Data in Brief 43 (2022) 108426

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Transcriptome profiling dataset of different developmental stage flowers of soybean (*Glycine max*)

Eszter Virág[a,b,*], Géza Hegedűs[a,c], Barbara Kutasy[d], Kincső Decsi[d]

[a] *EduCoMat Ltd., 8360 Keszthely, Iskola u. 12/A., Hungary*
[b] *Department of Molecular Biotechnology and Microbiology, Institute of Biotechnology, Faculty of Science and Technology, University of Debrecen, 4132 Debrecen, Egyetem tér 1., Hungary*
[c] *Department of Information Technology and its Applications, Faculty of Information Technology, University Center Zalaegerszeg, University of Pannonia, 8900 Zalaegerszeg, Gasparich Márk u. 18., Hungary*
[d] *Department of Plant Physiology and Plant Ecology, Hungarian University of Agriculture and Life Sciences, Georgikon Campus Keszthely, 8360 Keszthely, Festetics Gy. u.7., Hungary*

ARTICLE INFO

Article history:
Received 14 May 2022
Revised 19 June 2022
Accepted 22 June 2022
Available online 27 June 2022

Keywords:
Glycine max
Soybean
Flower
Development
Transcriptome

ABSTRACT

The dynamic of flower development is a key agronomic characteristic affecting soybean yield. RNA-seq dataset of field-cultivated soybean flowers in four developmental stages including flower buds, and early, mature, and overblown stage flowers are reported in this paper. Gene Expression (Gex) library construction and Illumina NextSeq550 sequencing were carried out to produce 86 bp long forward reads. Reads were preprocessed and deposited in the National Center for Biotechnology Information Sequence Read Archive (NCBI SRA) database. These SRA depositions are under the BioProject accession: PRJNA807844. A reference transcriptome dataset was *de novo* assembled using these SRA reads. Annotation, differential expression, and gene set enrichment analyses were performed and deposited in the Mendeley Data.

* Corresponding author's email address and Twitter handle
E-mail address: virag@educomat.hu (E. Virág).

https://doi.org/10.1016/j.dib.2022.108426

Specifications Table

Subject	Plant Science: Plant Physiology
Specific subject area	Genome-wide expression profiling was performed and differentially expressed genes were determined during the floral development of soy plants (*Glycine max*).
Type of data	Table Database record Figure
How the data were acquired	Floral samples were collected from field-cultivated soybean plants during the period 10-16 June 2021 in Tata, Hungary. Approximately 50 mg of plant tissues were used to prepare Next Generation Sequencing (NGS) libraries. NextSeq550 sequencing was performed, to produce 15-16M 86 bp long reads in each sample, approximately. Reads were pre-processed and assembled. A transcriptome dataset was reconstructed and genome-wide expression profiles were determined using combined and separated read sets per all samples. Pairwise differential expression with gene set enrichment analysis (GSEA) and differentially expressed genes (DEGs) were annotated with gene ontology (GO) terms.
Data format	Raw Analyzed Filtered
Description of data collection	Four developmental stage flowers including flower buds, and early, mature and overblown flowers of soybean plants were collected from field populations during the period 10-16 June 2021 in Tata, Hungary. Plant materials were stored in DNA/RNA Shield (Zymo research) at -25°C until sequencing.
Data source location	• EduCoMat Ltd • Keszthely • Hungary
Data accessibility	The BioProject and sequence reads are available in National Center for Biotechnology Information (NCBI) database under the accessions: Repository name: Glycine max flowers raw sequence reads Data identification number: PRJNA807844 Direct link to dataset: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA807844 Repository name: RNA-Seq of Glycine max flower: stage 0 Data identification number: SRR18059506 Direct link to dataset: https://www.ncbi.nlm.nih.gov/sra/?term=SRR18059506 Repository name: RNA-Seq of Glycine max flower: stage 1 Data identification number: SRR18059505 Direct link to dataset: https://www.ncbi.nlm.nih.gov/sra/?term=SRR18059505 Repository name: RNA-Seq of Glycine max flower: stage 2 Data identification number: SRR18059504 Direct link to dataset: https://www.ncbi.nlm.nih.gov/sra/?term=SRR18059504 Repository name: RNA-Seq of Glycine max flower: stage 3 Data identification number: SRR18059503 Direct link to dataset: https://www.ncbi.nlm.nih.gov/sra/?term=SRR18059503 Dataset of transcriptome assembly, annotation, and DEGs are available in Mendeley data: Repository name: Transcriptome profiling dataset of different developmental stage flowers of soybean (Glycine max) Data identification number (DOI): DOI:10.17632/pv2vn2v6bd.2 Direct link to dataset: https://data.mendeley.com/dataset/pv2vn2v6bd/2 *Data in Brief_Virág et al.2022..xlsx including AnnotationTable, CountTable and GSEATable. (In an excel file on the separate worksheet)*

