RESEARCH ARTICLE

dEMBF: A Comprehensive Database of Enzymes of Microalgal Biofuel Feedstock

Namrata Misra[1,2], Prasanna Kumar Panda[1,2]*, Bikram Kumar Parida[2], Barada Kanta Mishra[1]

1 Academy of Scientific and Innovative Research, CSIR-Institute of Minerals and Materials Technology, Bhubaneswar, Odisha, India, 2 Bioresources Engineering Department, CSIR-Institute of Minerals and Materials Technology, Bhubaneswar, Odisha, India

* pkpanda@immt.res.in; pkrpan@gmail.com

Abstract

Microalgae have attracted wide attention as one of the most versatile renewable feedstocks for production of biofuel. To develop genetically engineered high lipid yielding algal strains, a thorough understanding of the lipid biosynthetic pathway and the underpinning enzymes is essential. In this work, we have systematically mined the genomes of fifteen diverse algal species belonging to Chlorophyta, Heterokontophyta, Rhodophyta, and Haptophyta, to identify and annotate the putative enzymes of lipid metabolic pathway. Consequently, we have also developed a database, dEMBF (Database of Enzymes of Microalgal Biofuel Feedstock), which catalogues the complete list of identified enzymes along with their computed annotation details including length, hydrophobicity, amino acid composition, subcellular location, gene ontology, KEGG pathway, orthologous group, Pfam domain, intron-exon organization, transmembrane topology, and secondary/tertiary structural data. Furthermore, to facilitate functional and evolutionary study of these enzymes, a collection of built-in applications for BLAST search, motif identification, sequence and phylogenetic analysis have been seamlessly integrated into the database. dEMBF is the first database that brings together all enzymes responsible for lipid synthesis from available algal genomes, and provides an integrative platform for enzyme inquiry and analysis. This database will be extremely useful for algal biofuel research. It can be accessed at http://bbprof.immt.res.in/embf.

Introduction

With the irreversible depletion of petroleum resources, renewable biofuels are sustainable alternative to meet the global energy needs. Microalgae as a rich source of lipid, especially triacylglycerols (TAGs) have emerged as a potential biofuel feedstock due to several distinct advantages over other starch-based or lignocellulosic plant species, such as higher photosynthetic efficiency and higher biomass production rate. Besides, microalgae can be grown on non-arable land using wastewater, thus not competing with agri-resources and mitigating CO_2 emissions efficiently [1, 2]. However, to make biofuel production from microalgae a cost-competitive

process, the oil content in oleaginous algae needs to be significantly improved through genetic engineering techniques [3, 4]. It has been proposed that lipid biosynthesis can be increased by over expressing the rate-limiting enzymes of fatty acid biosynthesis pathway, of which acetyl-CoA carboxylase (ACCase) that catalyzes the first committed step of fatty acid synthesis viz., conversion of acetyl CoA to malonyl CoA plays a pivotal role [5, 6]. In addition, overexpression of the acyltransferases enzymes catalyzing the main regulatory steps involved in TAG biosynthesis, widely known as the Kennedy pathway, have also been determined as a potential approach to boost oil accumulation. For instance, overexpression of a type 2 diacylglycerol acyltransferases (DGAT) enzyme in the diatom *Phaeodactylum tricornutum* resulted in 35% increase in TAG content [7]. In another study, co-overexpression of multiple genes of the Kennedy pathway including glycerol-3 phosphate acyltransferase (GPAT), lysophosphatidyl acyltransferase (LPAT), phosphatidic acid phosphatase (PAP), diacylglycerol acyltransferase, glycerol-3 phosphate dehydrogenase (GPDH) and phospholipid:diacylglycerol acyltransferase (PDAT) in *Chlorella minutissima* resulted in a two-fold increase of TAG content [8]. Introduction of diacylglycerol acyltransferase 2 gene from *Brassica napus* to *Chlamydomonas reinhardtii* has also resulted in enhanced lipid production [9]. Together these studies indicate that understanding the regulation of microalgal lipid metabolism is absolutely essential for developing engineered microalgae with enhanced lipid production capabilities. [10]. While algal sequence data from genome assembly projects is rapidly increasing, the generated annotation for predicted sequences are usually limited and includes only user-defined function prediction with no detailed pathway, structure or genome-context information [11]. This limits our understanding of the overall lipid biosynthetic pathway in microalgae [12]. On contrary, the genes and enzymes involved in plant lipid biosynthetic pathway have been characterized extensively [13], and a number of biomass-related enzyme databases are also available to promote the development of transgenic biofuel crops [14–17]. Considering the importance of microalgae biofuel, paucity of information on algal lipid biosynthesis and unavailability of dedicated databases on enzymes underpinning the process, the present study was performed to identify a total of 289 enzymes responsible for lipid accumulation in fifteen sequenced microalgal species by using available homologous sequences from the model plant species, *Arabidopsis thaliana*. Functional annotation of the putative enzymes has also been improved by employing several bioinformatic tools to study metabolic pathways, ontology, subcellular location, secondary and tertiary structure, biophysical properties, cellular processes and protein family information. Furthermore, the emanated data are made publicly accessible through an open-access web-based database, dEMBF (database of Enzymes of Microalgal Biofuel Feedstock, http://bbprof.immt.res.in/embf). dEMBF is the first integrative platform that provides a complete list of enzymes putatively involved in lipid biosynthesis in microalgae. This database will certainly provide a roadmap for experimental as well as computational studies leading to identification of orthologous lipid synthesis enzymes in newly sequenced algal species and facilitate further R&D research aimed at attaining a sustainable and cost-effective biofuel production from microalgae.

Materials and Methods

Data sources

We analyzed a total of fifteen algal genomes belonging to diverse phylogenetic groups, namely Chlorophyta (*Chlorella variabilis, Chlamydomonas reinhardtii, Volvox carteri, Ostreococcus lucimarinus, Ostreococcus tauri, Micromonas pusilla CCMP1545, Micromonas sp. RCC299*, and *Bathycoccus prasinos*), Heterokontophyta (*Thalassiosira pseudonana, Phaeodactylum tricornutum, Ectocarpus siliculosus, Aureococcus anophagefferens, Nannochloropsis gaditana*),

Rhodophyta (*Cyanidioschyzon merolae*) and Haptophyta (*Emiliania huxleyi*). Proteomic sequences were retrieved mainly from the Phytozome (http://www.phtoozome.net) or from dedicated genome project websites for individual targeted species.

Enzyme prediction

The genome databases were queried by both keywords and sequence similarity BLASTp [18] search (E-value < 1e-5) using sequences of enzymes that are known to be involved in neutral lipid synthesis in Arabidopsis (Table 1). Subsequently, the successful hits were mapped to Uni-Prot ID [19], Enzyme commission (EC) number, Cluster of Orthologous groups (KOG) using KOGnitor [20], OrthoMCL [21] and Gene Ontology (GO) terms [22] using AmiGO [23], to remove any false positives. In addition, Pfam [24] was also employed to ensure that each candidate sequence shared the domain of the enzyme family to which it belongs. Finally, a complete set of 316 enzymes was collected from the studied algal species including Arabidopsis, for further detailed analysis of functional annotations as discussed below.

Functional annotations

Physico-chemical properties. The total number of amino acids, molecular weight, isoelectric point (pI), percentage of acidic/ basic amino acids, aliphatic index as well as GRAVY index was calculated using the Expasy's ProtParam server [25]. Hydropathy plot was generated using the BioEdit [26] software.

Secondary structure prediction. To predict secondary structure (the percentage of residues in helices, extended strands, and random coils) and transmembrane helix domains, MIN-NOU [27] and TMHMM [28] were used, respectively.

