Translational Bioinformatics Frameworks and AI Solutions for Multiomics Research

AgBioData Webinar
Aug 6, 2025

Dr. Trupti Joshi

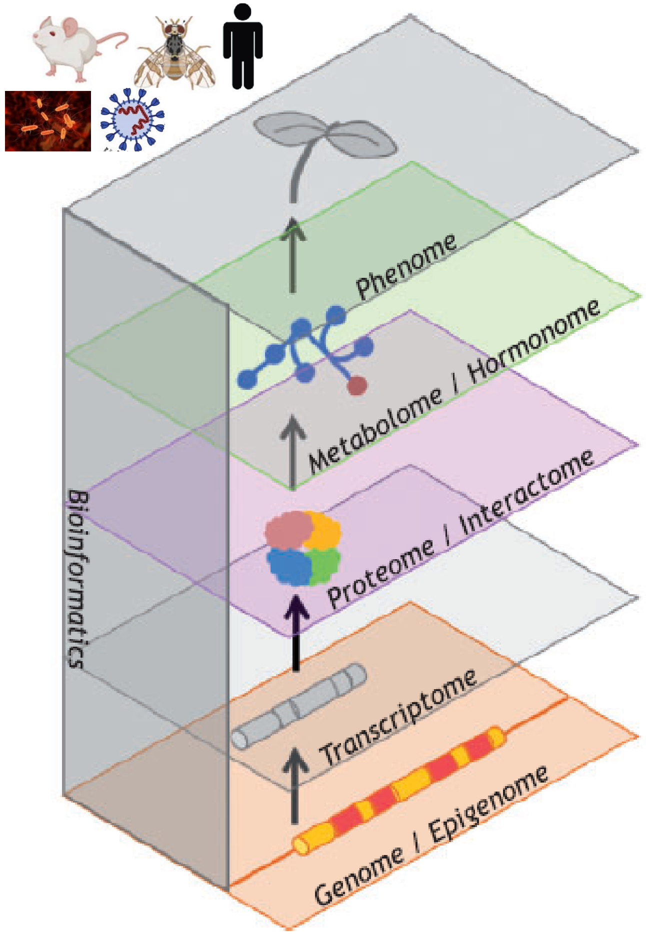
Senior Associate Dean for Informatics and Population Analytics
Professor - Department of Biomedical Sciences
Joan C Edwards School of Medicine
Marshall University

Adjunct Faculty
Department of Biomedical Informatics, Biostatistics and Medical Epidemiology
Core Faculty MUIDSI, DPST, LSC, IPG, EECS
University of Missouri - Columbia

Outline

- Introduction
- Translational Bioinformatics Frameworks and AI Solutions
    - Multiomics Data Framework Development
    - Multiomics Data Integration Tool Development
    - Multiomics Informatics Method Development

Genomics and Multiomics Datasets



Omics instances
Integrated database
Mutant lines
Natural variations
Metabolic map
Metabolome profiles
Hormonome profile
Proteome / modificome profiles
Subcellular localization
Interactome maps
Full-length cDNA clones, ESTs
Expression profiles
Non-coding RNA profiles
Co-expression network
Genome sequence, gene annotation
Re-sequencing
Focused gene family database (eg. Transcription factor)
DNA methylome
Chromatin epigenome

Informatics and Computational methods have rapidly become an essential component for all biological, biomedical, animal science and plant science studies!

The Model Has Changed!!

The Model of Generating and Consuming Data has Changed !!

Old Model: Few organizations are generating data, all others are consuming data

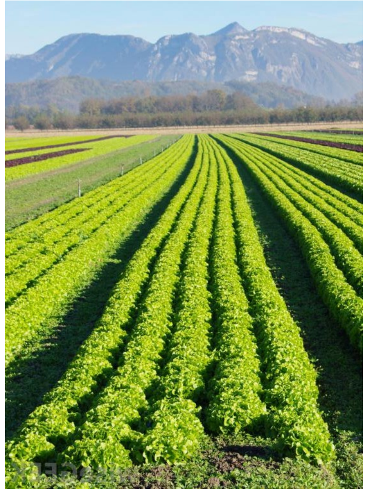
New Model: All of us are generating data, and all of us are consuming data

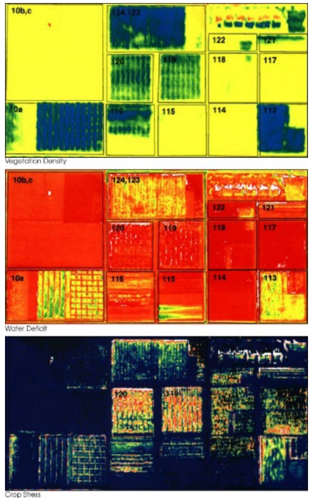
What’s Challenging?

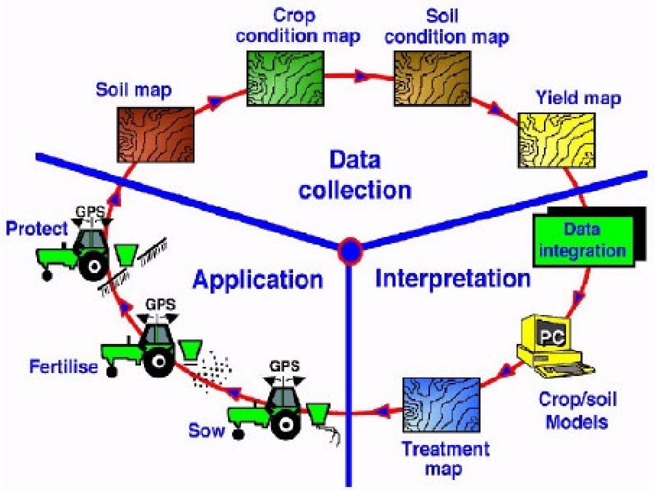
- It is the diversity of high-throughput technologies and the amount of generated data that need to be integrated together, to convert data into meaningful knowledge and making novel discoveries!
- Multiomics data are often individually scattered across different databases and repositories, making it difficult for its users to access them efficiently.
- Getting a comprehensive view from data integration needs advanced techniques.

Precision Agriculture

- Precision agriculture is one of many modern farming practices that make production more efficient.
- Decisions based on many Agronomic Data
  - Hybrid selections, Plant populations
  - Yield data , Soil data
  - Pesticide, Fungicides
  - Genotypes, Environments
  - Economics, Market Prices
- Data generated can be utilized and applied for selected crop breeding, improving yield, pest management, crop disease analytics, spraying decisions and many more.







Comparetti et al. 2015

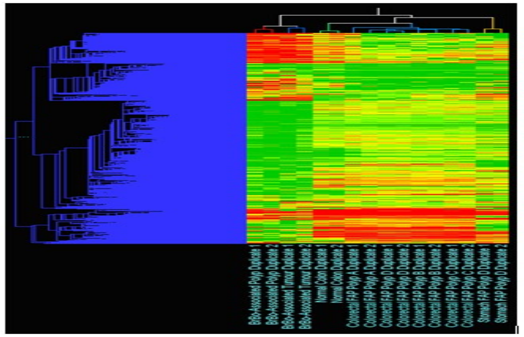
Modern Farming

Modern day farming and agriculture is as much about data, as it is about sowing and harvesting !!

BIG Data Challenges

Emerging Trends

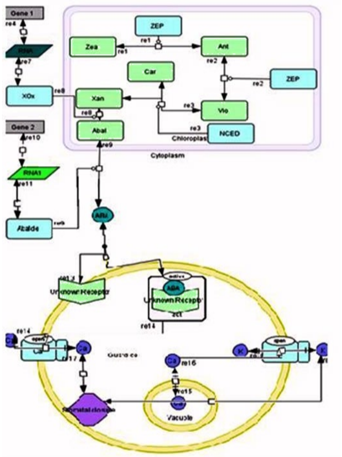
Joshi Lab
Enabling Translational Research



Data: Bulk & Single Cell

- Genomics & Epigenomics
- Transcriptomics
- Proteomics
- Metabolomics
- Phenomics and Others

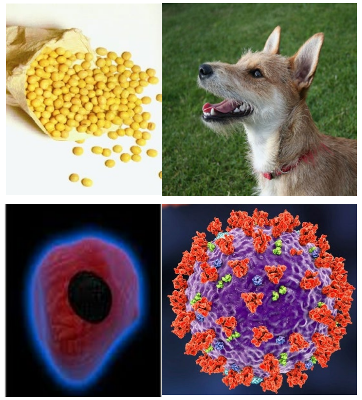
Data volume: huge
Data types: diverse
Data access: spread
Data complexity: high



Knowledge:

- Functions
- Pathways
- Regulation
- Networks
- Homology

Transform Data into Knowledge with Informatics Algorithms, Tools & Databases

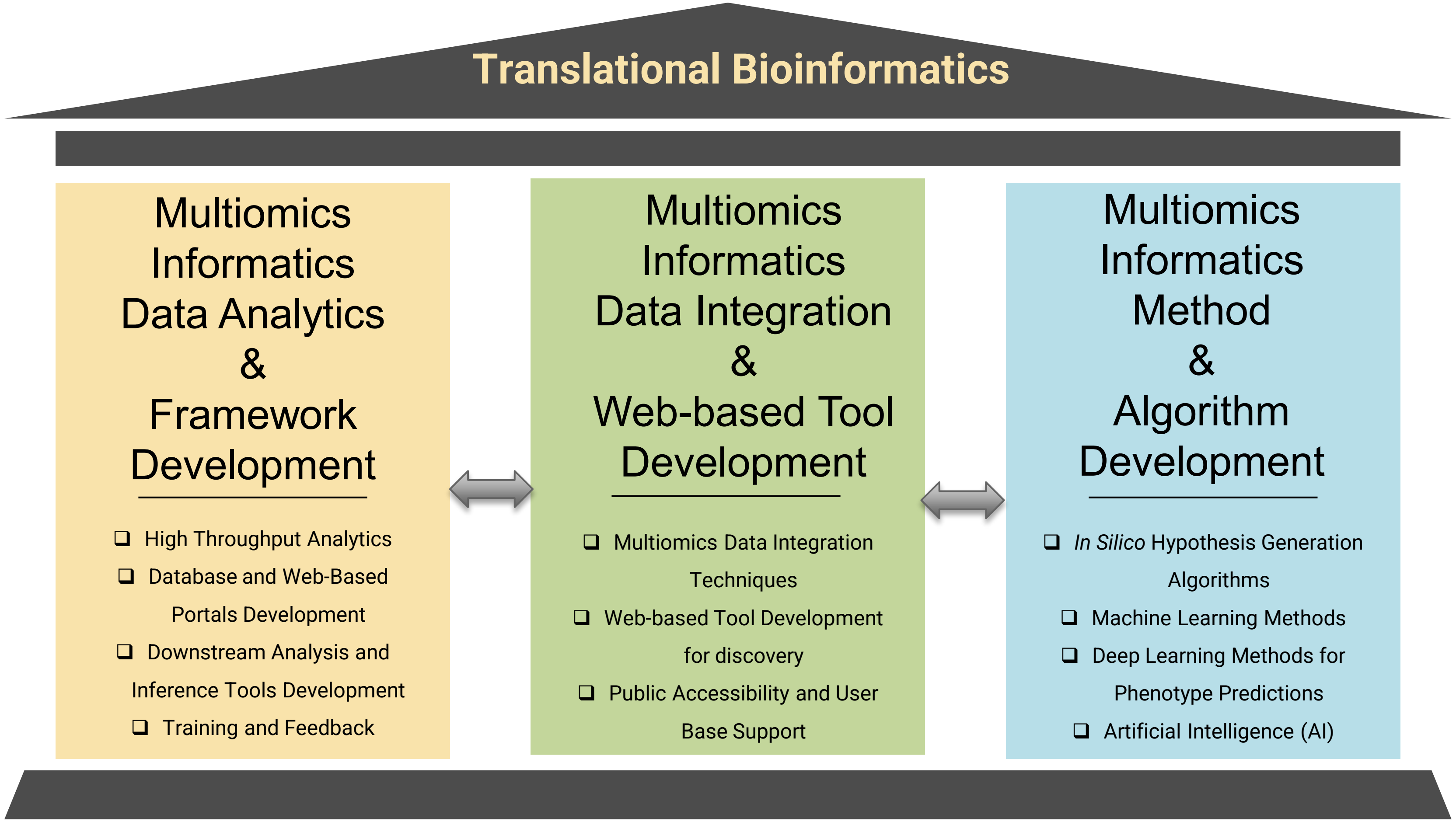


Phenotypes:

- Plants & Crops
- Animals & Pets
- Human & Diseases
- Microbes & Viruses

Precision Agriculture
Precision Medicine
Food Security
Genomic Epidemiology

Joshi Lab - 3 Essential Research Pillars



Translational Bioinformatics

Multiomics Informatics Data Analytics & Framework Development

- High Throughput Analytics
- Database and Web-Based Portals Development
- Downstream Analysis and Inference Tools Development
- Training and Feedback

Multiomics Informatics Data Integration & Web-based Tool Development

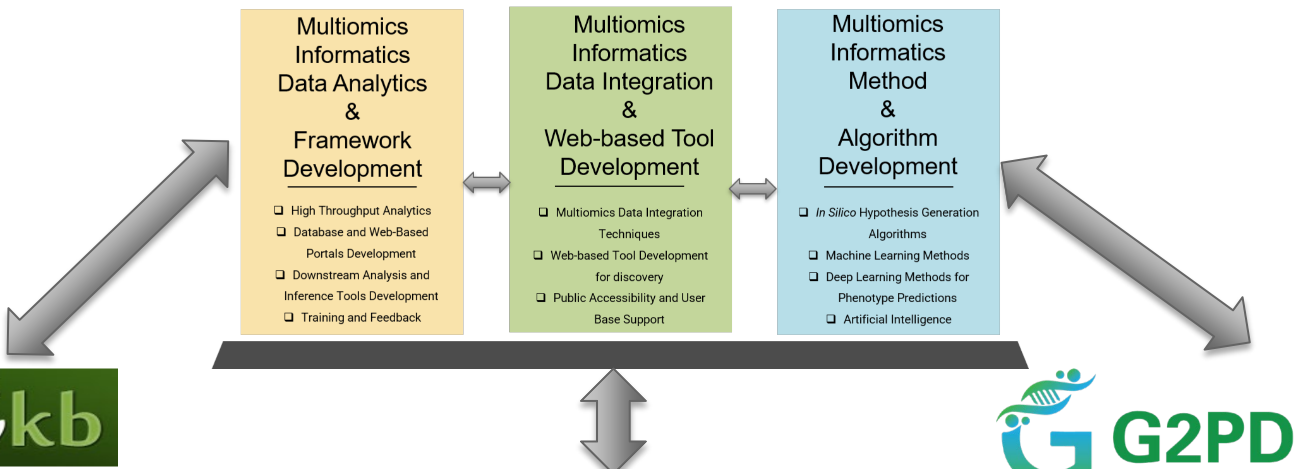
- Multiomics Data Integration Techniques
- Web-based Tool Development for discovery
- Public Accessibility and User Base Support

Multiomics Informatics Method & Algorithm Development

- *In Silico* Hypothesis Generation Algorithms
- Machine Learning Methods
- Deep Learning Methods for Phenotype Predictions
- Artificial Intelligence (AI)

Joshi Lab

Enabling Translational Bioinformatics Research



Multiomics Informatics Data Analytics & Framework Development

- ❑ High Throughput Analytics
- ❑ Database and Web-Based Portals Development
- ❑ Downstream Analysis and Inference Tools Development
- ❑ Training and Feedback

Multiomics Informatics Data Integration & Web-based Tool Development

- ❑ Multiomics Data Integration Techniques
- ❑ Web-based Tool Development for discovery
- ❑ Public Accessibility and User Base Support

Multiomics Informatics Method & Algorithm Development

- ❑ *In Silico* Hypothesis Generation Algorithms
- ❑ Machine Learning Methods
- ❑ Deep Learning Methods for Phenotype Predictions
- ❑ Artificial Intelligence

Soybean Knowledge Base

https://soykb.org/

KBCommons Knowledge Base

https://kbcommons.org/

Multiomics Data Integration

Allele Catalog Tool

GenVarX Tool

SNPViz Tool

AccuTool

MADis Tool

G2PDeep

https://g2pdeep.org/

IMPRes

https://impres.missouri.edu/impres

scPlantAnnotate
https://scplantannotate.missouri.edu/

scPlantAnnotate

Translational Bioinformatics Resources and AI Solutions

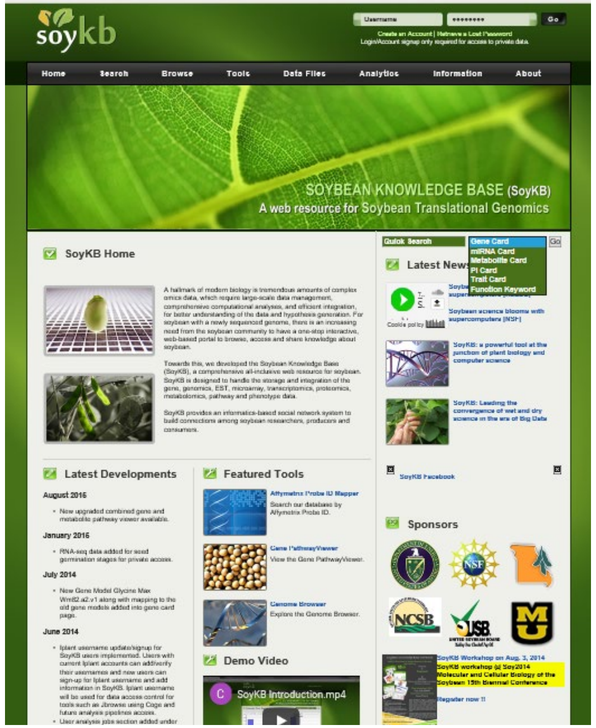
- Multiomics Data Framework Development
  - SoyKB and KBCommons
  - 3D Omics Suite
  - Cross-species and Comparative Multiomics Translation (CCMT) Tool

Soybean Knowledge Base (SoyKB)

- SoyKB is a comprehensive all-inclusive, “one-stop-shop” web resource bridging soybean translational genomics and molecular breeding

- SoyKB provides integration of multi-omics data types including
  - Epigenomics
  - Genomics
  - Transcriptomics
  - Proteomics
  - Metabolomics
  - Phenomics

- SoyKB provides access to analytics linked to XSEDE HPC and cloud data storage on Cyverse
- Has 1000+ registered users across academic and Industry all over the world



SoyKB Home

A hallmark of modern biology is tremendous amounts of complex omics data, which require large-scale data management, comprehensive computational analyses, and efficient integration, for better understanding of the data and hypothesis generation. For soybean with a newly sequenced genome, there is an increasing need from the soybean community to have a one-stop interactive, web-based portal to browse, access and share knowledge about soybean.

Towards this, we developed the Soybean Knowledge Base (SoyKB), a comprehensive all-inclusive web resource for soybean. SoyKB is designed to handle the storage and integration of the gene, genomics, EST, microarray, transcriptomics, proteomics, metabolomics, pathway and phenotype data.

SoyKB provides an informatics-based social network system to build connections among soybean researchers, producers and consumers.

Latest News

Soybean science blooms with supercomputers (NSF)

SoyKB: a powerful tool at the junction of plant biology and computer science

SoyKB: Leading the convergence of wet and dry science in the era of Big Data

Latest Developments

August 2016

- New upgraded combined gene and metabolite pathway viewer available.

January 2016

- RNA-seq data added for seed germination stages for private access.

July 2014

- New Gene Model Glycine Max Wm82.a2.v1 along with mapping to the old gene models added into gene card page.

June 2014

- Iplant username update/signup for SoyKB users implemented. Users with current Iplant accounts can add/verify their usernames and new users can sign-up for Iplant username and add information in SoyKB. Iplant username will be used for data access control for tools such as Jbrowse using Coge and future analysis pipelines access.
- User analysis jobs section added under

Featured Tools

Affymetrix Probe ID Mapper
Search our database by Affymetrix Probe ID.

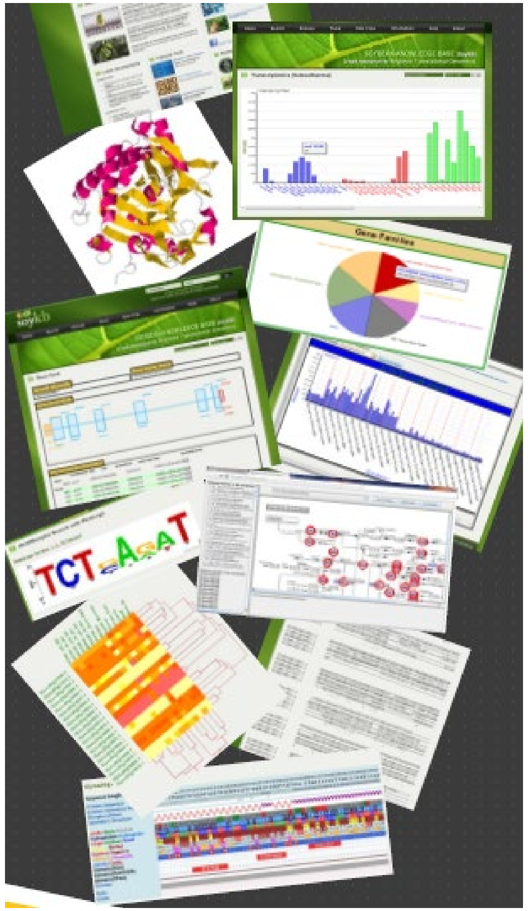
Gene PathwayViewer
View the Gene PathwayViewer.

Genome Browser
Explore the Genome Browser.

