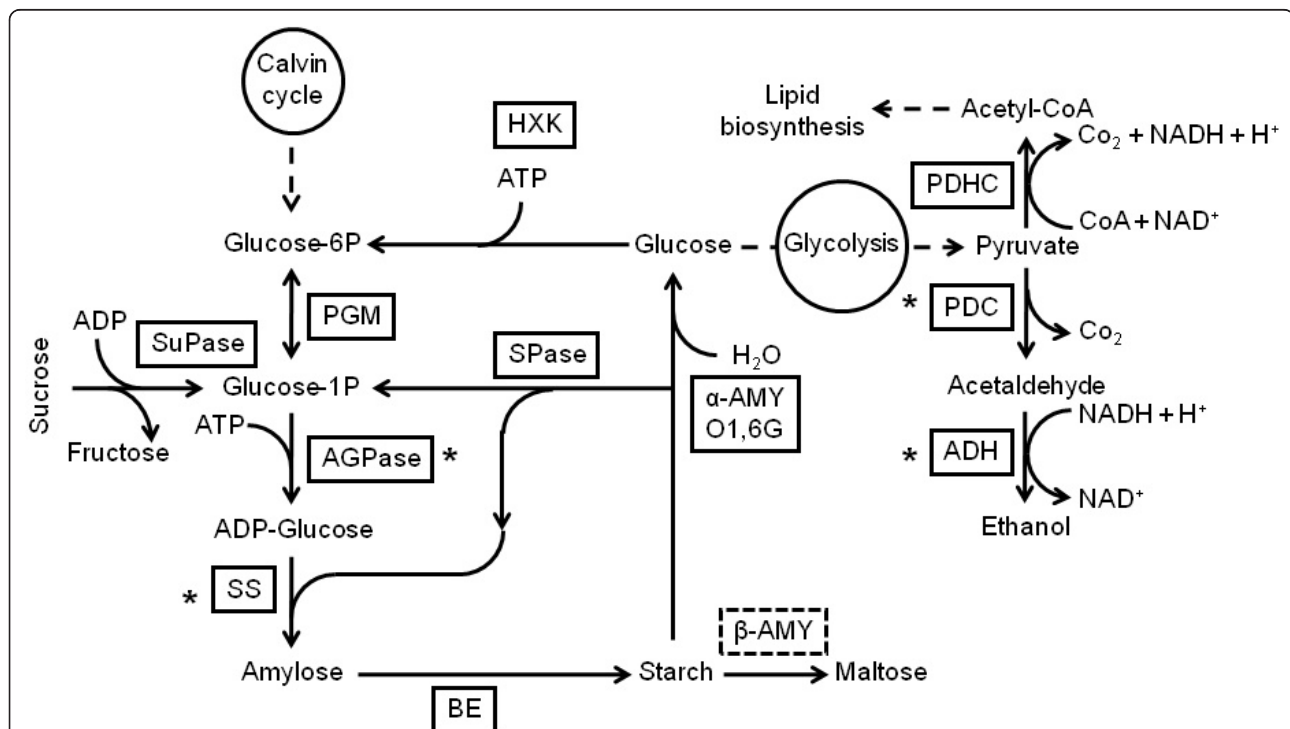


Figure 6 Starch biosynthesis and catabolism, and ethanol fermentation pathways reconstructed based on the *de novo* assembly and annotation of *D. tertiolecta* transcriptome. Identified and unidentified enzymes are shown in solid and dashed boxes, respectively. Enzymes involved in starch biosynthesis include: PGM, phosphoglucomutase (EC: 5.4.2.2); SuPase, sucrose phosphorylase (EC: 2.4.7.1); AGPase, ADP-glucose pyrophosphorylase (EC: 2.7.7.27); SS, starch synthase (EC: 2.4.1.21); BE, α -1,4-glucan branching enzyme (EC: 2.4.1.18); and HXK, hexokinase (2.7.1.1). Enzymes involved in starch catabolism include: α -AMY, α -amylase (EC: 3.2.1.1); O1, 6G, oligo-1,6-glucosidase (EC: 3.2.1.10); β -AMY, β -amylase (EC: 3.2.1.2); and SPase, starch phosphorylase (EC: 2.4.1.1). Enzymes involved in ethanol fermentation via pyruvate include: PDC, pyruvate decarboxylase (EC: 4.1.1.1); and ADH, alcohol dehydrogenase (EC: 1.1.1.1). Enzyme PDHC, Pyruvate dehydrogenase complex (EC 1.2.4.1, 2.3.1.12, 1.8.4.1), transforms pyruvate into acetyl-CoA which may then be used in the lipid biosynthesis pathway. Key enzymes are shown with an asterisk next to the boxes, and dashed arrows denote reaction(s) in which the enzymes are not shown.

synthesis pathway. We did not identify transcripts that code for β -amylase (β -AMY, EC: 3.2.1.2), which degrades starch into maltose. The phosphorytic degradation of starch in *D. tertiolecta* may involve starch phosphorylase (SPase, EC: 2.4.1.1), which mediates the transfer of glucose from the non-reducing end of an α -1,4-linked glucan to orthophosphate and generates G-1-P and a shorter glucan. Further investigations are warranted to determine the relative importance of these pathways in *D. tertiolecta*.