Value of the Data

- The presented genome-wide gene expression dataset contains numerical information on vegetative (leaf tissue) and generative tissue transcripts during the flowering ripening process of soy plants. Therefore, this dataset is useful to help understand the genetic background of this plant's flowering.
- The dataset gap filling in the field of soy flowering because there is no transcriptome analysis comparing the flower stages and leaf samples in the literature.
- Researchers specified for breeding and fundamental research of flowering may use this dataset and benefit.
- This dataset may contribute to understanding differences in physiological processes at different floral stages. With the use of identified transcript sequences, AnnotationTable, and functional information presented here, molecular biological experiments may be easier designed and developed.

1. Data Description

Soybeans are one of the major food crops in the Fabaceae family, capable of forming nitrogen-fixing symbioses with soil microorganisms and thus have been used in sustainable agricultural production for thousands of years. The genetic control of floral transition is a key agronomic factor affecting soybean yield [1]. Despite its important role in nutrition, little new research is known [2] mostly older scientific data on the genetic background of plant flowering regulation [3–5]. During soy cultivation, the treatment strategies to improve soybean yields are most effective if the developmental stages are well identified. These stages may be divided based on plant development into vegetative and reproductive stages. The vegetative stages are determined based on the appearance of fully-developed trifoliate leaves, the reproductive stages begin at flowering and include pod development, seed development, and plant maturation. Flowering maturation was investigated during the full flowering stage of soy (there is an open flower at one of the two uppermost nodes) and distinct flowers including buds (*Glycine max* flower: stage 0), early (*Glycine max* flower: stage 1), mature (*Glycine max flower*: stage 2) and overblown (*Glycine max* flower: stage 3). QuantSeq 3′ mRNA sequencing of these four samples (Fig. 1A-D) was performed to find certain classes of functionally important genes using differentially expressed genes. Using this method, differentially expressed genes were identified in the flower conferring tissue-specific functions. Illumina RNA-seq reads of the four distinct stages of flowers are deposited in the NCBI Sequence Read Archive (SRA) under the accession numbers: SRR18059506, SRR18059505, SRR18059504, SRR18059503. The BioProject can be found under the accession: PRJNA807844. To increase sequencing depths and specify the total number of genes found to be expressed, or differentially expressed we used a combined read set of the presented and earlier reported SRA reads of soybean [6] to create a reference transcript dataset containing 3964 contigs (see DOI:10.17632/pv2vn2v6bd.2), Glycine_max_flower_Trinity.fasta) GO annotation of the entire reference transcript dataset is presented in the AnnotationTable (see DOI:10.17632/pv2vn2v6bd.2), Glycine_max_flower_AnnotationTable.txt). DEGs of the four reproductive stages and one vegetative stage samples (Table 1) were determined based on the transcript abundancies presented in the CountTable (DOI:10.17632/pv2vn2v6bd.2, Glycine_max_flower_CountTable.txt). Distribution of pairwise transcripts using the samples *Glycine max* 525-1 leaf vs. *Glycine max* flower: stage 1 as reference and test are presented in a heatmap based on raw counts and multidimensional scaling (MDS) diagram in Figs. 2 and 3. Gene set enrichment analyses were performed to create a GSEA table (DOI:10.17632/pv2vn2v6bd.2, Glycine_max_flower_GSEA_Table.txt). Gene sets are groups of genes that are functionally related according to current knowledge and are determined as statistically significant differences between the investigated biological samples. This method was used to identify classes of transcripts that are over-represented using CountTable

Fig. 1. Floral samples were collected from all flowering stage plants of soybean. RNA-seq data of flower buds of *Glycine max* flower: stage 0 (A), early flowers of *Glycine max* flower: stage 1 (B), mature flowers of *Glycine max* flower: stage 2 (C) and overblown flowers of *Glycine max* flower: stage 3 (D) are presented.

Table 1
Samples used to create CountTable and determined DEGs.

