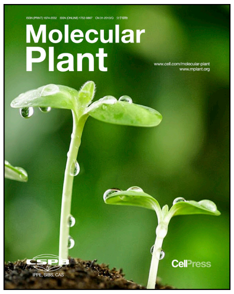


# SoyOmics: A deeply integrated database on soybean multi-omics

Yucheng Liu, Yang Zhang, Xiaonan Liu, Yanting Shen, Dongmei Tian, Xiaoyue Yang, Shulin Liu, Lingbin Ni, Zhang Zhang, Shuhui Song, Zhixi Tian

PII: S1674-2052(23)00075-8

DOI: https://doi.org/10.1016/j.molp.2023.03.011

Reference: MOLP 1527

To appear in: *MOLECULAR PLANT*

Received Date: 4 January 2023

Revised Date: 2 March 2023

Accepted Date: 19 March 2023

Please cite this article as: Liu Y., Zhang Y., Liu X., Shen Y., Tian D., Yang X., Liu S., Ni L., Zhang Z., Song S., and Tian Z. (2023). SoyOmics: A deeply integrated database on soybean multi-omics. Mol. Plant. doi: https://doi.org/10.1016/j.molp.2023.03.011.

# SoyOmics: A deeply integrated database on soybean multi-omics

Yucheng Liu[1†], Yang Zhang[2,3,4†], Xiaonan Liu[2,3,4†], Yanting Shen[1†], Dongmei Tian[2,3,4], Xiaoyue Yang[1], Shulin Liu[1], Lingbin Ni[1,4], Zhang Zhang[2,3,4*], Shuhui Song[2,3,4*], Zhixi Tian[1,4*]

[1] State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China

[2] National Genomics Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

[3] China National Center for Bioinformation, Beijing 100101, China

[4] University of Chinese Academy of Sciences, Beijing 100039, China

† Contributed equally to this work

* Corresponding authors (Zhang Zhang, email: zhangzhang@big.ac.cn; Shuhui Song, email: songshh@big.ac.cn; Zhixi Tian, email: zxtian@genetics.ac.cn)

**Dear Editor,**

As one of the most important crops to supply majority plant oil and protein for the whole world, soybean is facing an increasing global demand. The reference genome of accession “Williams82” opened the gate of genomics research in soybean (Schmutz et al., 2010). After that, vast multi-omics data were generated, thereby providing valuable resources for functional study and molecular breeding. Part of these data have been collected in different soybean databases (see details in Supplemental Table 1), such as Soybase (Grant et al., 2010) and SoyKB (Joshi et al., 2012), which made valuable efforts to facilitate the wide utility of these data. Nevertheless, these existing databases poorly tackled multi-omics data integration and interactivity for soybean, provoking tremendous challenges for researchers to deal with these big omics data, particularly considering the unprecedented rate of data growth (Yang et al., 2021). Thus, constructing an integrated multi-omics database for soybean that provides a one-stop solution for big data mining with friendly interactivity is highly desired.

Here, we collect the reported high-quality omics data, including assembly genomes, graph pan-genome, resequencing and phenotypic data of representative germplasms (Zhou et al., 2015; Fang et al., 2017; Liu et al., 2020), *de novo* assembled genomes of species of subgenus *Glycine* (Zhuang et al., 2022), transcriptomic and epigenomic data from different tissues, organs and accessions (Shen et al., 2014; Shen et al., 2018; Shen et al., 2019), knowledge of quantitative trait locus (QTL) and genome-wide association study (GWAS) (Grant et al., 2010), and construct an integrated soybean multi-omics database, named SoyOmics (https://ngdc.cncb.ac.cn/soyomics). By equipping with multiple analysis modules and toolkits, SoyOmics is of great utility to facilitate the global scientific community to fully use these big omics datasets for a wide range of soybean studies from fundamental functional investigation to molecular breeding.

