



APPLIED GENOMICS FOR ACCELERATED BREEDING OF SOYBEAN AND OTHER SPECIES

Mária Škrabišová, Ph.D.

Palacký University in Olomouc CZ | Faculty of Science | Department of Biochemistry

Plant and Animal Genome Conference PAG31, San Diego, CA, USA
Soybean Genomics, January 16, 2024

SOYBEAN APPLIED GENOMICS

Development of strategies to accelerate soybean breeding and improvement

- Genetic natural variation exploration
- Identification of causal genes
- Tools for applied genomics



Dr. Kristin Bilyeu
USDA-ARS
Plant Research Unit



Dr. Trupti Joshi
Dept. of Health Management
and Informatics



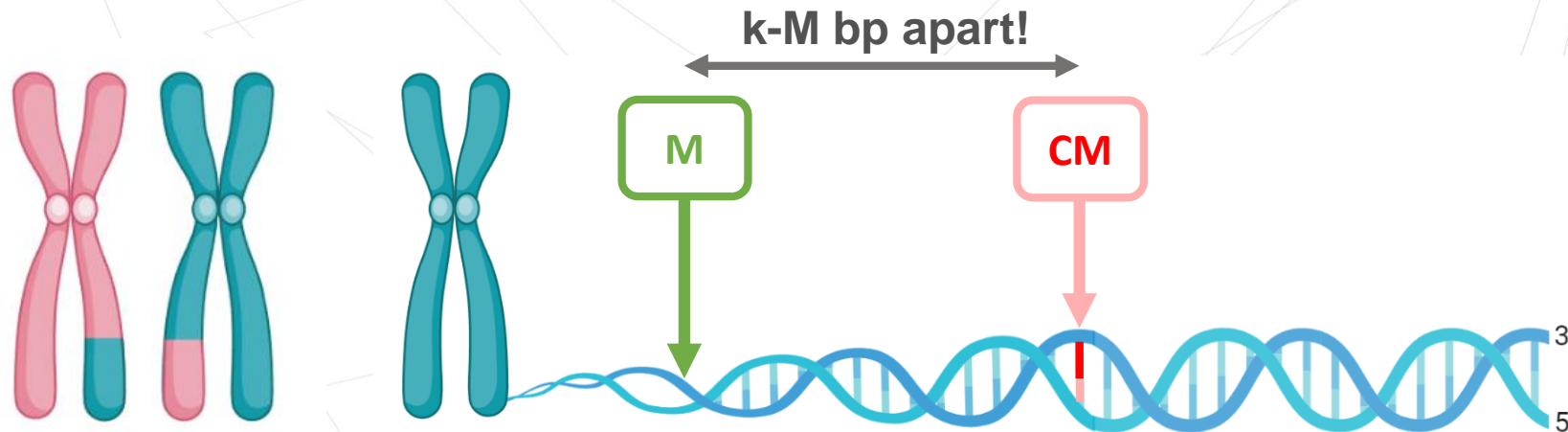
BREEDING IS NOT PRECISE ENOUGH YET!

Marker-assisted breeding

- A marker (**M**) can be located far from a gene that underlies a phenotype of interest

How to proceed? Identification of causal genes with **causative mutations (CMs)**