Table 1. List of enzymes involved in lipid biosynthesis in Arabidopsis that were used as query sequences for BLASTp sequence similarity searches on algal genomes. All sequences were retrieved from the Arabidopsis acyl-lipid metabolism website (http://aralip.plantbiology.msu.edu) See Reference [13].

Enzyme description	Enzyme name	EC No	UniProt ID
Fatty acid biosynthesis			
Homomeric Acetyl-CoA carboxylase	Homomeric ACCase	6.4.1.2	F4I1L3
Acetyl-CoA carboxylase alpha-carboxyltransferase	alpha-CT	6.4.1.2	Q9LD43
Acetyl-CoA carboxylase beta-carboxyltransferase	beta-CT	6.4.1.2	P56765
Biotin carboxylase	BC	6.3.4.14	O04983
Biotin carboxyl carrier protein	BCCP	6.4.1.2	Q9LLC1, Q42533
Malonyl-CoA-ACP Malonyltransacylase	MCMT/MCAT/FabD	2.3.1.39	Q8RU07
beta-ketoacyl-ACP Synthase I	KAS I/FabF	2.3.1.179	P52410
beta-ketoacyl-ACP Synthase II	KAS II/FabF	2.3.1.179	Q9C9P4
beta-ketoacyl-ACP Synthase III	KAS III/FabH	2.3.1.180	P49243
3-ketoacyl-ACP Reductase	KAR/FabG	1.1.1.100	P33207
3-hydroxyacyl-ACP Dehydratase	HAD/FabZ/FabA	4.2.1.59	Q9LX13, Q9SIE3
Enoyl-ACP Reductase	ENR/FabI	1.3.1.9	Q9SLA8
Triacylglycerol (TAG) assembly			
NAD-dependent Glycerol-3-phosphate dehydrogenase	GPDH	1.1.1.8	Q9SCX9, Q949Q0
Glycerol-3-phosphate acyltransferase	GPAT	2.3.1.15	Q8GWG0, Q43307
Lysophosphatidyl acyltransferase	LPAT/LPAAT	2.3.1.51	Q8GXU8, Q8L4Y2, Q8LG50, Q9LHN4, Q9SYC8
Phosphatidate phosphatase	PAP/PP	3.1.3.4	Q9FMN2, Q9SF47
Diacylglycerol acyltransferase Type 1	DGAT1	2.3.1.20	Q9SLD2
Diacylglycerol acyltransferase Type 2	DGAT2	2.3.1.20	Q9ASU1

doi:10.1371/journal.pone.0146158.t001

Subcellular location. Several prediction tools such as TargetP [29], ChloroP [30], Predotar [31], and WolfPsort [32] were employed for determining subcellular location.

Gene structure analysis. The exon-intron organizations of genes encoding the enzymes were determined by GeneWise program [33] through comparison of predicted coding sequence with corresponding genomic sequence.

Homology modeling of 3D structures. As no crystal structures of the predicted enzymes for microalgal species was found in the Protein Data Bank [34], we tried to model their 3D structure using MaxMod program [35]. Templates were selected based on crystal structures having more than 30% sequence identity. Ramachandran plot of the developed models were generated using the Procheck [36] program.

Database architecture

The dEMBF database runs on an Apache server (v. 2.2.17), where PHP v 5.3.5 was used for server side scripting while Java Script, AJAX, XHTML and CSS were used for client side scripting. Data was stored in a relational format using MySQL v 5.0.7 as the backend database, following basic normalization rules in order to reduce data redundancy and increase database efficiency.

Results

Annotation details of enzymes in dEMBF

The annotation detail page of dEMBF (Fig 1) displays multiple sequence and structural properties of the enzymes that has either been extracted manually from public resources or has been computed using a plethora of bioinformatics tools as described in the Methods section. Each sequence is annotated with information like symbol, gene name, enzyme class, organism, taxonomic identifier which is linked to NCBI taxonomic browser and organism lineage. The general information section furnishes information on chromosomal location, subcellular location, reaction, KEGG pathway, KEGG ortholog (KO) and KOG details. Similarly, the page also contains other predicted protein features such as gene ontology, physico-chemical properties, schematic representation of conserved domain, secondary structure, transmembrane topology, modeled 3D structure, intron-exon organization, amino acid and nucleotide sequences in fasta format, and cross references to external protein and gene databases. Particularly, the modeled 3D structure that has been built using homology modeling protocol along with details of the template employed and target-template alignment will be useful for users to study the structural conformation of enzymes in detail. To facilitate dynamic visualization of developed protein 3D conformation, JSmol applet (http://www.jmol.org) has been integrated. Furthermore, pre-generated Ramachandran plots for each modeled structure can be viewed using the "View Ramachandran plot" option.

Web-interface of dEMBF

The dEMBF database comprises of six major web interfaces, namely "Home", "Browse", "Search", "Tools", "Organisms" and "Resources". A schematic overview of dEMBF architecture is shown in Fig 2.

Home. The home page (Fig 3) contains a brief introduction to dEMBF and a site map detailing the outline of the database. Various convenient utilities are also available in the homepage to view and retrieve data from dEMBF. For instance with the "Search database" option, users can search for enzyme by name, symbol, UniProt Id, gene name, enzyme class, EC number, or organism name, directly from the homepage of database using auto complete text fields.
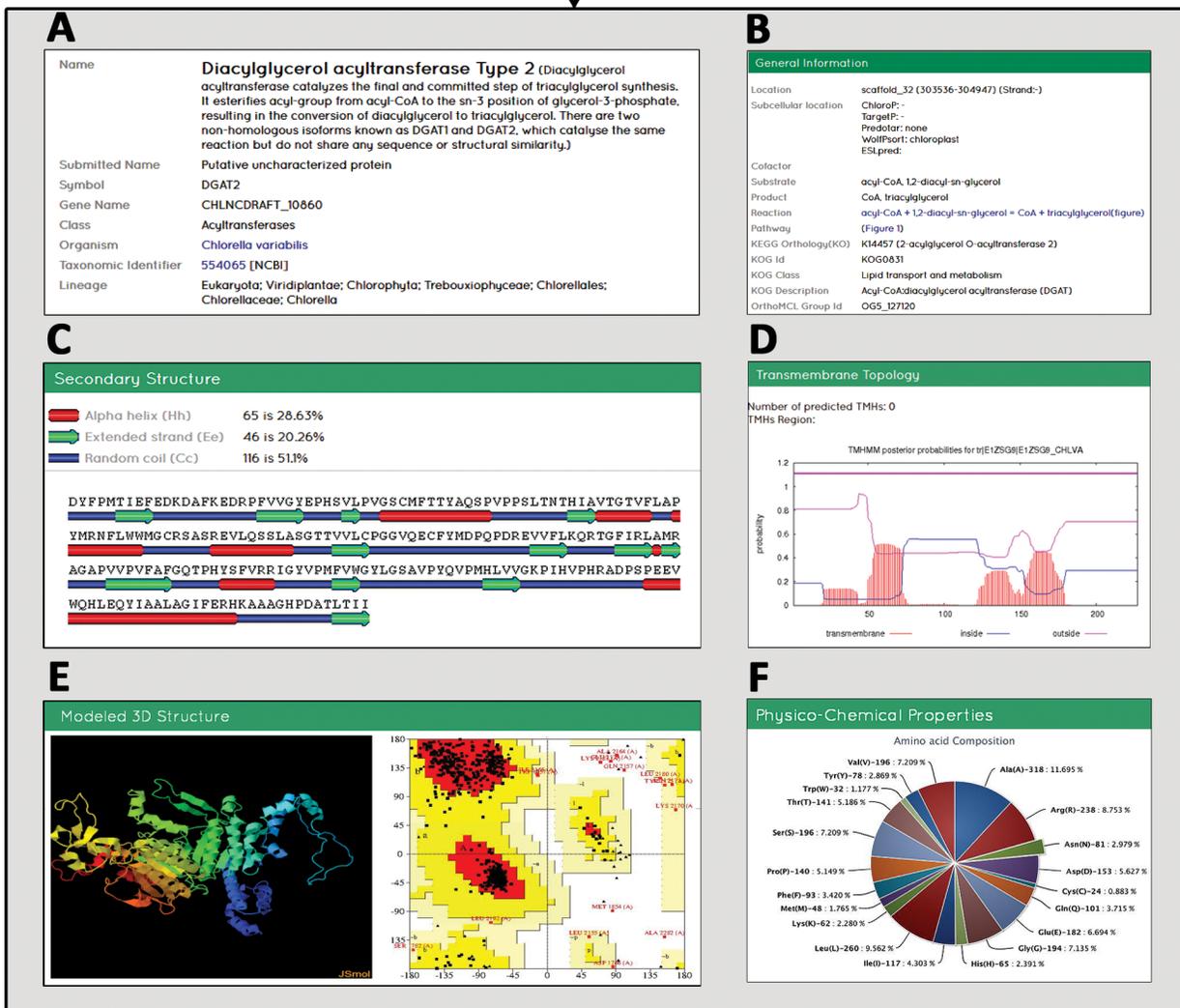
Fig 1. Demonstration of a typical example of annotation details for a lipid biosynthetic enzyme in dEMBF. By clicking on "Annotation details" tab, user will be navigated to a page displaying various information of an enzyme. (A) This field shows enzyme name, symbol, gene name, class, organism, taxonomic identifier and lineage details with hyperlinks to original sources. (B) The general information includes details on chromosomal location, subcellular location, reaction, pathway along with KOG and OrthoMCL ID. (C) An example of secondary structure determined by MINNOU tool. (D) An example of transmembrane topology predicted by TMHMM. (E) Modeled 3D structure and Ramachandran plot is provided. (F) Schematic representation of amino acid composition of an enzyme computed using ProtParam. Please note that not all fields are shown.