Pathways interactions, carbon partitioning and source-sink relationships

The metabolic pathways associated with biosynthesis and degradation of energy-rich molecules are closely linked. Starch catabolism provides the metabolites for biosynthesis of other energy rich products. Our KEGG pathway assignments revealed that *D. tertiolecta* has the genetic potential to link starch metabolism to ethanol fermentation through the glycolysis pathway (Figure 6) (Also see Additional files 5 and 6 for the pathway map and the complete set of identified enzymes involved in glycolysis, respectively). We identified transcripts coding for enzymes that catalyze the synthesis of ethanol from the intermediate metabolite, pyruvate in *D. tertiolecta*. These enzymes include pyruvate decarboxylase (PDC, EC: 4.1.1.1), which generates acetaldehyde and CO₂ from pyruvate, and alcohol dehydrogenase (ADH, EC: 1.1.1.1), which uses acetaldehyde and NADH + H⁺ to generate ethanol. Although ethanol production has been previously observed in marine microalgae [66], no reports exist in *D. tertiolecta* and existence of ethanol fermentation pathway raises the potential that this organism could be engineered to be an efficient converter of solar energy into ethanol.

Additionally, biosynthesis of starch can direct the flow of metabolites away from lipid biosynthesis and conversely starch degradation provides the metabolites for production of energy rich molecules (i.e, lipids, and ethanol). We identified 20 transcripts that code for a pyruvate dehydrogenase complex (PDHC) (EC: 1.2.4.1, 2.3.1.12, 1.8.1.4) that transforms pyruvate into acetyl-CoA through pyruvate decarboxylation. Acetyl-CoA may then be used in the fatty acid synthesis pathway. A blockage of starch synthesis has been shown to increase the accumulation of lipids in several starchless mutants of microalgae [30,67]. Disruption of genes related to starch degradation or over expression of genes involved in starch synthesis have successfully resulted in increased starch content in microalgae and *Arabidopsis thaliana* [45,65].

Concerted production and accumulation of energy rich molecule in microalgae also depends upon the integration of precursor supplying pathways (i.e. sources)

with synthesizing machineries (i.e. sinks). The accumulation of large quantities of lipids in microalgae requires a continuous supply of acetyl-CoA and NADPH. The pathways supplying these precursors lie outside of the fatty acid synthetic machinery, and it has been suggested that they are unique to oleaginous microorganisms [68]. The key supplier of acetyl-CoA for fatty acid synthesis in oleaginous microorganisms is considered to be ATP: citrate lyase (ACL, EC: 2.3.3.8), which catalyzes the formation of acetyl-CoA and oxaloacetate by cleaving citrate using an ATP molecule [68]. The formation of NADPH as an essential reductant for fatty acid synthesis has been mainly attributed to malate dehydrogenase (MDH, EC: 1.1.1.40), which uses malate and NADP⁺ to generate pyruvate, CO₂ and NADPH [68]. Citrate and malate are intermediates of tricarboxylic acid (TCA) cycle and pyruvate metabolism, respectively. Interestingly, we identified numerous sequences, 76 and 11, in our transcriptome library coding for ACL and MDH, respectively. The integration of these enzymes with fatty acid biosynthesizing machinery ensures the direct flow of acetyl-CoA into fatty acids, which are then used as precursors of TAG synthesis. Genetic manipulations that increase the availability of precursors for fatty acid and starch synthesis, through up-regulation/over-expression of related genes identified here, could be promising approaches to increase the yield of biofuel precursors in microalgae.

Conclusions

This study presents the first next-generation sequencing effort and transcriptome annotation of a non-model marine microalgae that is relevant for biofuel production. Genes encoding key enzymes have been successfully identified and metabolic pathways involved in biosynthesis and catabolism of fatty acids, TAG, and starch in *D. tertiolecta* have been reconstructed. Identification of these genes and pathways is in agreement with the empirically observed capability of *D. tertiolecta* to synthesize and accumulate energy rich molecules, and adds to the current knowledge on the molecular biology and biochemistry of their production in microalgae. By providing insight into the mechanisms underpinning these metabolic processes, results can be used to direct efforts to genetically manipulate this organism to enhance the production of feedstock for commercial microalgae-biofuels.

