Data in Brief 42 (2022) 108182

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

RNA-seq datasets of field soybean cultures conditioned by Elice16Indures® biostimulator

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

[a] Department of Plant Physiology and Plant Ecology, Hungarian University of Agriculture and Life Sciences, Georgikon Campus Keszthely, Festetics Gy. u.7., Keszthely 8360, Hungary
[b] EduCoMat Ltd., Iskola u. 12/A., Keszthely 8360, Hungary
[c] Research Institute for Medicinal Plants and Herbs Ltd., 2011 Budakalász, Lupaszigeti út 4., Hungary
[d] Department of Molecular Biotechnology and Microbiology, Institute of Biotechnology, Faculty of Science and Technology, University of Debrecen, Egyetem tér 1., Debrecen 4132, Hungary

ARTICLE INFO

Article history:
Received 14 March 2022
Revised 5 April 2022
Accepted 8 April 2022
Available online 13 April 2022

Dataset link: Glycine max Raw sequence reads (Original data)

Keywords:
Organic farming
Plant conditioner
Soybean
RNA-seq
Glycine max
Transcriptome
Illumina sequencing

ABSTRACT

The herbal drug-containing plant conditioner Elice16Indures® may help elicit plant immune responses in field dicotyledonous cultures. Application of this conditioner is also allowed in organic farming and recommended its drone spraying application in small doses. In this way, even distribution and better yields may be reached leading to economical and safe plant growing. The high protein content soy is an important food both in animal and human aspects which ecological cultivation is gaining prominence over GMO technology in the European Union. We present RNA-seq datasets of control and Elice16Indures treated soybean plants cultivated in field conditions from 01/05/2020 to 20/07/2020. For RNA seq experiments six samples were collected from vegetative tissues two times during the vegetation cycle: before and in flowering after 48 h of drone exposure. The 86 bp long Illumina NextSeq 550 reads were preprocessed and deposited in the NCBI SRA database. De novo assembly of combined read sets was performed and transcripts were deposited in the NCBI TSA database. Data of functional analysis of anno-

* Corresponding author.
E-mail address: kutasy.barbara.julia@uni-mate.hu (B. Kutasy).

https://doi.org/10.1016/j.dib.2022.108182

tated transcripts are presented. The SRA and TSA datasets are under the Bioproject accession PRJNA778970. The presented datasets may help new strategies of ecological production of soy.

Specifications Table

Subject	Plant Science: Plant Physiology
Specific subject area	RNA-seq profiling as a response to herbal drug-containing plant conditioner, Elice16Indures exposure were performed and compared between control and treated soybean plants cultivated in organic farming fields.
Type of data	Table Database record Figure
How the data were acquired	Six samples of vegetative plant tissues (30 mg of plant leaves) were collected from Elice16Indures sprayed field plots which were in a 4-repetition block system. The collected samples were from dosages of 0 g/ha, 20 g/ha, and 240 g/ha, and the collection was performed two times during the vegetation cycle: 25 May and 10 July 2020 two days after drone-spraying of the agent. Samples were sequenced by using the Illumina NextSeq550 platform appearing 14.3-15M 86 bp single-end reads, approximately. Reads were assembled using combined read sets. Functional annotation was performed by BLAST determining gene ontology numbers, enzyme names, and enzyme codes.
Data format	Raw Analysed Filtered
Description of data collection	Four repetitions of the six samples were collected from the four field blocks randomly. Plant materials were collected in RNA-shield (Zymo Research, Irvine, US) preservative and stored at -25°C until sequencing. Sequencing was performed by a third party, Xenovea Ltd, Szeged, Hungary. Raw Illumina read datasets were processed by comprehensive bioinformatics analysis.
Data source location	• EduCoMat Ltd • Keszthely • Hungary
Data accessibility	The bio project, RNA-seq reads and transcriptome assembly are available in National Center for Biotechnology Information database under the accessions: Repository name: *Glycine max* Raw sequence reads Data identification number: PRJNA778970 Direct link to datasets: https://www.ncbi.nlm.nih.gov/search/all/?term=PRJNA778970 Repository name: RNA-seq of *Glycine max* T1.0 Data identification number: SRR16927693 Direct link to datasets: https://www.ncbi.nlm.nih.gov/sra/?term=SRR16927693 Repository name: RNA-seq of *Glycine max* T1.20 Data identification number: SRR16927694 Direct link to datasets: https://www.ncbi.nlm.nih.gov/sra/?term=SRR16927694 Repository name: RNA-seq of *Glycine max* T1.240 Data identification number: SRR16927695 Direct link to datasets: https://www.ncbi.nlm.nih.gov/sra/?term=SRR16927695