Demo Video

SoyKB Workshop on Aug. 3, 2014
SoyKB workshop @ Soy2014 Molecular and Cellular Biology of the Soybean 15th Biennial Conference

Register now !!



https://soykb.org

Joshi et al. Plant Genomics Databases 2016; NAR 2014; BMC Genomics 2012.

KBCommons Informatics Framework

- Knowledge Base Commons (KBCommons) comprehensive framework providing web-based access for
  - Storage, management, sharing, analysis, integration and visualization of genomics and multi-omics data.
  - Adopts, expands and customizes our SoyKB tools for other model organisms, crops, animals and biomedical diseases.
  - Provides seamless links to HPC and cloud resources and integration of in-house developed suites of graphical visualization tools.

homoSapiens Home

Advancement of next generation sequencing and high-throughput technologies has resulted in generation of multi-level of 'OMICS' data for many organisms. However, these data are often individually scattered across different repositories based on data type, making it difficult to integrate them. A hallmark of modern biology is tremendous amounts of complex omics data, which require large-scale data management, comprehensive computational analyses, and efficient integration, for better understanding of the data and hypothesis generation. For most biological organisms ranging from plants to animals to viruses and bacteria with sequenced genomes, there is an increasing need for the community to have a one-stop interactive, web-based portal to browse, access, analyze, integrate and share knowledge about genomics and molecular mechanisms, with ultimate links to phenotypes and traits.

Towards this, we have expanded our in-house developed Soybean Knowledge Base (SoyKB), a comprehensive all-inclusive web resource for soybean to other organisms via Knowledge Base Commons (KBCommons) without the need for reinventing the wheel.

KBCommons, is a platform that automates the process of establishing the database and making the suite of informatics tools and visualization capability easily available for other organisms via a dedicated web resource. It provides information for six entities including genes/proteins, microRNAs/sRNAs, metabolites, SNP, traits as well as plant introduction or strains/populations. It also incorporate several multi-omics datasets including transcriptomics, proteomics, metabolomics, epigenomics, molecular breeding, phenotype and other types.

KBCommons also provides a suite of tools such as the gene/metabolite pathway viewer, Protein Bio-Viewer, 3D protein structure viewer, heatmaps, scatter plots and hierarchical clustering. KBCommons also has suite for differential expression analysis of omics datasets including venn diagrams, volcano plots, function enrichment and gene modules. It will also incorporate access to Pegasus analytics workflows for analysis in future.

KBCommons provides an informatics-based social network system to build connections among scientists, researchers, producers and consumers.

https://kbcommons.org

Zeng et al. 2018. BIBM; 2019 BMC Genomics

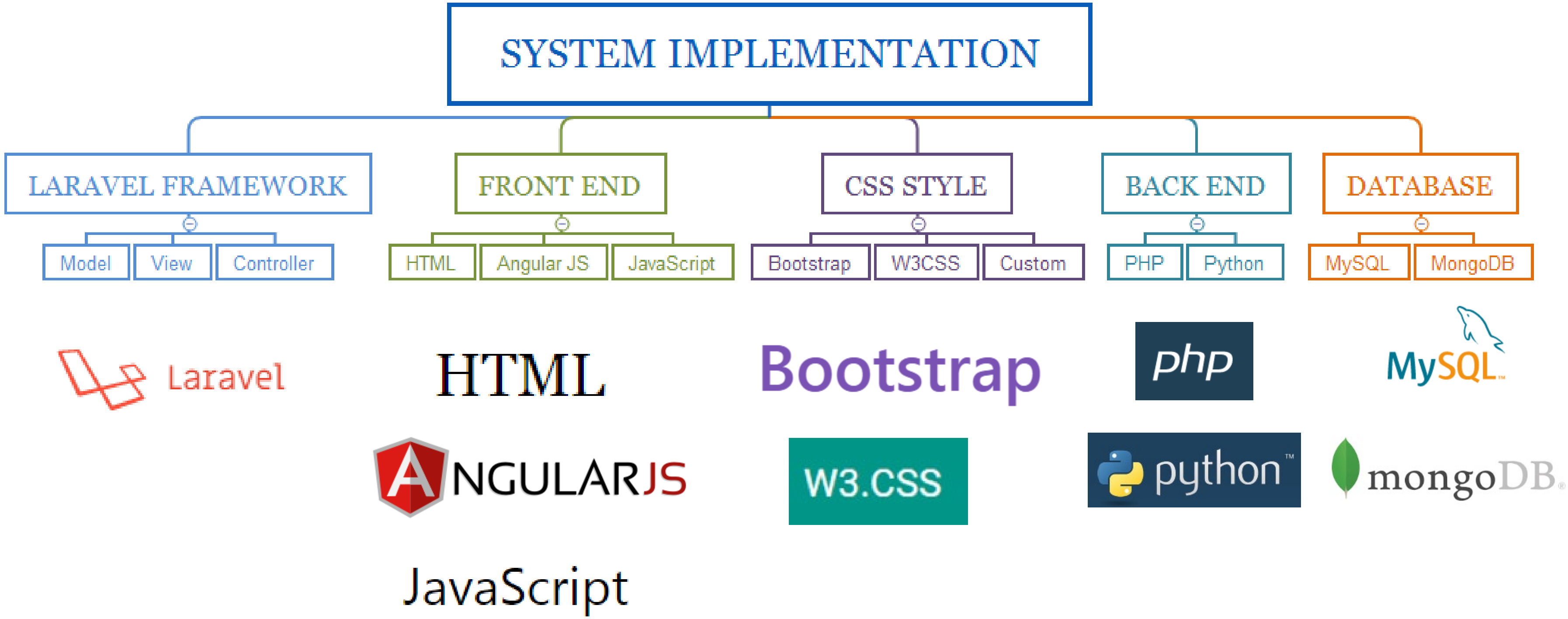
Browse KBCommons **Contribute to KB**

Add version to KB **Create new KB**

Available for various important species including:
- Plant & Crops (Maize, Arabidopsis, Rice, Camelina, Cuphea, Pennycress)
- Animal & Pets (Mouse, Rat, Dog, Worm)
- Human & Diseases (Cancer, Endometriosis)
- Microbes & Viruses (HIV, B. japonicum, SARS-CoV-2)

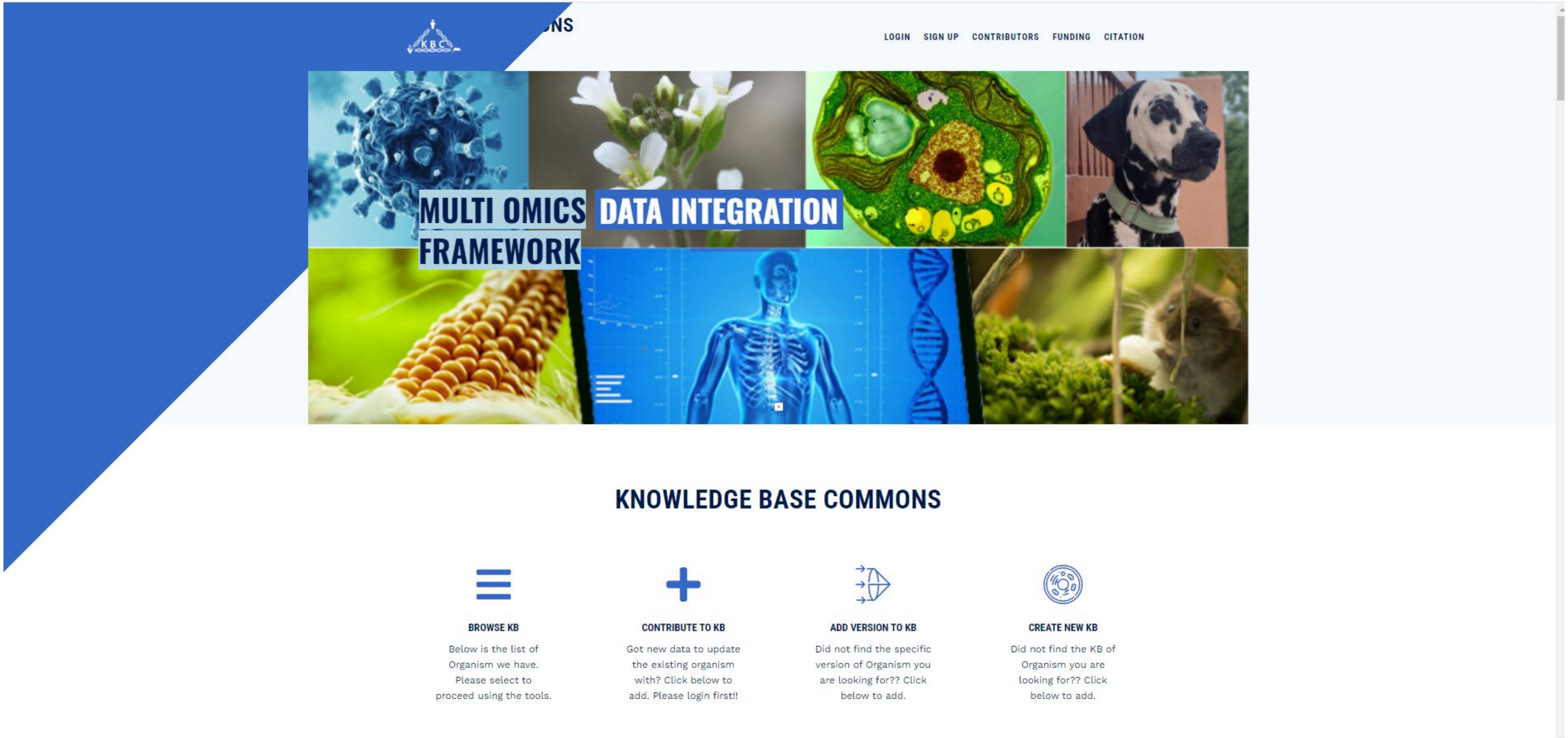
KBCommons Architecture

- KBCommons is implemented using Laravel-PHP framework, Javascript, PHP, MySQL and MongoDB databases.



- SYSTEM IMPLEMENTATION
  - LARAVEL FRAMEWORK
    - Model
    - View
    - Controller
  - FRONT END
    - HTML
    - Angular JS
    - JavaScript
  - CSS STYLE
    - Bootstrap
    - W3CSS
    - Custom
  - BACK END
    - PHP
    - Python
  - DATABASE
    - MySQL
    - MongoDB

KBCommons - V1 Website redesign currently ongoing !! - V2



KNOWLEDGE BASE COMMONS

BROWSE KB
Below is the list of Organism we have. Please select to proceed using the tools.

CONTRIBUTE TO KB
Got new data to update the existing organism with? Click below to add. Please login first!!

ADD VERSION TO KB
Did not find the specific version of Organism you are looking for?? Click below to add.

CREATE NEW KB
Did not find the KB of Organism you are looking for?? Click below to add.

3D Omics Studio

Differential Expression Suite being upgraded to 3D Omics Studio to support diverse multiomics data and now incorporate Upset Plots, Function and Pathways Enrichment Analysis (GO, KEGG, Reactome) and PPI Networks

Differential Expression

Select the organism, gene version, data type, dataset and method

Organism: MusMusculus

Version: GRCm38_83

DataType: RNA-Seq FPKM - Differential

Dataset: mod83_Notch_WT

Method: cuffdiff

Select multiple conditions

☑ Notch-Control_vs_WT-Control
☑ Notch-IL4_IL13_vs_WT-IL4_IL13
☐ Notch-IL4_IL13_vs_Notch-Control
☐ Notch-LPS_IFNY_vs_Notch-Control
☐ WT-LPS_IFNY_vs_WT-Control

Setup filtering parameters

Fold change method: log2(x): 2

Absolute log: 1

Filter Value: p-value / q-value

q-value: 0.05

Analysis Options: Both / Up-regulated / Down-regulated

Notch-Control_vs_WT-Control

GeneID	Locus	Sample_1	Sample_2	Status	Value_1	Value_2	Fold_change	Test_stat	P_value	Q_value	Significant
Bcl6b	chr11:70224127-70229798	Notch-Control	WT-Control	OK	0.54876	0.0983771	-2.47978	-1.20623	0.0002	0.00326563	yes
Rem1	chr2:152626950-152635198	Notch-Control	WT-Control	OK	1.32786	0.554531	-1.25976	-1.53958	0.00045	0.00656182	yes

2D Volcano

Fold Change: 1

Significance: 0.05

Color Selection

Downregulated: red

NotSignificant

Upset & Venn Plot

1. Select Plot: Upset / Venn

2. Select condition from the plot (optional).

Condition Legend

Code	Name
C1	Notch-Control_vs_WT-Control
C2	Notch-IL4_IL13_vs_WT-IL4_IL13
C3	Notch-LPS_IFNY_vs_WT-LPS_IFNY

Show 10 entries

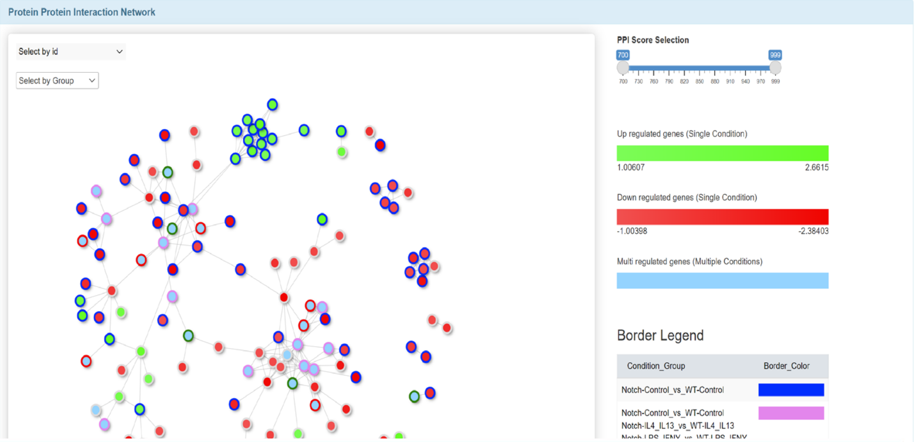
GeneID	Annotation
1500015A07Rik	RIKEN cDNA 1500015A07 gene
3632451O06Rik	
Aak1	AP2 associated kinase 1
Ada	adenosine deaminase

3D Omics Studio

Enrichment Analysis

Term Id	Term Name	Source	Term Size	Parents	P_value	Intersection Size	Precision	Recall	Significant
1901564	organonitrogen compound metabolic process	GO:BP	6589	GO:0006807,GO:0071704	0	1525	0.417	0.231	true
0048518	positive regulation of biological process	GO:BP	6729	GO:0008150,GO:0050789	0	1515	0.415	0.225	true

PPI Network



Up regulated genes (Single Condition)
1.00607 2.6615

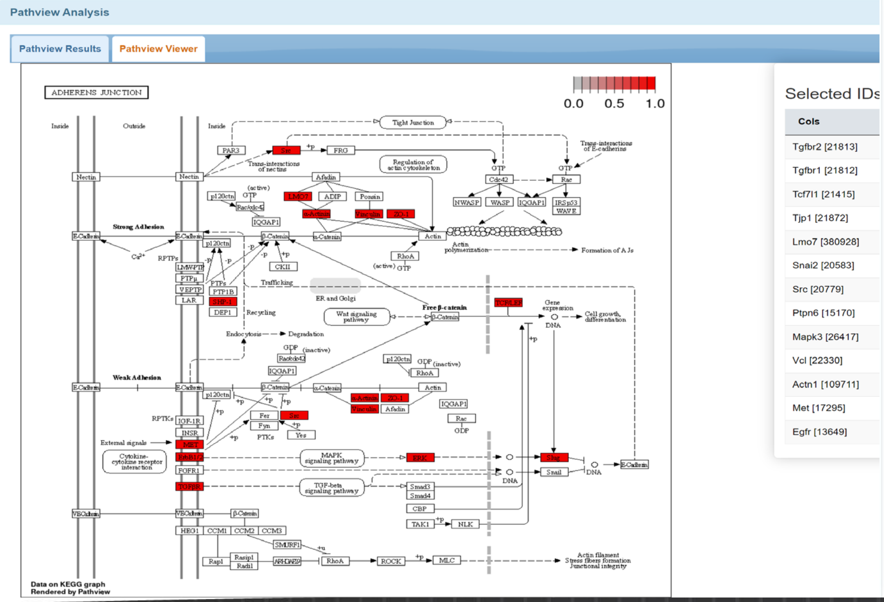
Down regulated genes (Single Condition)
-1.00398 -2.38403

Multi regulated genes (Multiple Conditions)

Border Legend

Condition_Group	Border_Color
Notch-Control_vs_WT-Control	🟦
Notch-Control_vs_WT-Control Notch-IL4_IL13_vs_WT-IL4_IL13	🟪

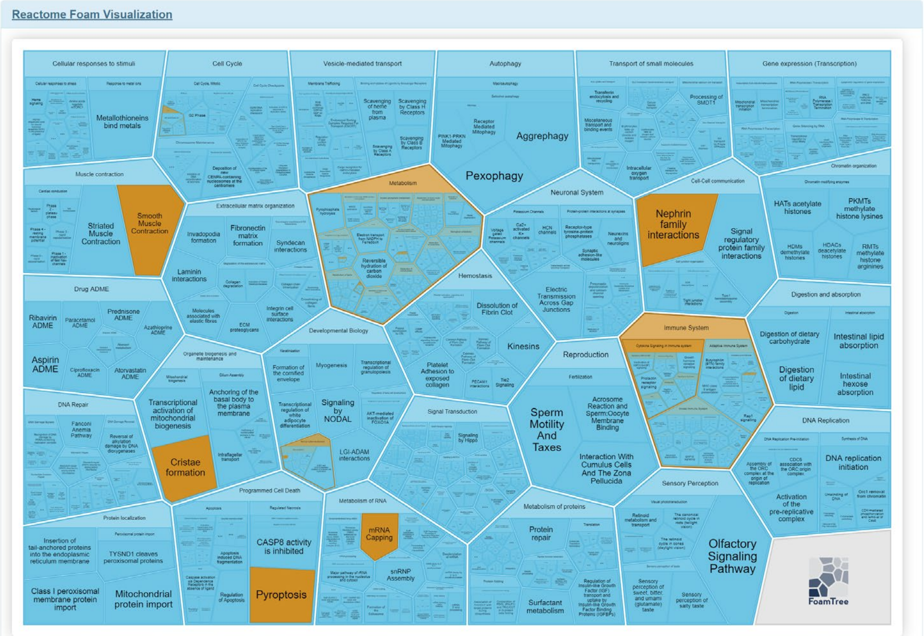
KEGG Pathways



Selected IDs

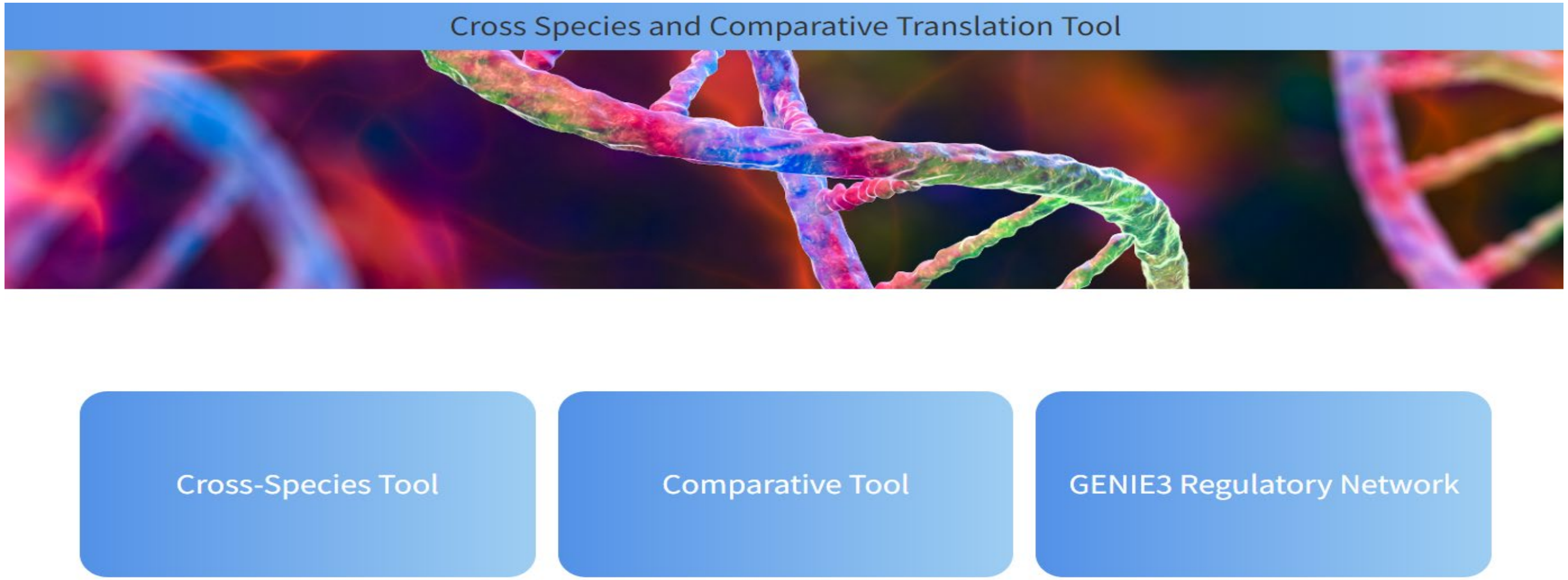
Cols
Tgfbr2 [21813]
Tgfbr1 [21812]
Tcf7l1 [21415]
Tjp1 [21872]
Lmo7 [380928]
Snai2 [20583]
Src [20779]
Ptpn6 [15170]
Mapk3 [26417]
Vcl [22330]
Actn1 [109711]
Met [17295]
Egfr [13649]