## An overview of SoyOmics

By integrating different multi-omics data, we develop six highly interactive basic modules in SoyOmics: Genome, Variome, Transcriptome, Phenome, Homology, and Synteny (Figure 1A). The Genome module embodies the information of 2,898 soybean germplasms and 27 *de novo* assembled genomes, providing users with open access to basic information of sequenced germplasms, assembled genomes and genes (Supplemental Figure 1). The Variome module organizes approximately 38 million SNPs and INDELs of the 2,898 soybean accessions, facilitating users to check the variation information and whole genome selective signals for any germplasm of interest (Supplemental Figure 2). The Transcriptome module contains two datasets of gene expression: one is from 27 tissues at different developmental stages from Williams82 and ZH13 accessions, respectively, and the other is from 9 tissues at different developmental stages from each of the 26 accessions used for pan-genome analysis. In this module, users can obtain gene expression profiles and gene orthologous information by specifying gene ID or functional description (Supplemental Figure 3). The Phenome module collects approximately 27 thousand records of 115 phenotypes with terms defined as controlled vocabularies that fall into 5 classes (including morphology, growth and development, biochemistry, biotic stress, and vigor) as well as 17 subclasses (Supplemental Figure 4). The Homology module displays the soybean pan-genome by characterizing 57,480 homologous gene groups. Users can specify any gene ID, homologous group ID or gene functional description to retrieve the homologous group of interest (Supplemental Figure 5). The Synteny module deposits approximately 550 thousand large-scale structural variations (SV) in the pan-genome, in which users can visualize and download the SVs and synteny blocks by setting a specific genomic region. Furthermore, the graph pan-genome is embedded and a SequenceTubeMap web service (https://github.com/vgteam/sequenceTubeMap) is deployed for visualization of pan-genome threads (or haplotypes) according to nodes made up by SVs (Supplemental Figure 6).

In addition, SoyOmics is designed to provide user-friendly search bar in each module and to cover as much as more possible substances. According to the searching category and inputting context, it features powerful search engine to provide comprehensive associated results with friendly links from one module to other modules (Supplemental Figure 7).

## Application toolkits

In addition to the six modules, we design several commonly easy-to-use toolkits, including *easyGWAS*, *ExpPattern*, *HapSnap*, *VersionMap*, *SoyArray*, and *SeqFetch* (Figure 1A). A BLAST module based on NCBI BLAST+ (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/) is also developed for sequence searching against genome, mRNA, CDS and protein sequences of the pan-genome accessions. The *easyGWAS* is a tool for quick-start GWAS analysis, providing friendly interface for parameters setting and algorithms selection and offering multiple high-quality outputs including Manhattan plot, QQ-plot and text result (Supplemental Figure 8). The *ExpPattern* is for conducting expression pattern analysis for a gene list against soybean tissues. It can generate expression heatmap, with options of whether to execute clustering or not (Supplemental Figure 9). Besides, the tspex (https://github.com/apcamargo/tspex/) is incorporated in the *ExpPattern* for advice of gene’s tissue-specificity. The *HapSnap* is designed for haplotype analysis for a genomic region. Users can refine the variations via selection of variation type and quality control.

The output includes haplotype frequency, haplotype vs. genotype, and linkage disequilibrium (Supplemental Figure 10). The *VersionMap* is capable to convert the genomic region between ZH13 (v2) and other *de novo* genomes of soybean, or gene ID between Williams82 (v2) and ZH13 (v2) (Supplemental Figure 11). The *SeqFetch* is developed to get the sequence for a specific genomic region, gene, mRNA, CDS, and/or protein from 29 soybean genomes (Supplemental Figure 12).

We also develop a toolkit named *SoyArray* by embedding the information of GenoBaits soybean array (Liu et al., 2022), in which users can search and download the marker information they are interested in. We also afford a function in the *SoyArray* to compare divergent sites between two germplasms based the makers from GenoBaits soybean array, which is helpful for parents' picking in genetic or breeding study (Supplemental Figure 13).

## Data mining using SoyOmics

As SoyOmics integrates a wide variety of soybean multi-omics data, it can be used for deep mining ranging from fundamental research to molecular breeding. Here we take a previously reported seed coat color causal gene, *G* (Wang et al., 2018), as an example. In SoyOmics, we can group germplasms by green or yellow seed coat colors (Figure 1B). According to the phenotype data, we can conveniently conduct GWAS analysis using the *easyGWAS* toolkit, and then identify a significant association signal that is located in the *G* gene, *SoyZH13_01G182000* (Figure 1C and 1D). According to the interested association genetic variant, users can get phenotype variations among different genotypes, such as the seed coat color (Figure 1E). By searching the candidate gene *SoyZH13_01G182000* from different modules, users can obtain a wealth of gene information including basic summary, functional annotation, homology in 29 soybean genomes and expression pattern in 28 tissues (Figure 1D, 1F, and 1G). Furthermore, users can also investigate functional annotations for any variant of interest (Figure 1H), linkage disequilibrium around the association genetic variant (Figure 1I), allele frequency in different populations (Figure 1J), and selection sweeps for the association regions by three different test methods (Figure 1K). Notably, the majority of charts generated in SoyOmics can be directly downloaded and edited.