NCBI accession number	Sample	Raw library size	Maturation stage	Group
SRR16927693	Glycine max 525-1 leaf	3,635,514	0	leaf
SRR18059506	Glycine max flower: stage 0	5,541,655	1	flower
SRR18059505	Glycine max flower: stage 1	4,308,133	2	flower
SRR18059504	Glycine max flower: stage 2	4,112,204	3	flower
SRR18059503	Glycine max flower: stage 3	5,672,747	4	flower

and AnnotationTable. The determined classes may have an association with biological functions like gene ontology terms, pathways, or chromosomal location or regulation. The GSA table contains the following statistics: ES: Reflects the degree to which a gene set is overrepresented at the top or bottom of a ranked list of genes; NES: By normalizing the enrichment score, GSEA accounts for differences in gene set size and correlations between gene sets and the expression dataset; FDR: The estimated probability that a gene set with a given NES represents a false positive finding; Nominal p-value: Estimates the statistical significance of the enrichment score for a single gene set [7]. The Fig. 4. summarizes the total workflow used in this study.

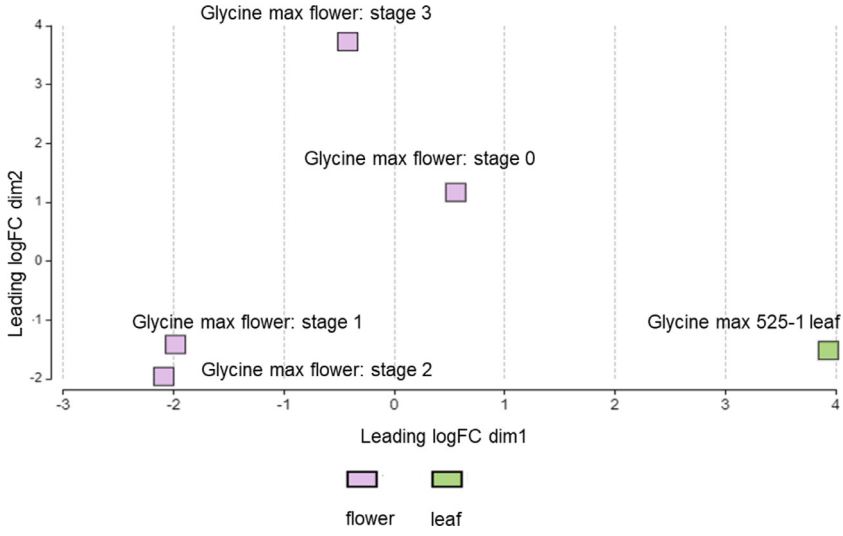

Fig. 2. MDS plot of vegetative and generative samples. The similarity between the samples, where the distances correspond to the leading log-fold change between each pair of samples. The leading log-fold change is the average (square root) of the largest absolute log-fold change between each sample pair.

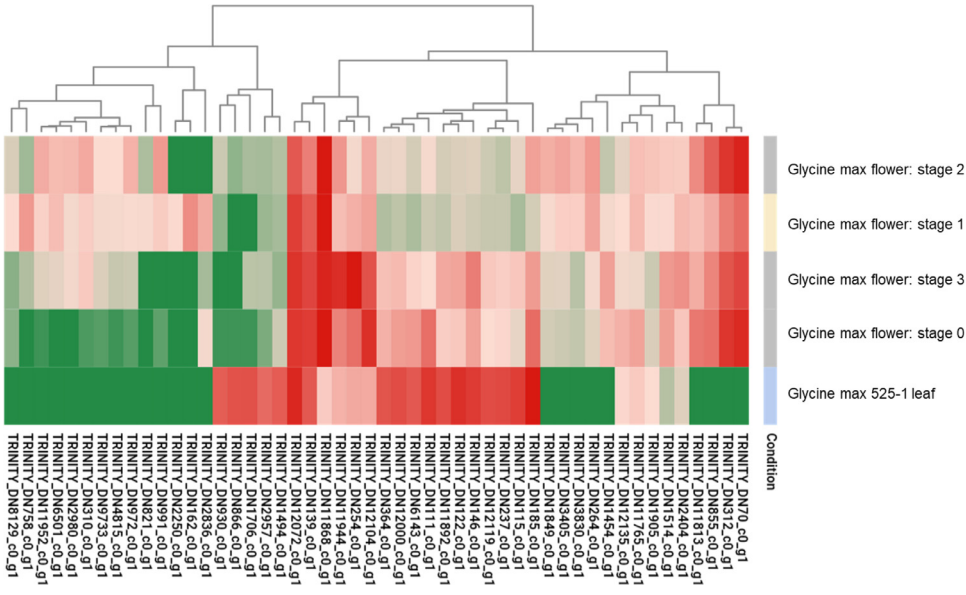

Fig. 3. Heatmap of differentiated genes where vegetative tissue (leaves) as reference and flower stage 1 as test condition were set. Flower Stage 0-3 are *Glycine max* flower: stage 0-3 samples and vegetative tissue leaves correspond to *Glycine max* 525-1 leaf sample. Annotation of transcript IDs see in AnnotationTable (Doi:10.17632/pv2vn2v6bd.2).

