Plant Communications
Correspondence

CellPress
Partner Journal

EupDB: An integrative and comprehensive functional genomics data hub for Euphorbiaceae plants

Dear Editor,

Euphorbiaceae is one of the largest plant families, with nearly 218 genera and 6745 species, and includes many economically and medicinally important species. Specifically, cassava (*Manihot esculenta*) is well known as the "king of starches" and is the fourth most important food source in the tropics; rubber tree (*Hevea brasiliensis*) is a principal source of natural rubber for over 50 000 industrial products worldwide; castor bean (*Ricinus communis*) is a globally important non-edible oilseed crop and the only commercial source of industrially valuable hydroxyl fatty acids; physic nut (*Jatropha curcas*) is currently the most promising biodiesel plant; and tung tree (*Vernicia fordii*) is an economically important woody tree and the sole source of tung oil, which is widely used as an industrial drying ingredient. These species have shown great promise for agriculture and industry, and their potential productivity and quality-related functional genomic elements need to be explored.

Over the past years, a large number of studies related to Euphorbiaceae plants have been performed and have generated a depth of sequencing data, including genome sequences, transcriptomes, population genetic variations, QTLs, and GWAS and epigenome data in individual species (Tang et al., 2016; Chen et al., 2020; Xu et al., 2021; Cao et al., 2022; Phumichai et al., 2022; Zhao et al., 2022; Cheng et al., 2023). However, limited efforts have been made to decipher the evolution and regulation of important economic traits through comparative genomic analyses, and a functional genomic database for multiple Euphorbiaceae species is not yet available. To fill this gap, we developed an integrative and comprehensive family-level database for Euphorbiaceae plants (EupDB: http://eupdb.liu-lab.com/) that includes multi-omics resources for six species (*M. esculenta*, *H. brasiliensis*, *R. communis*, *J. curcas*, *V. fordii*, and *Euphorbia lathyris*), as well as many web-based bioinformatics tools for omics data analysis. The overall workflow of EupDB is summarized in Figure 1A, and more details of the data collection and analysis processes are provided in the Supplemental Tables and online documentation.

EupDB includes 13 published genomes, 1004 transcriptomic datasets, 708 miRNAs and their corresponding targets, methylomes, large-scale population genetic variations, and GWAS data for important economic traits, representing the most comprehensive Euphorbiaceae database at present. We re-annotated and updated the protein-coding genes for all included species by searching against up-to-date databases (e.g., EggNOG, KEGG, GO, iTAK, and InterPro). The vast majority of genes (70.7%–89.9%) were re-annotated, and 54 kinds of transcription factors (TFs) and 24 kinds of transcription regulators (TRs) were predicted (Figure 1B). We also developed many useful bioinformatics tools for data analysis and visualization, including (1) an advanced search engine in the "Search" module, through which data can be retrieved by keywords, genomic location, TFs, and TRs (Figure 1C), (2) a Jbrowse genome and epigenome browser, (3) a web-based BLAST analysis server, and tools for (4) GO/KEGG enrichment analysis, (5) genomic synteny analysis, (6) pathway and interaction network prediction, (7) gene expression profiling, and (8) GWAS/QTL analysis. For the convenience of users, these data and tools have been divided into multiple modules, which are introduced below. In addition, a case study is provided for reference in File S1.

In the "Expression" module, we developed a user-friendly web interface to explore gene expression profiles, as well as DNA methylomes and miRNAs that affect gene expression regulation (Figure 1D). We integrated various transcriptome data and developed a unified pipeline to normalize all datasets and recalculate gene expression levels. In brief, we used Trim Galore for quality control, HISAT2 for read mapping, and StringTie for transcript assembly and gene expression quantification. Users can specify a list containing gene IDs and tissues/conditions of interest and obtain a corresponding gene expression matrix and heatmap. In addition, we collected and re-analyzed all currently available Euphorbiaceae-related DNA methylation data and provided Jbrowse for visualization. In addition, we collected miRNAs from the PmiREN database and the literature (Zeng et al., 2010; Xu et al., 2013), and we predicted their target genes with psRNATarget. For miRNA-mediated post-transcriptional regulation, users can select a given miRNA and obtain its corresponding target genes and binding sequences.

In developing the first family-level database for Euphorbiaceae plants, we aimed to construct an efficient platform for comparative analysis of genomes, genes, and pathways across different species. We therefore developed the "Syntenic" module, which includes searches for orthologous groups and syntenic genes, as well as a tool for gene structure visualization (Figure 1E). In total, 35 204 orthogroups were identified between Euphorbiaceae species and *Arabidopsis thaliana* using Orthofinder and were further confirmed by bidirectional sequence BLAST. We also performed genome-wide collinear block analysis to identify collinear blocks between species pairs. In this module, gene structures, homologs, and collinear blocks can easily be displayed and retrieved by users. On the basis of

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.