- Era of sequencing
- Association methods



Created with BioRender.com

GWAS FOR MORE PRECISE BREEDING

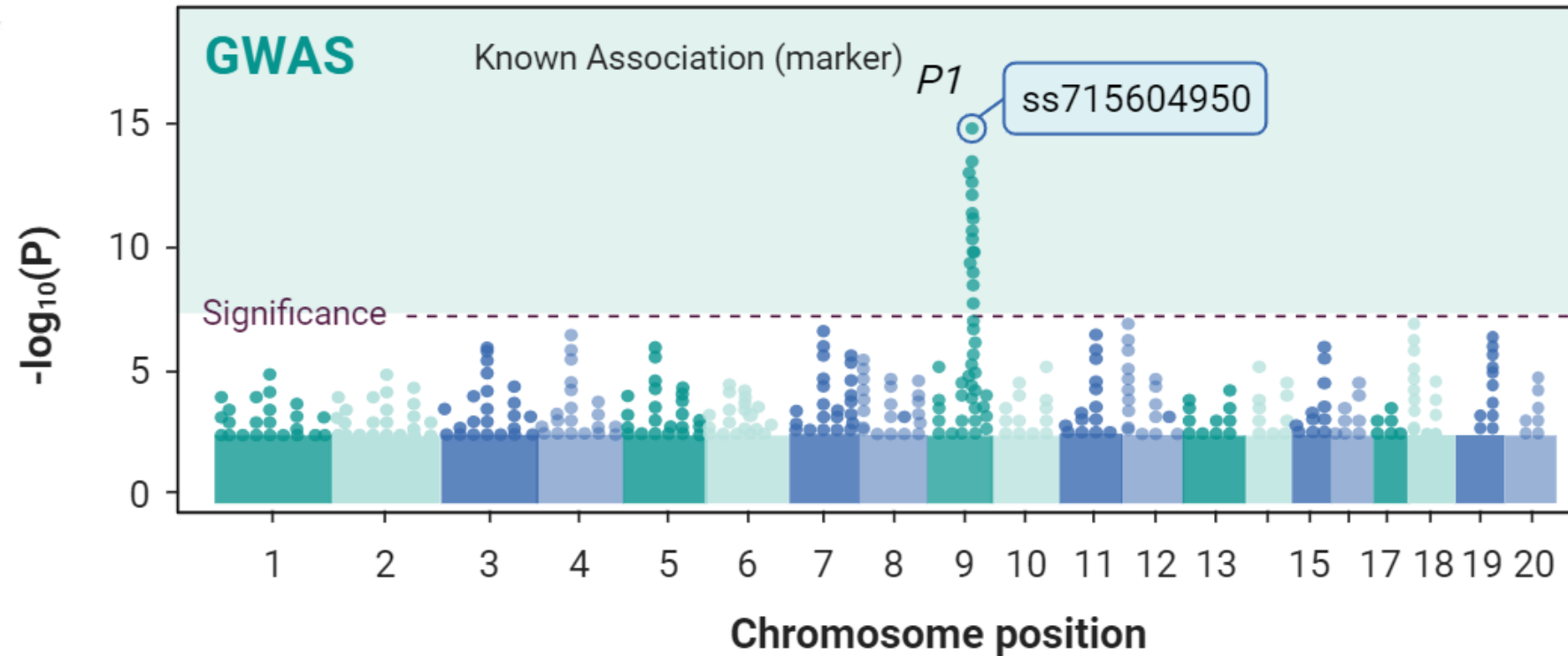
GWAS (Genome-wide association study)

- Associates a phenotype with a genomic locus > associated variant positions, tagging marker (TM)
- Assists in **identifying CM** and thus **could accelerate soybean breeding and improvement**
- Statistical method **dependent on data set size and quality of input data**

Inputs: genotype + phenotype

- **Phenotype:** quantitative or qualitative, proportional, disproportional, rare
- **Genotypic data:**
 - Low-density genotype = **genotyping data (limited representation of natural variation)**
 - Whole genome sequence = **resequenced data (\$\$\$)**

