Database, 2024, baae074
DOI: https://doi.org/10.1093/database/baae074
Original article

FatPlants: a comprehensive information system for lipid-related genes and metabolic pathways in plants

Chunhui Xu[1,2], Trey Shaw[2,3], Sai Akhil Choppararu[2,3], Yiwei Lu[2,3], Shaik Naveed Farooq[2,3], Yongfang Qin[2,3], Matt Hudson[2,3], Brock Weekley[2,3], Michael Fisher[2,3], Fei He[2,3], Jose Roberto Da Silva Nascimento[2,4], Nicholas Wergeles[2,3], Trupti Joshi[1,2,3,5], Philip D. Bates[6], Abraham J. Koo[4], Doug K. Allen[7,8], Edgar B. Cahoon[9], Jay J. Thelen[2,4], Dong Xu[1,2,3,*]

[1]Institute for Data Science and Informatics, University of Missouri, 22 Heinkel Building, Columbia, MO 65211, United States
[2]Christopher S. Bond Life Sciences Center, University of Missouri, 1201 Rollins St, Columbia, MO 65211, United States
[3]Department of Electrical Engineering and Computer Science, University of Missouri, Lafferre Hall, 416 S 6th St, Columbia, MO 65201, United States
[4]Department of Biochemistry, University of Missouri, Schweitzer Hall, 117, 503 S College Ave, Columbia, MO 65211, United States
[5]Department of Biomedical Informatics, Biostatistics and Medical Epidemiology, University of Missouri, CE707, Clinical Support and Education Building, 5 Hospital Dr. Columbia, MO, United States
[6]Institute of Biological Chemistry, Washington State University, 101D Plant Sciences Building, Pullman, WA 99164, United States
[7]Agriculture Research Service, United States Department of Agriculture, 975 N Warson Rd, St. Louis, MO 63132, United States
[8]Donald Danforth Plant Science Center, 975 N Warson Rd, St Louis, MO 63132, United States
[9]Department of Biochemistry and Center for Plant Science Innovation, University of Nebraska, 1901 Vine St, Lincoln, NE 68588, United States

[*]Corresponding author. Department of Electrical Engineering and Computer Science, University of Missouri, 201 Lafferre Hall, 416 S 6th St., Columbia, MO 65201, United States. E-mail: xudong@missouri.edu.

Citation details: Xu, C., Shaw, T., Choppararu, S. *et al.* FatPlants: a comprehensive information system for lipid-related genes and metabolic pathways in plants. *Database* (2024) Vol. 2024: article ID baae074; DOI: https://doi.org/10.1093/database/baae074

Abstract

FatPlants, an open-access, web-based database, consolidates data, annotations, analysis results, and visualizations of lipid-related genes, proteins, and metabolic pathways in plants. Serving as a minable resource, FatPlants offers a user-friendly interface for facilitating studies into the regulation of plant lipid metabolism and supporting breeding efforts aimed at increasing crop oil content. This web resource, developed using data derived from our own research, curated from public resources, and gleaned from academic literature, comprises information on known fatty-acid-related proteins, genes, and pathways in multiple plants, with an emphasis on *Glycine max, Arabidopsis thaliana*, and *Camelina sativa*. Furthermore, the platform includes machine-learning based methods and navigation tools designed to aid in characterizing metabolic pathways and protein interactions. Comprehensive gene and protein information cards, a Basic Local Alignment Search Tool search function, similar structure search capacities from AphaFold, and ChatGPT-based query for protein information are additional features.

Database URL: https://www.fatplants.net/

Introduction

Vegetable oils are an energy-dense renewable feedstock for chemicals and fuels and are an essential component of the human diet [1]. It is estimated that by 2050, the current vegetable oil production will need to double to meet societal needs [2, 3]. To date, increases in plant seed oil production through engineering or breeding have been reported but often failed to meet expectations [3–7]. Such efforts have often resulted in unintended consequences, including reduced seed shelf life and germination rate, and adverse effects on negatively impacted protein content. Lipid metabolism is a highly branched metabolic network that produces both membrane lipids and storage oils [2, 3, 8], and takes place across multiple organelles [9]. The regulatory nodes and metabolic bottlenecks [10, 11] that affect seed oil and protein accumulation are only partially characterized at the genetic and biochemical levels [8]. Hence, improving plant seed oil will require extensive effort. Such a challenge would benefit from a web portal equipped with analysis and visualization tools for fatty-acid-related proteins, which would comprehensively archive data and accelerate the process of knowledge discovery and crop design for biologists. Easy access to built-in analysis tools is also needed to empower researchers to develop and test hypotheses and design crops with value-added compositions.

Web resources are starting to emerge that have been developed to describe plant acyl-lipid metabolism or curate fatty-acid-related data, but frequently they are limited in scope and out of date. For example, Lipidbank [12], Seed Oil Fatty Acids Database [13], and LIPIDAT [14] are no longer maintained or updated; ARALIP [8], a widely used plant lipid-related protein

Received 28 August 2023; Revised 22 May 2024; Accepted 12 July 2024

database, focuses on *Arabidopsis thaliana* only; LIPIDMAPS [15] lacks integrated pathway knowledge; PlantFAdb [16] and Plant Lipid Databases [17] concentrate on the chemophysical properties and structures of lipids only. The growing research needs in plant lipids call for the development of a new platform that can provide comprehensive coverage of oilseed plants, genes, and knowledge in this area, and can continue to grow and improve with facile incorporation of community input.

To assist researchers in studying plant fatty acid metabolism efficiently, we developed a one-stop-shop web resource, FatPlants. Protein data has been manually curated and entered relevant to fatty acid metabolism in *Glycine max* (soybean), *A. thaliana* (Arabidopsis), and *Camelina sativa* (Camelina) from Uniprot [18], TAIR [19], SoyKB [20], KBcommons [21] LIPIDMAPS [15, 22], PlantFAdb [16], CamRegBase [23], and ARALIPS [8]. Molecular information on the fatty acid composition, chemical structures, and chemophysical properties from OPSIN [24] provides an in-depth description. For each protein record, general annotations from UniProt [18], including postmodification regions or sites, have been collected. For each specific species, we have included the cross-linked identifiers for different databases and the external links so that users can easily redirect to those databases. Sequences, annotation, and description are provided together with the structure information of those fatty-acid-related proteins.

Following data curation, we established a user-friendly searchable database augmented with visualization tools. FatPlants offers a suite of analysis features, including sequence or structure similarity searches. Users can submit a protein sequence to our database and obtain a list of similar proteins. Alternatively, the structure similarity method allows users to provide a protein sequence, and FatPlants returns proteins with the most analogous 3D structures based on AlphaFold API [25]. Functional analysis is facilitated by mapping fatty-acid-related proteins to pathway databases. We have manually converted images of fatty-acid-related pathways from academic literature into interactive graphs using machine learning, enabling users to explore protein or gene elements of lipid metabolism in depth. This feature allows regular updates with the latest fatty-acid-related pathways from recent literature. FatPlants provides links to protein–protein interaction (PPI) and Gene Ontology (GO) enrichment networks for fatty-acid-related genes. A unique feature we have integrated into FatPlants is the utilization of the ChatGPT API, enabling users to obtain specific protein information interactively. In essence, FatPlants serves as a comprehensive platform for plant fatty-acid-related data, knowledge, and analysis, with user-friendly search and analysis tools to facilitate understanding of the underlying biological frameworks.

Materials and methods

Data acquisition and curation

Our datasets were collected from three primary data sources, described in Fig. 1: ARALIP centered data, searchable