Reactome Foam View



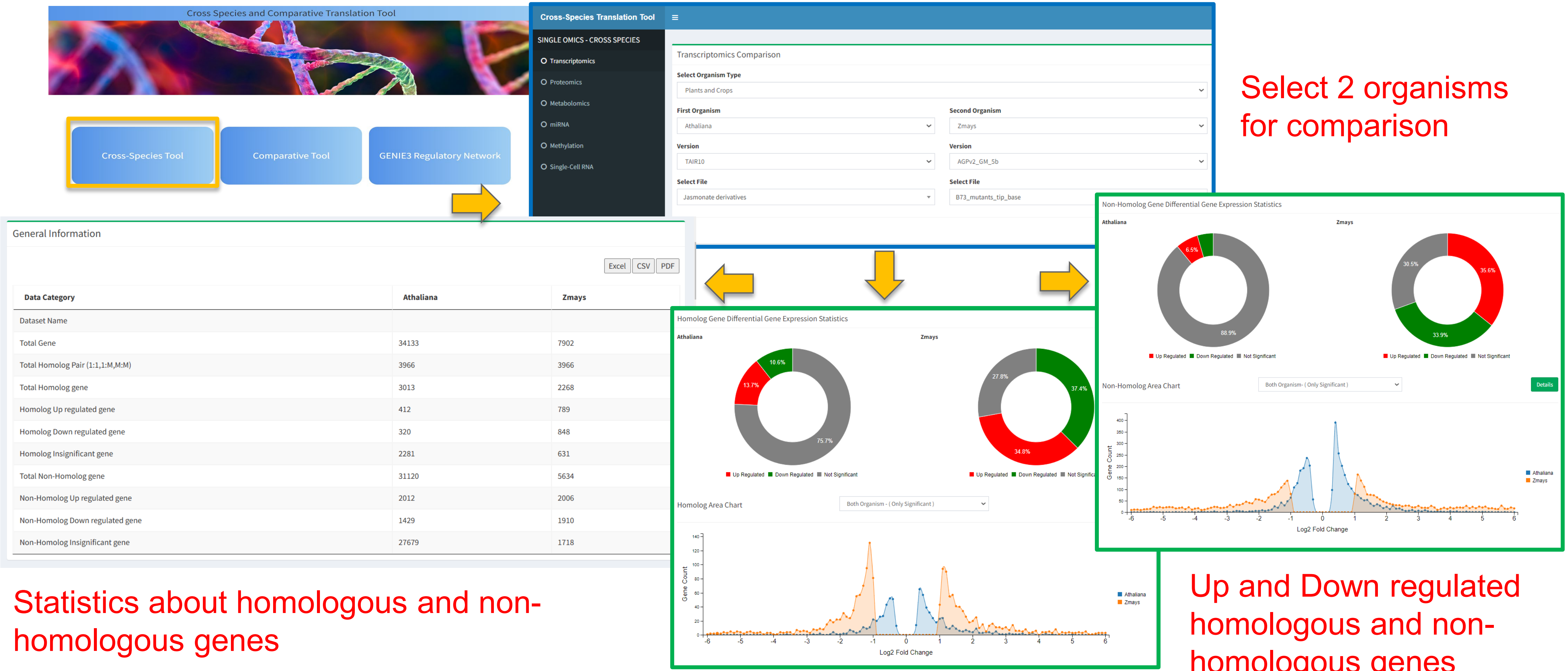
Cross-species and Comparative Multiomics Translation (CCMT) Tool

Tool in KBCommons for comparing multiomics data between species and/or between various multiomics data types within species and generating gene regulatory networks



Cross Species and Comparative Translation Tool

CCMT Tool – Cross-Species



Transcriptomics Comparison

Select Organism Type: Plants and Crops

First Organism	Second Organism
Athaliana	Zmays
Version: TAIR10	Version: AGPv2_GM_5b
Select File: Jasmonate derivatives	Select File: B73_mutants_tip_base

General Information

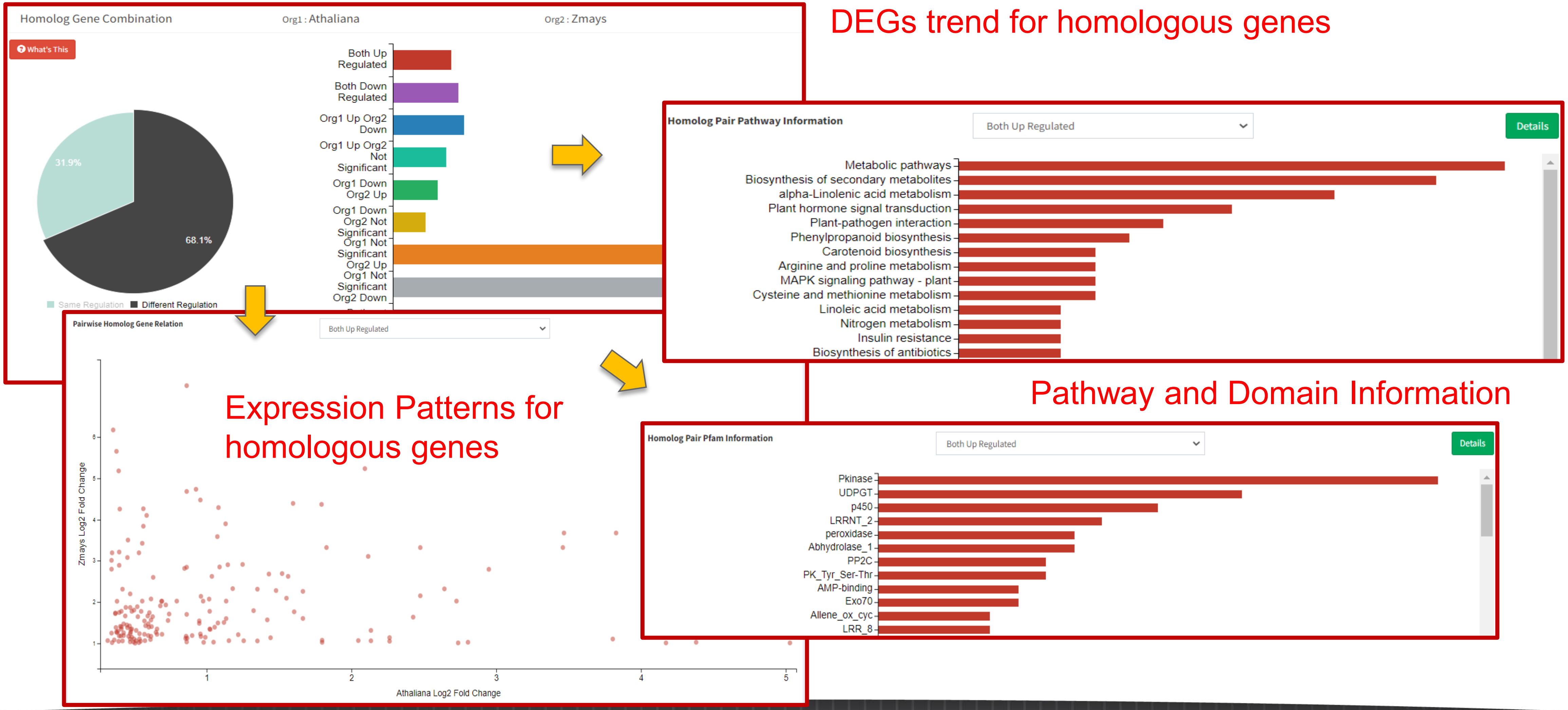
Data Category	Athaliana	Zmays
Dataset Name		
Total Gene	34133	7902
Total Homolog Pair (1:1,1:M,M:M)	3966	3966
Total Homolog gene	3013	2268
Homolog Up regulated gene	412	789
Homolog Down regulated gene	320	848
Homolog Insignificant gene	2281	631
Total Non-Homolog gene	31120	5634
Non-Homolog Up regulated gene	2012	2006
Non-Homolog Down regulated gene	1429	1910
Non-Homolog Insignificant gene	27679	1718

Select 2 organisms for comparison

Statistics about homologous and non-homologous genes

Up and Down regulated homologous and non-homologous genes

CCMT Tool – Cross-Species

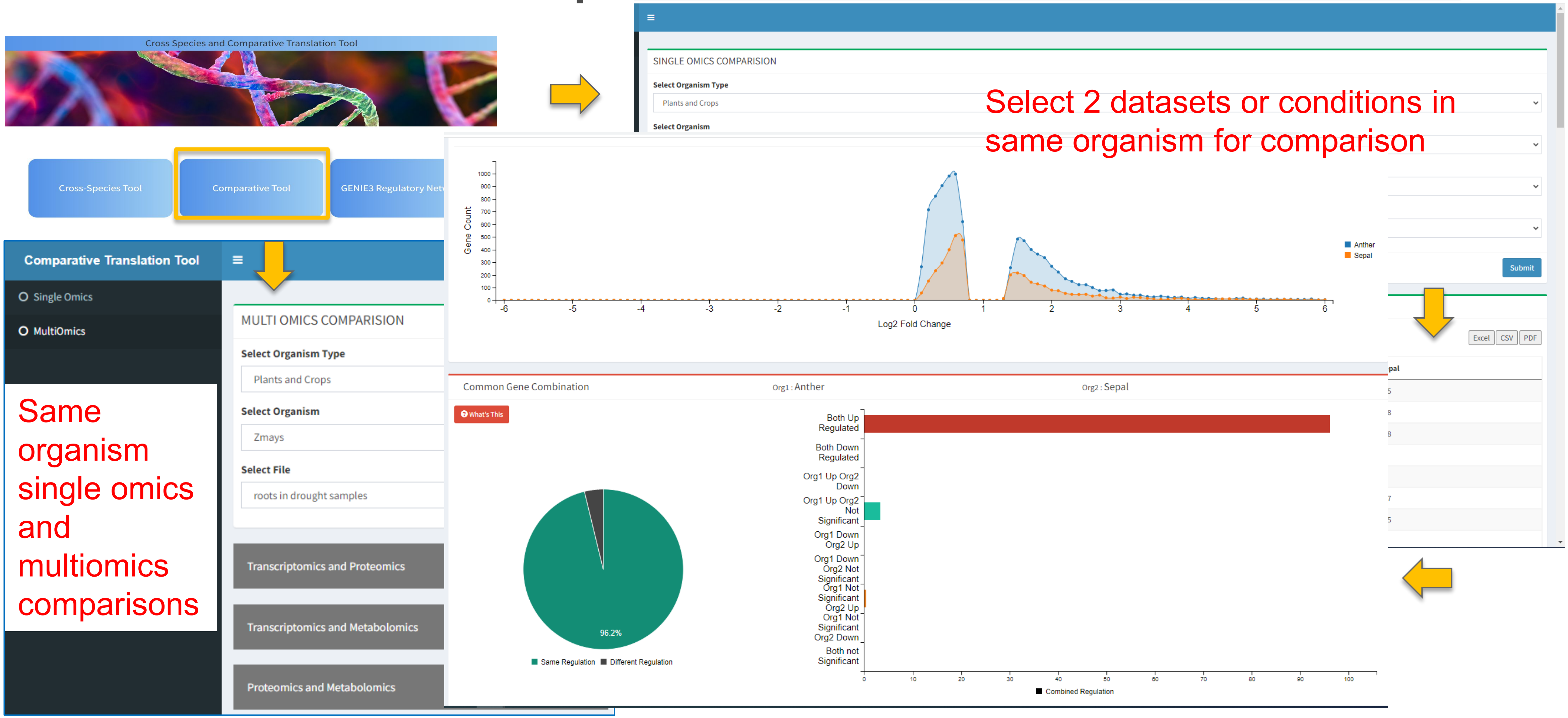


DEGs trend for homologous genes

Expression Patterns for homologous genes

Pathway and Domain Information

CCMT Tool - Comparative



SINGLE OMICS COMPARISION

Select Organism Type: Plants and Crops

Select Organism

MULTI OMICS COMPARISION

Select Organism Type: Plants and Crops

Select Organism: Zmays

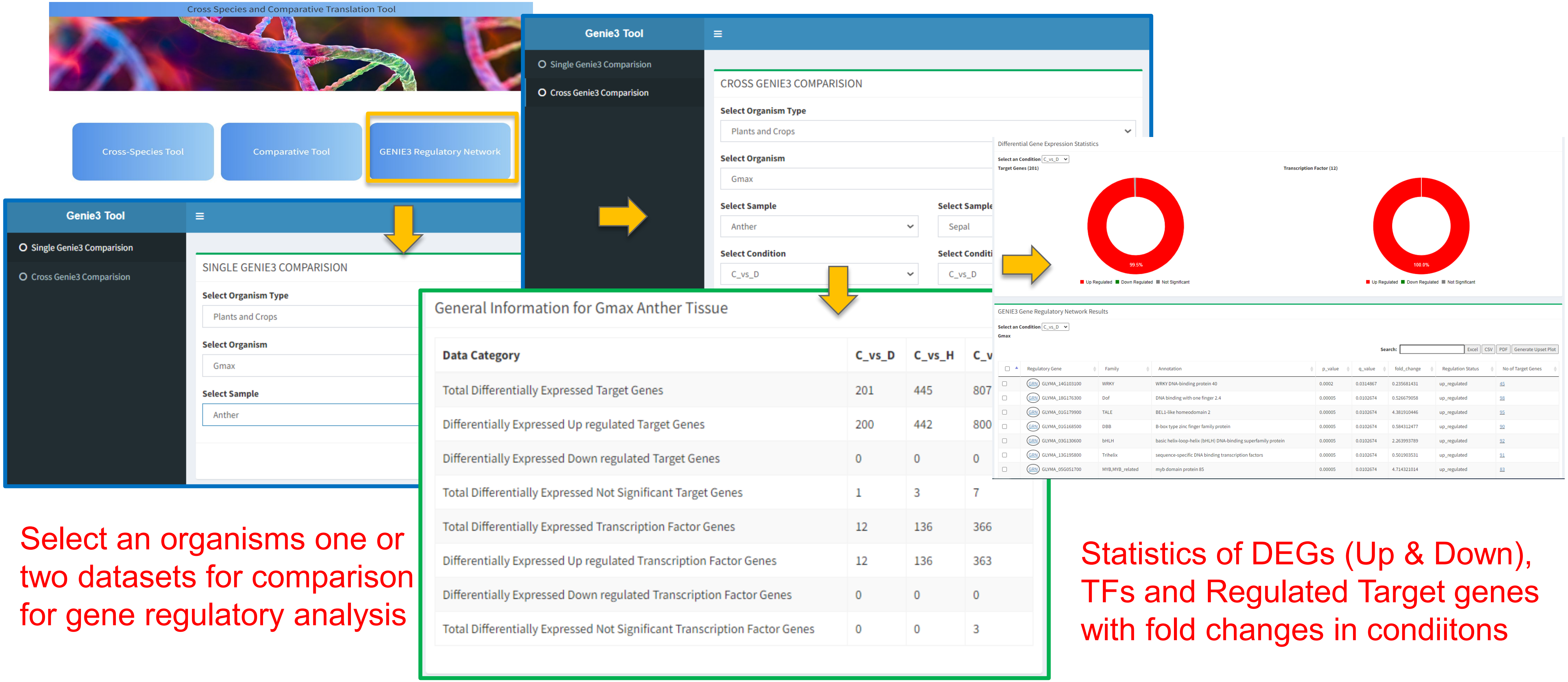
Select File: roots in drought samples

Common Gene Combination — Org1 : Anther — Org2 : Sepal

Select 2 datasets or conditions in same organism for comparison

Same organism single omics and multiomics comparisons

CCMT Tool – Regulatory Network



SINGLE GENIE3 COMPARISION

Select Organism Type
Plants and Crops

Select Organism
Gmax

Select Sample
Anther

CROSS GENIE3 COMPARISION

Select Organism Type
Plants and Crops

Select Organism
Gmax

Select Sample
Anther

Select Sample
Sepal

Select Condition
C_vs_D

Select Conditi
C_vs_D

General Information for Gmax Anther Tissue

Data Category	C_vs_D	C_vs_H	C_v
Total Differentially Expressed Target Genes	201	445	807
Differentially Expressed Up regulated Target Genes	200	442	800
Differentially Expressed Down regulated Target Genes	0	0	0
Total Differentially Expressed Not Significant Target Genes	1	3	7
Total Differentially Expressed Transcription Factor Genes	12	136	366
Differentially Expressed Up regulated Transcription Factor Genes	12	136	363
Differentially Expressed Down regulated Transcription Factor Genes	0	0	0
Total Differentially Expressed Not Significant Transcription Factor Genes	0	0	3

Differential Gene Expression Statistics

Select an Condition C_vs_D

Target Genes (201)

Transcription Factor (12)

GENIE3 Gene Regulatory Network Results

Select an Condition C_vs_D

Gmax

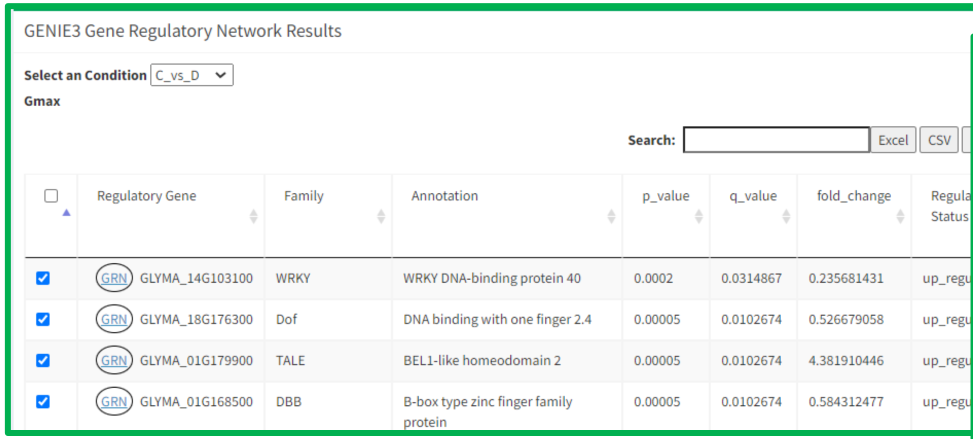
Search:

☐	Regulatory Gene	Family	Annotation	p_value	q_value	fold_change	Regulation Status	No of Target Genes
☐	GRN GLYMA_14G103100	WRKY	WRKY DNA-binding protein 40	0.0002	0.0314867	0.235681431	up_regulated	45
☐	GRN GLYMA_18G176300	Dof	DNA binding with one finger 2.4	0.00005	0.0102674	0.526679058	up_regulated	98
☐	GRN GLYMA_01G179900	TALE	BEL1-like homeodomain 2	0.00005	0.0102674	4.381910446	up_regulated	95
☐	GRN GLYMA_01G168500	DBB	B-box type zinc finger family protein	0.00005	0.0102674	0.584312477	up_regulated	90
☐	GRN GLYMA_03G130600	bHLH	basic helix-loop-helix (bHLH) DNA-binding superfamily protein	0.00005	0.0102674	2.263993789	up_regulated	92
☐	GRN GLYMA_13G195800	Trihelix	sequence-specific DNA binding transcription factors	0.00005	0.0102674	0.501903531	up_regulated	91
☐	GRN GLYMA_05G051700	MYB,MYB_related	myb domain protein 85	0.00005	0.0102674	4.714321014	up_regulated	83

Select an organisms one or two datasets for comparison for gene regulatory analysis

Statistics of DEGs (Up & Down), TFs and Regulated Target genes with fold changes in condititon

CCMT Tool – Regulatory Network



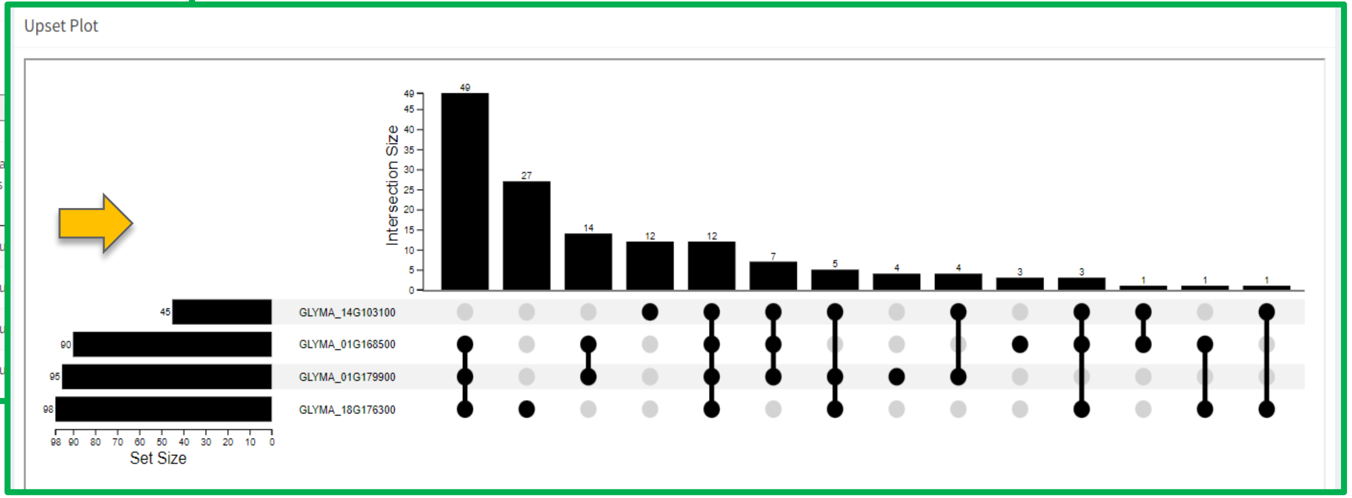
GENIE3 Gene Regulatory Network Results

Select an Condition C_vs_D

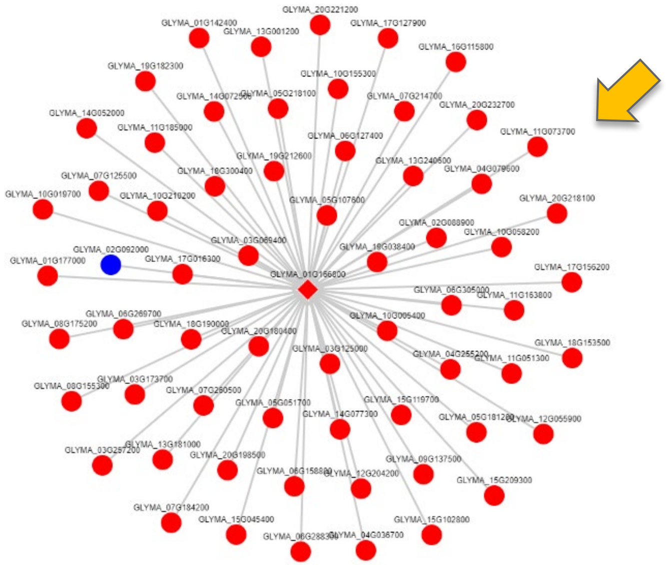
Gmax

Search: Excel CSV

☐	Regulatory Gene	Family	Annotation	p_value	q_value	fold_change	Regula Status
☑	GRN GLYMA_14G103100	WRKY	WRKY DNA-binding protein 40	0.0002	0.0314867	0.235681431	up_regu
☑	GRN GLYMA_18G176300	Dof	DNA binding with one finger 2.4	0.00005	0.0102674	0.526679058	up_regu
☑	GRN GLYMA_01G179900	TALE	BEL1-like homeodomain 2	0.00005	0.0102674	4.381910446	up_regu
☑	GRN GLYMA_01G168500	DBB	B-box type zinc finger family protein	0.00005	0.0102674	0.584312477	up_regu