In summary, SoyOmics features comprehensive integration of multi-omics datasets and provides user-friendly interfaces for soybean study. Comparing to other popular soybean databases, SoyOmics has significant advantages on multi-omics interaction, pan-genome scan and online analysis functionality (Supplemental Table 1), conforming well to the trend of omics database in the post-genomics era. Undoubtedly, soybean omics data are generated at increasing scales and rates, including resequencing data for more germplasms, transcriptome data from bulk, single-cell and spatial RNA-seq, epigenetic data from Hi-C, ATAC-seq or histone modification, etc. Therefore, future directions for SoyOmics mainly focus on continuous integration of these newly-generated omics data. In addition, artificial intelligence (AI)-based approaches for deep mining of these big data would provide valuable insights for a wide range of soybean studies, particularly for AI breeding in the era of big data. Towards this end, we would like to call for global collaborations to build SoyOmics as a valuable platform for the whole research community around the world.

## SUPPLEMENTAL INFORMATION

Supplemental information is available at Molecular Plant Online.

## FUNDING

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA24000000, XDA19050302, and XDA24040201), the Science and Technology Innovation 2030 - Major Project (2022ZD04017), the National Natural Science Foundation of China (32030021, and 32000475), the National Key Research and Development Program of China (2021YFF1001201), Taishan Scholars Program, and Xplorer Prize Award, the Youth Innovation Promotion Association of the Chinese Academy of Sciences (Y2021038).

## AUTHOR CONTRIBUTIONS

Z.T., and Z.Z. conceived this project. Z.T., S.S., and Z.Z. supervised this work. Y.L., Y.S., X.L., and Y.Z. designed the framework of database and wrote the pipelines. S.L., and X.Y. revised germplasm information and phenotype records. X.L. and Y.Z. constructed the database. L.N. developed the pipeline used in VersionMap module. D.T. built up the easyGWAS, VersionMap and graph-based genome module. Y.L., Y.Z., X.N., S.Y., and Z.T. wrote the manuscript. Z.T., S.S., and Z.Z. revised the manuscript.

All authors read and approved the final manuscript.

## ACKNOWLEDGMENTS

We thank a number of SoyOmics users for their kind advice. No conflict of interest is declared.

## Figure Legends

**Figure 1. Overview of SoyOmics and its practice for data mining.**

(A) Framework of SoyOmics, including data source, module organization and application scenarios.

(B) Germplasms grouped by seed coat color.

(C) GWAS analysis with seed coat color. Threshold is set by $P$ = 1e-10.

(D) Gene structure and variation list of *SoyZH13_01G182000*.

(E) Seed coat color grouped by genotypes of soy1873072. Multiple comparison is conducted by Student-Newman-Keuls (SNK) test, with significant level equal to 0.01.

(F) Homologous gene number of *SoyZH13_01G182000* in 29 soybean genomes.

(G) Expression of *SoyZH13_01G182000* in 28 tissues.

(H) Change consequence on mRNA and present knowledge of variation soy1873072.

(I) Linkage disequilibrium heatmap around soy1873072. The left chart shows heatmap of all variation pairs in the block, and the right shows heatmap of variation pairs with $r^2 \geq 0.9$.

(J) Genotype frequency distribution of *G* gene in *Soja*, landrace and cultivar.

(K) Selective test signal of genomic region Chr01:54.5-57.4 Mb. Green dash line means the significant threshold. Blue triangle shows the location of *SoyZH13_01G182000*.

## Reference

Fang, C., Ma, Y., Wu, S., Liu, Z., Wang, Z., Yang, R., Hu, G., Zhou, Z., Yu, H., Zhang, M., et al. (2017). Genome-wide association studies dissect the genetic networks underlying agronomical traits in soybean. Genome Biol. **18**:1-14.