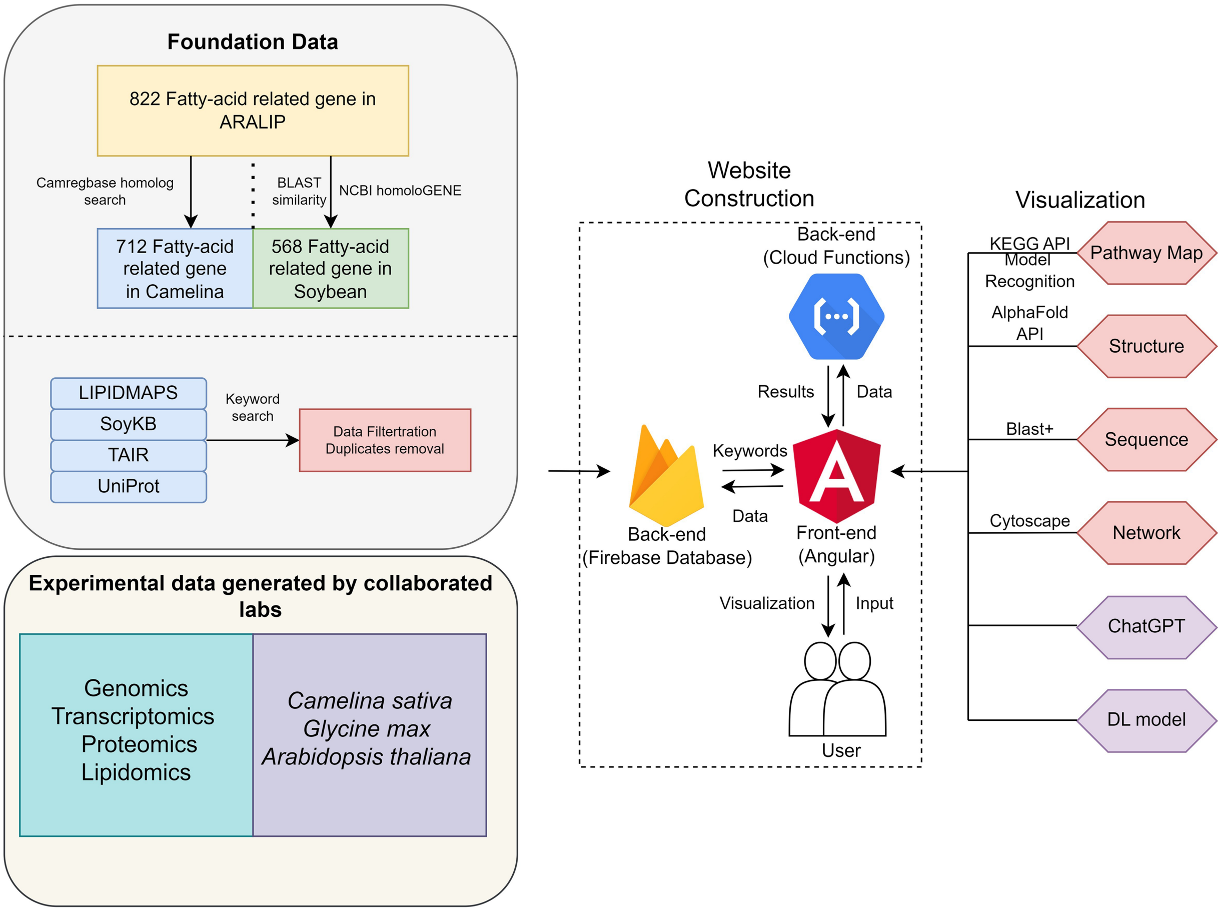


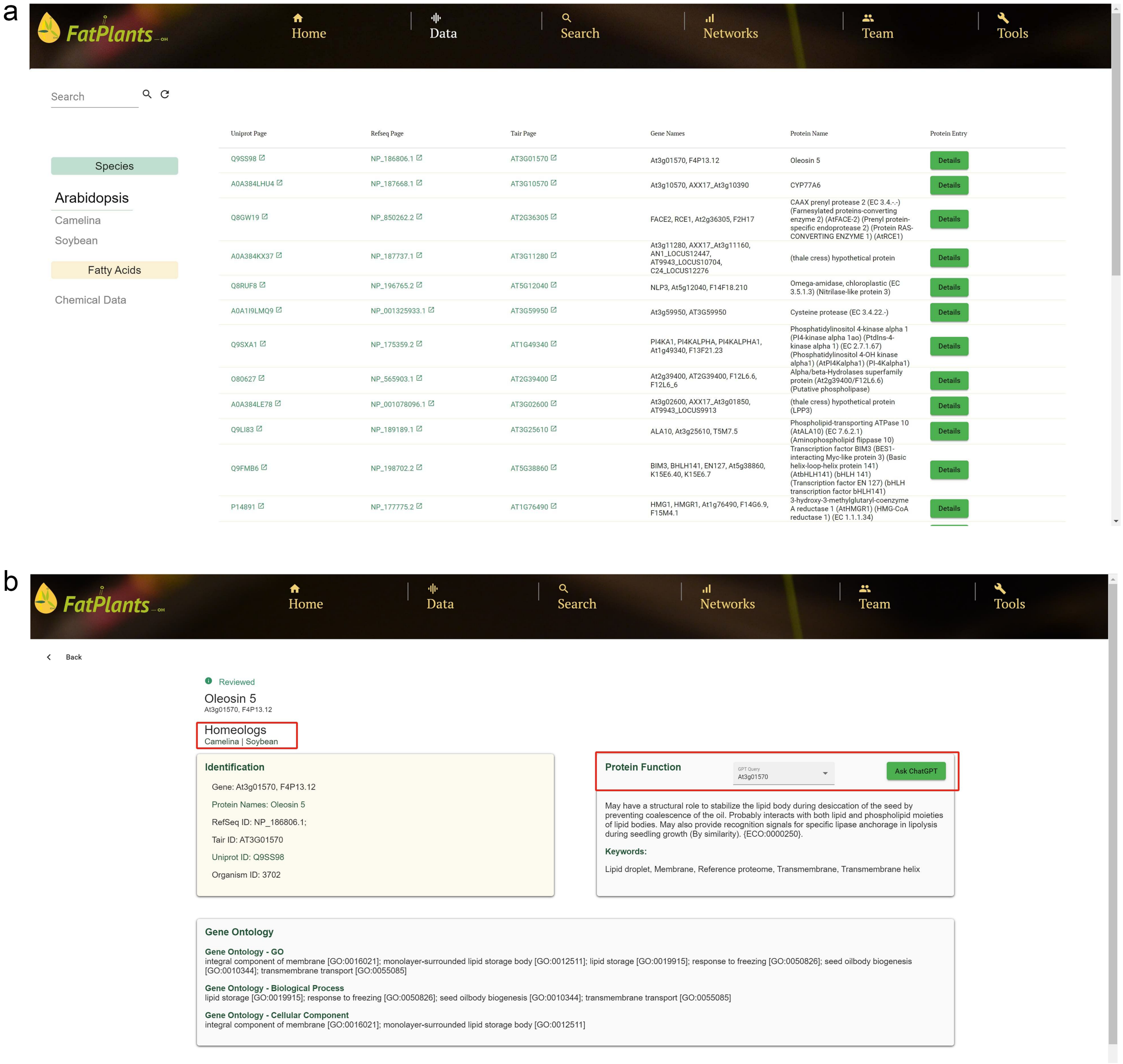
Figure 1. The data schema and functionalities of FatPlants.

Table 1. A summary of acyl-lipid metabolism data collection from different databases

	Arabidopsis thaliana	*Camelina sativa*	*Glycine max*
ARALIP centered	822	712	6602
UniProt	1559	223	422
TAIR	1718	N/A	N/A
SoyKB	N/A	N/A	389
LIPIDMAPS	2447	N/A	N/A
CamRegBase	N/A	9810	N/A
Total (Filtered)	6546 (3440)	10 845 (8581)	7413 (5606)

The total number represents the raw data we collected from the database source, and the filtered number shows how many proteins are left after our filtration schema. The N/A indicates that species are unavailable in the specific database.

data from protein databases, and in-house and published experimental data. The ARALIP data containing fatty acid-centric enzyme/gene data from Arabidopsis was utilized by searching homologs in Camelina and soybean, resulting in 712 genes in Camelina and 568 in soybean. We searched for proteins in the UniProt database with keywords including 'lipid' and 'fatty acid' in three species: G. max, A. thaliana, and C. sativa, and conducted the same keyword search in the TAIR database. For Arabidopsis, we collected 2447 fatty-acid-related proteins of Arabidopsis from LIPIDMAPS. Regarding physical and chemical properties data, we have collected 495 entries from PlantFAdb. The keywords 'lipid' and 'fatty acid' were used to search for genes with UniProt and the soybean Gene Model V9.00 in SoyKB. To perform data filtration, we mapped all the identifiers to UniProt ID and removed