(*continued on next page*)

Repository name: RNA-seq of *Glycine max* T2.0
Data identification number: SRR16927696
Direct link to datasets:
https://www.ncbi.nlm.nih.gov/sra/?term=SRR16927696
Repository name: RNA-seq of *Glycine max* T2.20
Data identification number: SRR16927697
Direct link to datasets:
https://www.ncbi.nlm.nih.gov/sra/?term=SRR16927697

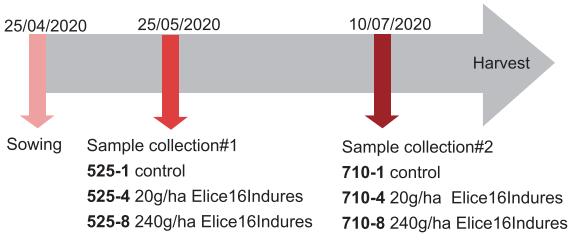
Repository name: RNA-seq of *Glycine max* T2.240
Data identification number: SRR16927698
Direct link to datasets:
https://www.ncbi.nlm.nih.gov/sra/?term=SRR16927698
Repository name: *Glycine max*, transcriptome shotgun assembly
Data identification number: GJRQ00000000
Direct link to datasets:
https://www.ncbi.nlm.nih.gov/nuccore/GJRQ00000000
AnnotationTable and CountTable as Supplementary 1-2. (In an excel file on separate worksheet) in Mendeley Data:
https://data.mendeley.com/datasets/d2yypjh2hr/1

Value of the Data

- The size of the areas involved in organic farming in the European Union is growing dynamically due to favorable support conditions. Organic farming requires appropriate plant conditioning agents to help develop the plant's natural adaptive capacity. The investigated plant-based conditioner (containing herbal extracts) can be used in organic cultivation, with EU permission. Data from RNA-seq may contribute to understanding the physiological effects of herbal extracts.
- The high protein content, soy is a functional food that plays a prominent role in both animal and human feed. Cultivating this plant under ecological conditions can produce a high-quality raw material free of genetic modification and residues. As a cultivation technology development that fits into organic farming, Elice16Indures can help farmers to grow soybeans more economically and environmentally friendly.
- Sustainable agricultural production, sees organic farming as an alternative to GM soybeans that are grown in a huge area of the world. In the future, organic farming will need robotizing preparations that strengthen the physiological condition of plants. Our dataset can help investigate the effects of plant-based roborating preparations, thereby developing new generation plant conditioners.
- Our dataset can be used for transcriptomic analysis of soy plants both genome-wide and individual genes. The information obtained can be a starting point for elucidating the cellular mechanisms of action of plant conditioners of similar composition that can be used for organic food production. On the other hand, transcriptomic data contribute to wider the information and research of soy.