Grant, D., Nelson, R.T., Cannon, S.B., and Shoemaker, R.C. (2010). SoyBase, the USDA-ARS soybean genetics and genomics database. Nucleic Acids Res. **38**:D843-D846.

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., Zhou, G.-A., Zhang, H., Liu, Z., and Shi, M. (2020). Pan-genome of wild and cultivated soybeans. Cell 182, 162-176. e113.

Liu, Y., Liu, S., Zhang, Z., Ni, L., Chen, X., Ge, Y., Zhou, G., and Tian, Z. (2022). GenoBaits Soy40K: a highly flexible and low-cost SNP array for soybean studies. Sci. China Life Sci. **65**:1898-1901.

Joshi, T., Patil, K., Fitzpatrick, M.R., Franklin, L.D., Yao, Q., Cook, J.R., Wang, Z., Libault, M., Brechenmacher, L., and Valliyodan, B. (2012). Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. BMC genom. **13 Suppl1**:S15.

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., et al. (2010). Genome sequence of the palaeopolyploid soybean. Nature **463**:178-183.

Shen, Y., Du, H., Liu, Y., Ni, L., Wang, Z., Liang, C., and Tian, Z. (2019). Update soybean Zhonghuang 13 genome to a golden reference. Sci. China Life Sci. **62**:1257-1260.

Shen, Y., Zhang, J., Liu, Y., Liu, S., Liu, Z., Duan, Z., Wang, Z., Zhu, B., Guo, Y.-L., and Tian, Z. (2018). DNA methylation footprints during soybean domestication and improvement. Genome Biol. **19**:128.

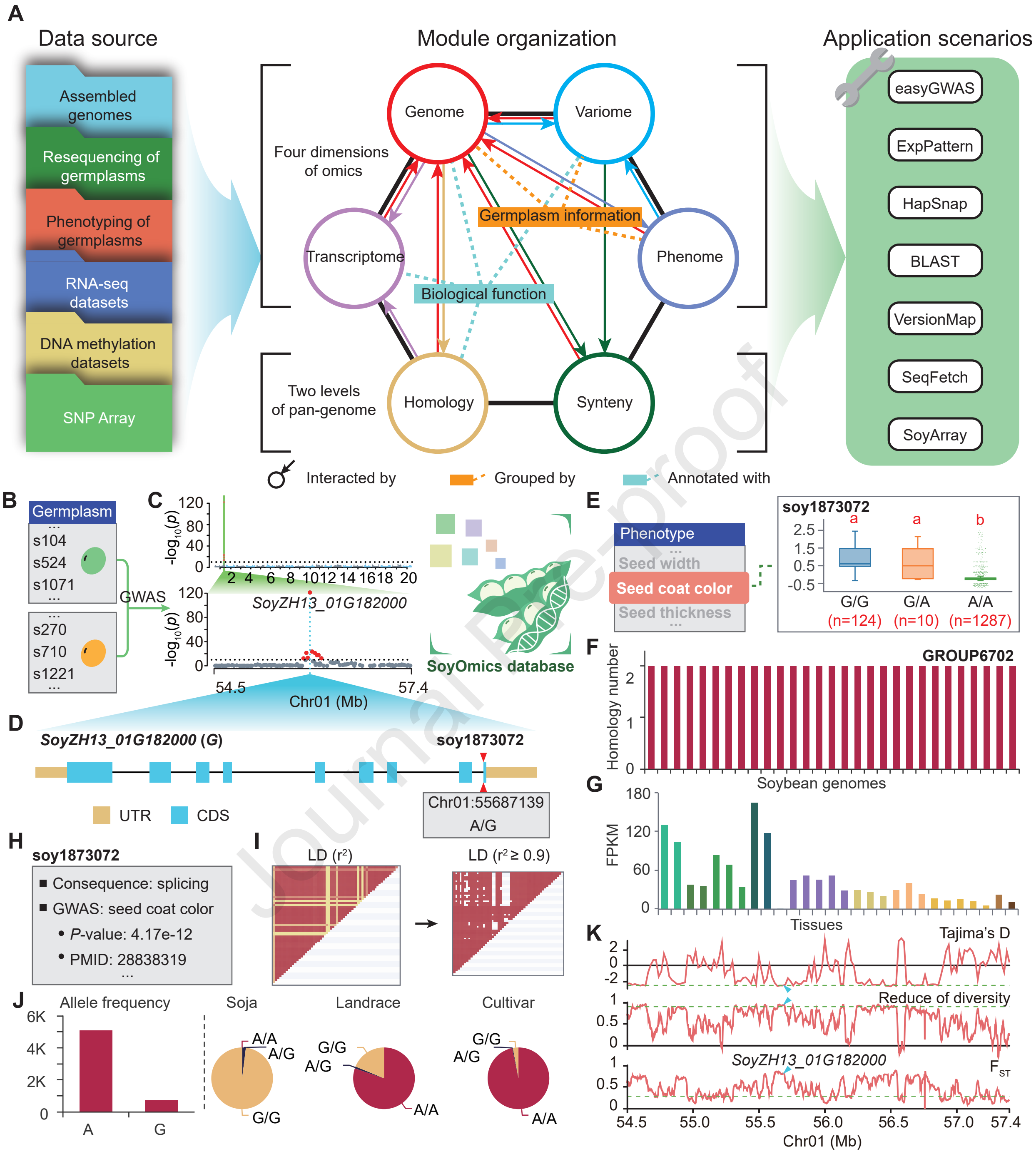
Shen, Y., Zhou, Z., Wang, Z., Li, W., Fang, C., Wu, M., Ma, Y., Liu, T., Kong, L.A., Peng, D.L., et al. (2014). Global dissection of alternative splicing in paleopolyploid soybean. Plant Cell **26**:996-1008.