Uniprot Page	Refseq Page	Tair Page	Gene Names	Protein Name	Protein Entry
Q9SS98	NP_186806.1	AT3G01570	At3g01570, F4P13.12	Oleosin 5	Details
A0A384LHU4	NP_187668.1	AT3G10570	At3g10570, AXX17_At3g10390	CYP77A6	Details
Q8GW19	NP_850262.2	AT2G36305	FACE2, RCE1, At2g36305, F2H17	CAAX prenyl protease 2 (EC 3.4.-.-) (Farnesylated proteins-converting enzyme 2) (AtFACE-2) (Prenyl protein-specific endoprotease 2) (Protein RAS-CONVERTING ENZYME 1) (AtRCE1)	Details
A0A384KX37	NP_187737.1	AT3G11280	At3g11280, AXX17_At3g11160, AN1_LOCUS12447, AT9943_LOCUS10704, C24_LOCUS12276	(thale cress) hypothetical protein	Details
Q8RUF8	NP_196765.2	AT5G12040	NLP3, At5g12040, F14F18.210	Omega-amidase, chloroplastic (EC 3.5.1.3) (Nitrilase-like protein 3)	Details
A0A1I9LMQ9	NP_001325933.1	AT3G59950	At3g59950, AT3G59950	Cysteine protease (EC 3.4.22.-)	Details
Q9SXA1	NP_175359.2	AT1G49340	PI4KA1, PI4KALPHA, PI4KALPHA1, At1g49340, F13F21.23	Phosphatidylinositol 4-kinase alpha 1 (PI4-kinase alpha 1ao) (PtdIns-4-kinase alpha 1) (EC 2.7.1.67) (Phosphatidylinositol 4-OH kinase alpha1) (AtPI4Kalpha1) (PI-4Kalpha1)	Details
O80627	NP_565903.1	AT2G39400	At2g39400, AT2G39400, F12L6.6, F12L6_6	Alpha/beta-Hydrolases superfamily protein (At2g39400/F12L6.6) (Putative phospholipase)	Details
A0A384LE78	NP_001078096.1	AT3G02600	At3g02600, AXX17_At3g01850, AT9943_LOCUS9913	(thale cress) hypothetical protein (LPP3)	Details
Q9LI83	NP_189189.1	AT3G25610	ALA10, At3g25610, T5M7.5	Phospholipid-transporting ATPase 10 (AtALA10) (EC 7.6.2.1) (Aminophospholipid flippase 10)	Details
Q9FMB6	NP_198702.2	AT5G38860	BIM3, BHLH141, EN127, At5g38860, K15E6.40, K15E6.7	Transcription factor BIM3 (BES1-interacting Myc-like protein 3) (Basic helix-loop-helix protein 141) (AtbHLH141) (bHLH 141) (Transcription factor EN 127) (bHLH transcription factor bHLH141)	Details
P14891	NP_177775.2	AT1G76490	HMG1, HMGR1, At1g76490, F14G6.9, F15M4.1	3-hydroxy-3-methylglutaryl-coenzyme A reductase 1 (AtHMGR1) (HMG-CoA reductase 1) (EC 1.1.1.34)	Details

Reviewed

Oleosin 5
At3g01570, F4P13.12

Homeologs
Camelina | Soybean

Identification

Gene: At3g01570, F4P13.12

Protein Names: Oleosin 5

RefSeq ID: NP_186806.1;

Tair ID: AT3G01570

Uniprot ID: Q9SS98

Organism ID: 3702

Protein Function

GPT Query: At3g01570

May have a structural role to stabilize the lipid body during desiccation of the seed by preventing coalescence of the oil. Probably interacts with both lipid and phospholipid moieties of lipid bodies. May also provide recognition signals for specific lipase anchorage in lipolysis during seedling growth (By similarity). {ECO:0000250}.

Keywords:

Lipid droplet, Membrane, Reference proteome, Transmembrane, Transmembrane helix

Gene Ontology

Gene Ontology - GO
integral component of membrane [GO:0016021]; monolayer-surrounded lipid storage body [GO:0012511]; lipid storage [GO:0019915]; response to freezing [GO:0050826]; seed oilbody biogenesis [GO:0010344]; transmembrane transport [GO:0055085]

Gene Ontology - Biological Process
lipid storage [GO:0019915]; response to freezing [GO:0050826]; seed oilbody biogenesis [GO:0010344]; transmembrane transport [GO:0055085]

Gene Ontology - Cellular Component
integral component of membrane [GO:0016021]; monolayer-surrounded lipid storage body [GO:0012511]

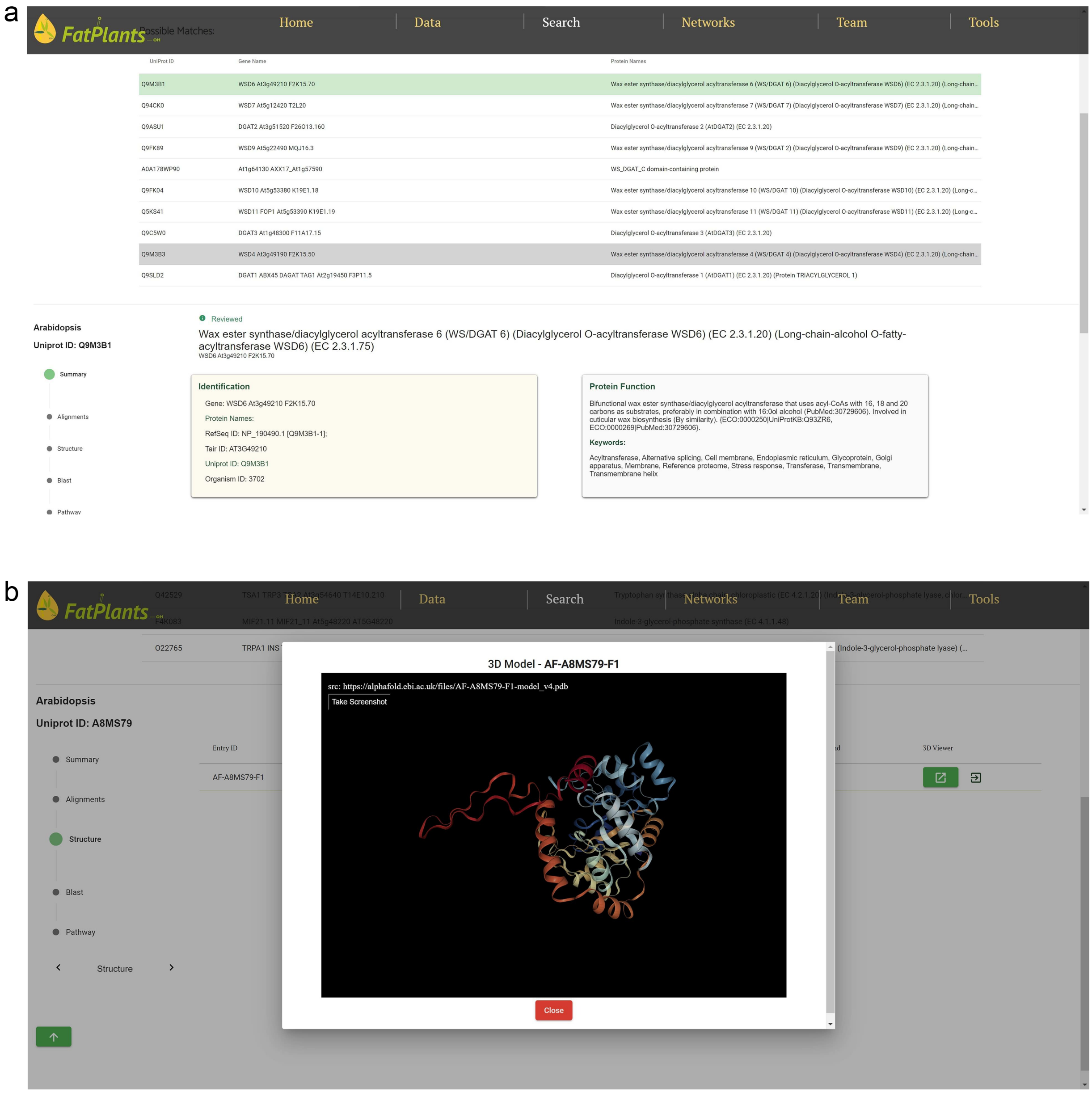
Figure 2. FatPlants data browse page. (a) The main data browse table. (b) The information card page for a selected protein with homologs and ChatGPT features.

redundant and unannotated proteins. A total of 3440 fatty-acid-related proteins were obtained for Arabidopsis and 5606 for soybean (Table 1). Fatty-acid-related protein data for Camelina could not be collected due to a lack of annotation. Therefore, a homology search against CamRegBase was performed to find the fatty-acid-related proteins of Camelina by using Arabidopsis data.

The protein list was used to retrieve the structure data from the Research Collaboratory for Structural Bioinformatics Protein Data Bank [26]. The PPI data were collected from the STRING database [27] and visualized in networks by direct or indirect interactions with intermediate nodes. The GO hierarchical annotations were retrieved from the GO [28] database and enriched and visualized in the network. Fatty-acid-related proteins were mapped to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database. In addition, a collection of fatty-acid-related pathway pictures from the literature [29–33] were visualized as an interactive map using our in-house machine-learning image understanding tool [34].