1. Data Description

The demand for GM-free soy is rapidly increasing involving high impact on organic production [1–4]. Therefore, biostimulator products that may protect plants from a broad range of pathogens (by activating the plant immune system) are of major agricultural importance [5–7]. Shallow RNA-sequencing [8] for gene expression profiling as a response to the application of the biostimulator Elice16Indures (Liposome formulation product of Elice16 family, https://gynki.hu/en/rimph-botanicals/products/) are presented here. Illumina RNA-seq reads of



Data	Accession number	seq#
525-1	SRR16927693	14.915.888
525-4	SRR16927694	15.182.823
525-8	SRR16927695	14.292.341
710-1	SRR16927696	14.724.434
710-4	SRR16927697	14.760.776
710-8	SRR16927698	14.473.178
Combined assembly	GJRQ00000000	8308

Fig. 1. Timeline of sample collection of field soybean. The sample marks are indicated with data deposition numbers. Sample marks were as follows: 525, collected on 25, May 2020; 710, collected on 10 July, 2020; 1, 0 g/ha; 4, 20 g/ha and 8, 240 g/ha Elice16Indures treatment. Combined assembly means the reference *Glycine max* transcriptome dataset.

Table 1
Statistics of contig lengths of the deposited TSA data, GJRQ00000000.

Contig length	Stats based on all transcripts	Stats based on the longest isoform per gene
N10	464	455
N20	389	383
N30	354	348
N40	327	323
N50	306	303
Median	294	292
Average	322.19	318. 27

low and high dosage treated, in different plant ripeness stages and in two-time points of field cultivated soy are deposited in the NCBI Sequence Read Archive (SRA). Reads are under the accession numbers: SRR16927693 (control in first treatment time point); SRR16927694 (lower dose in first treatment time point); SRR16927695 (higher dose in first treatment time point); SRR16927696 (control in second treatment time point); SRR16927697 (lower dose in second treatment time point); SRR16927698 (higher dose in second treatment time point). Experimental design is presented in Fig. 1. De novo assembly was performed using these SRA datasets combined to perform a reference *Glycine max* Transcriptome Shotgun Assembly (TSA) that has been deposited at DBJ/EMBL/GenBank under the accession GJRQ00000000. The version described in this paper is the first version GJRQ01000000. Statistic of shallow RNA-Seq contig lengths is summarized in Table 1. Functional annotation of whole 8308 transcripts of combined transcriptome dataset was performed by BLAST determining gene ontology numbers, enzyme names, and enzyme codes. Functional annotation was summarized in the AnnotationTable and presented in Supplementary 1. Statistics of annotation processing are presented in Fig. 2A–D. To determine differences in read abundances of the six samples the CountTable was created aligning the SRA reads to the TSA data that may use for further gene expression experiments. The CountTable is presented in the Supplementary 2. Based on the CountTable numerical data of transcripts were determined and presented in venn diagrams in Fig. 3.

2. Experimental Design, Materials and Methods

2.1. Plant materials

Glycine max cv. ES Director plants were cultured in field conditions. Samples were taken from both untreated plots and plots treated with Elice16Indures plant conditioner at doses of 0 g/ha, 20 g/ha and 240 g/ha, applied two times (25 May, 2020 and 10 July, 2020), in four repetition two days after treatments. Sample collection and storage were performed as described earlier