The accumulation of biofuel precursors and discovery of genes associated with their biosynthesis and metabolism in *D. tertiolecta* is intriguing and worthy of further investigation. The sequences and pathways produced here present the genetic framework required for further studies. Quantitative transcriptomics in concert with physiological and biochemical analysis in *D. tertiolecta*

under conditions that stimulate production and accumulation of biofuel precursors are needed to provide insight into the ways these pathways are regulated and linked.

Methods

D. tertiolecta culturing, harvesting and RNA extraction

D. tertiolecta (UTEX LB 999) was obtained from the Culture Collection of Algae at the University of Texas. Cells were cultured in 1 L flasks filled with 750 ml of Erdschreiber's medium (UTEX), modified to have different concentrations of nitrogen and salinity. Reactors were operated at room temperature in batch mode and exposed to fluorescent light (32 Watts, Cool White) at a photosynthetic photon flux density (PPFD) of 135 $\mu\text{mol-photon m}^{-2} \text{ s}^{-1}$, with a 14/10 h light/dark cycle. Gas flow rate was 200 ml min^{-1} of house air and controlled using a mass flow controller (Cole-Parmer Instrument Company, IL, USA). The air stream was passed over activated carbon and filtered through a 0.2 μm filter before being flushed into the reactors. All cultures were mixed by an orbital shaker at 200 rpm.

Cells were cultured under various growth conditions and at different phases of the growth cycle. These various growth conditions and phases were chosen to stimulate production and accumulation of lipids or starch, to induce expression of genes involved with lipids or starch biosynthesis, and to maximize the diversity of expressed genes [23,28-30,69]. The growth conditions included: nitrogen limited cultures (10 mg L^{-1} , N) with salinity levels similar to that of seawater (i.e. 0.5 M NaCl) harvested during stationary phase (sample A), nitrogen sufficient cultures (100 mg L^{-1} , N) with salinity of 0.5 M NaCl, harvested during exponential growth (sample B), and stationary (samples C) phases, and salt-stressed cultures (100 mg L^{-1} , N) with elevated salt concentration (1.5 M NaCl) harvested during the stationary phase (sample D). Cell growth was monitored in duplicate reactors under each condition by measuring the changes of optical density in the culture medium at 730 nm (OD_{730}), using a spectrophotometer (HP 8453, Hewlett Packard, CA, USA). Cells were harvested by centrifugation (RC-6 Plus, ThermoScientific, DE, USA), at 17,000 g , for 5 min, at 4°C. The supernatant was discarded and cell pellets were immediately frozen in liquid nitrogen and stored at -80°C until further analysis.

Total RNA from cells was extracted and purified using RNeasy Plant Mini Kit (Qiagen, CA, USA) with the following modifications for cell lysis. Cell pellets were re-suspended in buffer RLT (1×10^5 cells μl^{-1}), transferred to a 2-ml screw cap tube containing 300 mg of glass beads (0.5 mm, baked at 500°C for 4 h), and lysed and homogenized by agitation in a bead-beater (Mini-Bead-beater-16, BioSpec Products, OK, USA) at the maximum

speed (3450 oscillations/min) for 40 s. Bead-beating was repeated four times for each sample. The residual genomic DNA contamination was removed during the RNA cleanup using the optional on-column DNase I digestion as instructed by the manufacturer (Qiagen, CA, USA). The integrity of the purified total RNA was assessed using formaldehyde agarose gel electrophoresis, and RNA quantity was determined by NanoDrop spectrophotometer measurements (ThermoScientific, DE, USA). Because microalgae have high levels of pigments, polysaccharides, and glycolproteins that could interfere with cDNA synthesis and are difficult to remove using spin-column purification methods, total RNA was further purified using lithium-chloride precipitation as previously described [21].

Synthesis of cDNA and library construction

Synthesis of full-length double-stranded cDNA (ds-cDNA) from total RNA was performed using SMARTer PCR cDNA Synthesis Kit (Clontech, CA, USA) according to the manufacturer's instructions, with the exception of using a modified CDS primer (5'-AAGCAGTG GTATCAACGCAGAGTACGTGCAGTT TTTTTTTT TTTTTTVN-3'). This modified primer included a recognition site (GTGCAG), on the 5' end of the poly T tail for restriction enzyme *BsgI*. This restriction site was then used to eliminate the presence of poly (A:T) tail in the cDNA samples. These homopolymers could result in too strong of a light signal and thus produce sequencing reads of low quality when using the Genome Sequencer FLX with Titanium reagents.