Database and web interface implementation

FatPlants provides a user-friendly interface for data access and retrieval. It is implemented by a frontend Single-Page-Application architecture using Angular 10.0. The application interacts with users dynamically to update the current web page. In the backend, we have developed a document-oriented database based on Firestore 9.1.3. As shown in Fig. 1, the entire dataset is stored in Firebase with extensive authenti-



Possible Matches:

UniProt ID	Gene Name	Protein Names
Q9M3B1	WSD6 At3g49210 F2K15.70	Wax ester synthase/diacylglycerol acyltransferase 6 (WS/DGAT 6) (Diacylglycerol O-acyltransferase WSD6) (EC 2.3.1.20) (Long-chain...
Q94CK0	WSD7 At5g12420 T2L20	Wax ester synthase/diacylglycerol acyltransferase 7 (WS/DGAT 7) (Diacylglycerol O-acyltransferase WSD7) (EC 2.3.1.20) (Long-chain...
Q9ASU1	DGAT2 At3g51520 F26O13.160	Diacylglycerol O-acyltransferase 2 (AtDGAT2) (EC 2.3.1.20)
Q9FK89	WSD9 At5g22490 MQJ16.3	Wax ester synthase/diacylglycerol acyltransferase 9 (WS/DGAT 2) (Diacylglycerol O-acyltransferase WSD9) (EC 2.3.1.20) (Long-chain...
A0A178WP90	At1g64130 AXX17_At1g57590	WS_DGAT_C domain-containing protein
Q9FK04	WSD10 At5g53380 K19E1.18	Wax ester synthase/diacylglycerol acyltransferase 10 (WS/DGAT 10) (Diacylglycerol O-acyltransferase WSD10) (EC 2.3.1.20) (Long-c...
Q5KS41	WSD11 FOP1 At5g53390 K19E1.19	Wax ester synthase/diacylglycerol acyltransferase 11 (WS/DGAT 11) (Diacylglycerol O-acyltransferase WSD11) (EC 2.3.1.20) (Long-c...
Q9C5W0	DGAT3 At1g48300 F11A17.15	Diacylglycerol O-acyltransferase 3 (AtDGAT3) (EC 2.3.1.20)
Q9M3B3	WSD4 At3g49190 F2K15.50	Wax ester synthase/diacylglycerol acyltransferase 4 (WS/DGAT 4) (Diacylglycerol O-acyltransferase WSD4) (EC 2.3.1.20) (Long-chain...
Q9SLD2	DGAT1 ABX45 DAGAT TAG1 At2g19450 F3P11.5	Diacylglycerol O-acyltransferase 1 (AtDGAT1) (EC 2.3.1.20) (Protein TRIACYLGLYCEROL 1)

Arabidopsis

Uniprot ID: Q9M3B1

Reviewed

Wax ester synthase/diacylglycerol acyltransferase 6 (WS/DGAT 6) (Diacylglycerol O-acyltransferase WSD6) (EC 2.3.1.20) (Long-chain-alcohol O-fatty-acyltransferase WSD6) (EC 2.3.1.75)

WSD6 At3g49210 F2K15.70

Identification

Gene: WSD6 At3g49210 F2K15.70

Protein Names:

RefSeq ID: NP_190490.1 [Q9M3B1-1];

Tair ID: AT3G49210

Uniprot ID: Q9M3B1

Organism ID: 3702

Protein Function

Bifunctional wax ester synthase/diacylglycerol acyltransferase that uses acyl-CoAs with 16, 18 and 20 carbons as substrates, preferably in combination with 16:0ol alcohol (PubMed:30729606). Involved in cuticular wax biosynthesis (By similarity). {ECO:0000250|UniProtKB:Q93ZR6, ECO:0000269|PubMed:30729606}.

Keywords:

Acyltransferase, Alternative splicing, Cell membrane, Endoplasmic reticulum, Glycoprotein, Golgi apparatus, Membrane, Reference proteome, Stress response, Transferase, Transmembrane, Transmembrane helix

Q42529 TSA1 TRP3 TSA2 At3g54640 T14E10.210 Tryptophan synthase alpha chain, chloroplastic (EC 4.2.1.20) (Indole-3-glycerol-phosphate lyase, chlor...

F4K083 MIF21.11 MIF21_11 At5g48220 AT5G48220 Indole-3-glycerol-phosphate synthase (EC 4.1.1.48)

O22765 TRPA1 INS (Indole-3-glycerol-phosphate lyase) (...

Arabidopsis

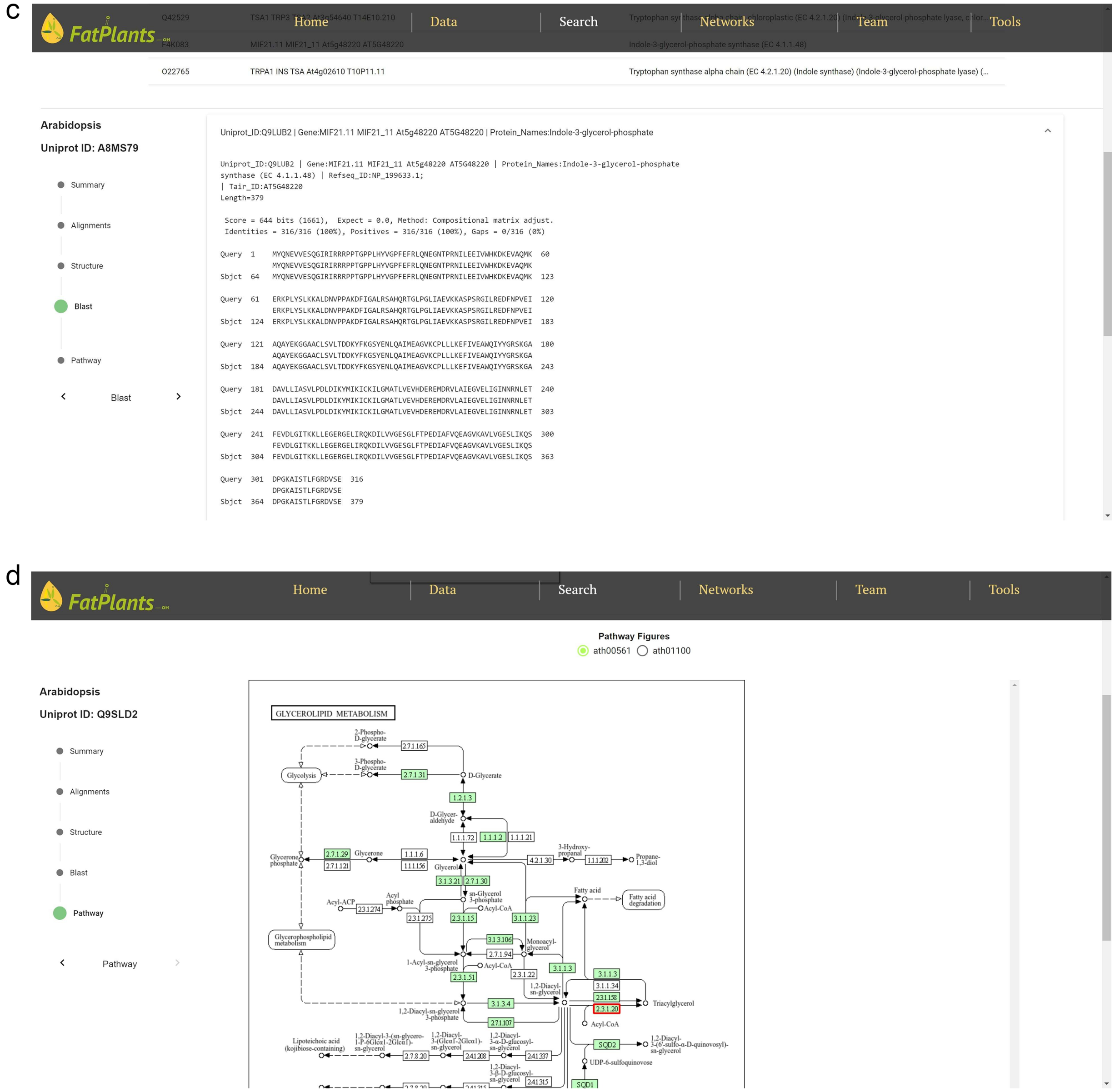
Uniprot ID: A8MS79

Entry ID	3D Viewer
AF-A8MS79-F1	

3D Model - **AF-A8MS79-F1**

src: https://alphafold.ebi.ac.uk/files/AF-A8MS79-F1-model_v4.pdb

Figure 3. One-stop search page of FatPlants. (a) The summary result page with all the candidates (in this case, we use name DGAT). (b) The structure result page, which is retrieved from Alphafold API. (c) The BLASTP result page. (d) KEGG pathway mapping page with the target gene highlighted in the red box.



Q42529	TSA1 TRP3 TSA2 At3g54640 T14E10.210	Tryptophan synthase alpha chain, chloroplastic (EC 4.2.1.20) (Indole-3-glycerol-phosphate lyase, chlor...
F4K083	MIF21.11 MIF21_11 At5g48220 AT5G48220	Indole-3-glycerol-phosphate synthase (EC 4.1.1.48)
O22765	TRPA1 INS TSA At4g02610 T10P11.11	Tryptophan synthase alpha chain (EC 4.2.1.20) (Indole synthase) (Indole-3-glycerol-phosphate lyase) (...