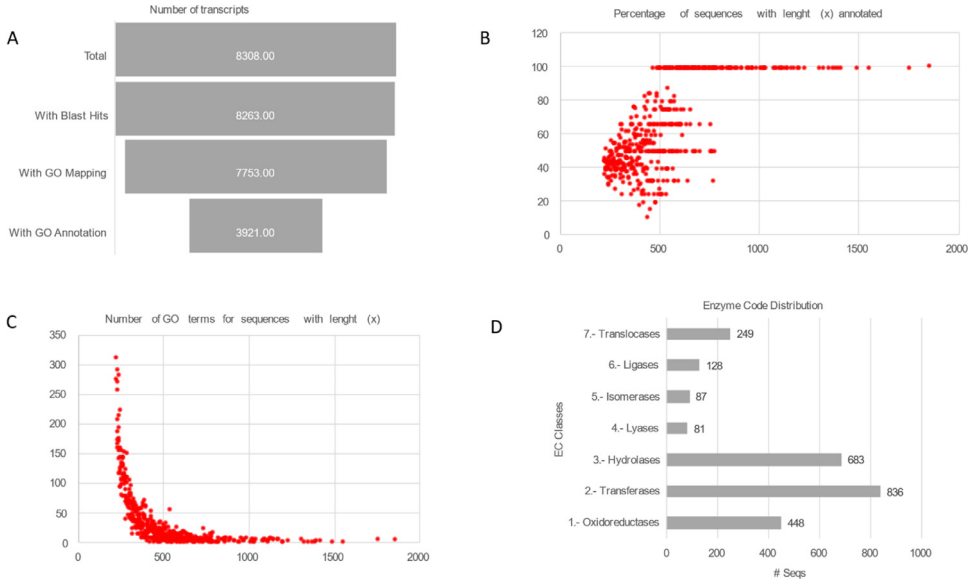


Fig. 2. Statistics of annotation processing. Marks: annotated transcript numbers of the functional analysis process (A), percentage of annotated sequences as a function of transcript lengths (B), number of GO terms as a function of transcript lengths (C), enzyme code distribution as a function of transcript number (D).

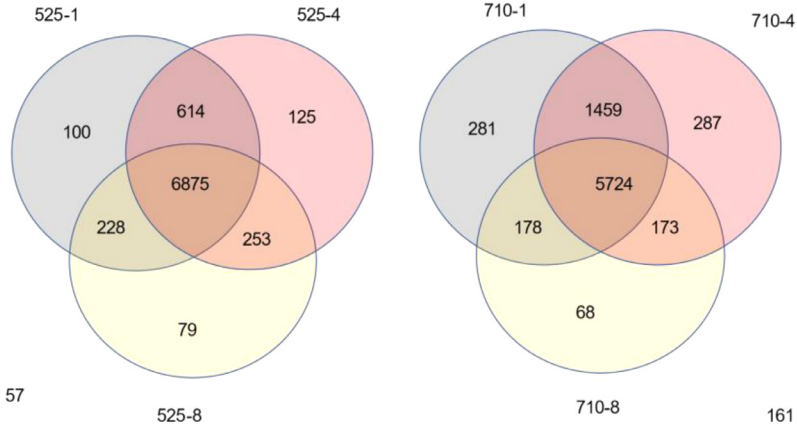


Fig. 3. Venn diagrams of numerical data of transcript distributions in treatments investigated in two time points (525, 25 May, 2020 and 710, 10 July, 2020). Numbers outside the sets are numbers of transcripts without abundances.

by Hegedűs et al. [5]. The four repetitions of each sample were pooled and sequenced by third party Xenovea Ltd, Szeged, Hungary. The marks of samples were used as indicated at the Fig. 1.

2.2. NGS Library preparation and sequencing

NGS libraries were constructed by using QuantSeq 3‘mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen GmbH, Wien, 510 Austria) according to the manufacturer’s protocol. Diluted samples (dilution to 1.8 pM) were sequenced using NextSeq 500/550 High Output v2 Kit (75-cycle)

on the NextSeq550 platform (Illumina, San Diego, CA, USA) to produce 1 × 86 bp single-end reads. Using QuantSeq the 3' end of poly(A) RNA may be pinpointed obtaining accurate information about the 3' UTR.

2.3. Pre-processing and assembly

Reads were pre-processed (removing adapters and contamination sequences) using Trimmomatic software [9]. During this step, low quality bases, short and low-quality reads were filtered out. Transcriptome assembly with cleaned and combined read sets (*Glycine max* combined) was performed by using Trinity and Bowtie2 [10,11].