Full-length cDNA templates were then amplified by long-distance PCR using the Advantage 2 PCR Kit (Clontech, CA, USA). To ensure that the PCR products were not over amplified, the optimal number of PCR cycles was determined according to manufacturer's guidelines. The PCR reactions were then chased [70] to maximize the quantities of fully double-stranded cDNA products and quality was verified by agarose gel electrophoresis. Replicate PCR reactions were performed for each library then pooled and purified using the QIAquick PCR Purification kit (Qiagen, CA, USA). The amplified cDNA libraries were quantified using NanoDrop (ThermoScientific, DE, USA), and equal amounts of PCR products from sample libraries B, C, and D were pooled to construct a new library "P" which was used along with library A for normalization.

To enhance gene discovery, the proportion of transcripts (i.e. expressed genes) that are highly abundant in the sample were reduced before sequencing. Equal amount of samples (1.6 μg) from cDNA libraries A and P were normalized using the Trimmer cDNA Normalization Kit (Evrogen, Moscow, Russia), and re-amplified using the 5' PCR primer II A (SMARTer PCR cDNA

Synthesis Kit), according to the manufacturer's instructions. The normalized cDNA samples were then purified as described above (QIAquick PCR Purification Kit), and digested using the *BspI* enzyme (New England Biolabs, MA, USA) to remove the residual poly (A:T) tails. Following digestion, an aliquot of restriction digest solution was evaluated on 10% TBE polyacrylamide gels (Invitrogen, CA, USA), to verify that the appropriate fragment size (49 bp) had been cleaved. Finally, digested, normalized cDNA libraries were purified (QIAquick PCR Purification Kit), and the quality of final samples was verified using agarose gel electrophoresis.

454 sequencing

Sequencing of cDNA samples was performed by Roche-454 Life Sciences (Branford, CT, USA) using the Genome Sequencer FLX with Titanium Chemistry. Each of the A and P samples was sequenced on the half of a PicoTiter Plate according to the manufacturer's instructions (Roche, IN, USA). All sequencing reads were deposited into the Short Read Archive (SRA) of the National Center for Biotechnology Information (NCBI), and can be accessed under the accession number SRA023642.

Sequence analysis, and assembly

Sequencing data obtained from samples A and P were pooled and subjected to a publicly available sequence cleaning and validation software, SeqClean [71], to account for size-selection, overall low complexity analysis, and to remove poly (A:T) regions, and adapters. In addition, a comprehensive ribosomal RNA database, Silva [72], containing regularly updated, high quality sequences of eukaryotic rRNAs were incorporated into the cleaning pipeline of the SeqClean to remove ribosomal RNA sequences. Following the sequence trimming and size selection (>100 bp), the reads were assembled using the Newbler v2.3 provided by Roche-454 Life Sciences (Branford, CT). Assembly parameters were used as default values with minimum quality score of 20, and overlap identity as 90% over 40 bp length to detect pairwise alignments. Assembly computations were duplicated to make sure that the results were reproducible.

Unique sequence mapping, functional annotation, and pathway assignments

Following the assembly, unique transcripts (isotigs) and singletons were compared to NCBI's non-redundant (nr) database using BLASTx algorithm [73], with a cut-off E-value of $\leq 10^{-6}$. Resulting top 10 blast hits were fed into publicly available Blast2GO software (v.2.4.4) [33] in order to retrieve associated gene ontology (GO) terms describing biological processes, molecular functions, and

cellular components [74]. By using specific gene identifiers and accession numbers, Blast2GO produces all GO annotations as well as corresponding enzyme commission numbers (EC) for sequences with an E-value equal to or less than 10^{-6} . To compare the enzyme-coding genes identified in this work with those identified in the model microalgae with sequenced genome, we used BLASTx algorithm with an E-value threshold of 10^{-6} to align the transcript sequences annotated as enzymes related to the production of biofuel precursors against the sequences of associated enzymes in *Volvox carteri*, and *Chlamydomonas reinhardtii*.

To determine metabolic pathways, Kyoto Encyclopedia of Genes and Genomes (KEGG) mapping was used [35]. The sequences with corresponding ECs obtained from Blast2GO were mapped to the KEGG metabolic pathway database. To further enrich the pathway annotation and to identify the BRITE functional hierarchies, sequences were also submitted to the KEGG Automatic Annotation Server (KAAS) [34], and the single-directional best hit information method was selected. KAAS annotates every submitted sequence with KEGG orthology (KO) identifiers, which represents an ortholog group of genes directly linked to an object in the KEGG pathways and BRITE functional hierarchy [34,75], and thus incorporates different types of relationships that exist in biological systems (i.e. genetic and environmental information processing, cellular processes, and organismal systems). The graphical KEGG Markup Language pathway editor (KGML-ED) was used to draw the fatty acid catabolism and glycolysis pathways [76]. Computationally processed assembly outputs and annotations are hosted at the corresponding author's website http://www.eng.yale.edu/peccialab/microalgae_sequences.html for public access.