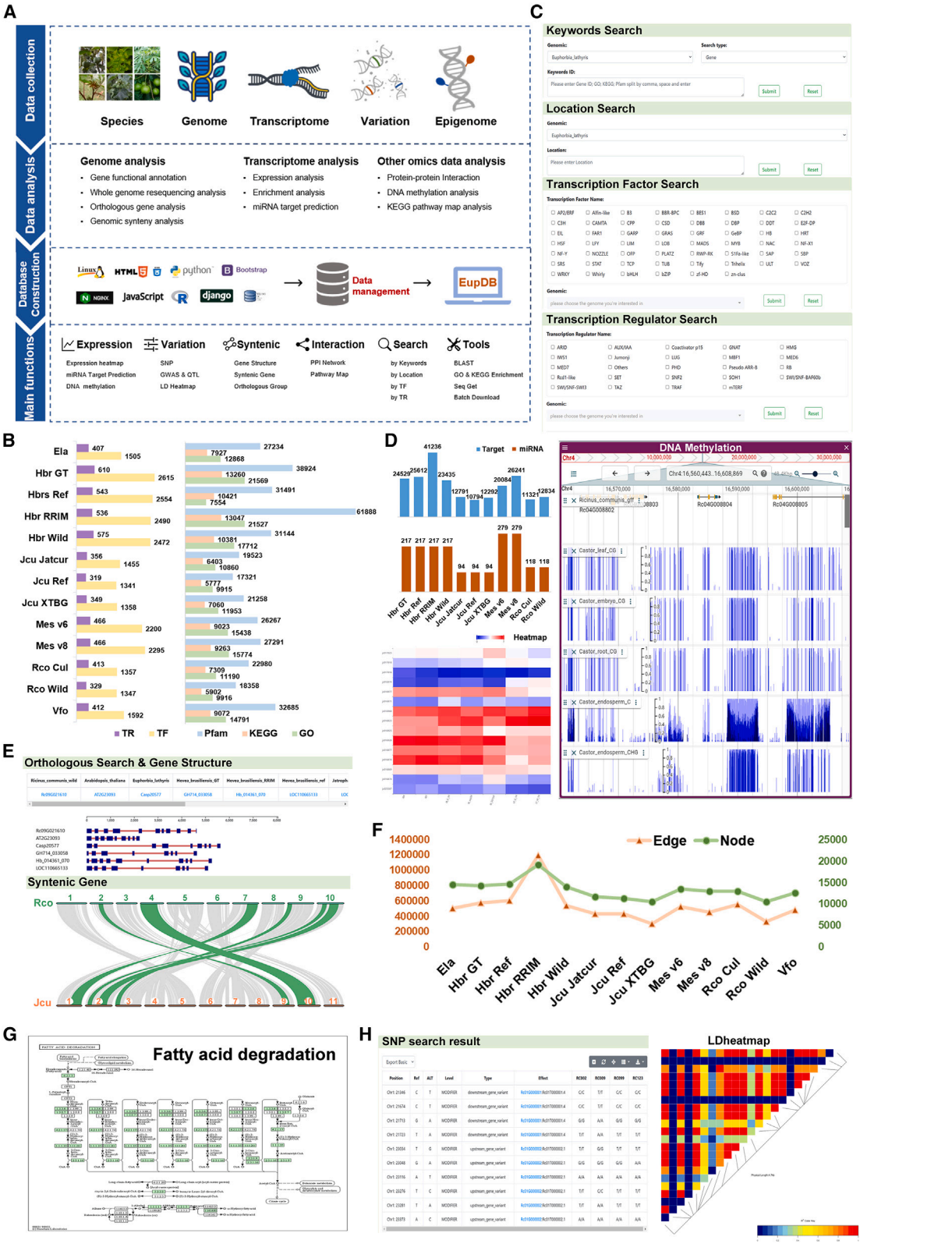

Figure 1. Structure of EupDB, an integrative and comprehensive functional genomics data hub for Euphorbiaceae plants

(A) Workflow of EupDB with data collection, data analysis, database construction, and main functions.

(B) Statistics of different types of functional annotations in EupDB, including transcription regulator (TR), transcription factor (TF), Pfam, GO, and KEGG.

(C) Query interface with rich options in the "Search" module.

(legend continued on next page)

the "Syntenic" module, known protein interactions (retrieved from the STRING database) and important pathways were established for Euphorbiaceae plants from related information in the model plant *Arabidopsis*, and they are stored and presented in the "Interaction" module. We constructed protein interaction networks for Euphorbiaceae plants, which contained 10 529 to 19 157 nodes and 301 621 to 1 199 071 edges (Figure 1F) and can be viewed in an interactive mode on the "PPI network" page. We also generated 19 maps of important KEGG metabolic pathways, including starch synthesis, fatty acid synthesis, lipid metabolism, and other biosynthetic pathways. These pathways are displayed on the "Pathway" page, where related genes are highlighted in green and shown in a table (Figure 1G).