doi:10.1371/journal.pone.0146158.g001
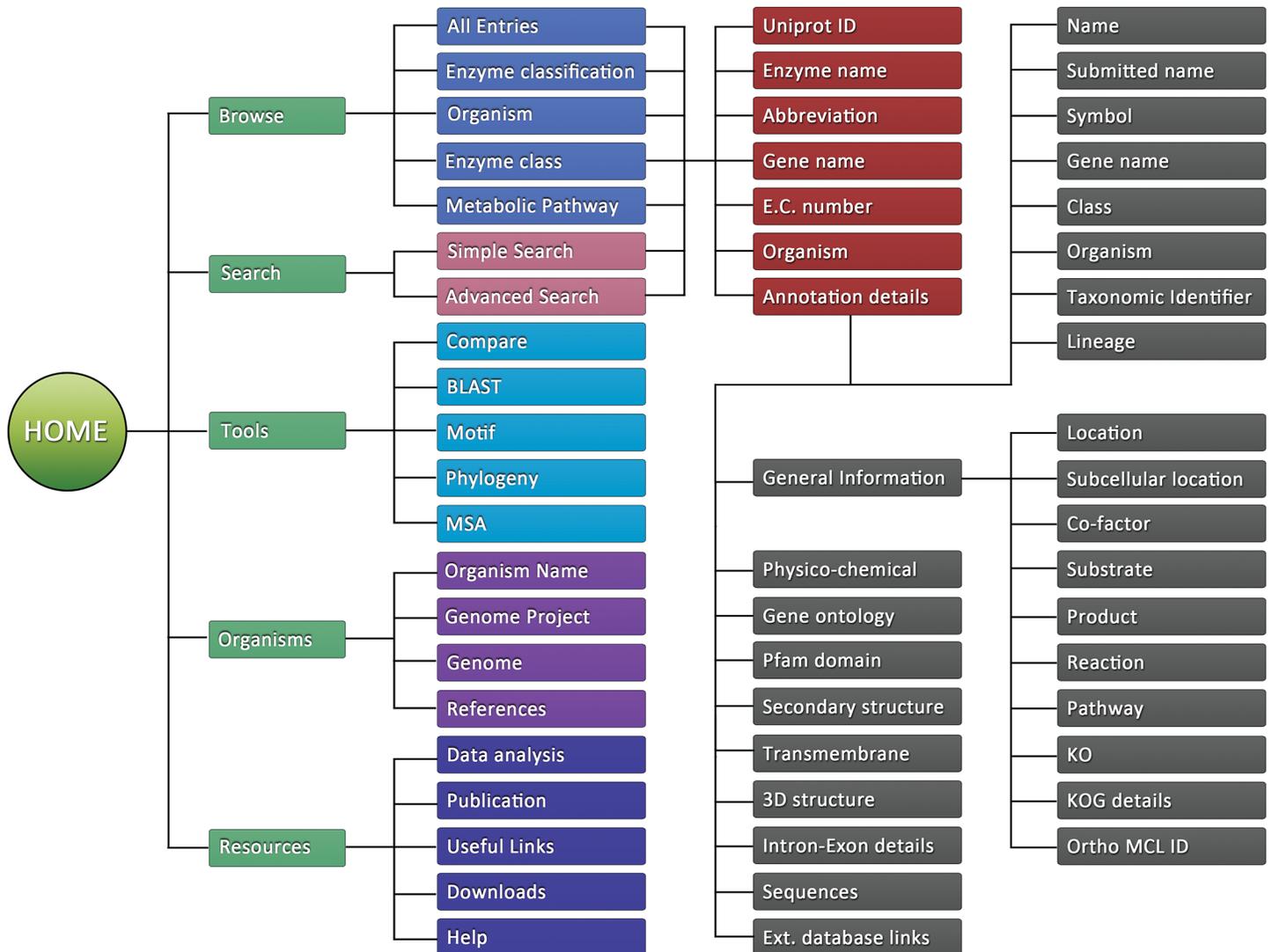
Fig 2. A schematic representation of architecture of dEMBF.

Likewise, the "Metabolic Pathway Browser" greatly facilities users to browse detailed information of any particular enzyme by just clicking on the enzyme name that has been manually mapped onto the lipid biosynthetic pathway. The "Database Summary" provides a complete list of the total number of lipid biosynthetic enzymes in various algal species currently present in dEMBF. Links to some of the important tools of dEMBF such as "BLAST", "Compare" and "Phylogeny", is also provided.