Additional material

Additional file 1: Comparison of *D. tertiolecta* transcriptome assembly outputs obtained using Newbler v2.3 and 2.5.

Additional file 2: A phylogenetic tree inferring the evolutionary relationship between *D. tertiolecta* and model microalgae *Volvox carteri*, and *Chlamydomonas reinhardtii*. The tree was generated using 18 S rRNA gene sequences of *D. tertiolecta*, *Volvox carteri*, and *Chlamydomonas reinhardtii* extracted from the NCBI database. Sequences were aligned using ClustalX and bootstrapping was performed in ClustalX with 100 iterations and values were displayed on the branch edges. The tree was visualized and published in Mega5. The distance bar represents 0.01 base changes/base. The tree was rooted with *Methanobacterium congolense* (NCBI Accession Number: AF233586.1).

Additional file 3: Candidate genes identified based on KEGG orthology (KO) annotation of the *D. tertiolecta* transcriptome.

Additional file 4: β -oxidation pathway of *D. tertiolecta* based on the annotation of transcriptome and KEGG pathway assignment.

Additional file 5: Glycolysis pathway of *D. tertiolecta* based on the annotation of transcriptome and KEGG pathway assignment.

Additional file 6: Enzymes involved in glycolysis identified by annotation of the *D. tertiolecta* transcriptome.

Acknowledgements

We thank two anonymous reviewers for their insightful suggestions on the manuscript. This research was supported by the Connecticut Center for Advanced Technologies under a Fuel Diversification Grant and by the National Science Foundation, Grant #0854322, BZH was supported by a joint postdoctoral fellowship from the Yale Climate and Energy Institute and Yale Institute for Biospheric Studies. We acknowledge the Yale University Biomedical High Performance Computing Center and the NIH Grant# RR19895, which funded the instrumentation.

Author details

¹Department of Chemical and Environmental Engineering, Yale University, New Haven, CT 06511, USA. ²Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

Authors' contributions

HRY and JP conceived and designed the experiments. HRY conducted the reactor-based experiments and prepared the cDNA libraries. HRY and BZH performed the bioinformatic analysis. KB was instrumental in bioinformatic analysis. HRY drafted the manuscript, constructed the tables and figures, and BZH and JP contributed to the final version. All authors have read and approved the final manuscript.