Upset Plot

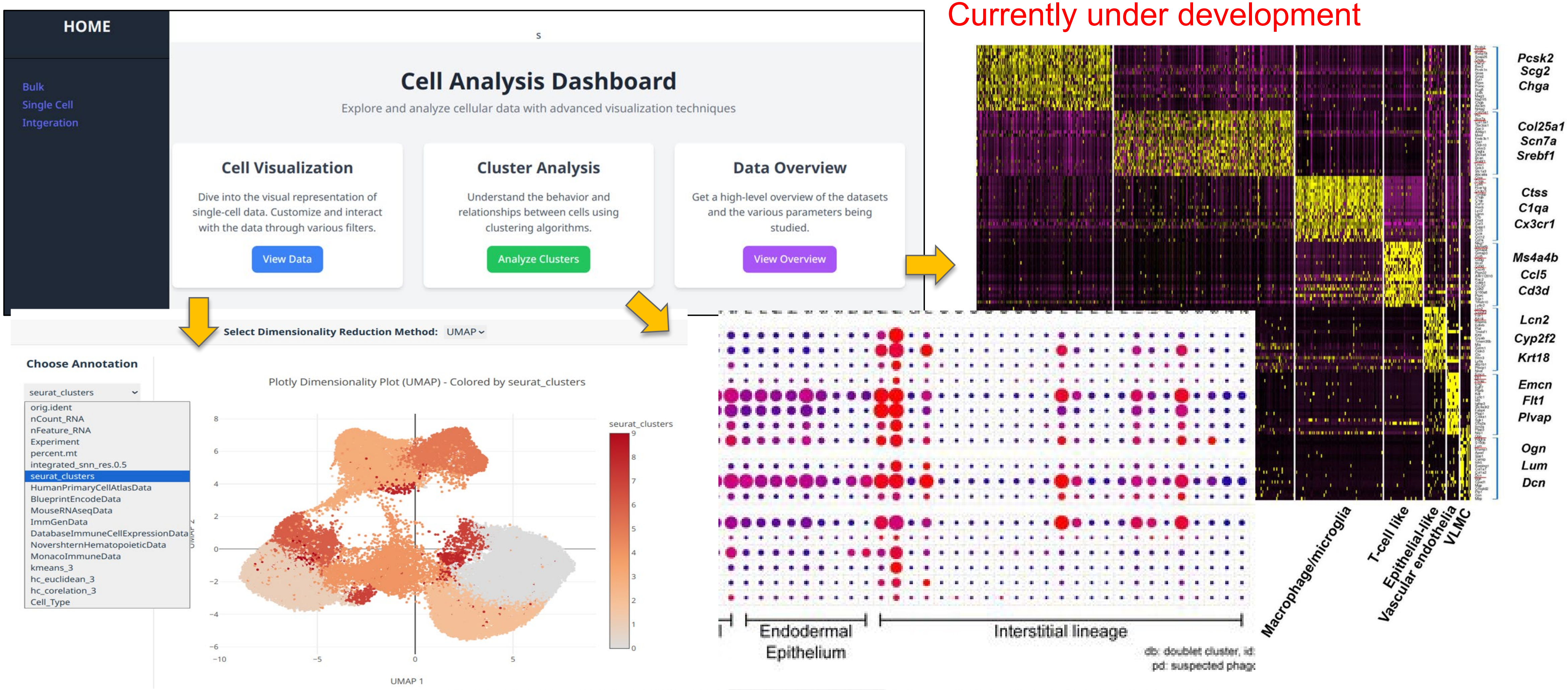


- ● Up Regulated Target Gene
- ● Down Regulated Target Gene
- ● Target Gene
- ◆ Up Regulated Transcription Factor
- ◆ Down Regulated Transcription Factor
- ◆ Transcription Factor
- ▲ Up Regulated Target Gene And Transcription Factor
- ▲ Down Regulated Target Gene And Transcription Factor
- ▲ Down Regulated Target Gene And Transcription Factor

Select 1 or many TFs to see overlap and network of regulated target genes

Single Cell Data Visualization in KBCommons

Currently under development



HOME

Bulk
Single Cell
Intgeration

Cell Analysis Dashboard

Explore and analyze cellular data with advanced visualization techniques

Cell Visualization

Dive into the visual representation of single-cell data. Customize and interact with the data through various filters.

View Data

Cluster Analysis

Understand the behavior and relationships between cells using clustering algorithms.

Analyze Clusters

Data Overview

Get a high-level overview of the datasets and the various parameters being studied.

View Overview

Select Dimensionality Reduction Method: UMAP

Choose Annotation

seurat_clusters

- orig.ident
- nCount_RNA
- nFeature_RNA
- Experiment
- percent.mt
- integrated_snn_res.0.5
- seurat_clusters
- HumanPrimaryCellAtlasData
- BlueprintEncodeData
- MouseRNAseqData
- ImmGenData
- DatabaseImmuneCellExpressionData
- NovershternHematopoieticData
- MonacoImmuneData
- kmeans_3
- hc_euclidean_3
- hc_corelation_3
- Cell_Type

Plotly Dimensionality Plot (UMAP) - Colored by seurat_clusters

Translational Bioinformatics Resources and AI Solutions

- Multiomics Data Integration Tool Development
  - Web Based Tools in SoyKB and KBCommons

– SNPViz v2.0

– AccuTool

– Allele Catalog

– GenVarX

- Phenotype Distribution Tool

- Protein Sequence Logo Tool

- MADis

NGS Resequencing Data in Soybean

Soybean Curated Panels	Soy775	Soy1066	Soy2939 ★
Number of Positions	37,207,330	38,302,167	44,301,648
Number of Positions (SNPs)	31,694,543	32,524,427	37,351,027
Number of Positions (Indels)	5,738,668	5,523,044	6,691,310
Number of Samples	770	1,066	2,939

Utilizes the NGS resequencing data generated over a decade ago from prior USB investments and enables newer discoveries by making it more accessible and interpretable for the community

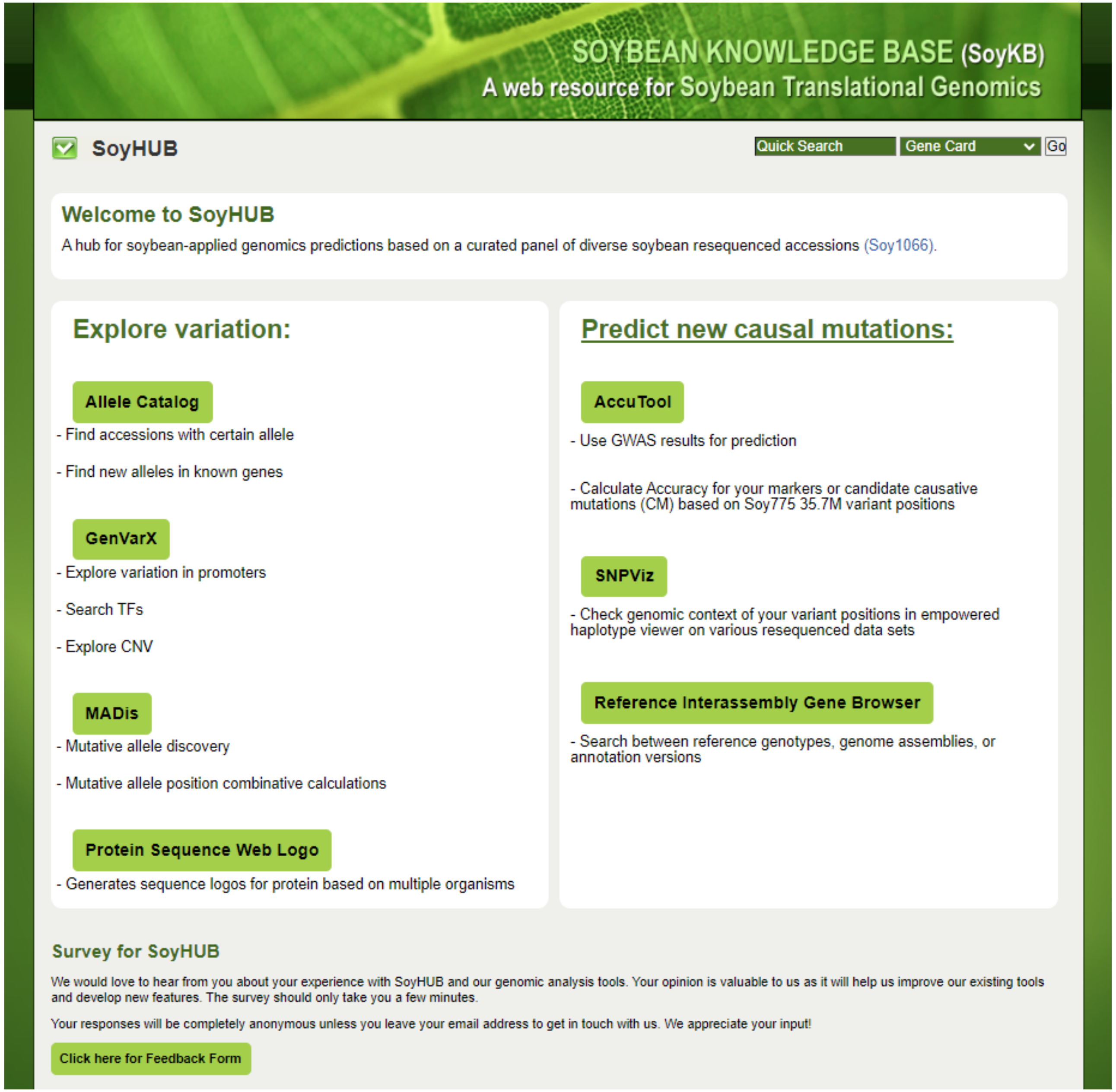
Curated panels upgraded with newer accessions data from our collaborators

Community can submit their accessions of interest for yearly upgrades in future

SoyHub in SoyKB

https://soykb.org/soyhub.php/

A “HUB” for collection of soybean applied genomics prediction tools based on the soybean curated panels (Soy775, Soy1066 and Soy2939) including diverse resequenced accessions.



SOYBEAN KNOWLEDGE BASE (SoyKB)
A web resource for Soybean Translational Genomics

SoyHUB

Welcome to SoyHUB

A hub for soybean-applied genomics predictions based on a curated panel of diverse soybean resequenced accessions (Soy1066).

Explore variation:

Allele Catalog
- Find accessions with certain allele
- Find new alleles in known genes

GenVarX
- Explore variation in promoters
- Search TFs
- Explore CNV

MADis
- Mutative allele discovery
- Mutative allele position combinative calculations

Protein Sequence Web Logo
- Generates sequence logos for protein based on multiple organisms

Predict new causal mutations:

AccuTool
- Use GWAS results for prediction
- Calculate Accuracy for your markers or candidate causative mutations (CM) based on Soy775 35.7M variant positions

SNPViz
- Check genomic context of your variant positions in empowered haplotype viewer on various resequenced data sets

Reference Interassembly Gene Browser
- Search between reference genotypes, genome assemblies, or annotation versions

Survey for SoyHUB

We would love to hear from you about your experience with SoyHUB and our genomic analysis tools. Your opinion is valuable to us as it will help us improve our existing tools and develop new features. The survey should only take you a few minutes.

Your responses will be completely anonymous unless you leave your email address to get in touch with us. We appreciate your input!

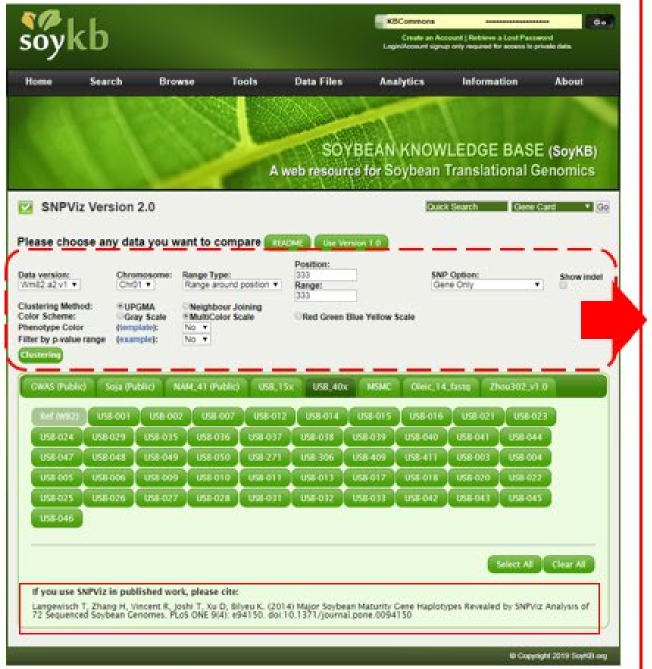
Tools and Methods Developed for Soybean (and beyond)

- **SNPViz**
  http://soykb.org/SNPViz2/
- **AccuTool**
  https://soykb.org/AccuTool/
- **Soybean Allele Catalog**
  https://soykb.org/SoybeanAlleleCatalogTool/
- **Soybean Genomic Variation Explorer (GenVarX)**
  https://soykb.org/SoybeanGenVarX/
- **Soybean Phenotype Distribution Tool**
  https://soykb.org/SoybeanPhenoDistTool
- **Soy Protein Sequence Logo Tool**
  https://soykb.org/SoybeanProteinSequenceLogos/
- **Soybean Multiple Alleles Discovery (MADis) Tool**
  https://soykb.org/SoybeanMADisTool/

SNPViz v2.0 for Haplotype Analysis and Allele Discovery

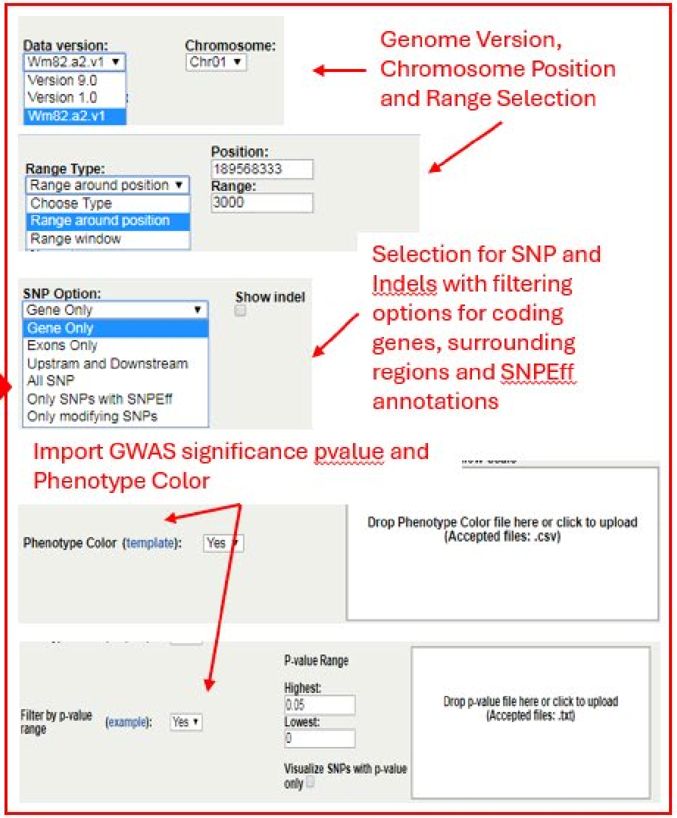
Langewisch et al. Plos One 2014

Zeng et al. BIBM 2020, IJDMB 2021



If you use SNPViz in published work, please cite:
Langewisch T, Zhang H, Vincent R, Joshi T, Xu D, Bilyeu K. (2014) Major Soybean Maturity Gene Haplotypes Revealed by SNPViz Analysis of 72 Sequenced Soybean Genomes. PLoS ONE 9(4): e94150. doi:10.1371/journal.pone.0094150

SNPViz v2.0 for haplotype analysis and visualization of phylogeny and haplotype to discovery allelic variations.

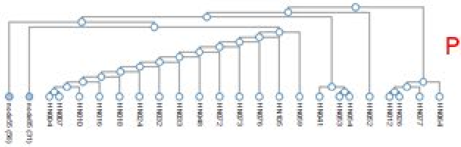


Data version: Wm82.a2.v1
Chromosome: Chr01
Range Type: Range around position
Position: 189568333
Range: 3000
SNP Option: Gene Only
Phenotype Color (template): Yes
Filter by p-value range (example): Yes
Highest: 0.05
Lowest: 0

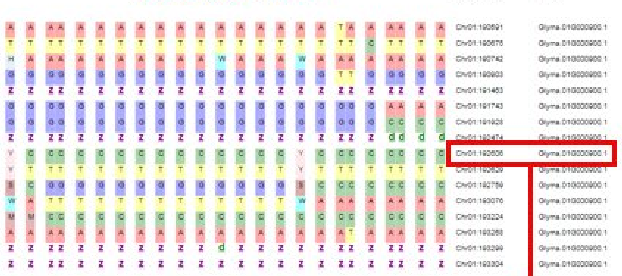
Genome Version, Chromosome Position and Range Selection

Selection for SNP and Indels with filtering options for coding genes, surrounding regions and SNPEff annotations

Import GWAS significance pvalue and Phenotype Color

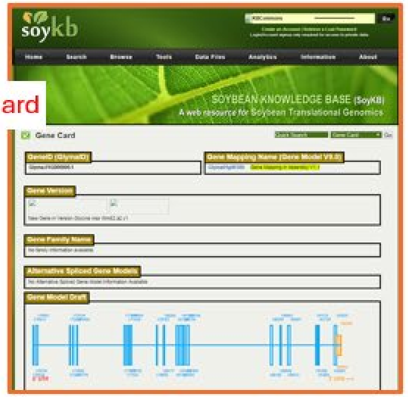


Phylogeny Tree

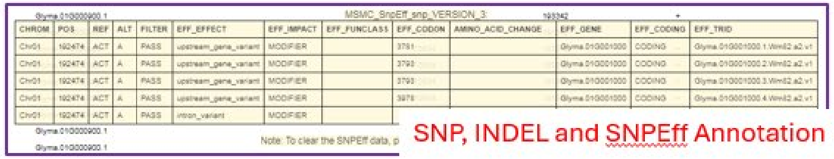


Haplotype Table

Annotation Table



Gene Card

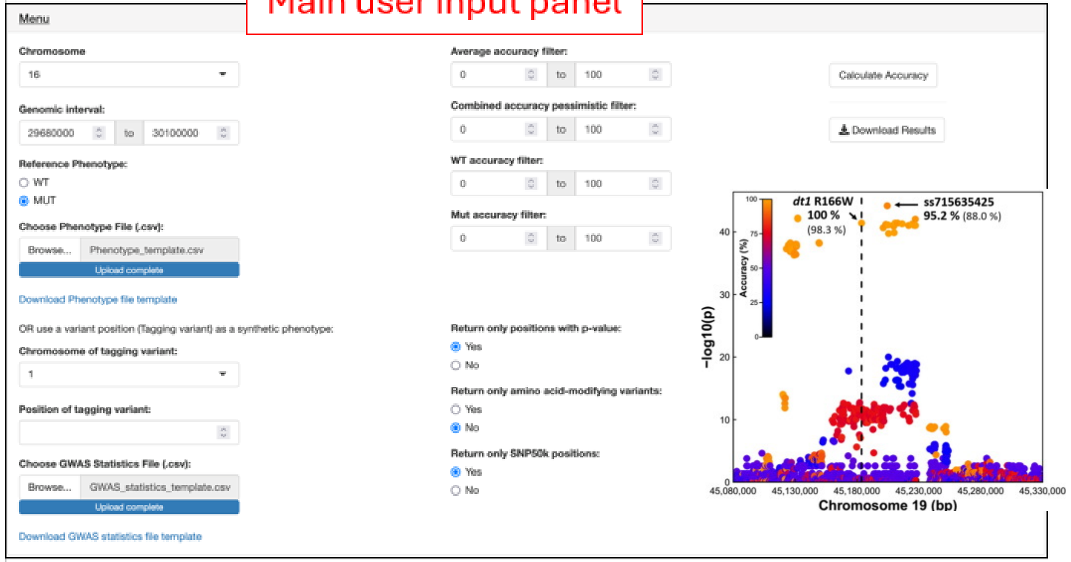


SNP, INDEL and SNPEff Annotation

Tool is designed to utilize Soy Curated Panels to enable efficient identification of soybean genes that control phenotypes by combining GWAS results and haplotype analysis

Soybean Accuracy Tool (AccuTool) for Synthetic Phenotype to Causative Mutation Analysis

Main user input panel



Menu

Chromosome: 16

Genomic interval: 29680000 to 30100000

Reference Phenotype:
○ WT
● MUT

Choose Phenotype File (.csv): Browse... Phenotype_template.csv
Upload complete

Download Phenotype file template

OR use a variant position (Tagging variant) as a synthetic phenotype:

Chromosome of tagging variant: 1

Position of tagging variant:

Choose GWAS Statistics File (.csv): Browse... GWAS_statistics_template.csv
Upload complete

Download GWAS statistics file template

Average accuracy filter: 0 to 100

Combined accuracy pessimistic filter: 0 to 100

WT accuracy filter: 0 to 100

Mut accuracy filter: 0 to 100

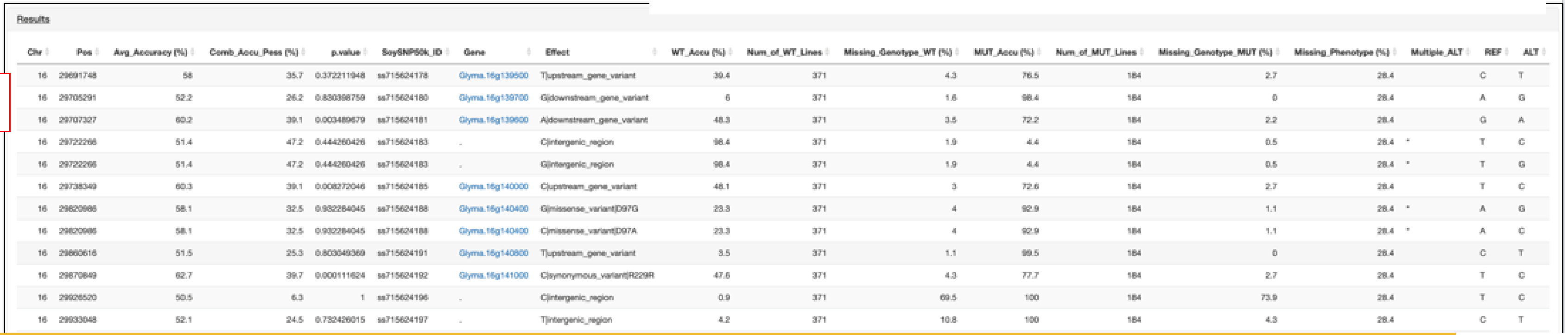
Return only positions with p-value:
● Yes
○ No

Return only amino acid-modifying variants:
○ Yes
● No

Return only SNP50k positions:
● Yes
○ No

- AccuTool is a web-based product of the implementation of the Synthetic Phenotype Association Study (SPAS) approach developed to enhance Genome-Wide Association Study (GWAS).
- It allows calculation of the accuracy by associating wildtype (WT) and mutative (MUT) alleles and phenotypes to understand the impact of the mutation in order to capture causative mutations.
- Through this methodology, more causative mutations of novel genes can be discovered effectively to evaluate the GWAS associations.