Arabidopsis

Uniprot ID: A8MS79

Uniprot_ID:Q9LUB2 | Gene:MIF21.11 MIF21_11 At5g48220 AT5G48220 | Protein_Names:Indole-3-glycerol-phosphate

```
Uniprot_ID:Q9LUB2 | Gene:MIF21.11 MIF21_11 At5g48220 AT5G48220 | Protein_Names:Indole-3-glycerol-phosphate
synthase (EC 4.1.1.48) | Refseq_ID:NP_199633.1;
| Tair_ID:AT5G48220
Length=379

 Score = 644 bits (1661),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 316/316 (100%), Positives = 316/316 (100%), Gaps = 0/316 (0%)

Query  1    MYQNEVVESQGIRIRRRPPTGPPLHYVGPFEFRLQNEGNTPRNILEEIVWHKDKEVAQMK  60
            MYQNEVVESQGIRIRRRPPTGPPLHYVGPFEFRLQNEGNTPRNILEEIVWHKDKEVAQMK
Sbjct  64   MYQNEVVESQGIRIRRRPPTGPPLHYVGPFEFRLQNEGNTPRNILEEIVWHKDKEVAQMK  123

Query  61   ERKPLYSLKKALDNVPPAKDFIGALRSAHQRTGLPGLIAEVKKASPSRGILREDFNPVEI  120
            ERKPLYSLKKALDNVPPAKDFIGALRSAHQRTGLPGLIAEVKKASPSRGILREDFNPVEI
Sbjct  124  ERKPLYSLKKALDNVPPAKDFIGALRSAHQRTGLPGLIAEVKKASPSRGILREDFNPVEI  183

Query  121  AQAYEKGGAACLSVLTDDKYFKGSYENLQAIMEAGVKCPLLLKEFIVEAWQIYYGRSKGA  180
            AQAYEKGGAACLSVLTDDKYFKGSYENLQAIMEAGVKCPLLLKEFIVEAWQIYYGRSKGA
Sbjct  184  AQAYEKGGAACLSVLTDDKYFKGSYENLQAIMEAGVKCPLLLKEFIVEAWQIYYGRSKGA  243

Query  181  DAVLLIASVLPDLDIKYMIKICKILGMATLVEVHDEREMDRVLAIEGVELIGINNRNLET  240
            DAVLLIASVLPDLDIKYMIKICKILGMATLVEVHDEREMDRVLAIEGVELIGINNRNLET
Sbjct  244  DAVLLIASVLPDLDIKYMIKICKILGMATLVEVHDEREMDRVLAIEGVELIGINNRNLET  303

Query  241  FEVDLGITKKLLEGERGELIRQKDILVVGESGLFTPEDIAFVQEAGVKAVLVGESLIKQS  300
            FEVDLGITKKLLEGERGELIRQKDILVVGESGLFTPEDIAFVQEAGVKAVLVGESLIKQS
Sbjct  304  FEVDLGITKKLLEGERGELIRQKDILVVGESGLFTPEDIAFVQEAGVKAVLVGESLIKQS  363

Query  301  DPGKAISTLFGRDVSE  316
            DPGKAISTLFGRDVSE
Sbjct  364  DPGKAISTLFGRDVSE  379
```

Pathway Figures

● ath00561 ○ ath01100

Arabidopsis

Uniprot ID: Q9SLD2

Figure 3. (Continued)

cation and a dynamic log system. FatPlants is deployed on Firebase. All backend functions associated with the Linux environment or outside APIs were implemented as Google Cloud Functions to accelerate the response time and reduce server latency. For feature development, the JavaScript library of Cytoscape [35] was used to visualize all network data and the Linux version of Basic Local Alignment Search Tool (BLAST) to build a sequence search function. In an innovative approach to enhance the accessibility of information, we integrated the ChatGPT API into our platform. This allows users to interactively retrieve specific protein information using natural language queries, thereby simplifying the process of data mining. Previously developed tools for structural prediction and pathway image recognition were incorporated to enable lipid characterization [25, 34]. In addition, to provide a smooth user experience of usage, FatPlants was validated on different browsers, such as Google Chrome, Edge, and Safari. It is also suitable for iOS and Android mobile devices.

Results

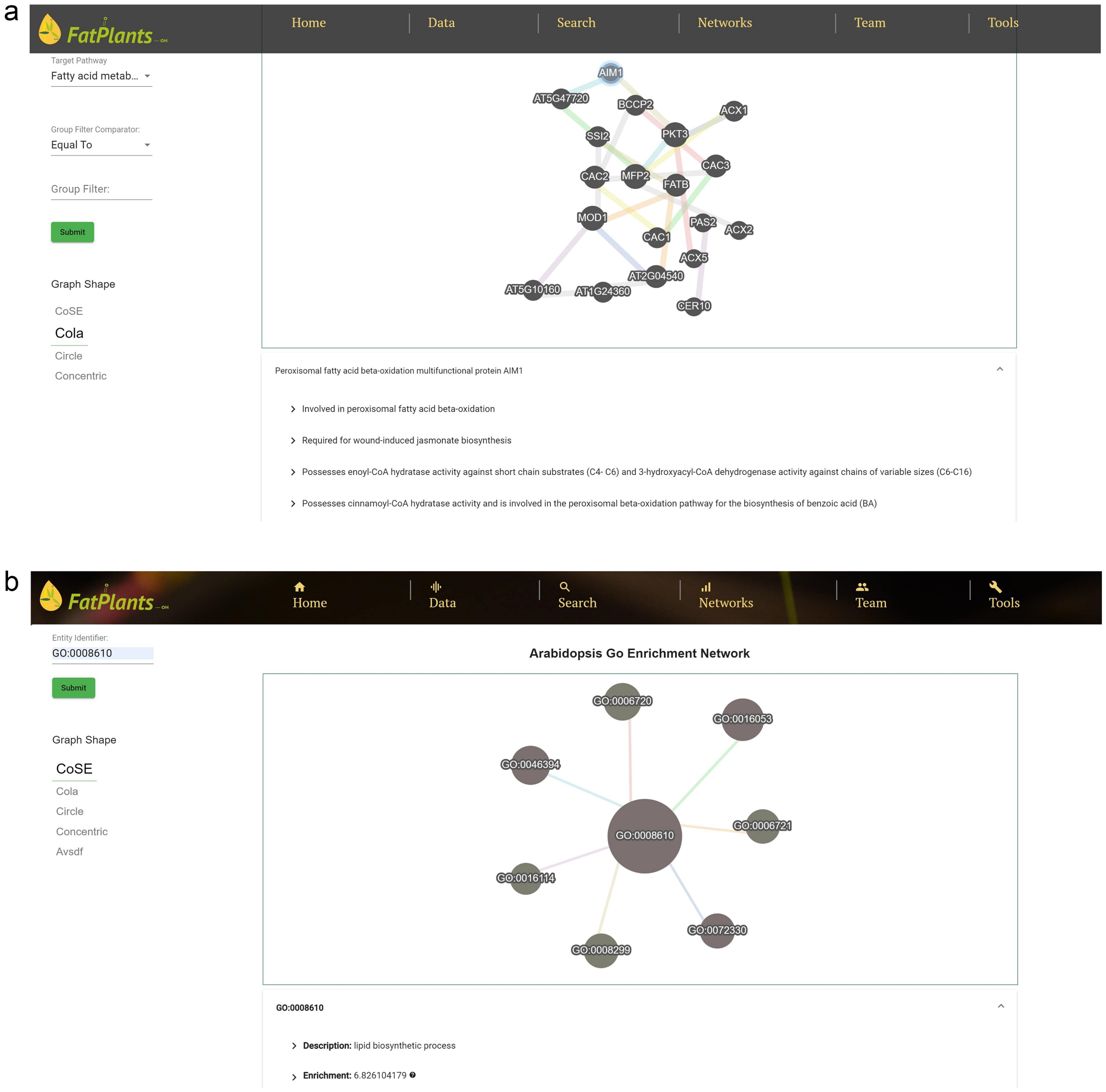
As an overview of the main content at our site, we include 2341 acyl-lipid metabolism proteins for *A. thaliana*, 1232 for *G. max*, and 623 for *C. sativa*. These data have extensive information about their properties, functions, descriptions, and modification domains. Chemical information is provided for a total of 495 fatty acids. Twelve PPI networks of Arabidopsis and 10 GO-enrichment networks can be visualized based on different metabolic pathways. Currently, 15 auto-recognized pathways have been retrieved from the latest published papers. Since the search function is linked to the KEGG and Protein Data Bank databases, users can study additional data via FatPlants.