2.4. Functional annotation

Functional annotation and Gene Ontology (GO) analyses were carried out using OmicsBox.BioBam (https://www.biobam.com/omicsbox/) [12], as follows: Sequences were blasted against NCBI nr (non-redundant) *Viridiplantae* database (downloaded in 2021) applying blastn configuration locally. To retrieving GO terms associated with the 10 Hits obtained by the Blast search, GO mapping and annotation were performed. GeneBank identifiers (gi) and the primary blast Hit identifiers were used to retrieve UniProt IDs making use of a mapping file from PIR (Non-redundant Reference Protein Database) including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept and PDB. Accessions were searched directly in the db x ref table of the GO database. Blasted sequences were searched directly in the gene-product table of the GO database. GO annotation were specified according to GO terms of molecular function, cellular component and biological process. The annotation of entire transcriptome is presented in the AnnotationTable (Supplementary 2). Statistics of annotation processing are summarized in Fig. 2 such as annotated transcript numbers (Fig. 2A), percentage of annotated sequences (Fig. 2B), number of GO terms (Fig. 2C) and enzyme code distribution for sequence number (Fig. 2D).

2.5. Read mapping to estimate transcript abundances

To estimate transcript abundances each sample reads were aligned to the combined transcriptome (TSA: GJRQ00000000). The number of reads for each feature are presented in the CountTable (see Supplementary 2). This process was performed by using the HTseq [13] package and Bowtie2 [10]. Transcript number distribution across the samples were determined by an in-house software for classification implementation for the 8 subsets defined by the 3 basic sets (in each time points).

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.

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Data Availability

Glycine max Raw sequence reads (Original data) (NCBI).

CRediT Author Statement

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

Acknowledgments

The work was founded by the KFI_16-1-2017-0457 - Development and production of a plant-based pesticide-plant conditioner for use in organic farming - project of the Hungarian Government. We express our thanks to József Péter Pallos, executive director, RIMPH Ltd. for the project financiering, and to Zsófia Thomas-Nyári, for the project supervision and administration.

References

[1] M. Krön, U. Bittner, Danube soya–improving European GM-free soya supply for food and feed, OCL 22 (5) (2015) D509, doi:10.1051/ocl/2015050.
[2] M. Klaiss, et al., Organic soybean production in Switzerland, OCL 27 (2020) 64, doi:10.1051/ocl/2020059.
[3] N. Prasetiaswati, et al., Economic feasibility of soybean cultivation using bio-pesticide technology, in: Proceedings of the IOP Conference Series: Earth and Environmental Science, IOP Publishing, 2021, doi:10.1088/1755-1315/756/1/012081.
[4] W. Wangiyana, N. Farida, Application bio-fertilizers to increase yields of zero-tillage soybean of two varieties under different planting distances in dry season on vertisol land of Central Lombok, Indonesia, in: Proceedings of the AIP Conference Proceedings, AIP Publishing LLC, 2019, doi:10.1063/1.5141296.
[5] G. Hegedűs, et al., Transcriptome datasets of β-Aminobutyric acid (BABA)-primed mono-and dicotyledonous plants, *Hordeum vulgare* and *Arabidopsis thaliana*, Data Brief (2022) 107983, doi:10.1016/j.dib.2022.107983.
[6] N.R. Chalfoun, et al., Elicitor-based biostimulant PSP1 protects soybean against late season diseases in field trials, Front. Plant Sci. 9 (2018) 763, doi:10.3389/fpls.2018.00763.
[7] O.I. Yakhin, et al., Biostimulants in plant science: a global perspective, Front. Plant Sci. 7 (2017) 2049, doi:10.3389/fpls.2016.02049.
[8] P. Moll, et al., QuantSeq 3′ mRNA sequencing for RNA quantification, Nat. Methods 11 (12) (2014) p. i-iii, doi:10.1038/nmeth.f.376.
[9] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.
[10] M.G. Grabherr, et al., Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data, Nat. Biotechnol. 29 (7) (2011) 644, doi:10.1038/nbt.1883.
[11] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (4) (2012) 357–359, doi:10.1038/nmeth.1923.
[12] B. Bioinformatics, S. Valencia, OmicsBox-bioinformatics made easy. March, 2019. 3: p. 2019. https://www.biobam.com/omicsbox/.
[13] S. Anders, P.T. Pyl, W. Huber, HTSeq—a python framework to work with high-throughput sequencing data, Bioinformatics 31 (2) (2015) 166–169, doi:10.1093/bioinformatics/btu638.