We constructed the "Variation" module (Figure 1H) to explore associations between genetic variations and key economic traits. We collected population resequencing data from 45 castor bean, 43 cassava, and six physic nut core germplasms (Li et al., 2021; Xu et al., 2021) and identified 3 569 884, 27 032 954, and 345 371 high-quality SNP loci in the castor bean, cassava, and physic nut populations, respectively. These SNPs were projected to the corresponding genome to estimate their effects on gene structure and function using snpEff. We also integrated GWAS and QTL analysis (Garcia-Oliveira et al., 2020; Xu et al., 2021; Phumichai et al., 2022; Cheng et al., 2023) into this module to identify candidate loci or genes that may be closely linked to key agronomic traits. In addition, we provided the linkage SNP loci associated with 10 agronomic traits in castor bean, 36 agronomic traits in cassava, and one agronomic trait in rubber tree. Specific candidate SNPs associated with important agronomic traits (e.g., seed size, starch content, and pest and disease resistance) were identified. Overall, this module enables users to obtain all SNP information at a specific genomic location or within a region, search GWAS and QTL information related to different agronomic traits, and calculate and visualize pairwise linkage disequilibrium within a given region.

In summary, we present the first family-level database for Euphorbiaceae functional genomics, EupDB, in which the latest and most abundant data, including genomes, transcriptomes, population resequencing data, genetic variations, GWAS results, and QTL loci, are integrated and analyzed. EupDB provides excellent resources and powerful web-based tools for comparative investigation of genes or important pathways in different genomes and also enables in-depth exploration of the potential roles of genes or loci in evolution, adaptation, and regulation of key agronomic traits. In the future, we will update EupDB by adding and analyzing new omics data, such as genomes, pangenomes, metabolomes, phenomes, and epigenomes, and develop more user-friendly interfaces and useful tools such as gene-editing modules. We believe that the modules and powerful tools in EupDB will not only facilitate comparative and functional genome research in the Euphorbiaceae family but also provide a comprehensive platform for genetic improvement of important economic traits such as starch, rubber, and seed-oil quality and yield.

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

FUNDING

This research was funded by the National Natural Science Foundation of China (grant nos. 31970609, 31970341, 32261143461); the Youth Innovation Promotion Association of CAS (2020389); the "Top Talents Program in Science and Technology" from Yunnan Province; the Yunnan Young & Elite Talents Project (YNWR-QNBJ-2020-286); the program for Digitalization, Development and Application Of Biotic Resources (202002AA100007); and a Startup Fund from Xishuangbanna Tropical Botanical Garden. The funders had no role in study design, data collection, analysis and interpretation, or preparation of the manuscript. Publication costs were funded by the National Natural Science Foundation of China (grant no. 31970609).

AUTHOR CONTRIBUTIONS

C.L. and W.X. conceived and supervised the project. C.L. and Jiazhi Liu designed the database structure. Jiazhi Liu and Y.L. collected and analyzed the datasets, built the database, and wrote the original draft. C.L., W.X., and Jing Li reviewed and edited the manuscript and the database. W.C. helped with database construction. B.P. helped with data collection. A.L. and Z.-F.X. helped with the project conception process. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS

No conflict of interest is declared.

Received: March 25, 2023
Revised: July 7, 2023
Accepted: September 1, 2023
Published: September 4, 2023

Jiazhi Liu[1,2,8], Yan Li[1,2,8], Jing Li[1], Wen Chen[3], Bangzhen Pan[1], Aizhong Liu[4], Zeng-Fu Xu[5,6], Wei Xu[7,] and Changning Liu[1,*]*

[1]CAS Key Laboratory of Tropical Plant Resources and Sustainable Use, Yunnan Key Laboratory of Crop Wild Relatives Omics, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming 650223, China
[2]College of Life Sciences, University of Chinese Academy of Sciences, Beijing 100049, China
[3]Key Laboratory of Vascular Biology and Translational Medicine, Medical School, Hunan University of Chinese Medicine, Changsha 410208, China
[4]Key Laboratory for Forest Resources Conservation and Utilization in the Southwest Mountains of China, Ministry of Education, Southwest Forestry University, Kunming 650224, China
[5]State Key Laboratory for Conservation and Utilization of Subtropical Agro-Bioresources, College of Forestry, Guangxi University, Nanning 530003, China
[6]Key Laboratory of National Forestry and Grassland Administration on Cultivation of Fast-Growing Timber in Central South China, College of Forestry, Guangxi University, Nanning 530003, China