Web interface and usage

FatPlants offers a user-friendly web interface, enabling users to conveniently browse, search, and retrieve data on fatty-acid-related proteins. Six functional header menus are situated on the top navigation bar—'Home', 'Data', 'Search', 'Networks', and 'Tool'—designed to facilitate easy access to the database. The 'Home' page provides a concise overview of our database and its three primary functions. Users can explore the principal datasets via the 'Browse' menu. On the main data page (Fig. 2a), FatPlants offers a selection panel for users to switch between species and fatty acids. Leveraging the Angular framework, we developed an instant filter search function within data tables. Users can search for any protein by submitting identifiers, gene names, or gene descriptions. Each protein is linked to its corresponding database using unique identifiers. For every specific protein, we provide a detailed information page encompassing key identifiers, functional annotation, functional sequence domain, and the protein function description. The 'Ask ChatGPT' button offers an additional avenue for users seeking in-depth knowledge about a specific protein. Furthermore, the 'Homologs' section indicates related homologous proteins in other species (Fig. 2b).

One-stop search

To accommodate the possibility that proteins might have multiple identifiers (UniProt ID, RefSeq ID, etc.), we have constructed an internal identifier mapping dictionary. This dictionary incorporates seven classes of widely used IDs: UniProt ID, Protein Name, Gene Symbol, EMBL ID, EnsemblPlants



Target Pathway: Fatty acid metab...

Group Filter Comparator: Equal To

Group Filter:

Submit

Graph Shape: CoSE, Cola, Circle, Concentric

Peroxisomal fatty acid beta-oxidation multifunctional protein AIM1

- Involved in peroxisomal fatty acid beta-oxidation
- Required for wound-induced jasmonate biosynthesis
- Possesses enoyl-CoA hydratase activity against short chain substrates (C4- C6) and 3-hydroxyacyl-CoA dehydrogenase activity against chains of variable sizes (C6-C16)
- Possesses cinnamoyl-CoA hydratase activity and is involved in the peroxisomal beta-oxidation pathway for the biosynthesis of benzoic acid (BA)

Entity Identifier: GO:0008610

Submit

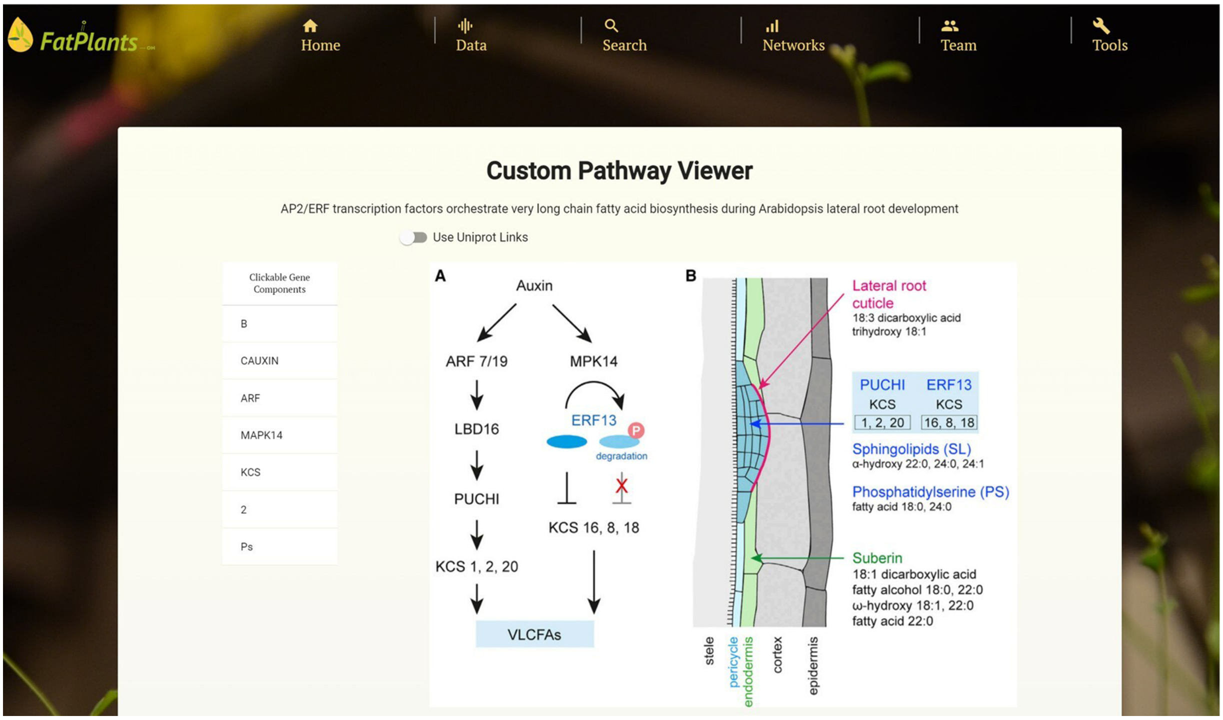
Graph Shape: CoSE, Cola, Circle, Concentric, Avsdf

Arabidopsis Go Enrichment Network

GO:0008610

- **Description:** lipid biosynthetic process
- **Enrichment:** 6.826104179

Figure 4. The FatPlants network viewer. (a) PPI network related to the fatty acid metabolism pathway. (b) GO enrichment network which includes gibberellin-related terms.

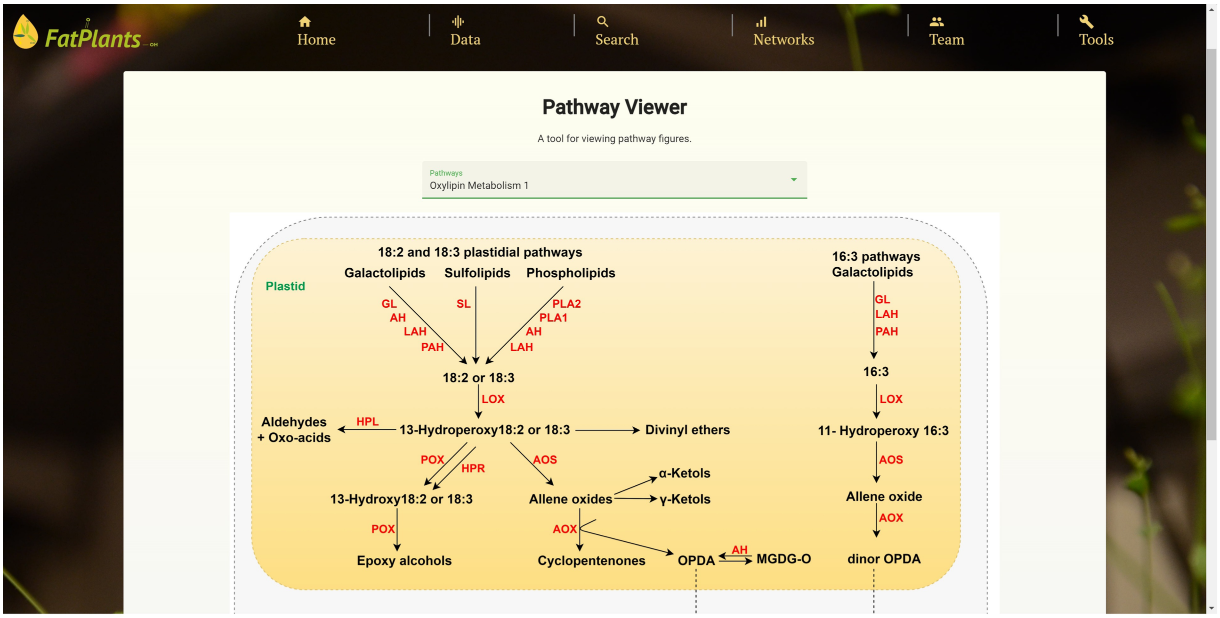


Custom Pathway Viewer

AP2/ERF transcription factors orchestrate very long chain fatty acid biosynthesis during Arabidopsis lateral root development

○ Use Uniprot Links

Figure 5. An example of Custom Pathway Viewer on FatPlants. The original graph is from 'Expression of sets of VLCFA biosynthetic genes is regulated by AP2/ERF transcription' [31].