Received: 2 November 2010 Accepted: 14 March 2011

Published: 14 March 2011

References

1. Peaking of world oil production: Impacts, mitigation, and risk management. [http://www.netl.doe.gov].
2. Chisti Y: **Biodiesel from microalgae**. *Biotechnology Advances* 2007, **25**:294-306.
3. Kurano N, Sasaki T, Miyachi S: **Carbon dioxide and microalgae**. In *Advances In Chemical Conversions For Mitigating Carbon Dioxide*. Volume 114. Amsterdam: Elsevier Science Publ B V; 1998:55-63.
4. NREL: **A look back at the US Department of Energy's aquatic species program: biodiesel from algae**, report NREL/TP-580-24190. National Renewable Energy Labs; 1998.
5. Yun Y-S, Lee SB, Park JM, Lee C-I, Yang J-W: **Carbon dioxide fixation by algal cultivation using wastewater nutrients**. *Journal of Chemical Technology and Biotechnology* 1997, **69**:451-455.
6. Dinh LTT, Guo Y, Mannan S: **Sustainability evaluation of biodiesel production using multicriteria decision-making**. *Environmental Progress and Sustainable Energy* 2009, **28**:38-46.
7. Radakovits R, Jinkerson RE, Darzins A, Posewitz MC: **Genetic engineering of algae for enhanced biofuel production**. *Eukaryotic Cell* 2010, **9**:486-501.
8. Benamotz A, Tornabene TG, Thomas WH: **Chemical profile of selected species of microalgae with emphasis on lipids**. *Journal of Phycology* 1985, **21**:72-81.
9. Hoshida H, Ohira T, Minematsu A, Akada R, Nishizawa Y: **Accumulation of eicosapentaenoic acid in *Nannochloropsis* sp in response to elevated CO₂ concentrations**. *Journal of Applied Phycology* 2005, **17**:29-34.
10. Li YQ, Horsman M, Wang B, Wu N, Lan CQ: **Effects of nitrogen sources on cell growth and lipid accumulation of green alga *Neochloris oleoabundans***. *Applied Microbiology and Biotechnology* 2008, **81**:629-636.
11. Livne A, Sukenik A: **Lipid-synthesis and abundance of acetyl CoA carboxylase in *Isochrysis galbana* (Prymnesiophyceae) following nitrogen starvation**. *Plant and Cell Physiology* 1992, **33**:1175-1181.
12. Rodolfi L, Zittelli GC, Bassi N, Padovani G, Biondi N, Bonini G, Tredici MR: **Microalgae for oil: Strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor**. *Biotechnology and Bioengineering* 2009, **102**:100-112.
13. Sriharan S, Bagga D, Nawaz M: **The effects of nutrients and temperature on biomass, growth, lipid production, and fatty-acid composition of *Cyclotella cryptica* Reimann, Lewin, And Guillard**. *Applied Biochemistry and Biotechnology* 1991, **28**:9:317-326.
14. Tsuzuki M, Ohnuma E, Sato N, Takaku T, Kawaguchi A: **Effects Of CO₂ concentration during growth on fatty-acid composition in microalgae**. *Plant Physiology* 1990, **93**:851-856.
15. Miller R, Wu GX, Deshpande RR, Vieler A, Gartner K, Li XB, Moellering ER, Zauner S, Cornish AJ, Liu BS, et al: **Changes in Transcript Abundance in *Chlamydomonas reinhardtii* following Nitrogen Deprivation Predict Diversion of Metabolism**. *Plant Physiology* 2010, **154**:1737-1752.
16. Moellering ER, Benning C: **RNA Interference Silencing of a Major Lipid Droplet Protein Affects Lipid Droplet Size in *Chlamydomonas reinhardtii***. *Eukaryotic Cell* 2010, **9**:97-106.
17. Hu Q, Sommerfeld M, Jarvis E, Ghriradi M, Posewitz M, Seibert M, Darzins A: **Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances**. *Plant Journal* 2008, **54**:621-639.
18. Cheung F, Win J, Lang JM, Hamilton J, Vuong H, Leach JE, Kamoun S, Levesque CA, Tisserat N, Buell CR: **Analysis of the *Pythium ultimum* transcriptome using Sanger and Pyrosequencing approaches**. *BMC Genomics* 2008, **9**(1):542.
19. Parchman TL, Geist KS, Grahn JA, Benkman GW, Buerkle CA: **Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery**. *BMC Genomics* 2010, **11**:180.
20. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing**. *Molecular Ecology* 2008, **17**:1636-1647.
21. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx**. *Bmc Genomics* 2009, **10**(219):1-18.
22. Takagi M, Yoshida T: **Effect of salt concentration on intracellular accumulation of lipids and triacylglyceride in marine microalgae *Dunaliella* cells**. *Journal of Bioscience and Bioengineering* 2006, **101**:223-226.
23. Goyal A: **Osmoregulation in *Dunaliella*, Part II: Photosynthesis and starch contribute carbon for glycerol synthesis during a salt stress in *Dunaliella tertiolecta***. *Plant Physiology and Biochemistry* 2007, **45**:705-710.
24. Goyal A, Brown AD, Gimmler H: **Regulation of salt-induced starch degradation in *Dunaliella tertiolecta***. *Journal of Plant Physiology* 1987, **127**:77-96.
25. Ben-Amotz A, Polle JEW, Rao DVS: **The alga *Dunaliella***. Enfield, NH: Science Publishers; 2009.
26. Hosseini Tafreshi A, Shariati M: **Dunaliella biotechnology: methods and applications**. *Journal of Applied Microbiology* 2009, **107**:14-35.
27. Barzegari A, Hejazi MA, Hosseinzadeh N, Eslami S, Aghdam EM, Hejazi MS: **Dunaliella as an attractive candidate for molecular farming**. *Molecular Biology Reports* 2010, **37**:3427-3430.
28. Hirokawa T, Hata M, Takeda H: **Correlation between the starch level and the rate of starch synthesis during the developmental cycle of *Chlorella ellipsoidea***. *Plant and Cell Physiology* 1982, **23**:813-820.
29. Ramazanov A, Ramazanov Z: **Isolation and characterization of a starchless mutant of *Chlorella pyrenoidosa* STL-PI with a high growth rate, and high protein and polyunsaturated fatty acid content**. *Phycological Research* 2006, **54**:255-259.
30. Wang ZT, Ullrich N, Joo S, Waffenschmidt S, Goodenough U: **Algal lipid bodies: stress induction, purification, and biochemical characterization in wild-type and starchless *Chlamydomonas reinhardtii***. *Eukaryotic Cell* 2009, **8**:1856-1868.
31. Wang H, Zhang HM, Wong YH, Voolstra C, Ravasi T, Bajic VB, Quan PY: **Rapid transcriptome and proteome profiling of a non-model marine invertebrate, *Bugula neritina***. *Proteomics* 2010, **10**:2972-2981.
32. Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS: **De novo characterization of a whitefly transcriptome and analysis of its gene expression during development**. *BMC Genomics* 2010, **11**.
33. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite**. *Nucleic Acids Research* 2008, **36**:3420-3435.
34. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M: **KAAS: an automatic genome annotation and pathway reconstruction server**. *Nucleic Acids Research* 2007, **35**:W182-W185.
35. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M: **KEGG: Kyoto Encyclopedia of Genes and Genomes**. *Nucleic Acids Research* 1999, **27**:29-34.
36. Scott SA, Davey MP, Dennis JS, Horst I, Howe CJ, Lea-Smith DJ, Smith AG: **Biodiesel from algae: challenges and prospects**. *Current Opinion in Biotechnology* 2010, **21**:277-286.
37. Griffiths G, Stobart A: **Fatty acid biosynthesis and triacylglycerol assembly**. *Handbook of Food Enzymology* CRC Press; 2002.