Browse. A number of browsing options are provided in dEMBF to allow users to navigate by specific criteria, such as selecting browse by "All Entries" for retrieving all enzymes present in the database or browse by "Enzyme Classification", "Organism", and "Enzyme Class", for specific enzymes of interest (Fig 4). On clicking the "Browse" option, user will be redirected to a page displaying all enzymes along with their respective accession ID, abbreviation, gene name, EC number, organism name and annotation details. The "Annotation details", option provides comprehensive sequence and structural properties of an enzyme (Fig 1), as discussed in the "Annotation details of enzymes in dEMBF" section of results. In addition, the "Metabolic

Fig 3. The dEMBF homepage displaying the different options available for the user. The major web interfaces of dEMBF, namely Browse, Search, Tools, Organisms and Resources are at the top right of the page. The "Metabolic Pathway Browser" is provided where users can select any enzymes (yellow color) of the lipid biosynthetic pathway to retrieve its detailed information. The "Database Summary" lists the total number of lipid biosynthetic enzymes currently present in dEMBF. An easy-to-use search field with multiple search criteria including enzyme name, symbol, UniProt ID, gene name, enzyme class, EC number, and organism name is also provided. Links to some of the important tools of dEMBF such as "Compare", "BLAST" and "Phylogeny" is available in the homepage.

doi:10.1371/journal.pone.0146158.g003

dEMBF | Database of Enzymes of Microalgal Biofuel Feedstock

Home Browse Search Tools Organisms Resources

All Entries

Enzyme Classification

Organisms

Enzyme Class

Metabolic Pathway Browser

A

All Entries

UNIPROT ID	ENZYME NAME	ABBREV.	GENE NAME	E.C.	ORGANISM NAME	ANNOTATION DETAILS
F4I1L3	Homomeric Acetyl-CoA carboxylase	Homomeric ACCase	At1g36180	6.4.1.2	Arabidopsis thaliana	●
F0YE78	Homomeric Acetyl-CoA carboxylase	Homomeric ACCase	AURANDRAFT_65524	6.4.1.2	Aureococcus anophagefferens	●
F0YJA4	Homomeric Acetyl-CoA carboxylase	Homomeric ACCase	AURANDRAFT_38836	6.4.1.2	Aureococcus anophagefferens	●
K8EIK1	Homomeric Acetyl-CoA carboxylase	Homomeric ACCase	Bathy08g01850	6.4.1.2	Bathycoccus prasinos	●

B

Enzyme Classification

ENZYME NAME	ABBREV.	E.C.	ALL ENTRIES
Homomeric Acetyl-CoA carboxylase	Homomeric ACCase	6.4.1.2	●
Acetyl-CoA carboxylase alpha-carboxyltransferase	alpha-CT	6.4.1.2	●
Biotin carboxylase	BC	6.3.4.14	●
Biotin carboxyl carrier protein	BCCP	6.4.1.2	●
Acetyl-CoA carboxylase beta-carboxyltransferase	beta-CT	6.4.1.2	●
3-hydroxyacyl-ACP Dehydratase	HAD/FabZ/FabA	4.2.1.59	●

C

Organisms

ORGANISM NAME	GENOME PROJECT	GENOME	REFERENCES	ALL ENTRIES
Arabidopsis thaliana cv. Aa-0. Taxon ID: 3702	The Arabidopsis Information Resource (Project Id-TAIR), Arabidopsis genome Initiative	●	●	●
Aureococcus anophagefferens CCMP1984. Taxon ID: 44056	US DOE Joint Genome Institute (Project Id-4000238)	●	●	●
Bathycoccus prasinos RCC 1105. Taxon ID: 41875	Genoscope (Project Id--)	●	●	●
Chlamydomonas reinhardtii CC-503. Taxon ID: 3055	US DOE Joint Genome Institute (Project Id-2662209)	●	●	●

D

Enzyme Class

ENZYME CLASS	TOTAL NO. OF ENTRIES	ALL ENTRIES
Ligases	39	●
Lyases	17	●
Acyltransferases	188	●
Oxidoreductases	55	●
Hydrolases	17	●

E

Acetyl-CoA
Homomeric ACCase
α-CT
β-CT
BC
BCCP
Malonyl-CoA
Acyl-CoA
MCMT
Malonyl-ACP
KAS I
KASII
KAS III
3-Ketoacyl-ACP KAR
Acyl-ACP
3-Hydroxyacyl-ACP
ENR
HAD
Trans-Enoyl-ACP
Acyl-CoA
DHAP
GPDH
GPAT
Glycerol-3-phosphate
Acyl-CoA
Lysophosphatidic acid
LPAT
Phosphatidic acid
PAP
Acyl-CoA
Diacylglycerol
DGAT 1
DGAT 2
Triacylglycerol

Fig 4. Screenshots of the dEMBF browse interface. A user can browse the database using five different browsing options including (A) Browse by "All Entries". (B) Browse by "Enzyme Classification". (C) Browse by "Organism". (D) Browse by "Enzyme Class". (E) "Metabolic Pathway Browser".

doi:10.1371/journal.pone.0146158.g004

Pathway Browser" is a dynamic browsing interface where the lipid biosynthesis enzymes have been linked to its information details.

Search. The "Search" function permits users to perform a simple search and advanced search in the database (Fig 5). The "Simple Search" option provides search queries for the enzyme name, symbol, gene name, organism, enzyme class, EC number, KOG ID, Gene Ontology, Pfam ID, UniProt ID and gene ID. The "Advanced Search" allows users to combine multiple search criteria in order to locate specific enzymes of interest more precisely.

Tools. A number of web-based tools have been integrated in dEMBF to facilitate further analysis of the enzymes. A brief description of these tools is as follows:

BLAST. The standalone NCBI's BLAST software was integrated as a part of the dEMBF tools. Users can perform a BLAST search of a query sequence either against the entire dEMBF database or against each individual enzyme to identify homologous sequences (Fig 6a). A wide range of E-values are available to control search sensitivity. The BLAST results are displayed

Fig 5. Screenshots of the dEMBF search interface. (A) The "Simple Search" provides search queries for the enzyme name, symbol, gene name, organism, enzyme class, EC number, KOG ID, Gene Ontology, Pfam ID, UniProt ID and gene ID. (B) The "Advanced Search" allows users to input multiple search queries simultaneously to retrieve specific enzymes of interest. (C) An example of a simple search result queried by the enzyme name, biotin carboxylase.