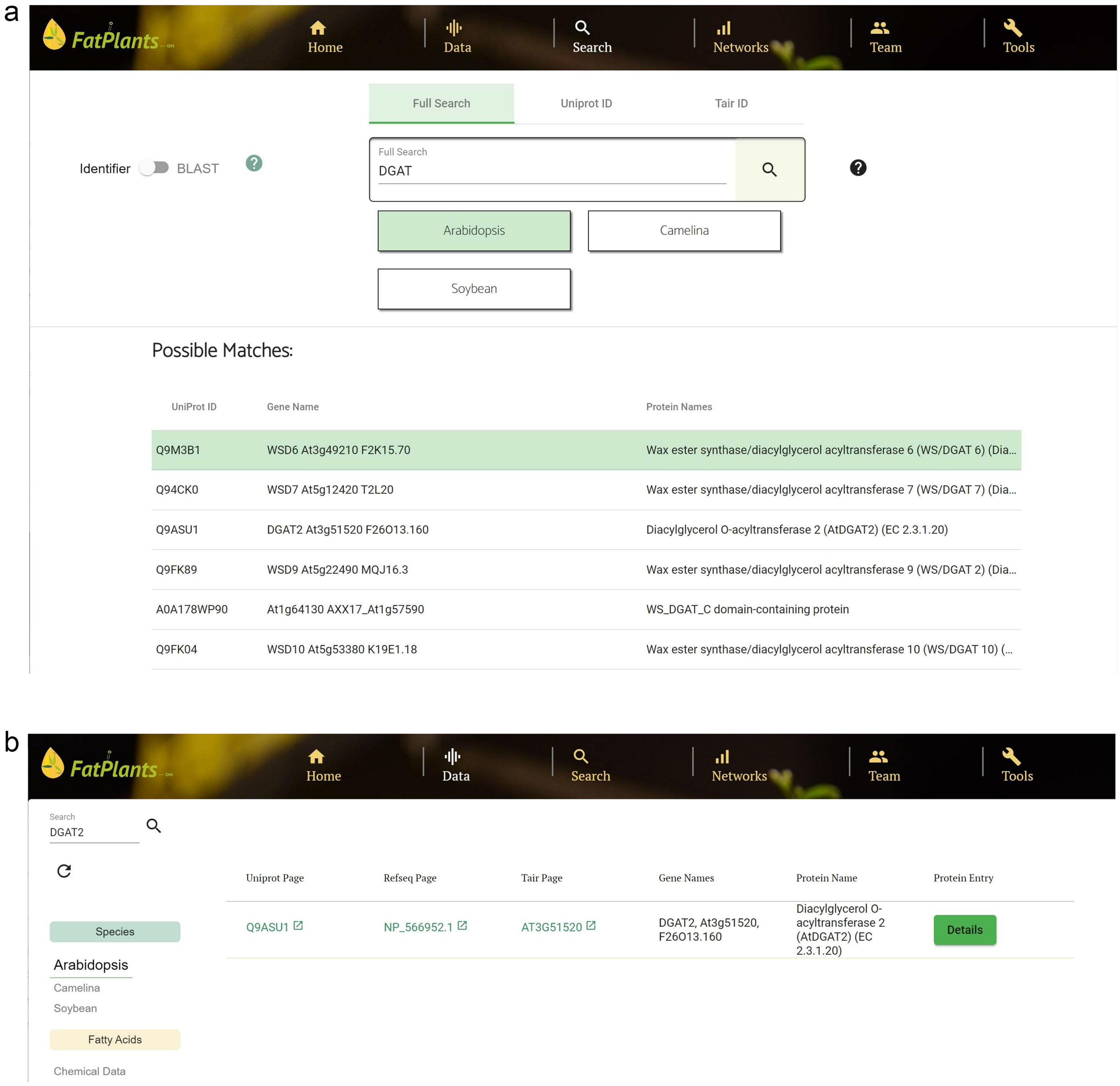
Pathway Viewer

A tool for viewing pathway figures.

Pathways
Oxylipin Metabolism 1

Figure 6. An example of a manually drafted pathway (oxylipin metabolism pathway).

ID, STRING ID, and Locus ID. Any identifier entered will automatically link to a specific protein in our database. The core analysis features of FatPlants include a one-stop-search function based on a sequence similarity search, similar to BLASTP, and a structure similarity search algorithm utilizing the AlphaFold API. Users can effortlessly search for a given protein against the FatPlants fatty-acid-related protein database to find similar sequence or structure results and visualize them in a 3D model (Fig. 3). Moreover, a pathway mapping function is available through the KEGG API. As depicted in Fig. 3a, the one-stop search function accepts both sequences and identifiers as input. The default page displays a summary result, including the most structurally similar 3D model, identifier list, and sequence. Users can toggle between three different result types from the side panel. Figure 3b presents a structure similarity result table generated by Alphafold API. The Blast results display all candidate matches from the FatPlants data collection (Fig. 3c). On the



Full Search: DGAT

Possible Matches:

UniProt ID	Gene Name	Protein Names
Q9M3B1	WSD6 At3g49210 F2K15.70	Wax ester synthase/diacylglycerol acyltransferase 6 (WS/DGAT 6) (Dia...
Q94CK0	WSD7 At5g12420 T2L20	Wax ester synthase/diacylglycerol acyltransferase 7 (WS/DGAT 7) (Dia...
Q9ASU1	DGAT2 At3g51520 F26O13.160	Diacylglycerol O-acyltransferase 2 (AtDGAT2) (EC 2.3.1.20)
Q9FK89	WSD9 At5g22490 MQJ16.3	Wax ester synthase/diacylglycerol acyltransferase 9 (WS/DGAT 2) (Dia...
A0A178WP90	At1g64130 AXX17_At1g57590	WS_DGAT_C domain-containing protein
Q9FK04	WSD10 At5g53380 K19E1.18	Wax ester synthase/diacylglycerol acyltransferase 10 (WS/DGAT 10) (...

Search: DGAT2

Arabidopsis
Camelina
Soybean

Uniprot Page	Refseq Page	Tair Page	Gene Names	Protein Name	Protein Entry
Q9ASU1	NP_566952.1	AT3G51520	DGAT2, At3g51520, F26O13.160	Diacylglycerol O-acyltransferase 2 (AtDGAT2) (EC 2.3.1.20)	Details

Figure 7. A use case for searching diacylglycerol O-acyltransferase 2, with gene symbol DGAT2. (a) Match results using 'DGAT' as the input on the 'Search' page. (b) Search result for 'DGAT2' on the 'Data' page by selecting the target species (Arabidopsis).

pathway mapping result page, graphs depict all pathways involving the input protein (marked by the red boxes), as provided by the KEGG API (Fig. 3d).

Network viewers

The proteins in FatPlants can also be visualized in the context of PPI networks based on the STRING database [27]. We present all PPI networks in terms of their locations within the metabolic network. Figure 4a provides a PPI example in the fatty acid metabolism category. Users can easily browse the PPI network by selecting the desired pathways and clicking a network node to explore the protein's functional description in the bottom table. The fatty-acid-related protein GO enrichment network can be visualized through an enrichment network page to capture the enrichment connection between ontology terms. Users can search any specific protein using different identifiers to retrieve the ontology information. An example of a lipid biosynthetic process involving seven other GO terms enrichment (monocarboxylic acid biosynthetic process, isoprenoid metabolic process, organic acid biosynthetic process, carboxylic acid biosynthetic process, terpenoid biosynthetic process, isoprenoid biosynthetic process, and terpenoid metabolic process) is presented in Fig. 4b as an example.

Custom pathway viewer

Within the custom pathway viewer page, users can manually submit pathway graphs from fatty-acid-related research papers. Leveraging our in-house machine learning image understanding tool [34], these submitted pathway graphs are transformed into interactive pathway maps, where genes/proteins are linked to entries in FatPlants. We currently showcase 15 graphs as trial datasets [29–33]. Figure 5 provides an example of this tool's functionality. Protein elements that can be interacted with are highlighted in red when hovered over, and all recognized proteins are cataloged in a table on the page. Users can access detailed information from the FatPlants database or the comprehensive protein records in the UniProt database. This tool enables FatPlants to integrate the latest fatty acid pathway research, capturing key interactions with crucial proteins. In addition to the machine learning-based pathway graphs, we have a set of manually drafted pathway graphs. It presents a graphical representation inspired by ARALIP [8] (Fig. 6).

A use case example

Diacylglycerol O-acyltransferase 2 (gene symbol DGAT2) is involved in triacylglycerol synthesis. It catalyzes the acylation of the sn-3 hydroxy group of sn-1,2-diacylglycerol using acyl-CoA. To find related information on this gene, a user can perform a partial search on the 'Search' page by entering 'DGAT' to obtain a list of hits, as shown in Fig. 7a. The user can select a hit of interest to explore more information, such as protein structure and similar sequences. The user can also search for DGAT2 on the 'Data' page by selecting the target species (Arabidopsis in this case), which leads to a unique hit, as shown in Fig. 7b.

Conclusions and future work

FatPlants is a comprehensive and systematic fatty-acid-related protein database resource. It can help users understand plant oil synthesis and breeders improve oil content. Users can also leverage AI assistance to gain deeper insights into specific proteins. FatPlants provides several network-based data representations and visualization tools to explore fatty-acid-related protein functions and relationships. By integrating different tools, the one-stop search can help users retrieve the corresponding information efficiently and comprehensively.