38. Yap CY, Chen F: **Polyunsaturated fatty acids: Biological significance, biosynthesis, and production by microalgae and microalgae-like organisms.** In *Algae and their Biotechnological Potential*. Edited by: Chen F, Jiang Y. Dordrecht, Netherlands: Kluwer Academic Publishers; 2000:1-32.
39. Evans RW, Kates M, Ginzburg M, Ginzburg BZ: **Lipid-composition of halotolerant algae, *Dunaliella parva lerche* and *Dunaliella tertiolecta*.** *Biochimica et Biophysica Acta* 1982, **712**:186-195.
40. Volkman JK, Jeffrey SW, Nichols PD, Rogers GI, Garland CD: **Fatty acid and lipid composition of 10 species of microalgae used in mariculture.** *Journal of Experimental Marine Biology and Ecology* 1989, **128**:219-240.
41. Ohlrogge J, Browse J: **Lipid biosynthesis.** *Plant Cell* 1995, **7**:957-970.
42. Jaworski JG, Clough RC, Barnum SR: **A cerulenin insensitive short chain 3-ketoacyl-acyl carrier protein synthase in *Spinacia oleracea* leaves.** *Plant Physiology* 1989, **90**:41-44.
43. Gurr MI, Harwood JL: **Lipid biochemistry.** Oxford, UK: Blackwell Science Ltd; 2002.
44. Schultz DJ, Ohlrogge JB: **Metabolic engineering of fatty acid biosynthesis.** In *Lipid Biotechnology*. Edited by: Gardner HW, Kuo TM. New York: Marcel Dekker AG; 2002.
45. Verwoert I, Vanderlinden KH, Walsh MC, Nijkamp HJJ, Stuitje AR: **Modification of *Brassica napus* seed oil by expression of the *Escherichia coli* FabH gene, encoding 3-ketoacyl-acyl carrier protein synthase-III.** *Plant Molecular Biology* 1995, **27**:875-886.
46. Radakovits R, Jinkerson RE, Darzins A, Posewitz MC: **Genetic engineering of algae for enhanced biofuel production.** *Eukaryotic Cell* 2010, **9**(4):486-501.
47. Khozin-Goldberg I, Cohen Z: **Unraveling algal lipid metabolism: Recent advances in gene identification.** *Biochimie*.
48. Courchesne NMD, Parisien A, Wang B, Lan CQ: **Enhancement of lipid production using biochemical, genetic and transcription factor engineering approaches.** *Journal of Biotechnology* 2009, **141**:31-41.
49. Coleman RA, Lee DP: **Enzymes of triacylglycerol synthesis and their regulation.** *Progress in Lipid Research* 2004, **43**:134-176.
50. Dahlqvist A, Stahl U, Lenman M, Banas A, Lee M, Sandager L, Ronne H, Szymne H: **Phospholipid: diacylglycerol acyltransferase: An enzyme that catalyzes the acyl-CoA-independent formation of triacylglycerol in yeast and plants.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**:6487-6492.
51. Padham AK, Hopkins MT, Wang TW, McNamara LM, Lo M, Richardson LGL, Smith MD, Taylor CA, Thompson JE: **Characterization of a plastid triacylglycerol lipase from *Arabidopsis*.** *Plant Physiology* 2007, **143**:1372-1384.
52. Wendel AA, Lewin TM, Coleman RA: **Glycerol-3-phosphate acyltransferases: Rate limiting enzymes of triacylglycerol biosynthesis.** *Biochimica Et Biophysica Acta-Molecular and Cell Biology of Lipids* 2009, **1791**:501-506.
53. Kroon JTM, Wei W, Simon WJ, Slabas AR: **Identification and functional expression of a type 2 acyl-CoA:diacylglycerol acyltransferase (DGAT2) in developing castor bean seeds which has high homology to the major triglyceride biosynthetic enzyme of fungi and animals.** *Phytochemistry* 2006, **67**:2541-2549.
54. Zou JT, Katavic V, Giblin EM, Barton DL, MacKenzie SL, Keller WA, Hu X, Taylor DC: **Modification of seed oil content and acyl composition in the brassicaceae by expression of a yeast sn-2 acyltransferase gene.** *Plant Cell* 1997, **9**:909-923.
55. Stahl U, Carlsson AS, Lenman M, Dahlqvist A, Huang BQ, Banas W, Banas A, Szymne S: **Cloning and functional characterization of a Phospholipid: Diacylglycerol acyltransferase from *Arabidopsis*.** *Plant Physiology* 2004, **135**:1324-1335.