doi:10.1371/journal.pone.0146158.g005
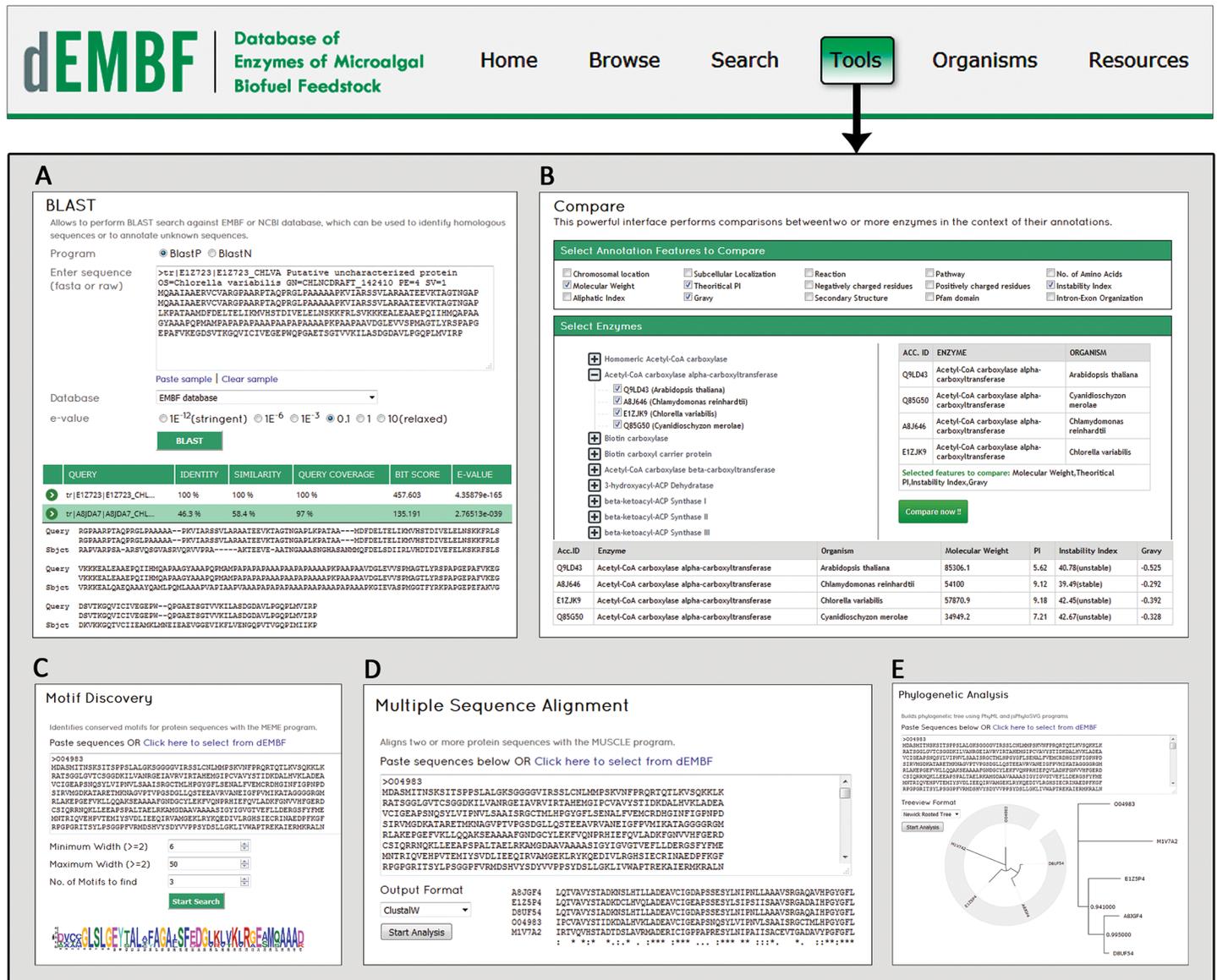
Fig 6. Screenshots of dEMBF analysis tools displaying their query pages and resulting outputs. (A) "BLAST" tool allow users to perform similarity search for protein or nucleotide sequences against NCBI, dEMBF database or against individual enzymes. A wide range of E-values are provided to control search sensitivity. BLAST results are sorted by percentage of identity, similarity, query coverage, bit score and E-value. (B) "Compare" tool to perform comparative analysis of enzyme between one or multiple algal species. Users can select various clickable annotation feature alongwith enzyme name and corresponding organisms between which comparisons is to be carried out. (C) "Motif" tool to identify conserved motifs in query sequences using the integrated MEME program. (D) "MSA" tool to align two or more protein sequences with the MUSCLE program. (E) "Phylogeny" tool to construct phylogenetic tree (Newick rooted tree or Circular tree) using PhyML and jsPhyloSVG.

doi:10.1371/journal.pone.0146158.g006

on the same page in a tabular format sorted by percentage of identity, similarity, query coverage, bit score and E-value. This interface is particularly useful for users to annotate the function of an unknown sequence.

Compare. The "Compare" tool ([Fig 6b](#)) allows user to perform comparative analysis of enzymes between one or multiple algal species. User has to select at least two enzymes from the same or different organism alongwith the annotation features based on which the comparison will be carried out. The results are displayed in a condensed tabular format.

dEMBF | Database of Enzymes of Microalgal Biofuel Feedstock

Home   Browse   Search   Tools   Organisms   Resources

ORGANISM NAME	GENOME PROJECT	GENOME	REFERENCES
Arabidopsis thaliana cv. Aa-0. Taxon ID: 3702	The Arabidopsis Information Res		
Aureococcus anophagefferens CCMP1984. Taxon ID: 44056	US DOE Joint Genome Institute (		
Bathycoccus prasinos RCC 1105. Taxon ID: 41875	Genoscope (Project Id--)		
Chlamydomonas reinhardtii CC-503. Taxon ID: 3055	US DOE Joint Genome Institute (		
Chlorella variabilis NC64A. Taxon ID: 554065	US DOE Joint Genome Institute (Project Id-4002944)		
Cyanidioschyzon merolae 10D. Taxon ID: 280699			
Ectocarpus siliculosus Ec. Taxon ID: 2880			
Emiliania huxleyi CCMP1 Taxon ID: 280463	-16965)		
Micromonas pusilla CCM Taxon ID: 564608	-4000188		
Micromonas sp. RCC299 Taxon ID: 296587	-4000189		
Nannochloropsis gaditar Taxon ID: 1093141	ject Id-17		
Ostreococcus lucimarinus CCE9901. Taxon ID: 436017	US DOE Joint Genome Institute (Project Id-3436080)		
Ostreococcus tauri OTTH0595. Taxon ID: 70448	US DOE Joint Genome Institute (Project Id-na)		
Phaeodactylum tricornutum CCAP 1055/1. Taxon ID: 556484	US DOE Joint Genome Institute (Project Id-3634397)		
Thalassiosira pseudonana CCMP1335. Taxon ID: 35128	US DOE Joint Genome Institute (Project Id-2662235)		
Volvox carteri Eve. Taxon ID: 3067	US DOE Joint Genome Institute (Project Id-3436681)		

Chlorella variabilis NC64A (Genome Details)

Topology	: Linear
Nuclear genome size	: 46.2 (Mb)
Number of chromosomes	: 12
Total GC Content	: 67(%)
Gene count	: 9791
Avg. protein length	: 456 (amino acids)
Avg. gene length	: 2.92 (kb)
Avg. transcript length	: 1.4 (kb)
Avg. gene density	: 4.7 (kb/gene)
Avg. number of exons per gene	: 7.3
Avg. number of introns per gene	: 0
Avg. exon length	: 170 (kb/gene)
Avg. intron length	: 109 (kb/gene)
Avg. coding sequence	: 29 (%)
Proteins	: 9780
GenBank	: ADIC00000000.1,

Chlorella variabilis NC64A

Super Kingdom: Eukaryota
Kingdom: Viridiplantae
Phylum: Chlorophyta
Class: Trebouxiophyceae
Order: Chlorellales
Family: Chlorellaceae
Genus: Chlorella
NCBI Taxon ID: 554065

Chlorella is a genus of about ten unicellular algal species that have small, spherical or ellipsoidal cells. They generally occur on soil and in freshwater. They have also been found as endosymbionts in invertebrate animals like Paramecium. They are non-flagellate cells but contain a vestigial flagellar apparatus. Reproduction is asexual and achieved by producing non motile autospores. They possess a firm polysaccharide cell wall which, in part, is made of sporopollenin like substance that occurs in the walls of the pollen grains of higher plants. It contains several essential nutrients and is a rich source of lutein, an important carotenoid. Chlorella species have been extensively studied and used in various practical applications in agriculture and biotechnology such as biofuels, sequestering CO2, producing molecules of high economic value or removing heavy metals from waste waters.

Emiliania huxleyi CCMP1516

Read BA, Kegel J, Klute MJ, Kuo A, Lefebvre SC, Maumus F, et al. Pan genome of the phytoplankton Emiliania underpins its global distribution.Nature. 2013 Jul 11;499(7457):209-13DOI: 10.1038/nature12221. PUBMED: 23760476

Sanchez Puerta MV, Bachvaroff TR, Delwiche CF.The complete plastid genome sequence of the haptophyte Emiliania huxleyi: a comparison to other plastid genomes.DNA Res. 2005;12(2):151-6.DOI: 10.1093/dnares/l2.2.151. PUBMED: 16303746

Sanchez Puerta MV, Bachvaroff TR, Delwiche CF.The complete mitochondrial genome sequence of the haptophyte Emiliania huxleyi and its relation to heterokonts.DNA Res. 2004 Feb 29;11(1):1-10.DOI: 10.1093/dnares/11.1.1. PUBMED: 15141941

Fig 7. Screenshot of the dEMBF organism page. This page shows the list of algal species analysed in this study alongwith information on their taxonomic lineage, genome project, genome details and references.

doi:10.1371/journal.pone.0146158.g007

Motif. Users can predict conserved motifs using the MEME [37] program integrated in dEMBF (Fig 6c). On submission of protein sequences, the database will redirect the query to MEME and after the completion of job, results in various pre-defined formats are made available for download.