Result section



Results

Chr	Pos	Avg_Accuracy (%)	Comb_Accu_Pess (%)	p.value	SoySNP50k_ID	Gene	Effect	WT_Accu (%)	Num_of_WT_Lines	Missing_Genotype_WT (%)	MUT_Accu (%)	Num_of_MUT_Lines	Missing_Genotype_MUT (%)	Missing_Phenotype (%)	Multiple_ALT	REF	ALT
16	29691748	58	35.7	0.372211948	ss715624178	Glyma.16g139500	T|upstream_gene_variant	39.4	371	4.3	76.5	184	2.7	28.4		C	T
16	29705291	52.2	26.2	0.830398759	ss715624180	Glyma.16g139700	G|downstream_gene_variant	6	371	1.6	98.4	184	0	28.4		A	G
16	29707327	60.2	39.1	0.003489679	ss715624181	Glyma.16g139600	A|downstream_gene_variant	48.3	371	3.5	72.2	184	2.2	28.4		G	A
16	29722266	51.4	47.2	0.444260426	ss715624183	.	C|intergenic_region	98.4	371	1.9	4.4	184	0.5	28.4	*	T	C
16	29722266	51.4	47.2	0.444260426	ss715624183	.	G|intergenic_region	98.4	371	1.9	4.4	184	0.5	28.4	*	T	G
16	29738349	60.3	39.1	0.008272046	ss715624185	Glyma.16g140000	C|upstream_gene_variant	48.1	371	3	72.6	184	2.7	28.4		T	C
16	29820986	58.1	32.5	0.932284045	ss715624188	Glyma.16g140400	G|missense_variant|D97G	23.3	371	4	92.9	184	1.1	28.4	*	A	G
16	29820986	58.1	32.5	0.932284045	ss715624188	Glyma.16g140400	C|missense_variant|D97A	23.3	371	4	92.9	184	1.1	28.4	*	A	C
16	29860616	51.5	25.3	0.803049369	ss715624191	Glyma.16g140800	T|upstream_gene_variant	3.5	371	1.1	99.5	184	0	28.4		C	T
16	29870849	62.7	39.7	0.000111624	ss715624192	Glyma.16g141000	C|synonymous_variant|R229R	47.6	371	4.3	77.7	184	2.7	28.4		T	C
16	29926520	50.5	6.3	1	ss715624196	.	C|intergenic_region	0.9	371	69.5	100	184	73.9	28.4		T	C
16	29933048	52.1	24.5	0.732426015	ss715624197	.	T|intergenic_region	4.2	371	10.8	100	184	4.3	28.4		C	T

**Biova et al.
Genes 2023**

Skrabisova et al. 2022 Journal of Advanced Research

AccuTool is designed to identity causative mutations by effectively evaluating GWAS hits by utilizing Soy Curated Panels

Allele Catalog Tool for Allele Discovery and Phenotype Improvement

Research article | Open access | Published: 09 November 2011

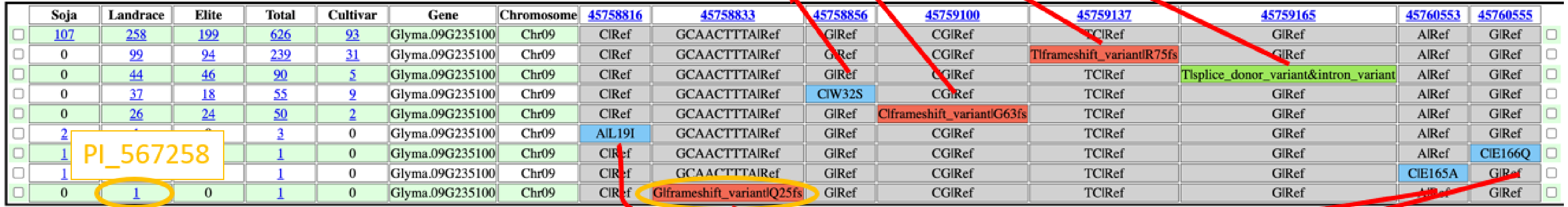
Loss-of-function mutations affecting a specific *Glycine max* R2R3 MYB transcription factor result in brown hilum and brown seed coats

Jason D Gillman, Ashley Tetlow, Jeong-Deong Lee, J Grover Shannon & Kristin Bilyeu

BMC Plant Biology 11, Article number: 155 (2011) | Cite this article

7489 Accesses | 57 Citations | Metrics

R Gene (Glyma.09G235100) Use Case

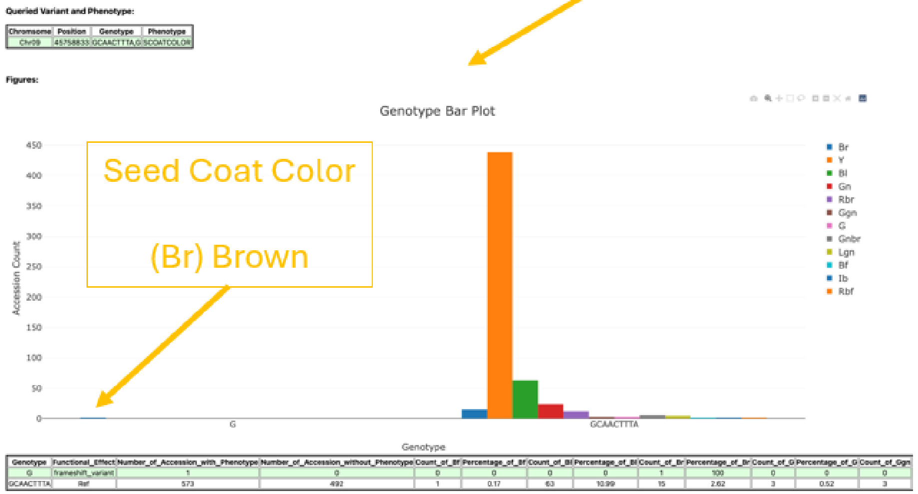


	Soja	Landrace	Elite	Total	Cultivar	Gene	Chromosome	45758816	45758833	45758856	45759100	45759137	45759165	45760553	45760555	
☐	107	258	199	626	93	Glyma.09G235100	Chr09	C|Ref	GCAACTTTA|Ref	G|Ref	CG|Ref	TC|Ref	G|Ref	A|Ref	G|Ref	☐
☐	0	99	94	239	31	Glyma.09G235100	Chr09	C|Ref	GCAACTTTA|Ref	G|Ref	CG|Ref	T|frameshift_variant|R75fs	G|Ref	A|Ref	G|Ref	☐
☐	0	44	46	90	5	Glyma.09G235100	Chr09	C|Ref	GCAACTTTA|Ref	G|Ref	CG|Ref	TC|Ref	T|splice_donor_variant&intron_variant	A|Ref	G|Ref	☐
☐	0	37	18	55	9	Glyma.09G235100	Chr09	C|Ref	GCAACTTTA|Ref	C|W32S	CG|Ref	TC|Ref	G|Ref	A|Ref	G|Ref	☐
☐	0	26	24	50	2	Glyma.09G235100	Chr09	C|Ref	GCAACTTTA|Ref	G|Ref	C|frameshift_variant|G63fs	TC|Ref	G|Ref	A|Ref	G|Ref	☐
☐	2	[illegible]	[illegible]	3	0	Glyma.09G235100	Chr09	A|L19I	GCAACTTTA|Ref	G|Ref	CG|Ref	TC|Ref	G|Ref	A|Ref	G|Ref	☐
☐	1	[illegible]	[illegible]	1	0	Glyma.09G235100	Chr09	C|Ref	GCAACTTTA|Ref	G|Ref	CG|Ref	TC|Ref	G|Ref	A|Ref	C|E166Q	☐
☐	1	[illegible]	[illegible]	1	0	Glyma.09G235100	Chr09	C|Ref	GCAACTTTA|Ref	G|Ref	CG|Ref	TC|Ref	G|Ref	C|E165A	G|Ref	☐
☐	0	1	0	1	0	Glyma.09G235100	Chr09	C|Ref	G|frameshift_variant|Q25fs	G|Ref	CG|Ref	TC|Ref	G|Ref	[illegible]	G|Ref	☐

PI_567258

Newly Discovered Alleles

- The Allele Catalog can be used to discover new alleles in genes
- The R gene is an R2R3 MYB transcription factor that shifts pigment production from brown to black.



Queried Variant and Phenotype:

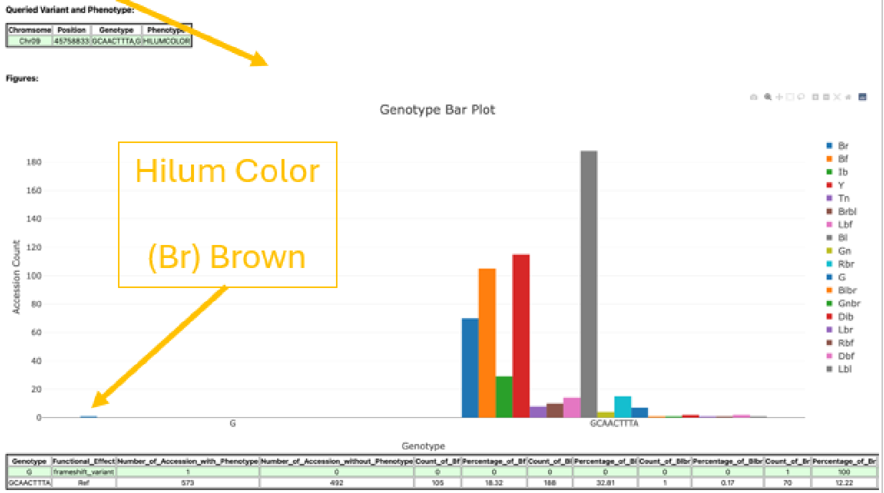
Chromosome	Position	Genotype	Phenotype
Chr09	45758833	GCAACTTTA,G	SCOATCOLOR

Figures:

Seed Coat Color

(Br) Brown

Genotype	Functional_Effect	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype	Count_of_Bf	Percentage_of_Bf	Count_of_Bl	Percentage_of_Bl	Count_of_Br	Percentage_of_Br	Count_of_G	Percentage_of_G	Count_of_Ggn
G	frameshift_variant	1	0	0	0	0	0	1	100	0	0	0
GCAACTTTA	Ref	573	492	1	0.17	63	10.99	15	2.62	3	0.52	3



Queried Variant and Phenotype:

Chromosome	Position	Genotype	Phenotype
Chr09	45758833	GCAACTTTA,G	HILUMCOLOR

Figures:

Hilum Color

(Br) Brown

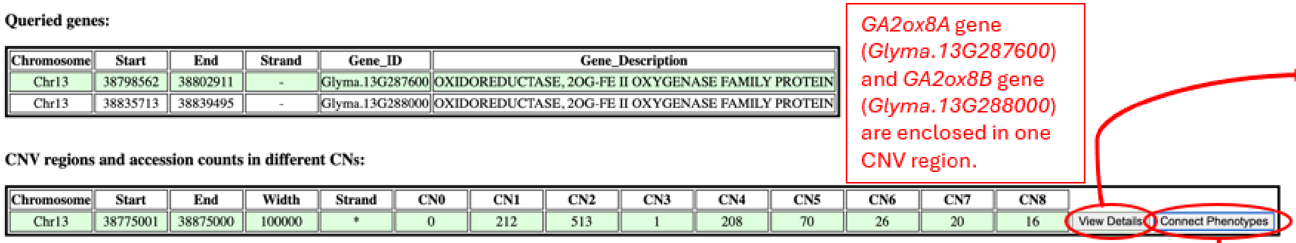
Genotype	Functional_Effect	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype	Count_of_Bf	Percentage_of_Bf	Count_of_Bl	Percentage_of_Bl	Count_of_Blbr	Percentage_of_Blbr	Count_of_Br	Percentage_of_Br
G	frameshift_variant	1	0	0	0	0	0	0	0	1	100
GCAACTTTA	Ref	573	492	105	18.32	188	32.81	1	0.17	70	12.22

Chan et al. 2023. BMC Genomics.

Tool is designed to enhance the discovery and visualization of new alleles and connect causative alleles to phenotypes by utilizing Soy Curated Panels

Genomic Variations Explorer (GenVarX) Use Case to Target Copy Number Variations Affecting Soybean Plant Height

GA2ox8 Gene (*Glyma.13G287600 and Glyma.13G288000*) Use Case



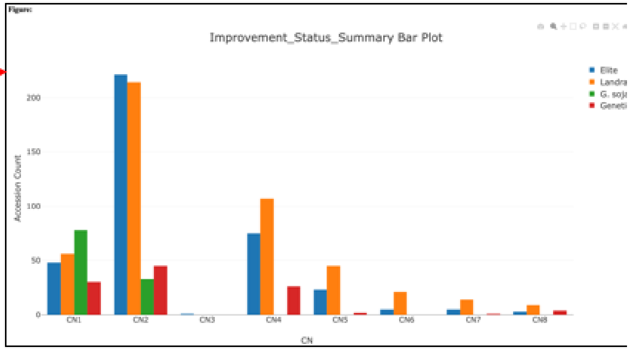
Queried genes:

Chromosome	Start	End	Strand	Gene_ID	Gene_Description
Chr13	38798562	38802911	-	Glyma.13G287600	OXIDOREDUCTASE, 2OG-FE II OXYGENASE FAMILY PROTEIN
Chr13	38835713	38839495	-	Glyma.13G288000	OXIDOREDUCTASE, 2OG-FE II OXYGENASE FAMILY PROTEIN

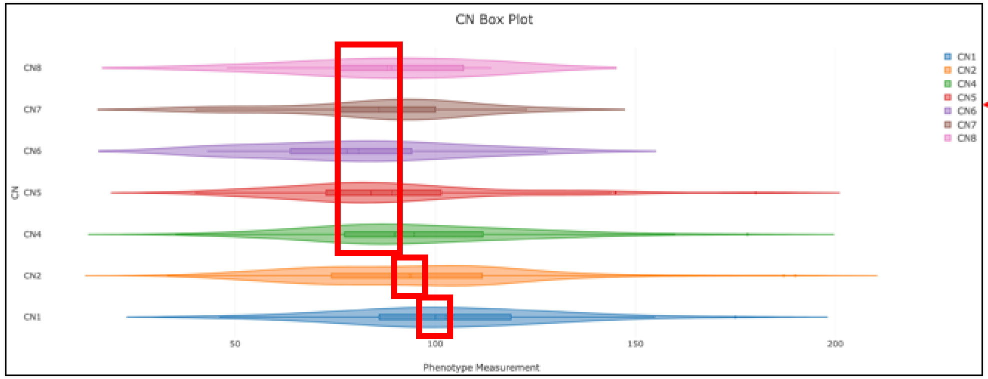
CNV regions and accession counts in different CNs:

Chromosome	Start	End	Width	Strand	CN0	CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8	
Chr13	38775001	38875000	100000	*	0	212	513	1	208	70	26	20	16	View Details Connect Phenotypes

GA2ox8A gene (*Glyma.13G287600*) and *GA2ox8B* gene (*Glyma.13G288000*) are enclosed in one CNV region.



The distribution of improvement status of soybean 1066 accessions in different copy numbers associate to the *GA2ox8* gene.



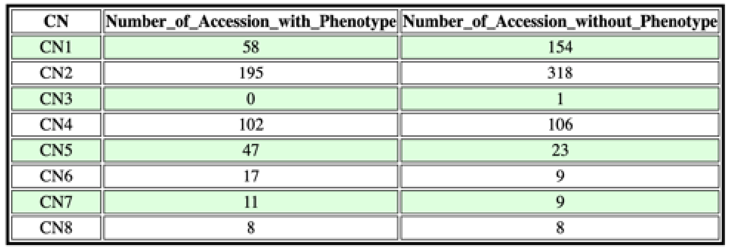
The distribution of plant height phenotype associated to the *GA2ox8* gene is plotted in a violin plot.

CN loss (CN1):
- 212 accessions

CN normal (CN2):
- 513 accessions

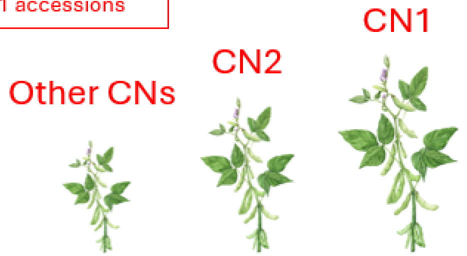
CN gain (CN3 – CN8):
- 341 accessions

Chan et al. 2023. Frontier in Genetics.



CN	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype
CN1	58	154
CN2	195	318
CN3	0	1
CN4	102	106
CN5	47	23
CN6	17	9
CN7	11	9
CN8	8	8

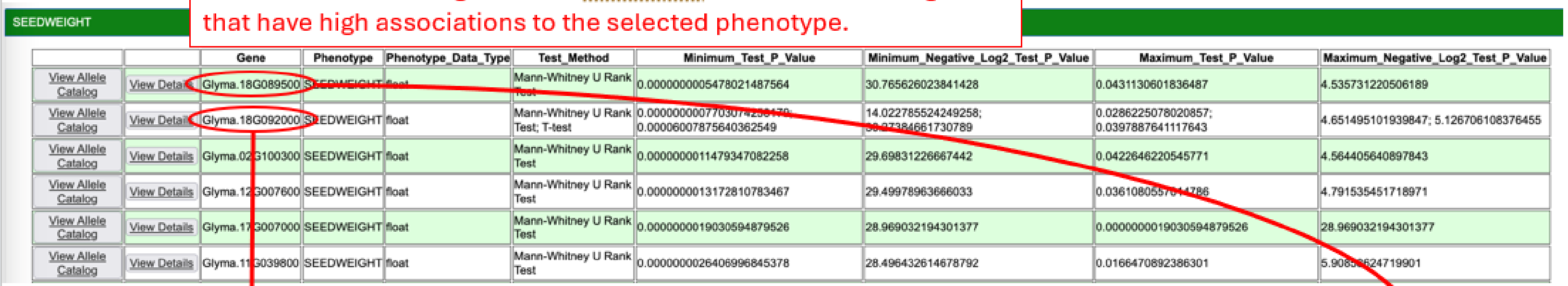
A summary table associated with the plant height phenotype plotted in the violin plot above. This table summarizes the accessions with or without phenotype data in counts.



Tool is designed to utilize Soy Curated Panels for discovery of variations in promoter regions and copy number variations (CNV) impact on phenotypes

Soybean Phenotype Distribution Tool for Allele Discovery from A Phenotype Perspective

Generate phenotype distributions and utilize a mixed of statistics methods to test for significance in order to roll out candidate genes that have high associations to the selected phenotype.