For future work, this data repository and a suite of visualization and analysis tools will be continuously updated with new data collected from oilseed research, particularly for important emerging crops such as Camelina and pennycress, two related Brassicaceae species that are not as well-developed as Arabidopsis. User feedback will guide new analysis or visualization tools to explore the fatty-acid-related protein data. To take advantage of our in-house image understanding tool, a Web-based pipeline will be developed for users to submit fatty-acid-related pathway figures. The pipeline will automatically parse the figures into pathway components and their relationships. In addition, we are implementing an internal API to collect the latest plant lipid publications on PubMed so that FatPlants can be updated accordingly. We will also use some large language models, such as ChatGPT, to help identify more relevant data/knowledge sources for FatPlants.

Acknowledgements

The authors would like to thank Chengzi Cao for her graphic design of website Logo, Ye Han for her technical assistance in data analysis, and Zhi Zhang, Joshua Thompson, and Jonah Marz for the prototype development of the website.

Conflict of interest

None declared.

Funding

This work was supported by the National Science Foundation (grant IOS-1829365), the Department of Energy (grant DE-SC0023142), and the National Institute of Health (grant R01LM013392).

Data Availability

The FatPlants database is publicly available at https://www.fatplants.net/.

References

1. Thelen JJ, Ohlrogge JBJME. Metabolic engineering of fatty acid biosynthesis in plants. *Metab Eng* 2002;**4**:12–21. https://doi.org/10.1006/mben.2001.0204
2. Carlsson AS, Yilmaz JL, Green AG *et al.* Replacing fossil oil with fresh oil - with what and for what? *Eur J Lipid Sci Technol* 2011;**113**:812–31. https://doi.org/10.1002/ejlt.201100032
3. Bates PD. Understanding the control of acyl flux through the lipid metabolic network of plant oil biosynthesis. *Biochim Biophys Acta* 2016;**1861**:1214–25. https://doi.org/10.1016/j.bbalip.2016.03.021
4. Fabre F, Planchon CJPS. Nitrogen nutrition, yield and protein content in soybean. *Plant Sci* 2000;**152**:51–58.
5. Nakasathien S, Israel DW, Wilson RF *et al.* Regulation of seed protein concentration in soybean by supra-optimal nitrogen supply. *Crop Sci* 2000;**40**:1277–84. https://doi.org/10.2135/cropsci2000.4051277x
6. Pipolo AE, Sinclair TR, Camara GMS. Protein and oil concentration of soybean seed cultured *in vitro* using nutrient solutions of differing glutamine concentration. *Ann Appl Biol* 2004;**144**:223–27. https://doi.org/10.1111/j.1744-7348.2004.tb00337.x
7. Hernandez-Sebastia C, Marsolais F, Saravitz C *et al.* Free amino acid profiles suggest a possible role for asparagine in the control of storage-product accumulation in developing seeds of low- and high-protein soybean lines. *J Exp Bot* 2005;**56**:1951–63. https://doi.org/10.1093/jxb/eri191
8. Li-Beisson Y, Shorrosh B, Beisson F *et al.* Acyl-lipid metabolism. *Arabidopsis Book* 2013;**11**:e0161. https://doi.org/10.1199/tab.0161
9. Allen DK. Assessing compartmentalized flux in lipid metabolism with isotopes. *Biochim Biophys Acta* 2016;**1861**:1226–42. https://doi.org/10.1016/j.bbalip.2016.03.017
10. Bates PD, Johnson SR, Cao X *et al.* Fatty acid synthesis is inhibited by inefficient utilization of unusual fatty acids for glycerolipid assembly. *Proc Natl Acad Sci* 2014;**111**:1204–09. https://doi.org/10.1073/pnas.1318511111
11. Allen DK, Young JD. Carbon and nitrogen provisions alter the metabolic flux in developing soybean embryos. *Plant Physiol* 2013;**161**:1458–75. https://doi.org/10.1104/pp.112.203299
12. Yasugi E, Watanabe K. LIPIDBANK for Web, the newly developed lipid database. *Tanpakushitsu Kakusan Koso* 2002;**47**:837–41.
13. Aitzetmüller K, Matthäus B, Friedrich H. A new database for seed oil fatty acids — the database SOFA. *Eur J Lipid Sci Technol* 2003;**105**:92–103. https://doi.org/10.1002/ejlt.200390022
14. Caffrey M, Hogan J. LIPIDAT: a database of lipid phase transition temperatures and enthalpy changes. DMPC data subset analysis.

Chem Phys Lipids 1992;**61**:1–109. https://doi.org/10.1016/0009-3084(92)90002-7

15. Cotter D, Maer A, Guda C *et al*. LMPD: LIPID MAPS proteome database. *Nucleic Acids Res* 2006;**34**:D507–510. https://doi.org/10.1093/nar/gkj122
16. Ohlrogge J, Thrower N, Mhaske V *et al*. PlantFAdb: a resource for exploring hundreds of plant fatty acid structures synthesized by thousands of plants and their phylogenetic relationships. *Plant J* 2018;**96**:1299–308. https://doi.org/10.1111/tpj.14102
17. Dormann P. Plant lipid databases. *Methods Mol Biol* 2021;**2295**:441–54.
18. UniProt C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–D489. https://doi.org/10.1093/nar/gkaa1100
19. Swarbreck D, Wilks C, Lamesch P *et al*. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res* 2008;**36**:D1009–1014. https://doi.org/10.1093/nar/gkm965
20. Joshi T, Fitzpatrick MR, Chen S *et al*. Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. *Nucleic Acids Res* 2014;**42**:D1245–1252. https://doi.org/10.1093/nar/gkt905
21. Zeng S, Lyu Z, Narisetti SRK *et al*. Knowledge Base Commons (KBCommons) v1. 1: a universal framework for multi-omics data integration and biological discoveries. *BMC genomics* 2019;**20**:1–16.
22. Fahy E, Sud M, Cotter D *et al*. LIPID MAPS online tools for lipid research. *Nucleic Acids Res* 2007;**35**:W606–612. https://doi.org/10.1093/nar/gkm324
23. Gomez-Cano F, Carey L, Lucas K *et al*. CamRegBase: a gene regulation database for the biofuel crop, *Camelina sativa*. *Database (Oxford)* 2020;**2020**:baaa075.
24. Lowe DM, Corbett PT, Murray-Rust P *et al*. Chemical name to structure: OPSIN, an open source solution. *J Chem Inf Model* 2011;**51**:739–53. https://doi.org/10.1021/ci100384d
25. Jumper J, Evans R, Pritzel A *et al*. Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**:583–89 .
26. Burley SK, Bhikadiya C, Bi C *et al*. RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* 2021;**49**:D437–D451. https://doi.org/10.1093/nar/gkaa1038
27. Szklarczyk D, Gable AL, Lyon D *et al*. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–D613. https://doi.org/10.1093/nar/gky1131
28. Ashburner M, Ball CA, Blake JA *et al*. Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–29. https://doi.org/10.1038/75556
29. Zhu L, He S, Liu Y *et al*. Arabidopsis FAX1 mediated fatty acid export is required for the transcriptional regulation of anther development and pollen wall formation. *Plant Mol Biol* 2020;**104**:187–201. https://doi.org/10.1007/s11103-020-01036-5
30. Bates PD, Stymne S, Ohlrogge J. Biochemical pathways in seed oil synthesis. *Curr Opin Plant Biol* 2013;**16**:358–64. https://doi.org/10.1016/j.pbi.2013.02.015
31. Guyomarc'h S, Boutte Y, Laplaze L. AP2/ERF transcription factors orchestrate very long chain fatty acid biosynthesis during *Arabidopsis* lateral root development. *Mol Plant* 2021;**14**:205–07. https://doi.org/10.1016/j.molp.2021.01.004
32. Zhang QY, Yu R, Xie LH *et al*. Fatty acid and associated gene expression analyses of three tree peony species reveal key genes for alpha-linolenic acid synthesis in seeds. *Front Plant Sci* 2018;**9**:106. https://doi.org/10.3389/fpls.2018.00106
33. Regmi A, Shockey J, Kotapati HK *et al*. Oil-producing metabolons containing DGAT1 use separate substrate pools from those containing DGAT2 or PDAT. *Plant Physiol* 2020;**184**:720–37. https://doi.org/10.1104/pp.20.00461
34. He F, Wang D, Innokenteva Y *et al*. Extracting molecular entities and their interactions from pathway figures based on deep learning. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 397–404, Niagara Falls, New York, Unite State, September 7 - 10, 2019, 2019.
35. Franz M, Lopes CT, Huck G *et al*. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* 2016;**32**:309–11. https://doi.org/10.1093/bioinformatics/btv557

Database, 2024, 00, baae074, DOI: https://doi.org/10.1093/database/baae074, Original article