MSA and Phylogeny. In addition to above tools, both multiple sequence alignment (Fig 6d) and phylogenetic (Fig 6e) tools are also provided in the "Tools" page of dEMBF. Alignment of two or more protein sequences is done by MUSCLE [38] and newick trees are built using PhyML [39] and jsPhyloSVG [40], the later being a java-independent function for viewing phylogenetic tree files online.

Organisms. The "Organism" page (Fig 7) displays the list of sequenced genomes analyzed in dEMBF, which comprises of fifteen microalgal species alongwith Arabidopsis as the

Fig 8. Screenshots of the dEMBF resource page. (A) "Data Analysis" displays a statistical overview of the data present in dEMBF. (B) "Publication" provides a list of research articles on algal lipid biosynthesis with hyperlinks to PubMed database. (C) "Useful Links" tab presents hyperlinks to other important bioinformatic resources such as algal genome project databases (JGI, ORCAE, CRIBI Genomics), Arabidopsis lipid gene database, metabolic pathway databases (KEGG, MetaCyc, BioCyc, ChlamyCyc, and AFAT). (D) "Download" tab allow users to export a copy of the entire dEMBF data.

reference plant species. This page includes the name of the organism, corresponding genome project database, genome details and related references, where each of the above fields are linked to further detailed information.

Resources. Given below are brief descriptions of the various utilities, available in the "Resources" page of dEMBF:

Data analysis: A statistical overview of the data present in dEMBF is provided (Fig 8a).
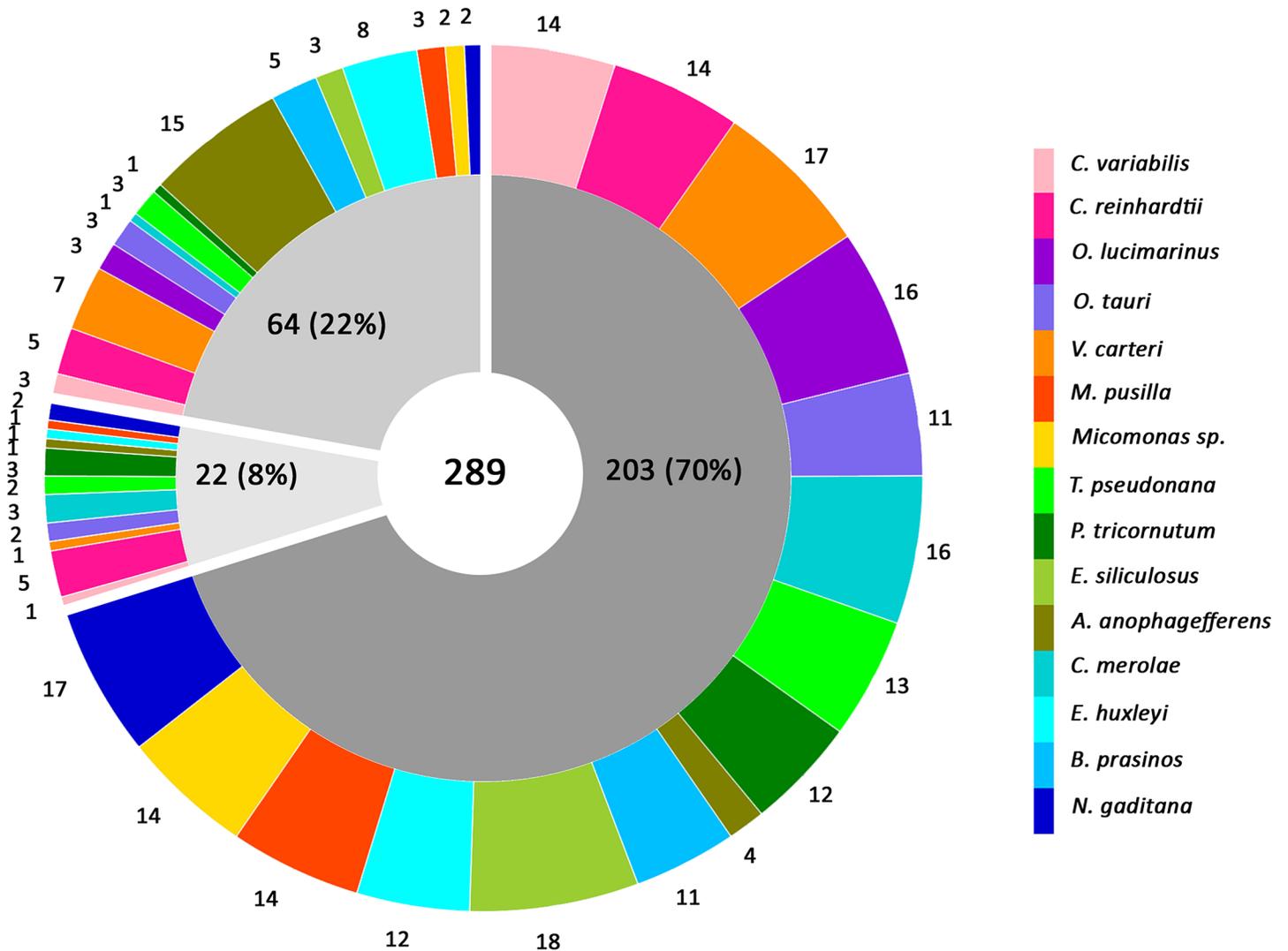
Fig 9. Chart showing the annotations of 289 enzymes present in dEMBF that are putatively involved in lipid biosynthesis in fifteen algal species. The dark grey sector indicates the total number of enzymes with functional annotations available from JGI database, but was further evaluated in this study for confirmation or assignment of any missing functional features. The light grey sector indicates the total number of previously uncharacterized enzymes from JGI, for which putative functions were predicted based on UniProt annotations. The medium grey sector indicates the number of enzymes for which no annotations were available in JGI as well as in UniProt. A putative function for each of these enzymes was predicted using various bioinformatics tools. The values inside the chart refer to the total number of enzymes while the values outside the chart indicate the distribution of the enzymes per organism. The names of the fifteen microalgae are indicated in the right panel with different color codes for each species.

doi:10.1371/journal.pone.0146158.g009

Publications: Recent research articles on algal lipid biosynthesis pathway have been compiled alongwith hyperlinks to PubMed for user references (Fig 8b).

Useful Links: External database links are provided to other bioinformatics resources such as algal genome project databases, Arabidopsis lipid gene database, metabolic pathway databases and protein databases (Fig 8c).