(D) Statistics on the numbers of miRNA target genes and miRNAs, an example of an expression heatmap, and a DNA methylation browser in the "Expression" module.
(E) Orthologous group search, gene structure visualization, and syntenic gene search in the "Syntenic" module.
(F) Statistics on the numbers of edges and nodes in the protein interaction networks of Euphorbiaceae plants.
(G) Fatty acid degradation—an example of a KEGG pathway map.
(H) SNPs and GWAS & QTL analysis in the "Variation" module.
Ela, *Euphorbia lathyris*; Hbr, *Hevea brasiliensis*; Jcu, *Jatropha curcas*; Mes, *Manihot esculenta*; Rco, *Ricinus communis*; Vfo, *Vernicia fordii*.

[7]Germplasm Bank of Wild Species, Yunnan Key Laboratory of Crop Wild Relatives Omics, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650201, Yunnan, China
[8]These authors contributed equally to this article.
***Correspondence: Wei Xu (xuwei@mail.kib.ac.cn), Changning Liu (liuchangning@xtbg.ac.cn)**
https://doi.org/10.1016/j.xplc.2023.100683

REFERENCES

Cao, Y., Li, Y., Wang, L., Zhang, L., and Jiang, L. (2022). Evolution and function of ubiquitin-specific proteases (UBPs): Insight into seed development roles in plants. Int. J. Biol. Macromol. **221**:796–805.

Chen, M.-S., Niu, L., Zhao, M.-L., Xu, C., Pan, B.-Z., Fu, Q., Tao, Y.-B., He, H., Hou, C., and Xu, Z.-F. (2020). De novo genome assembly and Hi-C analysis reveal an association between chromatin architecture alterations and sex differentiation in the woody plant *Jatropha curcas*. GigaScience **9**, giaa009. giaa009.

Cheng, H., Song, X., Hu, Y., Wu, T., Yang, Q., An, Z., Feng, S., Deng, Z., Wu, W., Zeng, X., et al. (2023). Chromosome-level wild *Hevea brasiliensis* genome provides new tools for genomic-assisted breeding and valuable loci to elevate rubber yield. Plant Biotechnol. J. **21**:1058–1072.

Garcia-Oliveira, A.L., Kimata, B., Kasele, S., Kapinga, F., Masumba, E., Mkamilo, G., Sichalwe, C., Bredeson, J.V., Lyons, J.B., Shah, T., et al. (2020). Genetic analysis and QTL mapping for multiple biotic stress resistance in cassava. PLoS One **15**, e0236674.

Li, C., Tian, D., Tang, B., Liu, X., Teng, X., Zhao, W., Zhang, Z., and Song, S. (2021). Genome Variation Map: a worldwide collection of genome variations across multiple species. Nucleic Acids Res. **49**:D1186–D1191.

Phumichai, C., Aiemnaka, P., Nathaisong, P., Hunsawattanakul, S., Fungfoo, P., Rojanaridpiched, C., Vichukit, V., Kongsil, P., Kittipadakul, P., Wannarat, W., et al. (2022). Genome-wide association mapping and genomic prediction of yield-related traits and starch pasting properties in cassava. Theor. Appl. Genet. **135**:145–171.

Tang, C., Yang, M., Fang, Y., Luo, Y., Gao, S., Xiao, X., An, Z., Zhou, B., Zhang, B., Tan, X., et al. (2016). The rubber tree genome reveals new insights into rubber production and species adaptation. Nat. Plants **2**:16073.

Xu, W., Cui, Q., Li, F., and Liu, A. (2013). Transcriptome-wide identification and characterization of microRNAs from castor bean (*Ricinus communis* L.). PLoS One **8**, e69995.

Xu, W., Wu, D., Yang, T., Sun, C., Wang, Z., Han, B., Wu, S., Yu, A., Chapman, M.A., Muraguri, S., et al. (2021). Genomic insights into the origin, domestication and genetic basis of agronomic traits of castor bean. Genome Biol. **22**:113–127.

Zeng, C., Wang, W., Zheng, Y., Chen, X., Bo, W., Song, S., Zhang, W., and Peng, M. (2010). Conservation and divergence of microRNAs and their functions in Euphorbiaceous plants. Nucleic Acids Res. **38**:981–995.

Zhao, W., Peng, J., Wang, F., Tian, M., Li, P., Feng, B., Yin, M., Xu, Y., Xue, J.-Y., Xue, J., et al. (2022). Integrating metabolomics and transcriptomics to unveil the spatiotemporal distribution of macrocyclic diterpenoids and candidate genes involved in ingenol biosynthesis in the medicinal plant Euphorbia lathyris L. Ind. Crop. Prod. **184**, 115096.