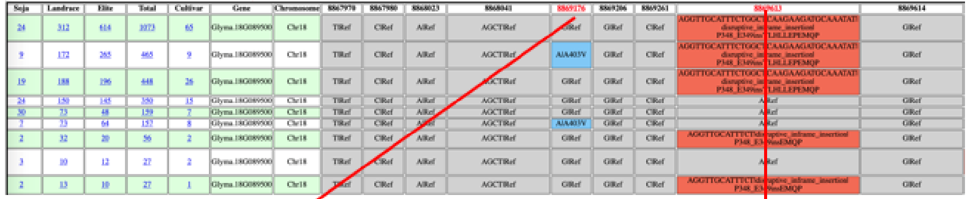


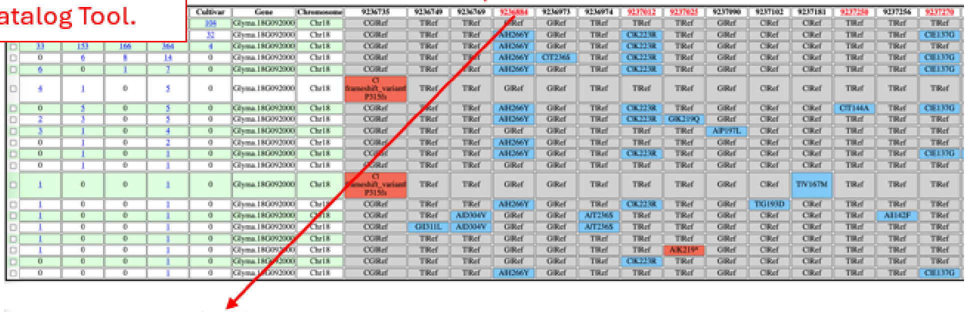
SEEDWEIGHT

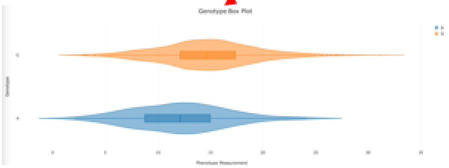
		Gene	Phenotype	Phenotype_Data_Type	Test_Method	Minimum_Test_P_Value	Minimum_Negative_Log2_Test_P_Value	Maximum_Test_P_Value	Maximum_Negative_Log2_Test_P_Value
View Allele Catalog	View Details	Glyma.18G089500	SEEDWEIGHT	float	Mann-Whitney U Rank Test	0.00000000005478021487564	30.765626023841428	0.0431130601836487	4.535731220506189
View Allele Catalog	View Details	Glyma.18G092000	SEEDWEIGHT	float	Mann-Whitney U Rank Test; T-test	0.0000000000770307[illegible]; 0.00006007875640362549	14.022785524249258; [illegible]27384661730789	0.0286225078020857; 0.0397887641117643	4.651495101939847; 5.126706108376455
View Allele Catalog	View Details	Glyma.02G100300	SEEDWEIGHT	float	Mann-Whitney U Rank Test	0.00000000011479347082258	29.69831226667442	0.0422646220545771	4.564405640897843
View Allele Catalog	View Details	Glyma.12G007600	SEEDWEIGHT	float	Mann-Whitney U Rank Test	0.00000000013172810783467	29.49978963666033	0.036108055[illegible]4786	4.791535451718971
View Allele Catalog	View Details	Glyma.17G007000	SEEDWEIGHT	float	Mann-Whitney U Rank Test	0.00000000019030594879526	28.969032194301377	0.00000000019030594879526	28.969032194301377
View Allele Catalog	View Details	Glyma.11G039800	SEEDWEIGHT	float	Mann-Whitney U Rank Test	0.00000000026406996845378	28.496432614678792	0.0166470892386301	5.908[illegible]24719901

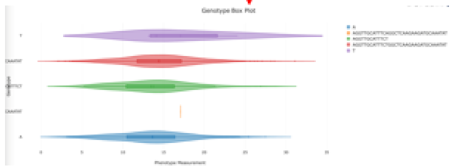
Order by minimum –log2(P-value) in descending order.

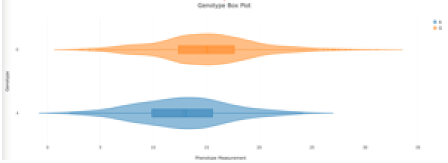
Pinpoint causative variant positions in Allele Catalog Tool.







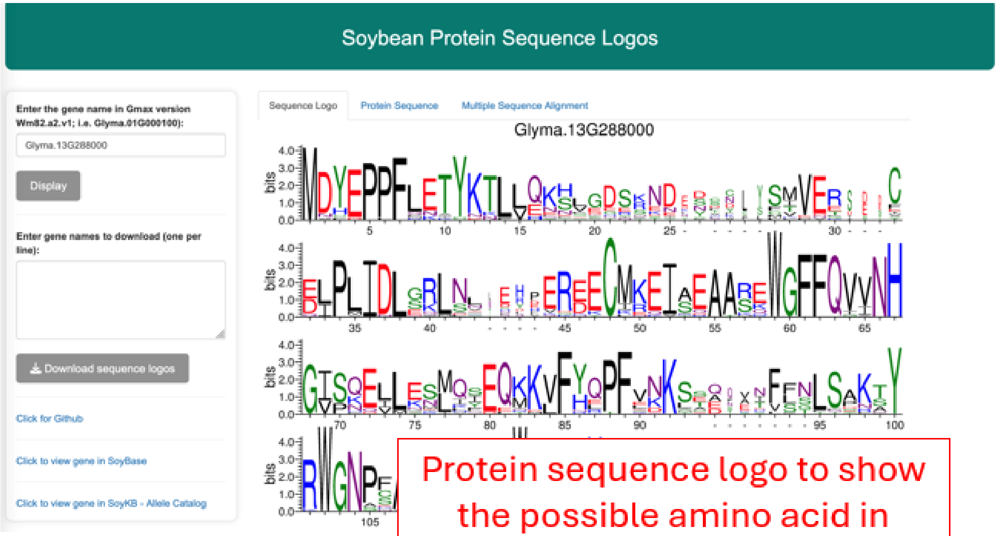




Phenotype distribution visualizations for researchers to understand the phenotype differences in different genotypes.

Tool is designed to easily identify the most important alleles for any given phenotype by ranking all identified novel alleles by using the Allele Catalog Tool

Soybean Protein Sequence Logos Tool for Cross Species Protein Discovery



Soybean Protein Sequence Logos

Enter the gene name in Gmax version Wm82.a2.v1; i.e. Glyma.01G000100):

Glyma.13G288000

Display

Enter gene names to download (one per line):

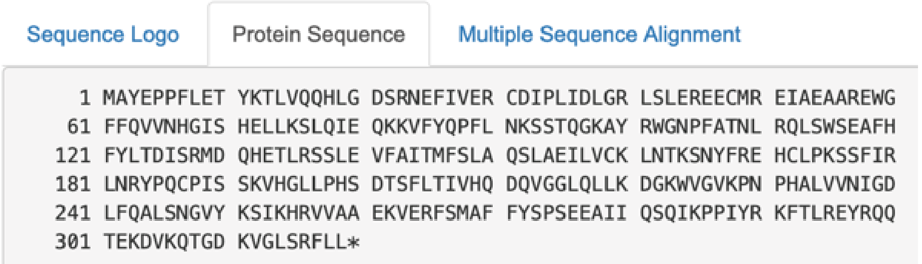
Download sequence logos

Click for Github

Click to view gene in SoyBase

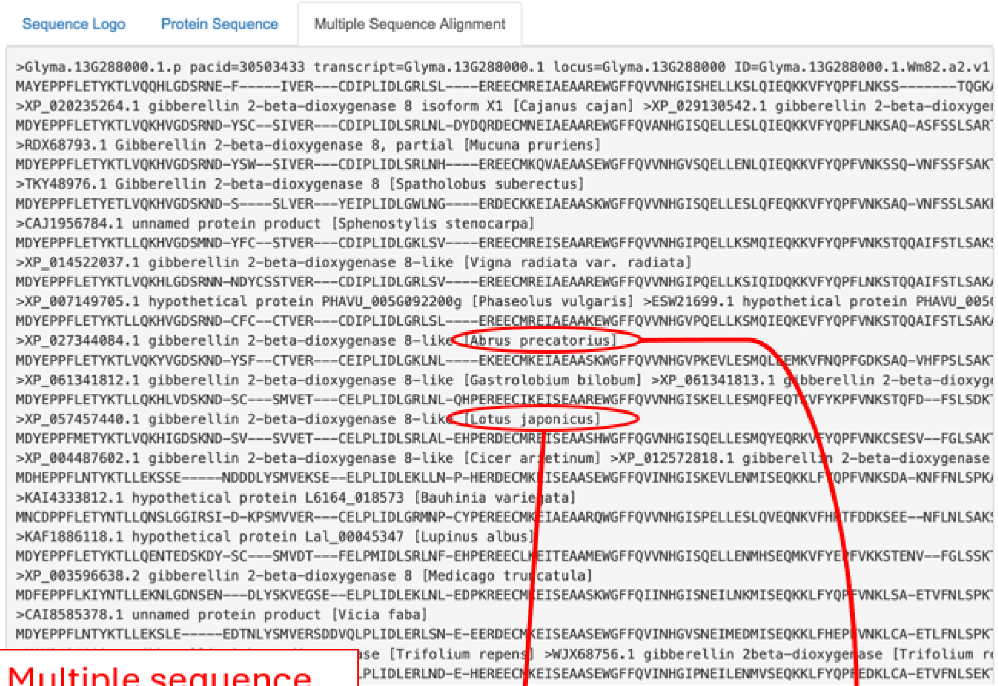
Click to view gene in SoyKB - Allele Catalog

Protein sequence logo to show the possible amino acid in each position of the sequence.



```
  1 MAYEPPFLET YKTLVQQHLG DSRNEFIVER CDIPLIDLGR LSLEREECMR EIAEAAREWG
 61 FFQVVNHGIS HELLKSLQIE QKKVFYQPFL NKSSTQGKAY RWGNPFATNL RQLSWSEAFH
121 FYLTDISRMD QHETLRSSLE VFAITMFSLA QSLAEILVCK LNTKSNYFRE HCLPKSSFIR
181 LNRYPQCPIS SKVHGLLPHS DTSFLTIVHQ DQVGGLQLLK DGKWVGVKPN PHALVVNIGD
241 LFQALSNGVY KSIKHRVVAA EKVERFSMAF FYSPSEEAII QSQIKPPIYR KFTLREYRQQ
301 TEKDVKQTGD KVGLSRFLL*
```

The exact protein sequence from the queried genes.

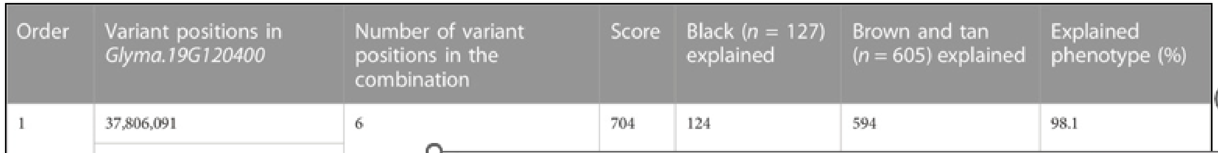


```
>Glyma.13G288000.1.p pacid=30503433 transcript=Glyma.13G288000.1 locus=Glyma.13G288000 ID=Glyma.13G288000.1.Wm82.a2.v1
MAYEPPFLETYKTLVQQHLGDSRNE-F-----IVER---CDIPLIDLGRLSL----EREECMREIAEAAREWGFFQVVNHGISHELLKSLQIEQKKVFYQPFLNKSS-------TQGK
>XP_020235264.1 gibberellin 2-beta-dioxygenase 8 isoform X1 [Cajanus cajan] >XP_029130542.1 gibberellin 2-beta-dioxyge
MDYEPPFLETYKTLVQKHVGDSRND-YSC--SIVER---CDIPLIDLSRLNL-DYDQRDECMNEIAEAAREWGFFQVANHGISQELLESLQIEQKKVFYQPFLNKSAQ-ASFSSLSAR
>RDX68793.1 Gibberellin 2-beta-dioxygenase 8, partial [Mucuna pruriens]
MDYEPPFLETYKTLVQKHVGDSRND-YSW--SIVER---CDIPLIDLSRLNH----EREECMKQVAEAASEWGFFQVVNHGVSQELLENLQIEQKKVFYQPFVNKSSQ-VNFSSFSAK
>TKY48976.1 Gibberellin 2-beta-dioxygenase 8 [Spatholobus suberectus]
MDYEPPFLETYETLVQKHVGDSKND-S----SLVER---YEIPLIDLGWLNG----ERDECKKEIAEAASKWGFFQVVNHGISQELLESLQFEQKKVFYQPFVNKSAQ-VNFSSLSAK
>CAJ1956784.1 unnamed protein product [Sphenostylis stenocarpa]
MDYEPPFLETYKTLLQKHVGDSMND-YFC--STVER---CDIPLIDLGKLSV----EREECMREISEAAREWGFFQVVNHGIPQELLKSMQIEQKKVFYQPFVNKSTQQAIFSTLSAK
>XP_014522037.1 gibberellin 2-beta-dioxygenase 8-like [Vigna radiata var. radiata]
MDYEPPFLETYKTLVQKHLGDSRNN-NDYCSSTVER---CDIPLIDLGRLSV----EREECMREIAEAAREWGFFQVVNHGIPQELLKSIQIDQKKVFYQPFLNKSTQQAIFSTLSAK
>XP_007149705.1 hypothetical protein PHAVU_005G092200g [Phaseolus vulgaris] >ESW21699.1 hypothetical protein PHAVU_005
MDYEPPFLETYKTLLQKHVGDSRND-CFC--CTVER---CDIPLIDLGRLSL----EREECMREIAEAAKEWGFFQVVNHGVPQELLKSMQIEQKEVFYQPFVNKSTQQAIFSTLSAK
>XP_027344084.1 gibberellin 2-beta-dioxygenase 8-like [Abrus precatorius]
MDYEPPFLETYKTLVQKYVGDSKND-YSF--CTVER---CEIPLIDLGKLNL----EKEECMKEIAEAASKWGFFQVVNHGVPKEVLESMQLEEMKVFNQPFGDKSAQ-VHFPSLSAK
>XP_061341812.1 gibberellin 2-beta-dioxygenase 8-like [Gastrolobium bilobum] >XP_061341813.1 gibberellin 2-beta-dioxyge
MDYEPPFLETYKTLLQKHLVDSKND-SC---SMVET---CELPLIDLGRLNL-QHPEREECIKEISEAAREWGFFQVVNHGISKELLESMQFEQTKVFYKPFVNKSTQFD--FSLSDK
>XP_057457440.1 gibberellin 2-beta-dioxygenase 8-like [Lotus japonicus]
MDYEPPFMETYKTLVQKHIGDSKND-SV---SVVET---CELPLIDLSRLAL-EHPERDECMREISEAASHWGFFQGVNHGISQELLESMQYEQRKVFYQPFVNKCSESV--FGLSAK
>XP_004487602.1 gibberellin 2-beta-dioxygenase 8-like [Cicer arietinum] >XP_012572818.1 gibberellin 2-beta-dioxygenase
MDHEPPFLNTYKTLLEKSSE-----NDDDLYSMVEKSE--ELPLIDLEKLLN-P-HERDECMKEISEAASEWGFFQVINHGISKEVLENMISEQKKLFYQPFVNKSDA-KNFFNLSPK
>KAI4333812.1 hypothetical protein L6164_018573 [Bauhinia variegata]
MNCDPPFLETYNTLLQNSLGGIRSI-D-KPSMVVER---CELPLIDLGRMNP-CYPEREECMKEIAEAARQWGFFQVVNHGISPELLESLQVEQNKVFHRTFDDKSEE--NFLNLSAK
>KAF1886118.1 hypothetical protein Lal_00045347 [Lupinus albus]
MDYEPPFLETYKTLLQENTEDSKDY-SC---SMVDT---FELPMIDLSRLNF-EHPEREECLKEITEAAMEWGFFQVVNHGISQELLENMHSEQMKVFYEPFVKKSTENV--FGLSSK
>XP_003596638.2 gibberellin 2-beta-dioxygenase 8 [Medicago truncatula]
MDFEPPFLKIYNTLLEKNLGDNSEN---DLYSKVEGSE--ELPLIDLEKLNL-EDPKREECMKEISEAASKWGFFQIINHGISNEILNKMISEQKKLFYQPFVNKLSA-ETVFNLSPK
>CAI8585378.1 unnamed protein product [Vicia faba]
MDYEPPFLNTYKTLLEKSLE-----EDTNLYSMVERSDDVQLPLIDLERLSN-E-EERDECMKEISEAASEWGFFQVINHGVSNEIMEDMISEQKKLFHEPFVNKLCA-ETLFNLSPK
ase [Trifolium repens] >WJX68756.1 gibberellin 2beta-dioxygenase [Trifolium r
LPLIDLERLND-E-HEREECMKEISEAASEWGFFQVINHGIPNEILENMVSEQKKLFYQPFEDKLCA-ETVFNLSEK
```

Multiple sequence alignment to enable protein discovery in other species.

Tool is designed to support identification and exploration of amino acid diversity in a protein across species to support biological inference of Allele Catalog results

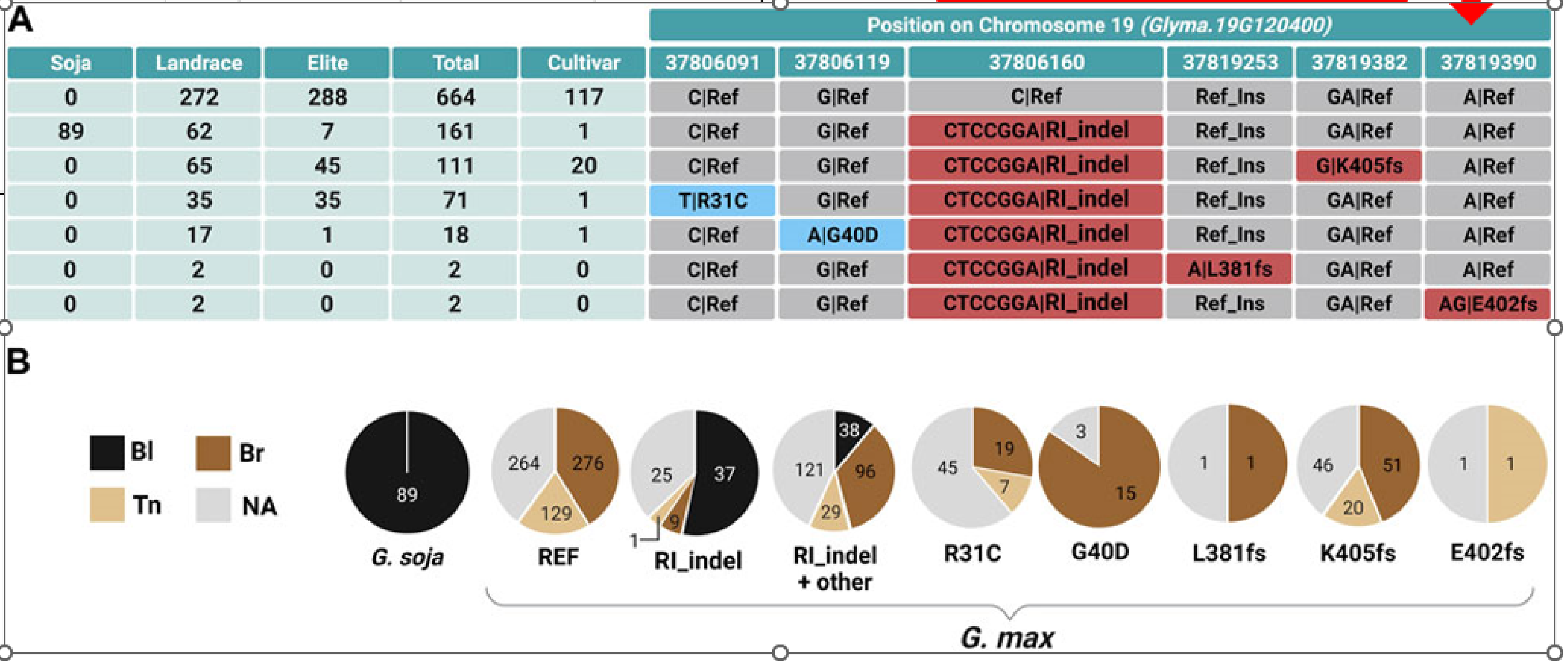
Multiple Allele Discovery (MADis) Tool Use Case to Understand Pod Colors



Order	Variant positions in *Glyma.19G120400*	Number of variant positions in the combination	Score	Black (*n* = 127) explained	Brown and tan (*n* = 605) explained	Explained phenotype (%)
1	37,806,091	6	704	124	594	98.1
	37,806,119					
	37,806,160					
	37,819,253					
	37,819,382					
	37,819,390					

MADis Tool Results

Allele Catalog Tool Results



A

Soja	Landrace	Elite	Total	Cultivar	Position on Chromosome 19 (*Glyma.19G120400*)					
					37806091	37806119	37806160	37819253	37819382	37819390
0	272	288	664	117	C|Ref	G|Ref	C|Ref	Ref_Ins	GA|Ref	A|Ref
89	62	7	161	1	C|Ref	G|Ref	CTCCGGA|RI_indel	Ref_Ins	GA|Ref	A|Ref
0	65	45	111	20	C|Ref	G|Ref	CTCCGGA|RI_indel	Ref_Ins	G|K405fs	A|Ref
0	35	35	71	1	T|R31C	G|Ref	CTCCGGA|RI_indel	Ref_Ins	GA|Ref	A|Ref
0	17	1	18	1	C|Ref	A|G40D	CTCCGGA|RI_indel	Ref_Ins	GA|Ref	A|Ref
0	2	0	2	0	C|Ref	G|Ref	CTCCGGA|RI_indel	A|L381fs	GA|Ref	A|Ref
0	2	0	2	0	C|Ref	G|Ref	CTCCGGA|RI_indel	Ref_Ins	GA|Ref	AG|E402fs

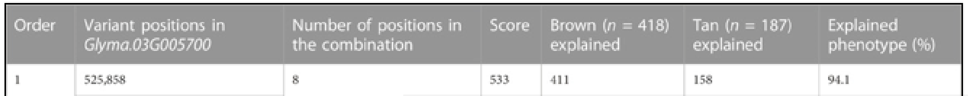
Skrabisova et al. 2024 Frontiers in Genetics

Allele distribution in Glyma.19G120400 in Soy1066 and its correlation with different pod-colored phenotypes.