Downloads: All protein and nucleotide sequences present in the dEMBF are available for download from "Download" page of Resources (Fig 8d).
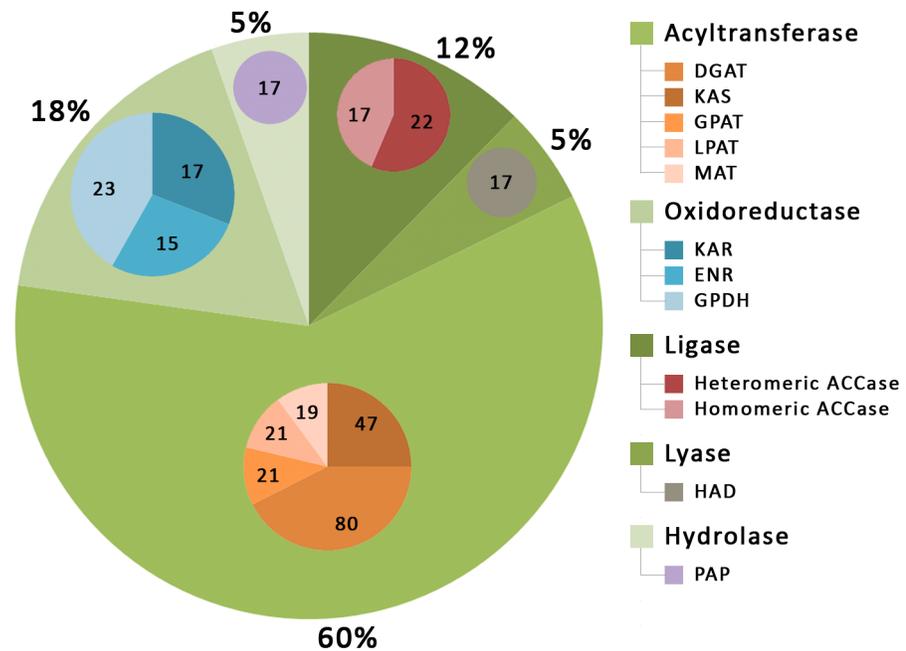
Fig 10. Distribution of sequences according to enzyme classification in the dEMBF database. The main chart shows the overall percentage of enzymes belonging to acyltransferase, oxidoreductase, ligase, lyase and hydrolase while the insert charts shows the total number of each enzymes (values indicated) belonging to a particular enzymes class.

Help: A detailed description on the use of various features of dEMBF database is provided in a Help file.

Discussion

After a thorough examination of the fifteen algal genomes, a total of 289 enzymes with putative roles in lipid synthesis were identified (S1 and S2 Tables). Sequence-structure information of these enzymes, together with the 27 well characterized homologous enzymes from Arabidopsis used as reference dataset in this study, are provided in the database. While previous studies have identified some key enzymes associated with lipid metabolic pathway in few algal species [41–49], the genomes of *C. variabilis*, *M. pusilla*, *Micromonas sp.*, *B. prasinos*, *T. pseudonana*, *P. tricornutum*, *E. siliculosus*, *A. anophagefferens* and *E. huxleyi* have been mined for the first time in this study to collate the entire repertoire of enzymes responsible for lipid accumulation in microalgae. In addition to genome mining, we have assigned pathways, gene ontology terms and cluster of orthologous (S3 Table), subcellular location, secondary and tertiary structure, biophysical properties, cellular processes and protein family terms to each of the enzymes. Consequently, we have improved the existing functional annotation of all 289 enzymes including 86 previously uncharacterized sequences for which a putative function in lipid biosynthesis has been determined (Fig 9). We observed that the analyzed algal genomes exhibited an overall comparable enzymatic makeup and each encode the major enzymes for lipid synthesis similar to Arabidopsis (S2 Table and S1 Fig). However, we found that four algal species viz., *C. variabilis*. *C. reinhardtii*, *V. carteri* and *C. merolae* contain both homomeric and heteromeric ACCase

enzyme, while the rest contain only the homomeric form of ACCase. This is in agreement to a previous published report, stating that the green (Chlorophyta) and red (Rhodophyta) algae with the exception of the green algal class Prasinophyceae (*O. lucimarinus*, *O. tauri*, *M. pusilla*, *Micromonas sp*. and *B. prasinos*) contain both homomeric and heteromeric ACCase while other algal species belonging to Heterokontophyta and Haptophyta lack heteromeric ACCase [42]. Furthermore, we found that the acyltransferases (60% of the total number of enzymes) is the most abundant enzyme class (Fig 10). The increased number of enzymes belonging to this class is probably significant considering that the three acyltransferases including GPAT, LPAT and DGAT catalyzes sequentially to acylate glycerol backbone, to ultimately produce TAG. These enzymes play a vital role in determining the acyl composition of glycerolipids and the final content of TAG [12]. In particular, relatively more number of DGAT (80 in number) followed by ACCase (39 in number) enzyme was observed in all algal genomes. The fact that ACCase catalyzes the initial rate limiting step of fatty acid biosynthesis by converting acetyl CoA to malonyl CoA while the DGAT enzyme drives the final step of TAG synthesis acylating diacylglycerol to TAG [4, 10], clearly reflects the high lipid accumulation capability of microalgae for biofuel production.

To our knowledge, dEMBF is the first comprehensive database on enzymes responsible for lipid accumulation in fifteen diverse algal species whose genome sequences are available. This work could be useful towards better understanding of fatty acid and TAG biosynthetic pathways in microalgae, besides facilitating the development of genetically engineered algal strains for a sustainable and economical viable biofuel production.

Supporting Information

S1 Fig. Genome-wide distribution of enzymes in the dEMBF database. Comparison of number of lipid biosynthesis enzymes in *Arabidopsis thaliana*, *Chlorella variabilis*, *Chlamydomonas reinhardtii*, *Ostreococcus lucimarinus*, *Ostreococcus tauri*, *Volvox carteri*, *Micromonas pusilla strain CCMP1545*, *Micromonas sp. strain RCC2999*, *Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, *Ectocarpus siliculosus*, *Aureococcus anophagefferens*, *Cyanidioschyzon merolae*, *Emiliania huxleyi*, *Bathycoccus prasinos* and *Nannochloropsis gaditana*. Enzymes are indicated with different colors as defined in the legend.
(TIF)

S1 Table. Distribution of enzymes (UniProt accession IDs) putatively involved in lipid biosynthesis in various algal species.
(DOCX)

S2 Table. Genome-wide comparative analysis of homologous lipid biosynthesis enzymes.
(DOCX)

S3 Table. Gene ontology classification and clusters of orthologous of lipid biosynthesis enzymes in the dEMBF database.
(DOCX)

Acknowledgments

Author Contributions

Conceived and designed the experiments: PKP BKM NM. Performed the experiments: NM BKP. Analyzed the data: PKP NM BKP. Contributed reagents/materials/analysis tools: BKM. Wrote the paper: NM PKP BKM.

References

1. Han SF, Jin WB, Tu RJ, Wu WM. Biofuel production from microalgae as feedstock: current status and potential. Crit Rev Biotechnol. 2015; 35:255–268. doi: 10.3109/07388551.2013.835301 PMID: 24641484

2. Medipally SR, Yusoff FM, Banerjee S, Shariff M. Microalgae as sustainable renewable energy feedstock for biofuel production. Biomed Res Int. 2015; doi: 10.1155/2015/519513

3. Radakovits R, Jinkerson RE, Darzins A, Posewitz MC. Genetic Engineering of Algae for Enhanced Biofuel Production. Eukaryot Cell. 2010; 9:486–501. doi: 10.1128/EC.00364-09 PMID: 20139239

4. Courchesne NM, Parisien A, Wang B, Lan CQ. Enhancement of lipid production using biochemical, genetic and transcription factor engineering approaches. J Biotechnol. 2009; 141:31–41. doi: 10.1016/j.jbiotec.2009.02.018 PMID: 19428728

5. Bhowmick GD, Koduru L, Sen R. Metabolic pathway engineering towards enhancing microalgal lipid biosynthesis for biofuel application—A review Renew Sust Energ Rev. 2015; 50: 1239–1253.

6. Lu J, Sheahan C, Fu P. Metabolic engineering of algae for fourth generation biofuels production. Energy Environ Sci. 2011; 4: 2451–2466.