Wang, M., Li, W., Fang, C., Xu, F., Liu, Y., Wang, Z., Yang, R., Zhang, M., Liu, S., Lu, S., et al. (2018). Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat. Genet. **50**:1435-1441.

Yang, Y., Saand, M.A., Huang, L., Abdelaal, W.B., Zhang, J., Wu, Y., Li, J., Sirohi, M.H., and Wang, F. (2021). Applications of multi-omics technologies for crop improvement. Front. Plant Sci. **12**:563953.

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., Yu, Y., Shu, L., Zhao, Y., Ma, Y., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat. Biotechnol. **33**:408-414.

Zhuang, Y., Wang, X., Li, X., Hu, J., Fan, L., Landis, J.B., Cannon, S.B., Grimwood, J., Schmutz, J., and Jackson, S.A. (2022). Phylogenomics of the genus *Glycine* sheds light on polyploid evolution and life-strategy transition. Nat. Plants **8**:233-244.


A
Data source
Assembled genomes
Resequencing of germplasms
Phenotyping of germplasms
RNA-seq datasets
DNA methylation datasets
SNP Array
Module organization
Four dimensions of omics
Genome
Variome
Transcriptome
Phenome
Germplasm information
Biological function
Two levels of pan-genome
Homology
Synteny
Interacted by
Grouped by
Annotated with
Application scenarios
easyGWAS
ExpPattern
HapSnap
BLAST
VersionMap
SeqFetch
SoyArray
B
Germplasm
...
s104
s524
s1071
...
...
s270
s710
s1221
...
GWAS
C
-log10(p)
120
80
40
0
2 4 6 8 10 12 14 16 18 20
SoyZH13_01G182000
54.5
57.4
Chr01 (Mb)
SoyOmics database
D
SoyZH13_01G182000 (G)
soy1873072
Chr01:55687139
A/G
UTR
CDS
E
Phenotype
...
Seed width
Seed coat color
Seed thickness
...
soy1873072
a
a
b
2.5
1.5
0.5
-0.5
G/G
G/A
A/A
(n=124)
(n=10)
(n=1287)
F
GROUP6702
Homology number
2
1
0
Soybean genomes
G
FPKM
180
120
60
0
Tissues
H
soy1873072
Consequence: splicing
GWAS: seed coat color
P-value: 4.17e-12
PMID: 28838319
...
I
LD (r²)
LD (r² ≥ 0.9)
J
Allele frequency
6K
4K
2K
0
A
G
Soja
A/A
A/G
G/G
Landrace
G/G
A/G
A/A
Cultivar
G/G
A/G
A/A
K
Tajima's D
2
0
-2
Reduce of diversity
1
0.5
0
SoyZH13_01G182000
F_ST
1
0.5
0
54.5
55.0
55.5
56.0
56.5
57.0
57.4
Chr01 (Mb)