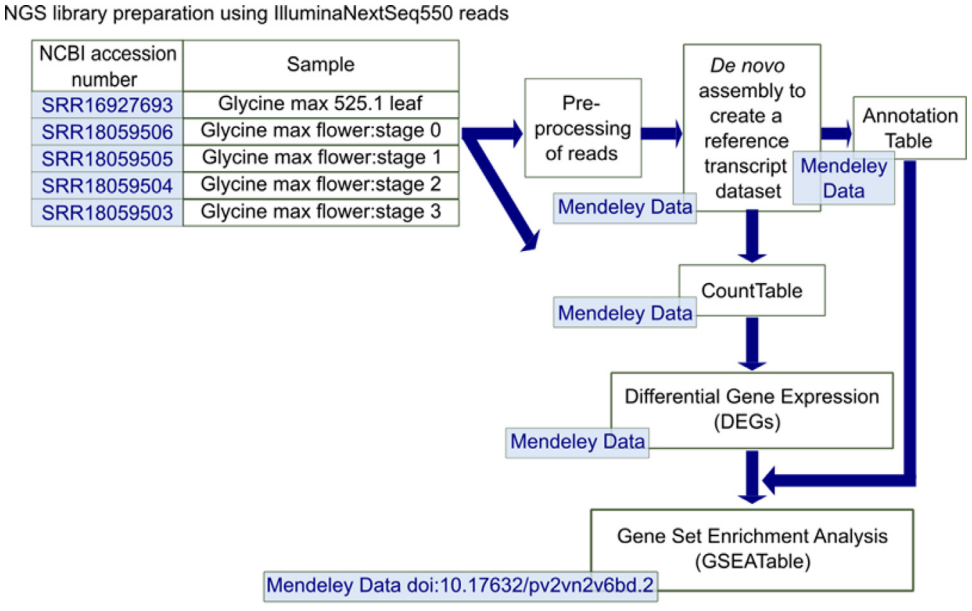

Fig. 4. The workflow of the used methodology of the presented dataset. The flowchart includes the investigated samples and experimental steps with output data accessibility.

2. Experimental Design, Materials and Methods

2.1. Plant materials

Glycine max cv. ES Director plants were cultivated in field conditions. Vegetative and generative samples were taken between 10-16 June 2021 in Tata, Hungary. Sample collection and storage were performed as described earlier by Decsi et. al [6]. The four repetitions of each sample were pooled and sequenced by third-party Xenovea Ltd, Szeged, Hungary.

2.2. Sequencing and bioinformatics

2.2.1. NGS library preparation

NGS libraries of floral samples were performed as described by Hegedűs et al. 2022 [8]. Briefly: QuantSeq 3‘mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen GmbH, Wien, 510 Austria) was applied. Libraries were diluted to 1.8 pM for 1 × 86 bp single-end sequencing with 75-cycle High Output v2 Kit on the NextSeq 550 Sequencing System (Illumina, San Diego, CA, USA) according to the manufacturer’s protocol.

2.2.2. Pre-processing of reads

Filtering of. fastq files including quality control and trimming were performed in a pre-processing step. The QC analysis was carried out by using FastQC software (v0.11.9) [9]. For all the libraries the Phred-like quality scores (Q scores) were set to >30. Poor quality reads were eliminated by using Trimmomatic software (v0.39) [10]. Contamination sequences and N’s were filtered out with a self-developed application GenoUtils as described earlier [11]. Reads passed of pre-processing step were further assembled.

2.2.3. De novo assembly and creating AnnotationTable

Full-length transcriptome assembly of cleaned and combined read sets of five samples (Glycine max flower: stage 0-3 samples and Glycine max 525-1 leaf sample) from shallow RNA-Seq data was performed by using Trinity (v2.13.2) and Bowtie2 (v2.4.5) [12,13]. In the case of Trinity minimum contig length, 250 and K mer coverage 20 were applied. AnnotationTable including functional annotation of the entire *de novo* transcriptome was performed with Gene Ontology (GO) analyses using OmicsBox.BioBam (v2.0) [14] as detailed by Decsi et al. 2022 [6]. In this step, due to the shallow sequencing, the Blastx-fast with a permissive expectation value of 1 was used

2.3. Determination of CountTable

RNA-Seq count data were identified using cleaned SRA reads. Transcript abundances were calculated and written into a CountTable data file. This process was performed by using the HTseq package (v2.0.0) and Bowtie2 (v2.4.5) [13,15].