- The RI_indel-only allele has higher chance in leading to black pod colors.
- The RI_indel-only allele is crucial for the production and accumulation of pigments responsible for the black color of pods.

Tool is designed to identify multiple alleles and more importantly causative mutations using Soy Curated Panels for phenotypes of interest

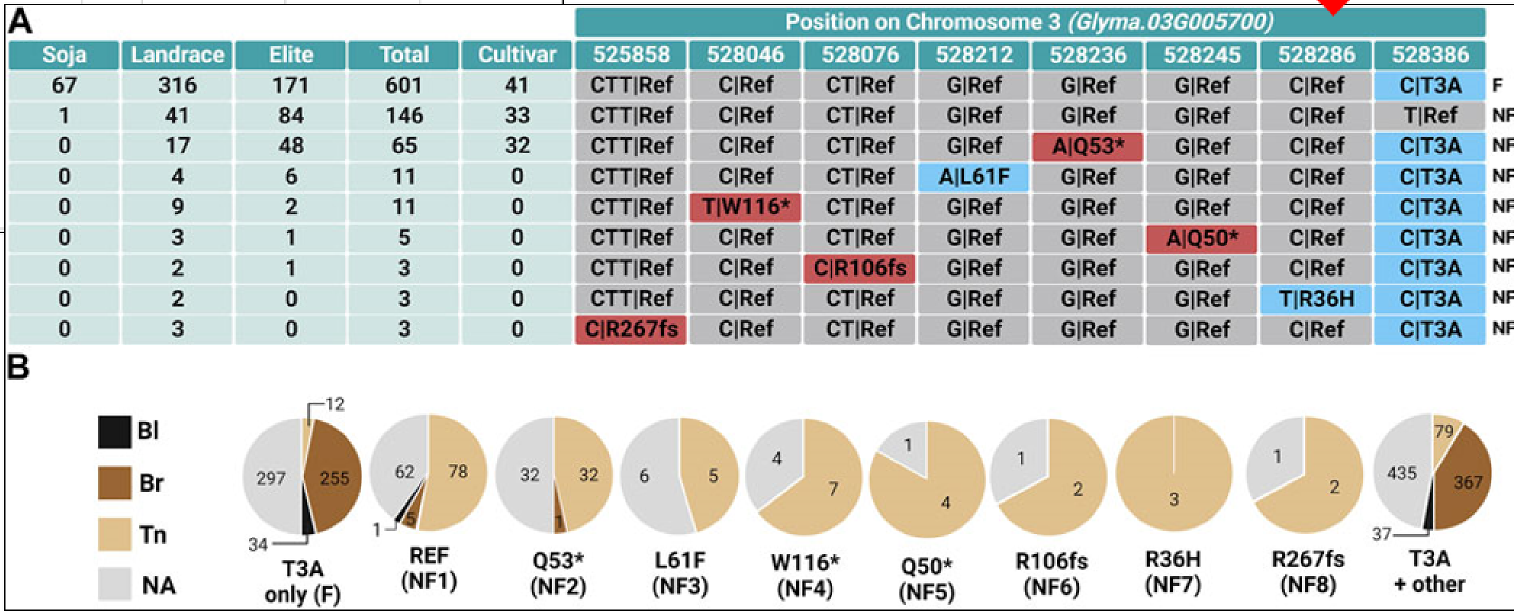
Multiple Allele Discovery (MADis) Tool Use Case to Understand Pod Colors



Order	Variant positions in *Glyma.03G005700*	Number of positions in the combination	Score	Brown (*n* = 418) explained	Tan (*n* = 187) explained	Explained phenotype (%)
1	525,858	8	533	411	158	94.1
	528,046					
	528,076					
	528,212					
	528,236					
	528,245					
	528,286					
	528,386					

MADis Tool Results

Allele Catalog Tool Results



A

					Position on Chromosome 3 *(Glyma.03G005700)*								
Soja	Landrace	Elite	Total	Cultivar	525858	528046	528076	528212	528236	528245	528286	528386	
67	316	171	601	41	CTT|Ref	C|Ref	CT|Ref	G|Ref	G|Ref	G|Ref	C|Ref	C|T3A	F
1	41	84	146	33	CTT|Ref	C|Ref	CT|Ref	G|Ref	G|Ref	G|Ref	C|Ref	T|Ref	NF
0	17	48	65	32	CTT|Ref	C|Ref	CT|Ref	G|Ref	A|Q53*	G|Ref	C|Ref	C|T3A	NF
0	4	6	11	0	CTT|Ref	C|Ref	CT|Ref	A|L61F	G|Ref	G|Ref	C|Ref	C|T3A	NF
0	9	2	11	0	CTT|Ref	T|W116*	CT|Ref	G|Ref	G|Ref	G|Ref	C|Ref	C|T3A	NF
0	3	1	5	0	CTT|Ref	C|Ref	CT|Ref	G|Ref	G|Ref	A|Q50*	C|Ref	C|T3A	NF
0	2	1	3	0	CTT|Ref	C|Ref	C|R106fs	G|Ref	G|Ref	G|Ref	C|Ref	C|T3A	NF
0	2	0	3	0	CTT|Ref	C|Ref	CT|Ref	G|Ref	G|Ref	G|Ref	T|R36H	C|T3A	NF
0	3	0	3	0	C|R267fs	C|Ref	CT|Ref	G|Ref	G|Ref	G|Ref	C|Ref	C|T3A	NF

Allele distribution in Glyma.03G005700 in Soy1066 and its correlation with different pod-colored phenotypes.

Skrabisova et al. 2024 Frontiers in Genetics

- The T3A missense mutation has higher chance leading to loss of pigment in soybean pods.

Tool computes a score for a combination of variant positions in a single candidate gene and based on the highest score, identifies the best number and combination of causative mutations.

Tools Expanded to Other Plants & Crops

- KBCommons and 3D Omics and CCMT -> All organisms
- SNPViz -> Arabidopsis, Maize, Mouse
- Allele Catalog -> Maize, Arabidopsis, Rice, Poplar
- GenVarX -> Arabidopsis, Rice
- Phenotype Distribution Tool -> Arabidopsis, Poplar
- Protein Sequence Logo Tool -> Arabidopsis, Sorghum
- MADis -> Maize

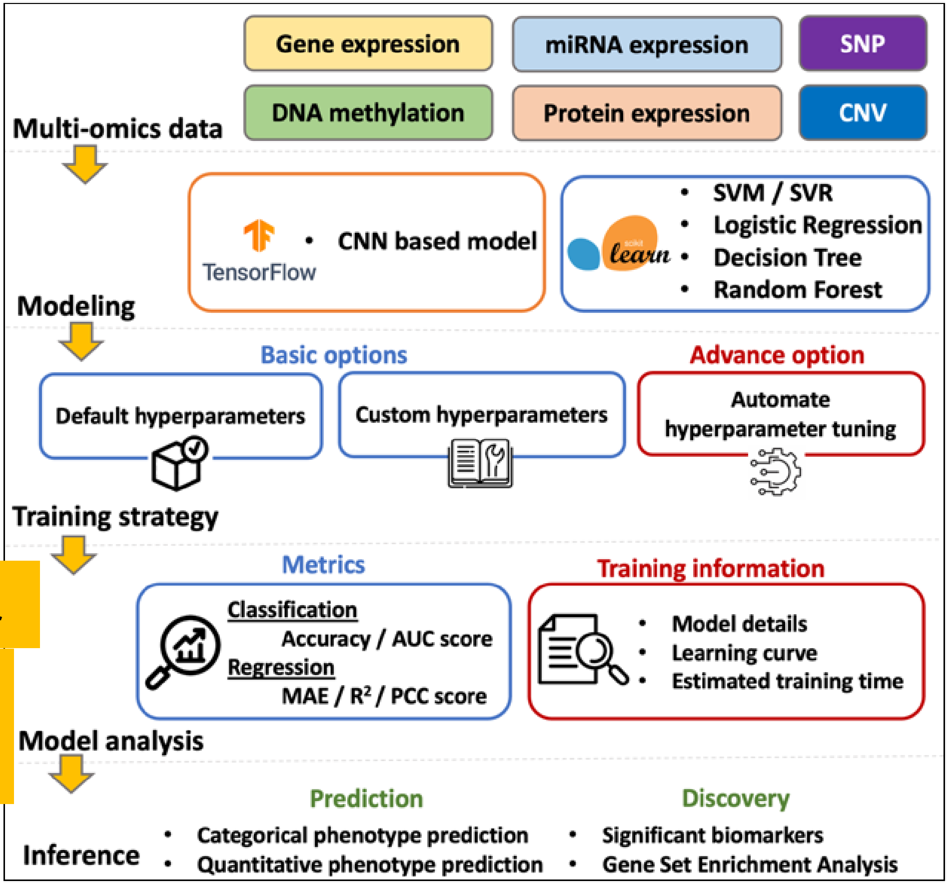
Translational Bioinformatics Resources and AI Solutions

- Multiomics Informatics Method Development
  - AI Solutions
    - G2PDeep – Phenotype Prediction using Deep Learning
    - scPlantAnnotate - scRNAseq annotation using LLMs

Genotype to Phenotype Deep Learning Framework (G2PDeep) for Genotype to Phenotype Prediction

G2PDeep

https://g2pdeep.org/



Multi-omics data
- Gene expression
- miRNA expression
- SNP
- DNA methylation
- Protein expression
- CNV

Modeling
- TensorFlow
  - CNN based model
- learn
  - SVM / SVR
  - Logistic Regression
  - Decision Tree
  - Random Forest

Training strategy

Basic options
- Default hyperparameters
- Custom hyperparameters

Advance option
- Automate hyperparameter tuning

Model analysis

Metrics
- Classification
  - Accuracy / AUC score
- Regression
  - MAE / R^2 / PCC score

Training information
- Model details
- Learning curve
- Estimated training time

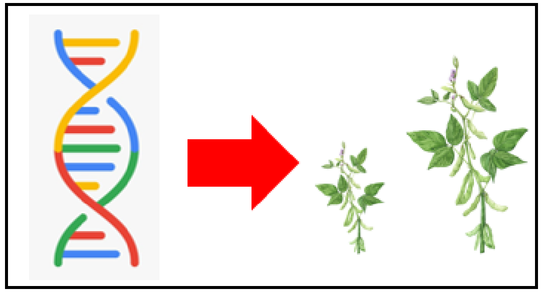
Inference

Prediction
- Categorical phenotype prediction
- Quantitative phenotype prediction

Discovery
- Significant biomarkers
- Gene Set Enrichment Analysis

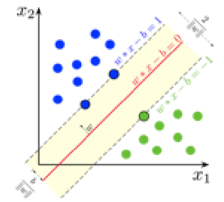
Zeng et al. 2021 NAR Web Server

Zeng et al. 2024 bioRxiv; 2025 Bioinformatics, Under Review

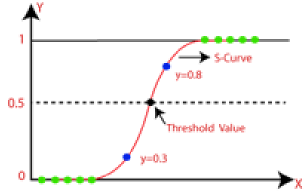


- Enable AI through G2PDeep framework for training multi-omics data to create different variant of models for crop phenotype predictions and discover causative mutations.

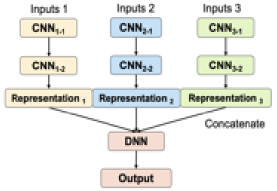
Logistic Regression



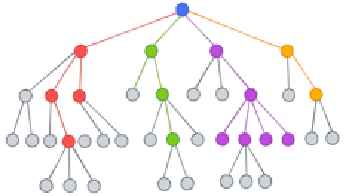
SVM/SVR Model



CNN Model



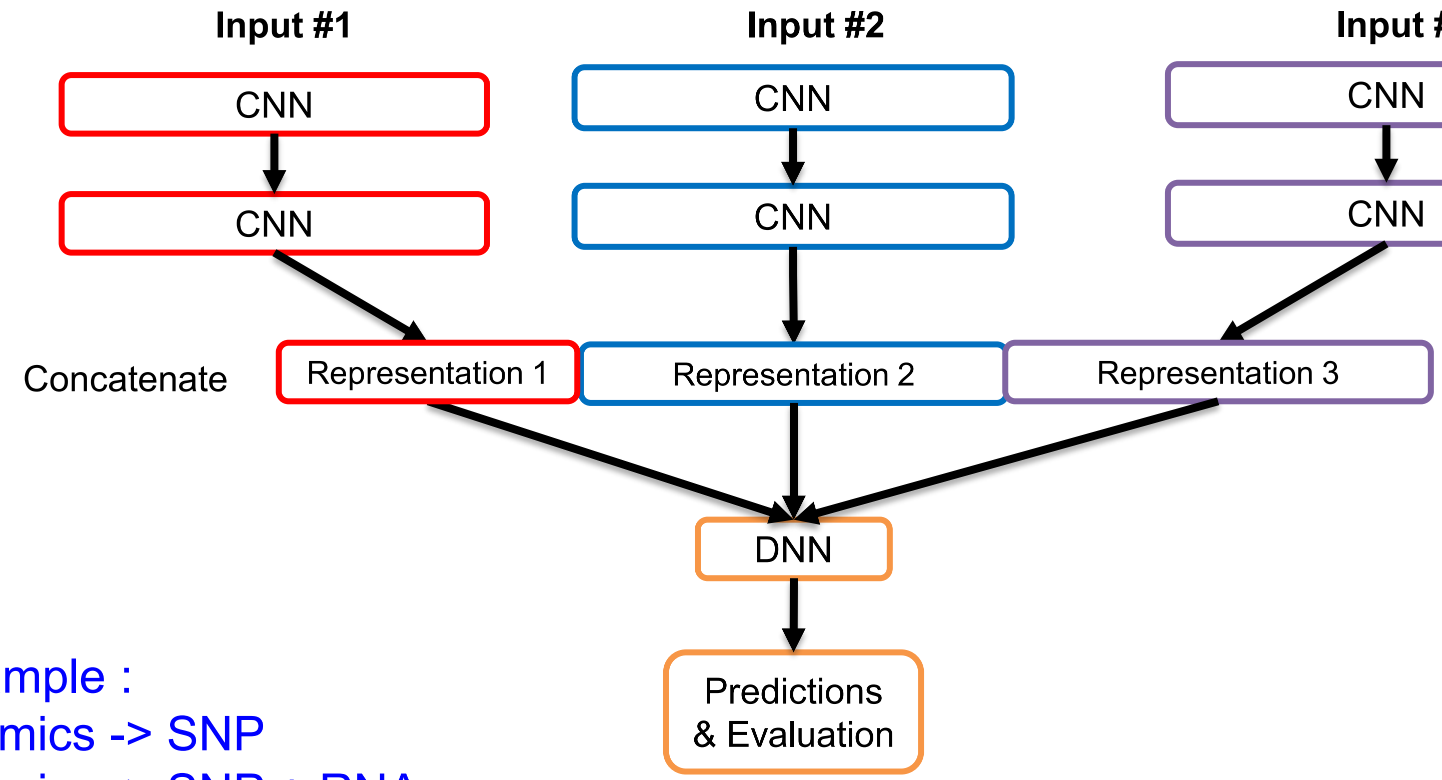
Decision Tree & Random Forest



G2PDeep is a comprehensive web-based platform to do deep learning based phenotype prediction and markers discovery using any (1-3) combinations of 6 diverse multiomics data types

G2PDeep v2 – Omics Combinations

Multi CNN-based model



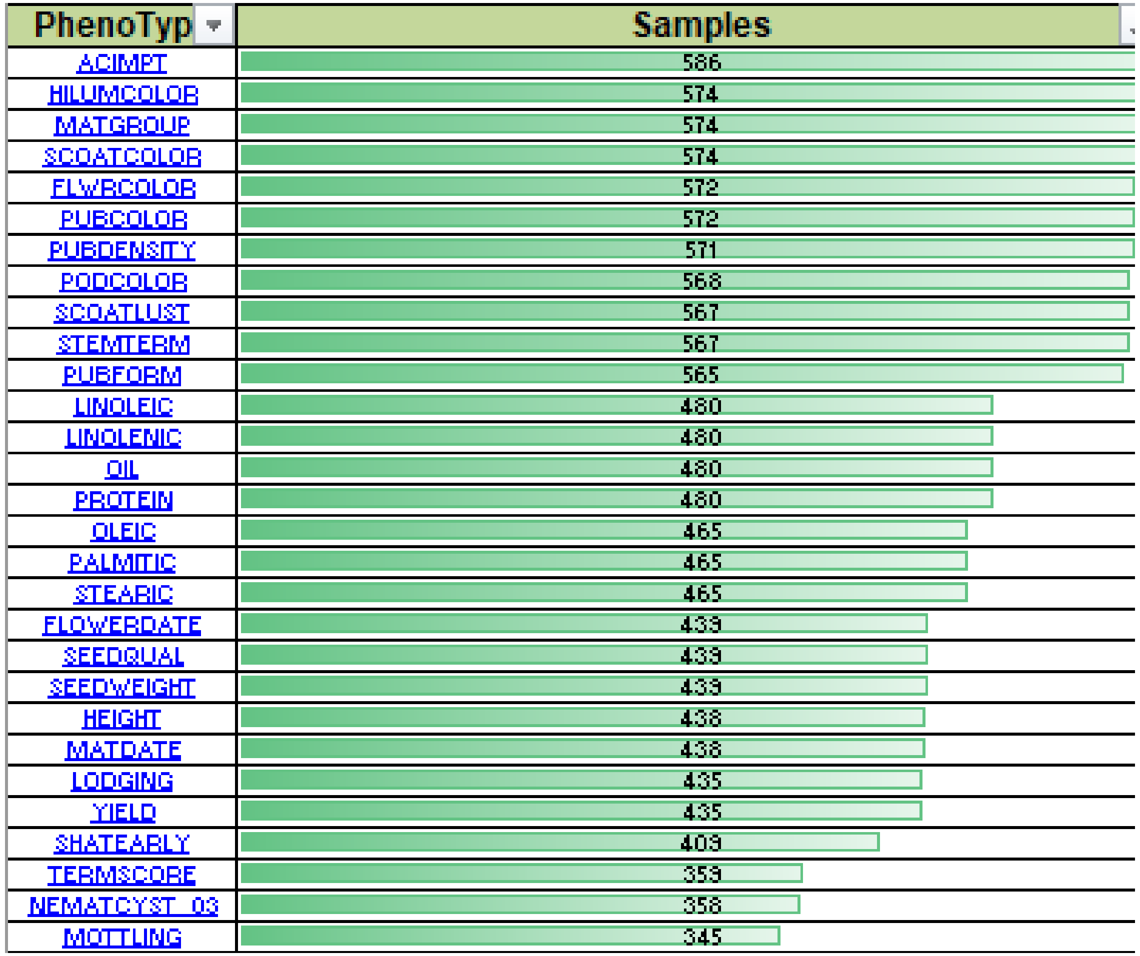
Zeng et al. 2021
NAR Web Server

Zeng et al. 2024
bioRxiv; 2025
Bioinformatics,
Under Review

Example :
1 Omics -> SNP
2 Omics -> SNP + RNAseq
3 Omics -> SNP + RNAseq + CNV

G2PDeep Application to Soybean

- Soybean Soy1066 curated panel (32M SNPs) and GRIN Phenotypic data overlap
- Bulk transcriptomics data available from Genome Sequence Archive (GSA) database in BIG Data Center (CRA009979) for 600+ accessions
- Together can be utilized for phenotypic predictions for diverse traits