7. Niu YF, Zhang MH, Li DW, Yang WD, Liu JS, Bai WB, et al. Improvement of neutral lipid and polyunsaturated fatty acid biosynthesis by overexpressing a type 2 diacylglycerol acyltransferase in marine diatom Phaeodactylum tricornutum. Mar Drugs. 2013; 11: 4558–4569. doi: 10.3390/md11114558 PMID: 24232669

8. Hsieh HJ, Su CH, Chien LJ. Accumulation of lipid production in Chlorella minutissima by triacylglycerol biosynthesis-related genes cloned from Saccharomyces cerevisiae and Yarrowia lipolytica. J Microbiol. 2012; 50: 526–534. doi: 10.1007/s12275-012-2041-5 PMID: 22752918

9. Ahmad I, Sharma AK, Daniell H, Kumar S. Altered lipid composition and enhanced lipid production in microalgae by introduction of brassica diacylglycerol acyltransferase 2. Plant Biotech J. 2015; 13: 540–50.

10. Hu Q, Sommerfeld M, Jarvis E, Ghirardi M, Posewitz M, Seibert M, et al. Microalgal triacylglycerols as feedstocks for biofuel production: perspectives and advances. Plant J. 2008; 54:621–639. doi: 10.1111/j.1365-313X.2008.03492.x PMID: 18476868

11. Reijnders MJ, van Heck RG, Lam CM, Scaife MA, dos Santos VA, Smith AG, et al. Green genes: bioinformatics and systems-biology innovations drive algal biotechnology. Trends Biotechnol. 2014; 32:617–626. doi: 10.1016/j.tibtech.2014.10.003 PMID: 25457388

12. Khozin-Goldberg I, Cohen Z. Unraveling algal lipid metabolism: Recent advances in gene identification. Biochimie. 2011; 93:91–100. doi: 10.1016/j.biochi.2010.07.020 PMID: 20709142

13. Beisson F, Koo AJ, Ruuska S, Schwender J, Pollard M, Thelen JJ, 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; 132:681–697. PMID: 12805597

14. Mao F, Yin Y, Zhou F, Chou WC, Zhou C, Chen H, et al. pDAWG: An Integrated Database for Plant Cell Wall Genes. Bioenerg Res. 2009; 2:209–216.

15. Ekstrom A, Taujale R, McGinn N, Yin Y. PlantCAZyme: a database for plant carbohydrate-active enzymes. Database.2014; doi: 10.1093/database/bau079

16. Childs KL, Konganti K, Buell CR. The Biofuel Feedstock Genomics Resource: a web-based portal and database to enable functional genomics of plant biofuel feedstock species. Database.2012; doi: 10.1093/database/bar061

17. Girke T, Lauricha J, Tran H, Keegstra K, Raikhel N. The Cell Wall Navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism. Plant Physiol. 2004; 136:3003–3008. PMID: 15489283

18. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990; 215:403–10. PMID: 2231712

19. UniProt C. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014; 42:D191–198. doi: 10.1093/nar/gkt1140 PMID: 24253303

20. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 2003; 4:41–55. PMID: 12969510

21. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic Acids Res. 2006; 34:D363–368. PMID: 16381887

22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–29. PMID: 10802651

23. Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S. AmiGO: online access to ontology and annotation data. Bioinformatics. 2009; 25:288–289. doi: 10.1093/bioinformatics/btn615 PMID: 19033274

24. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42:D222–230. doi: 10.1093/nar/gkt1223 PMID: 24288371

25. Wilkins MR, Gasteiger E, Bairoch A, Sanchez JC, Williams KL, Appel RD, et al. Protein identification and analysis tools in the ExPASy server. Methods Mol Biol. 1999; 112:531–552. PMID: 10027275

26. Hall TA. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Res. 1999; 41:95–98.

27. Cao B, Porollo A, Adamczak R, Jarrell M, Meller J. Enhanced recognition of protein transmembrane domains with prediction-based structural profiles. Bioinformatics. 2006; 22:303–309. PMID: 16293670

28. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001; 305:567–580. PMID: 11152613

29. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000; 300:1005–1016. PMID: 10891285

30. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. 1999; 8:978–984. PMID: 10338008

31. Small I, Peeters N, Legeai F, Lurin C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics. 2004; 4:1581–1590. PMID: 15174128

32. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007; 35:W585–587. PMID: 17517783

33. Birney E, Durbin R. Dynamite: a flexible code generating language for dynamic programming methods used in sequence comparison. Proc Int Conf Intell Syst Mol Biol. 1997; 5:56–64. PMID: 9322016

34. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–42. PMID: 10592235

35. Parida BK, Panda PK, Misra N, Mishra BK. MaxMod: a hidden Markov model based novel interface to MODELLER for improved prediction of protein 3D models. J Mol Model. 2015; 21:30. doi: 10.1007/s00894-014-2563-3 PMID: 25636267

36. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: A program to check the stereochemical quality of protein structures. J Appl Cryst. 1993; 26: 283–291.

37. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 2009; 37:W202–208. doi: 10.1093/nar/gkp335 PMID: 19458158

38. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. PMID: 15034147

39. Guindon S, Lethiec F, Duroux P, Gascuel O. PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res. 2005; 3:W557–559.

40. Smits SA, Ouverney CC. jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. Plos One. 2010; 5:e12267. doi: 10.1371/journal.pone.0012267 PMID: 20805892

41. Misra N, Panda PK, Parida BK, Mishra BK. Phylogenomic study of lipid genes involved in microalgal biofuel production-candidate gene mining and metabolic pathway analyses. Evol Bioinform.2012; 8:545–564.

42. Huerlimann R, Heimann K. Comprehensive guide to acetyl-carboxylases in algae. Crit Rev Biotechnol. 2013; 33:49–65. doi: 10.3109/07388551.2012.668671 PMID: 22524446

43. 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 1006–1017. PMID: 17416897

44. Chen JE, Smith AG. A look at diacylglycerol acyltransferases (DGATs) in algae. J Biotechnol. 2012; 162:28–39. doi: 10.1016/j.jbiotec.2012.05.009 PMID: 22750092

45.   Boyle NR, Page MD, Liu B, Blaby IK, Casero D, Kropat J, et al. Three acyltransferases and nitrogen-responsive regulator are implicated in nitrogen starvation-induced triacylglycerol accumulation in Chlamydomonas. J Biol Chem. 2012; 287:15811–25. doi: 10.1074/jbc.M111.334052 PMID: 22403401

46.   Msanne J, Xu D, Konda AR, Casas-Mollano JA, Awada T, Cahoon EB, et al. Metabolic and gene expression changes triggered by nitrogen deprivation in the photoautotrophically grown microalgae Chlamydomonas reinhardtii and Coccomyxa sp. C-169. Phytochemistry. 2012; 75:50–59. doi: 10.1016/j.phytochem.2011.12.007 PMID: 22226037

47.   Wagner M, Hoppe K, Czabany T, Heilmann M, Daum G, Feussner I, et al. Identification and characterization of an acyl-CoA:diacylglycerol acyltransferase 2 (DGAT2) gene from the microalga O. tauri. Plant Physiol Biochem. 2010; 48: 407–16. doi: 10.1016/j.plaphy.2010.03.008 PMID: 20400321

48.   Guihéneuf F, Leu S, Zarka A, Khozin-Goldberg I, Khalilov I, Boussiba S. Cloning and molecular characterization of a novel acyl-CoA:diacylglycerol acyltransferase 1-like gene (PtDGAT1) from the diatom Phaeodactylum tricornutum. FEBS J. 2011; 278:3651–66. doi: 10.1111/j.1742-4658.2011.08284.x PMID: 21812932

49.   Radakovits R, Jinkerson RE, Fuerstenberg SI, Tae H, Settlage RE, Boore JL, et al. Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropis gaditana. Nat Commun. 2012; 3:686. doi: 10.1038/ncomms1688 PMID: 22353717