2.4. Determination of DEGs

Determination of DEGs was performed by pairwise differential expression analysis without replicates using RNA-seq count data applying the software package NOISeq (v2.40.0, Bioconductor project) [16]. Briefly: NOISeq generates a null or noise distribution of numerical changes by contrasting absolute expression differences (D) and multiple change differences (M) considering all of the sample genes under the same conditions. This reference distribution is used to evaluate whether the M and D values were calculated under two conditions for a given gene that is likely to be part of the noise or represent a true differential expression [17,18].

2.5. Determination of GSEATable

The gene set enrichment analysis was performed according to the GSEA computational method defining sets of genes as statistically significant and showing differences between two biological states consistently [7]. The GSEATable was performed by using OmicsBox.BioBam (v2.0).

Ethics Statements

Not relevant for the data.

CRediT Author Statement

Eszter Virág: Conceptualization, Software, Supervision, Writing – original draft; **Géza Hegedűs:** Software, Investigation; **Barbara Kutasy:** Validation, Visualization; **Kincső Decsi:** Visualization, Validation, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The information provided in this article was supported by EduComat Ltd., Hungary.

We express our thanks to Plant-Art Research Ltd., Hungary, and Xenovea Ltd., Hungary to provide field samples and perform NGS sequencing.

References

[1] C.-H. Jung, C.E. Wong, M.B. Singh, P.L. Bhalla, Comparative genomic analysis of soybean flowering genes, PLoS One 7 (6) (2012) e38250.
[2] X. Lin, B. Liu, J.L. Weller, J. Abe, F. Kong, Molecular mechanisms for the photoperiodic regulation of flowering in soybean, J. Integr. Plant Biol. 63 (6) (2021) 981–994.
[3] F. Kong, H. Nan, D. Cao, Y. Li, F. Wu, J. Wang, S. Lu, X. Yuan, E.R. Cober, J. Abe, A new dominant gene E9 conditions early flowering and maturity in soybean, Crop Sci. 54 (6) (2014) 2529–2535.
[4] J. Cockram, H. Jones, F.J. Leigh, D. O'Sullivan, W. Powell, D.A. Laurie, A.J. Greenland, Control of flowering time in temperate cereals: genes, domestication, and sustainable productivity, J. Exp. Bot. 58 (6) (2007) 1231–1244.
[5] S. Watanabe, K. Harada, J. Abe, Genetic and molecular bases of photoperiod responses of flowering in soybean, Breeding Sci. 61 (5) (2012) 531–543.
[6] K. Decsi, B. Kutasy, M. Kiniczky, G. Hegedűs, E. Virág, RNA-seq datasets of field soybean cultures conditioned by Elice16Indures® biostimulator, Data in Brief (2022) 108182.
[7] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. 102 (43) (2005) 15545–15550.
[8] G. Hegedűs, Á. Nagy, K. Decsi, B. Kutasy, E. Virág, Transcriptome datasets of β-Aminobutyric acid (BABA)-primed mono-and dicotyledonous plants, Hordeum vulgare and Arabidopsis thaliana, Data Brief (2022) 107983.
[9] S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger, S. Wingett, FastQC, 2010.
[10] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120.
[11] K.K. Mátyás, G. Hegedűs, J. Taller, E. Farkas, K. Decsi, B. Kutasy, N. Kálmán, E. Nagy, B. Kolics, E. Virág, Different expression pattern of flowering pathway genes contribute to male or female organ development during floral transition in the monoecious weed Ambrosia artemisiifolia L.(Asteraceae), PeerJ 7 (2019) e7421.
[12] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, Nat. Biotechnol. 29 (7) (2011) 644.
[13] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (4) (2012) 357–359.
[14] B. Bioinformatics, S. Valencia, OmicsBox-Bioinformatics made easy, March 3 (2019) 2019.
[15] S. Anders, P.T. Pyl, W. Huber, HTSeq—a Python framework to work with high-throughput sequencing data, Bioinformatics 31 (2) (2015) 166–169.
[16] S. Tarazona, F. García, A. Ferrer, J. Dopazo, A. Conesa, NOIseq: a RNA-seq differential expression method robust for sequencing depth biases, EMBnet. J. 17 (B) (2011) 18–19.
[17] S. Tarazona, F. García-Alcalde, J. Dopazo, A. Ferrer, A. Conesa, Differential expression in RNA-seq: a matter of depth, Genome Res. 21 (12) (2011) 2213–2223.
[18] S. Tarazona, P. Furió-Tarí, D. Turrà, A.D. Pietro, M.J. Nueda, A. Ferrer, A. Conesa, Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package, Nucleic Acids Res. 43 (21) (2015) e140-e140.