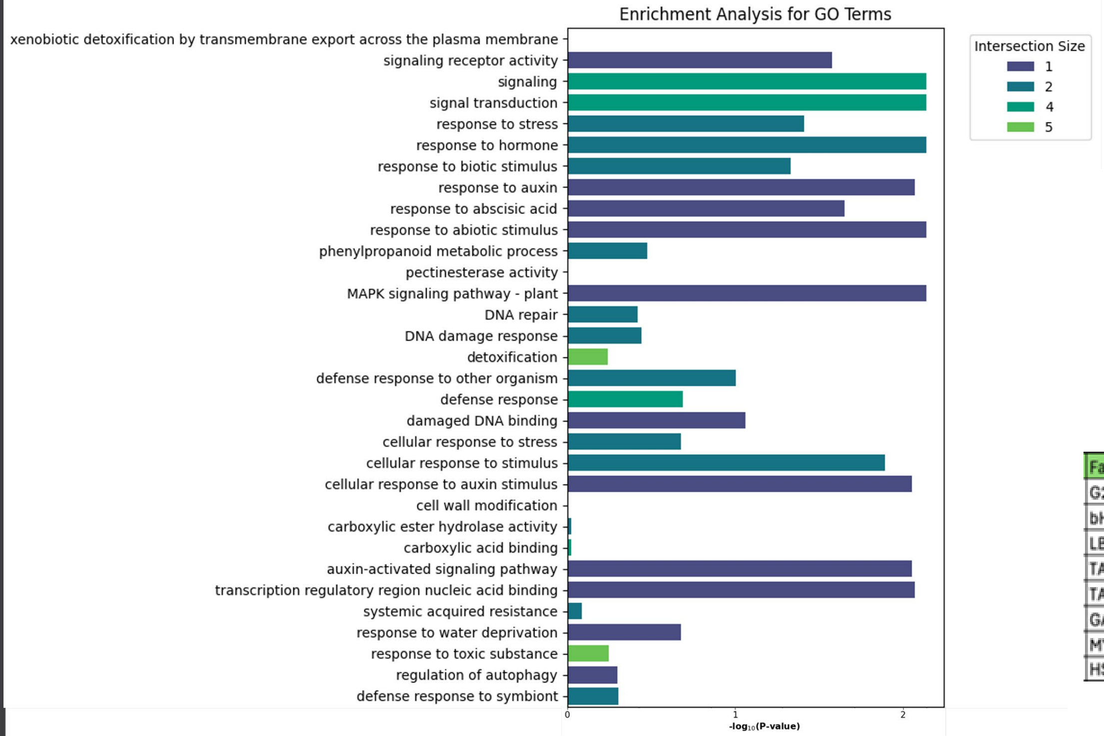


PhenoTyp	Samples
ACIMPT	586
HILUMCOLOR	574
MATGROUP	574
SCOATCOLOR	574
FLWRCOLOR	572
PUBCOLOR	572
PUBDENSITY	571
PODCOLOR	568
SCOATLUST	567
STEMTERM	567
PUBFORM	565
LINOLEIC	480
LINOLENIC	480
OIL	480
PROTEIN	480
OLEIC	465
PALMITIC	465
STEARIC	465
FLOWERDATE	439
SEEDQUAL	439
SEEDWEIGHT	439
HEIGHT	438
MATDATE	438
LODGING	435
YIELD	435
SHATEARLY	409
TERMSCORE	359
NEMATCYST_03	358
MOTTLING	345

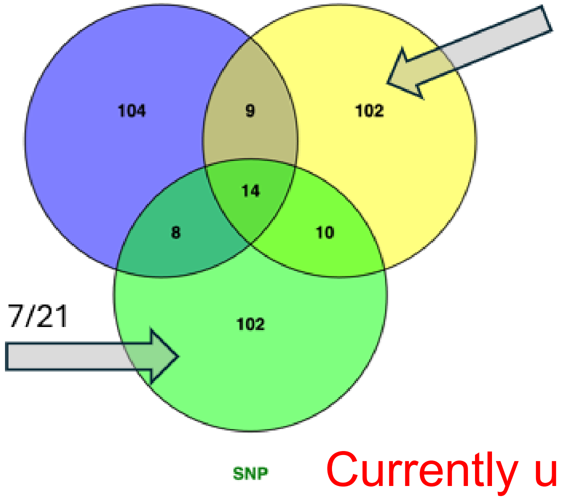
Soybean Cyst Nematode (SCN) Resistance Prediction in Soybean

Omic	Accessions	Features
SNPs	268	4162197
CNV	268	1092

Omics	F1 (Score)	Accuracy
SNP	0.7127	0.89
CNV	0.6845	0.86
SNP | CNV	0.7693	0.79

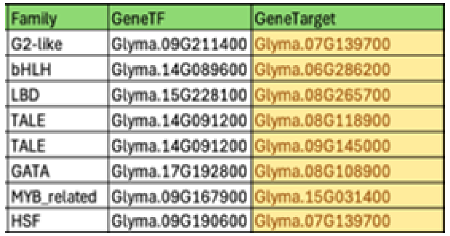


$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2\frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{2\mathrm{tp}}{2\mathrm{tp} + \mathrm{fp} + \mathrm{fn}}.$$



New candidate gene

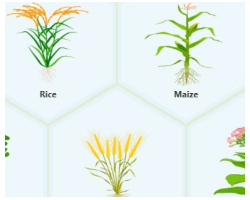
7/21



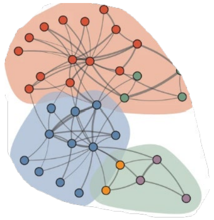
Family	GeneTF	GeneTarget
G2-like	Glyma.09G211400	Glyma.07G139700
bHLH	Glyma.14G089600	Glyma.06G286200
LBD	Glyma.15G228100	Glyma.08G265700
TALE	Glyma.14G091200	Glyma.08G118900
TALE	Glyma.14G091200	Glyma.09G145000
GATA	Glyma.17G192800	Glyma.08G108900
MYB_related	Glyma.09G167900	Glyma.15G031400
HSF	Glyma.09G190600	Glyma.07G139700

Currently under evaluation

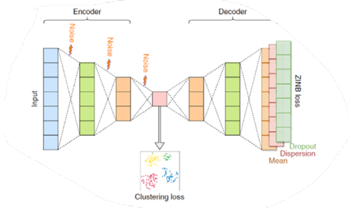
scPLantAnnotate - scRNAseq annotations using LLMs



Marker-gene based methods



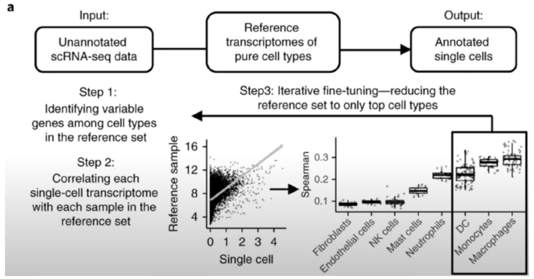
Scanpy



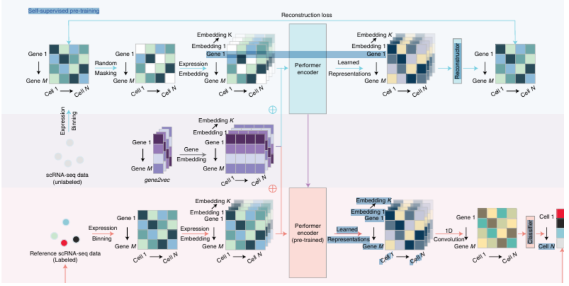
scDeepClust

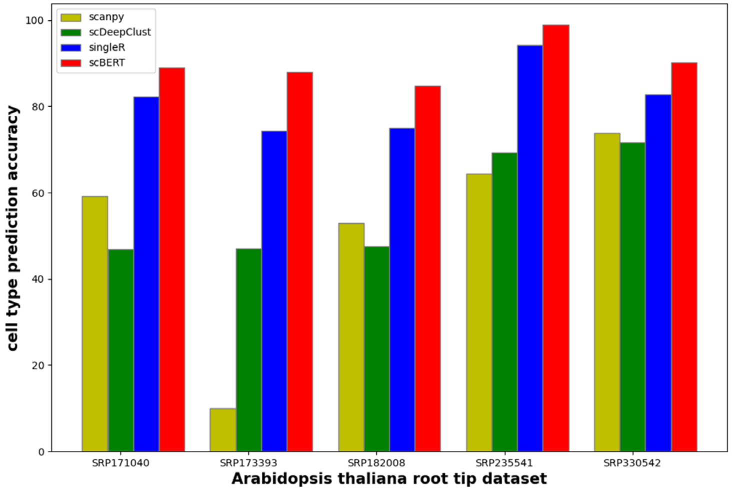
Reference based methods

singleR

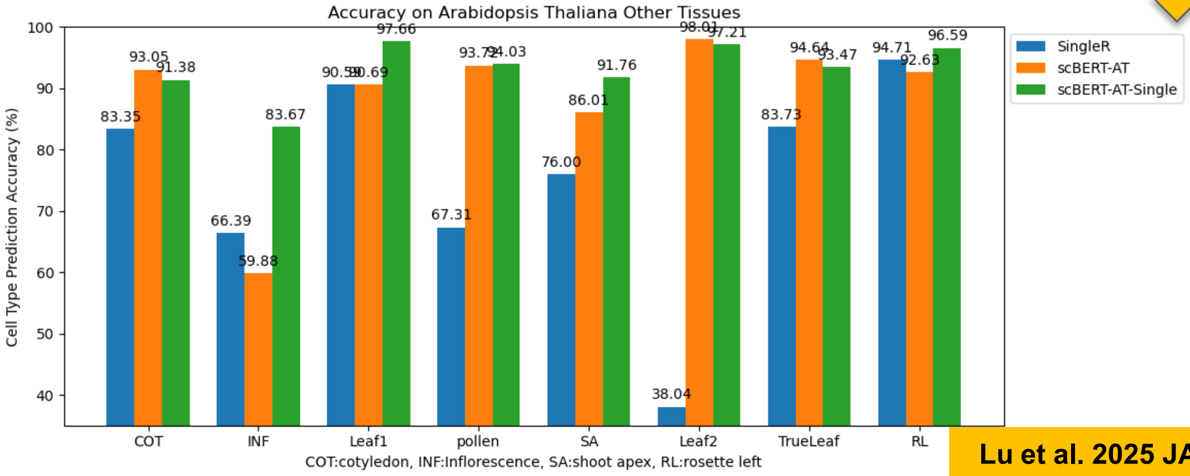


scBERT



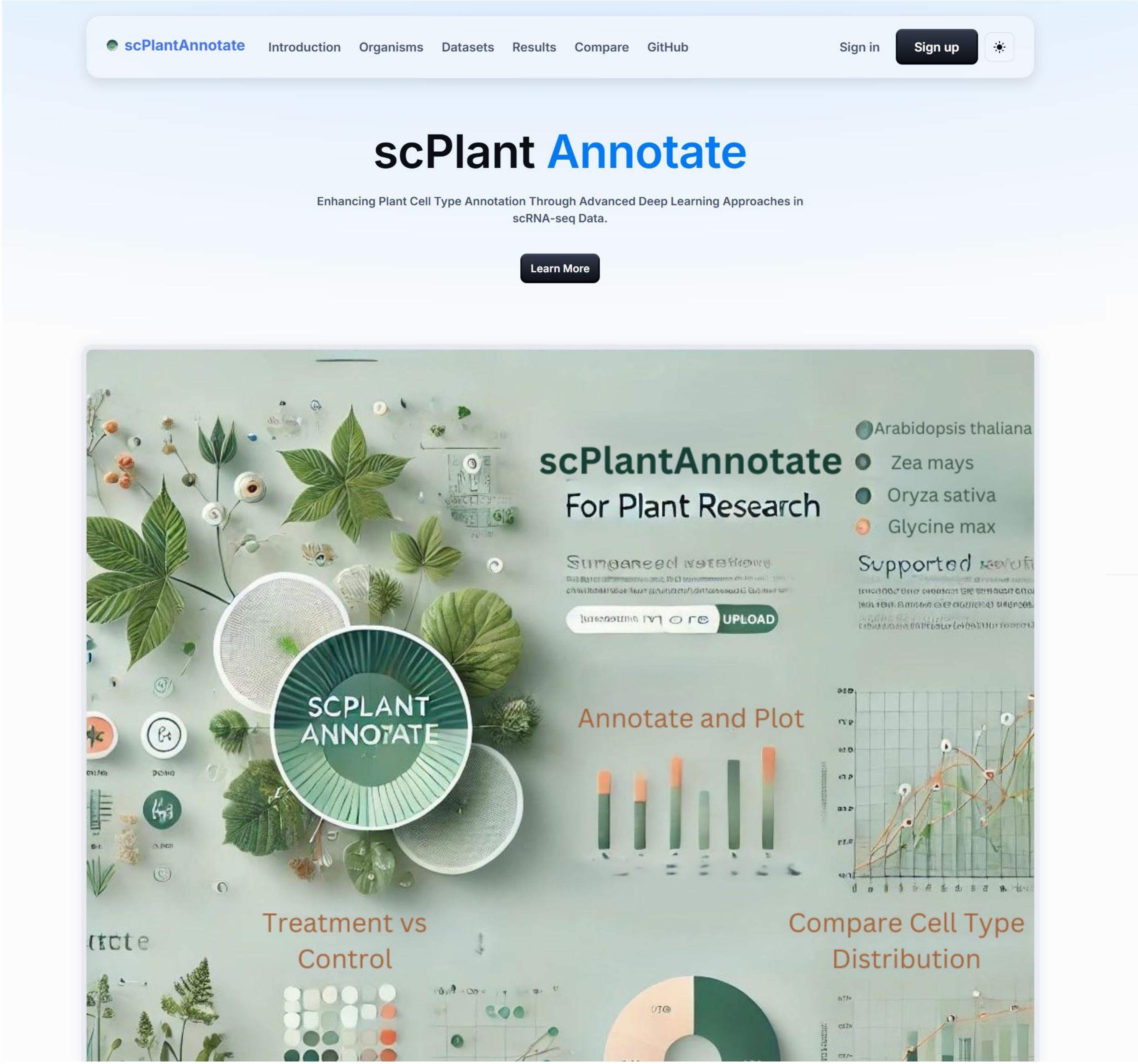


- Use large language models (LLMs) for predictions.
- scBERT-AT-Single is ONE single model for Arabidopsis.
- Trained on 28 datasets across all tissues, 1.2M cells. 50 cell types.



- Currently predictions being done using scRNAseq datasets from maize, rice, soybean available in scPlantDB and from our collaborators.

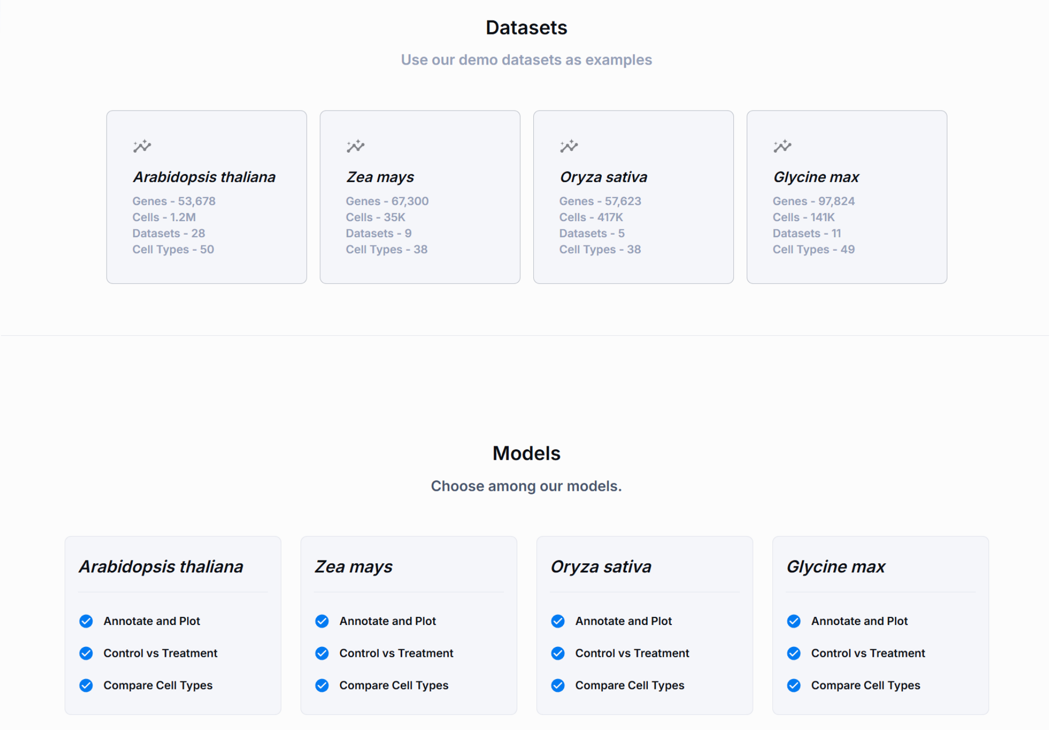
Lu et al. 2025 JAR Under Review



scPlant Annotate

Enhancing Plant Cell Type Annotation Through Advanced Deep Learning Approaches in scRNA-seq Data.

https://scplantannotate.missouri.edu/



Datasets

Use our demo datasets as examples

Arabidopsis thaliana	*Zea mays*	*Oryza sativa*	*Glycine max*
Genes - 53,678	Genes - 67,300	Genes - 57,623	Genes - 97,824
Cells - 1.2M	Cells - 35K	Cells - 417K	Cells - 141K
Datasets - 28	Datasets - 9	Datasets - 5	Datasets - 11
Cell Types - 50	Cell Types - 38	Cell Types - 38	Cell Types - 49

Models

Choose among our models.

Arabidopsis thaliana	*Zea mays*	*Oryza sativa*	*Glycine max*
✓ Annotate and Plot	✓ Annotate and Plot	✓ Annotate and Plot	✓ Annotate and Plot
✓ Control vs Treatment	✓ Control vs Treatment	✓ Control vs Treatment	✓ Control vs Treatment
✓ Compare Cell Types	✓ Compare Cell Types	✓ Compare Cell Types	✓ Compare Cell Types

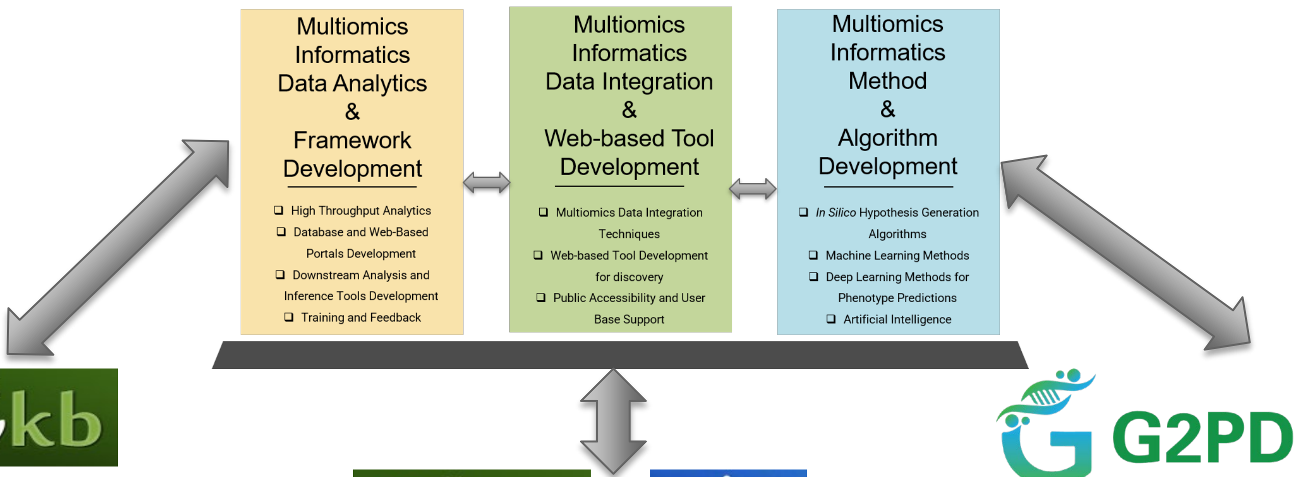
Lu et al. 2025 JAR
Under Review

Summary

- Translational Bioinformatics resources and AI solutions provide great opportunities to incorporate genomics and multiomics information into advancing research for a diverse set of applications including crop breeding.
- As the technology and computational fields advance, it is important to adopt and incorporate these advanced techniques and build applications to facilitate easy translation.
- Making the advanced AI solutions more accessible to the users, will facilitate easier adoption and biggest impacts!

Joshi Lab

Enabling Translational Bioinformatics Research



Multiomics Informatics Data Analytics & Framework Development

- ☐ High Throughput Analytics
- ☐ Database and Web-Based Portals Development
- ☐ Downstream Analysis and Inference Tools Development
- ☐ Training and Feedback

Multiomics Informatics Data Integration & Web-based Tool Development

- ☐ Multiomics Data Integration Techniques
- ☐ Web-based Tool Development for discovery
- ☐ Public Accessibility and User Base Support

Multiomics Informatics Method & Algorithm Development

- ☐ *In Silico* Hypothesis Generation Algorithms
- ☐ Machine Learning Methods
- ☐ Deep Learning Methods for Phenotype Predictions
- ☐ Artificial Intelligence

Soybean Knowledge Base

https://soykb.org/

KBCommons Knowledge Base

https://kbcommons.org/

Multiomics Data Integration

Allele Catalog Tool

GenVarX Tool

SNPViz Tool

AccuTool

MADis Tool

G2PDeep

https://g2pdeep.org/

IMPRes

https://impres.missouri.edu/impres

scPlantAnnotate

https://scplantannotate.missouri.edu/

Acknowledgements

SoyKB - KBCommons Team

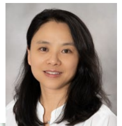
- **Yen On Chan (MUIDSI)**
- **Manish Immadi (CS)**
- **Sabin Dahal (CS)**

Bioinformatics Analytics Team

- Mohit Verma (Postdoc)
- Sameep Dhakal (MUIDSI)
- Preethi Induri (Senior Research Specialist II)

Method Development Team

- Mohammad Vahed (Scientist)
- Zhen Lyu (CS)
- **Chunyang Lu (CS)**



Recent Alumni

- **Shuai Zeng**

Key Collaborators

- Dr. Dong Xu (EECS)

USB Soybean Project Collaborators

- **Dr. Kristin Bilyeu**
- **Dr. Maria Skrabisova**
- **Jana Biova**
- **Anser Mahmood (Plant Sci)**

Multiomics Plant Project Collaborators

- Dr. Ron Mittler
- Dr. Ed Cahoon
- Dr. Marc Libault

Acknowledgements

MISSOURI DEPARTMENT OF HEALTH & SENIOR SERVICES

NSF

UNITED STATES OF AMERICA DEPARTMENT OF ENERGY

NIH

UNITED SOYBEAN BOARD™

Weblinks and Contact Details

- **Soybean Knowledge Base**
  https://soykb.org/
- **KBCommons Knowledge Base**
  https://kbcommons.org/
- **G2PDeep Web Server**
  https://g2pdeep.org/
- **scPlantAnnotate**
  https://scplantannotate.missouri.edu/

Contact : Joshitr@marshall.edu
Joshitr@missouri.edu