56. Ball SG, Morell MK: **From bacterial glycogen to starch: Understanding the biogenesis of the plant starch granule.** *Annual Review of Plant Biology* 2003, **54**:207-233.
57. Choi SP, Nguyen MT, Sim SJ: **Enzymatic pretreatment of *Chlamydomonas reinhardtii* biomass for ethanol production.** *Bioresource Technology* 2010, **101**:5330-5336.
58. Sanchez OJ, Cardona CA: **Trends in biotechnological production of fuel ethanol from different feedstocks.** *Bioresource Technology* 2008, **99**:5270-5295.
59. Lin CY, Chang CC, Hung CH: **Fermentative hydrogen production from starch using natural mixed cultures.** *International Journal of Hydrogen Energy* 2008, **33**:2445-2453.
60. Argun H, Kargi F: **Bio-hydrogen production from ground wheat starch by continuous combined fermentation using annular-hybrid bioreactor.** *International Journal of Hydrogen Energy* 2010, **35**:6170-6178.
61. Searchinger T, Heimlich R, Houghton RA, Dong FX, Elobeid A, Fabiosa J, Tokgoz S, Hayes D, Yu TH: **Use of US croplands for biofuels increases greenhouse gases through emissions from land-use change.** *Science* 2008, **319**:1238-1240.
62. Ball SG: **Regulation of starch biosynthesis.** *Advances in Photosynthesis* 1998, **7**:549-567.
63. Streb S, Egli B, Eicke S, Zeeman SC: **The debate on the pathway of starch synthesis: A closer look at low-starch mutants lacking plastidial phosphoglucomutase supports the chloroplast-localized pathway.** *Plant Physiology* 2009, **151**:1769-1772.
64. Kombrink E, Wober G: **Identification and subcellular-localization of starch-metabolizing enzymes in the green-alga *Dunaliella marina*.** *Planta* 1980, **149**:130-137.
65. Deschamps P, Haferkamp I, d'Hulst C, Neuhaus HE, Ball SG: **The relocation of starch metabolism to chloroplasts: when, why and how.** *Trends in Plant Science* 2008, **13**:574-582.
66. Smith AM: **Prospects for increasing starch and sucrose yields for bioethanol production.** *Plant Journal* 2008, **54**:546-558.
67. Hirano A, Ueda R, Hirayama S, Ogushi Y: **CO₂ fixation and ethanol production with microalgal photosynthesis and intracellular anaerobic fermentation.** *Energy* 1997, **22**:137-142.
68. Work VH, Radakovits R, Jinkerson RE, Meuser JE, Elliott LG, Vinyard DJ, Laurens LML, Dismukes GC, Posewitz MC: **Increased lipid accumulation in the *Chlamydomonas reinhardtii* sta7-10 starchless isoamylase mutant and increased carbohydrate synthesis in complemented strains.** *Eukaryotic Cell* 2010, **9**:1251-1261.
69. Ratledge C: **Fatty acid biosynthesis in microorganisms being used for single cell oil production.** *Biochimie* 2004, **86**:807-815.
70. Chen H, Jiang JG: **Osmotic responses of *Dunaliella* to the changes of salinity.** *Journal of Cellular Physiology* 2009, **219**:251-258.
71. SeqClean sequence trimming and validation tool. [<http://sourceforge.net/projects/seqclean/>].
72. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, Glockner FO: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.** *Nucleic Acids Research* 2007, **35**:7188-7196.
73. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215**:403-410.
74. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, *et al*: **Gene Ontology: tool for the unification of biology.** *Nature Genetics* 2000, **25**:25-29.
75. Mao XZ, Cai T, Olyarchuk JG, Wei LP: **Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary.** *Bioinformatics* 2005, **21**:3787-3793.
76. Klukas C, Schreiber F: **Dynamic exploration and editing of KEGG pathway diagrams.** *Bioinformatics* 2007, **23**:344-350.
77. Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, Extavour CG: **The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*.** *BMC Genomics* 2011, **12**:61.

doi:10.1186/1471-2164-12-148

Cite this article as: Rismani-Yazdi *et al*: 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.