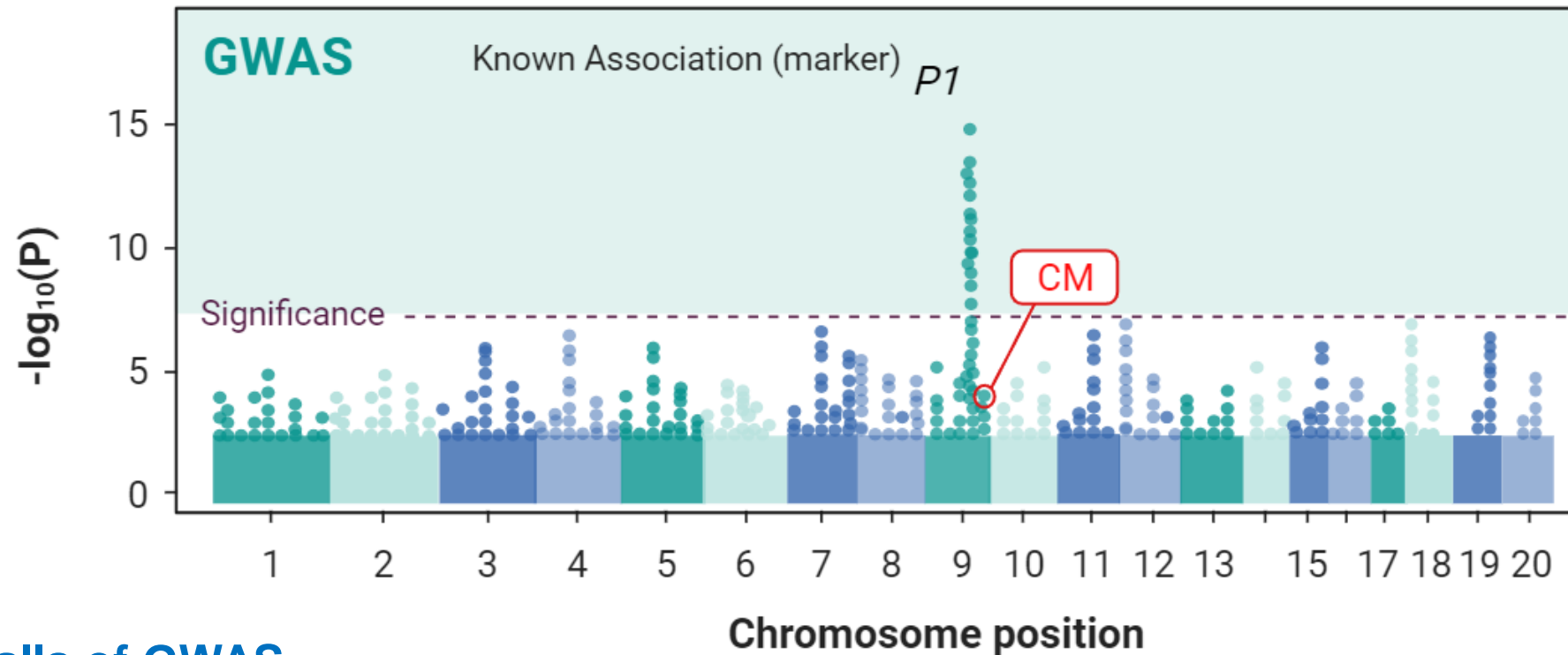
GWAS WITH GENOTYPING DATA



Tagging *P1* locus for loss of trichomes

- CM not present
- Standard GWAS follow-up: **guessing** genes based on protein annotation or TM vicinity
- Post-GWAS methodology (transcriptome, metabolome, ...)

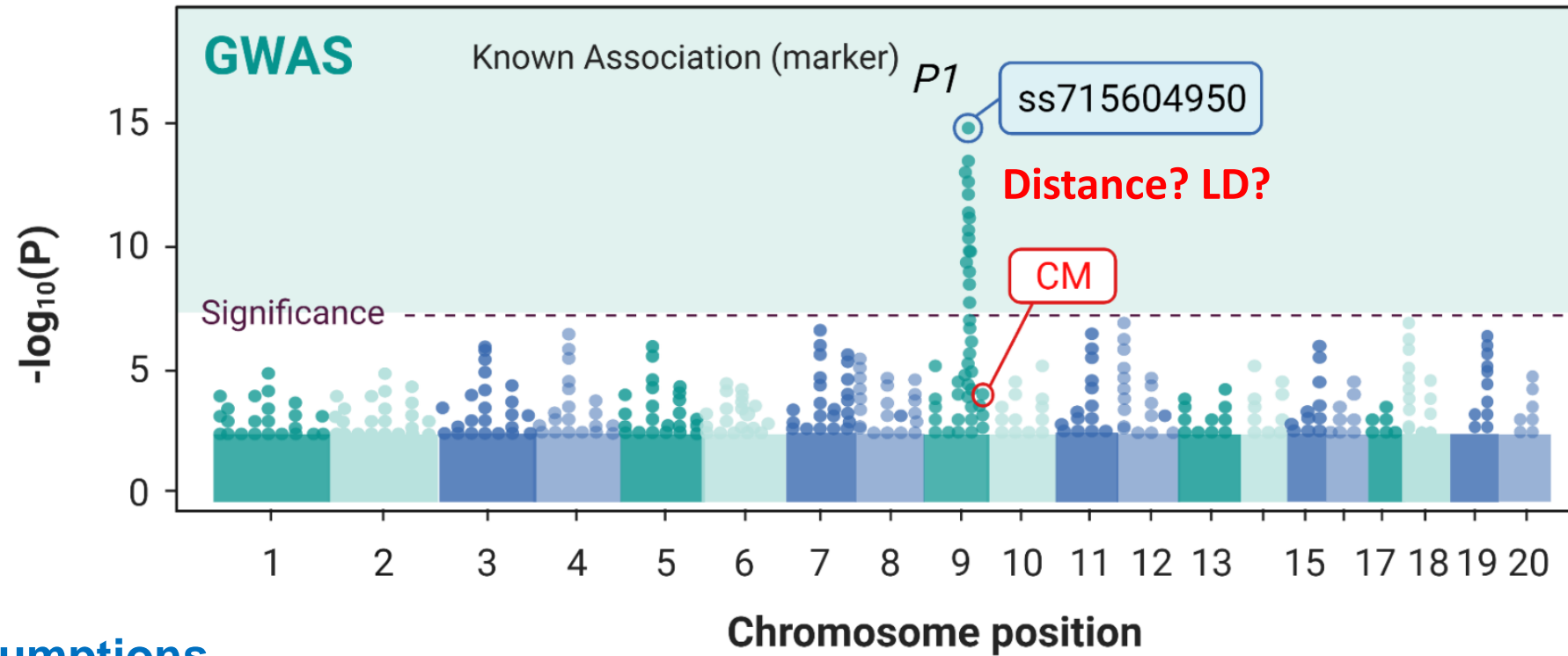
GWAS WITH RESEQUENCED DATA



Pitfalls of GWAS

- Complicated **genetic architecture** (large indels, etc.)
- Small data set size with limited phenotype information
- **Multidimensional collinearity** (chromosomal rearrangements, duplications, etc.)
- **Complexity** of traits (multiple CMs, multiple alleles, etc.)

GWAS FOR MORE PRECISE BREEDING

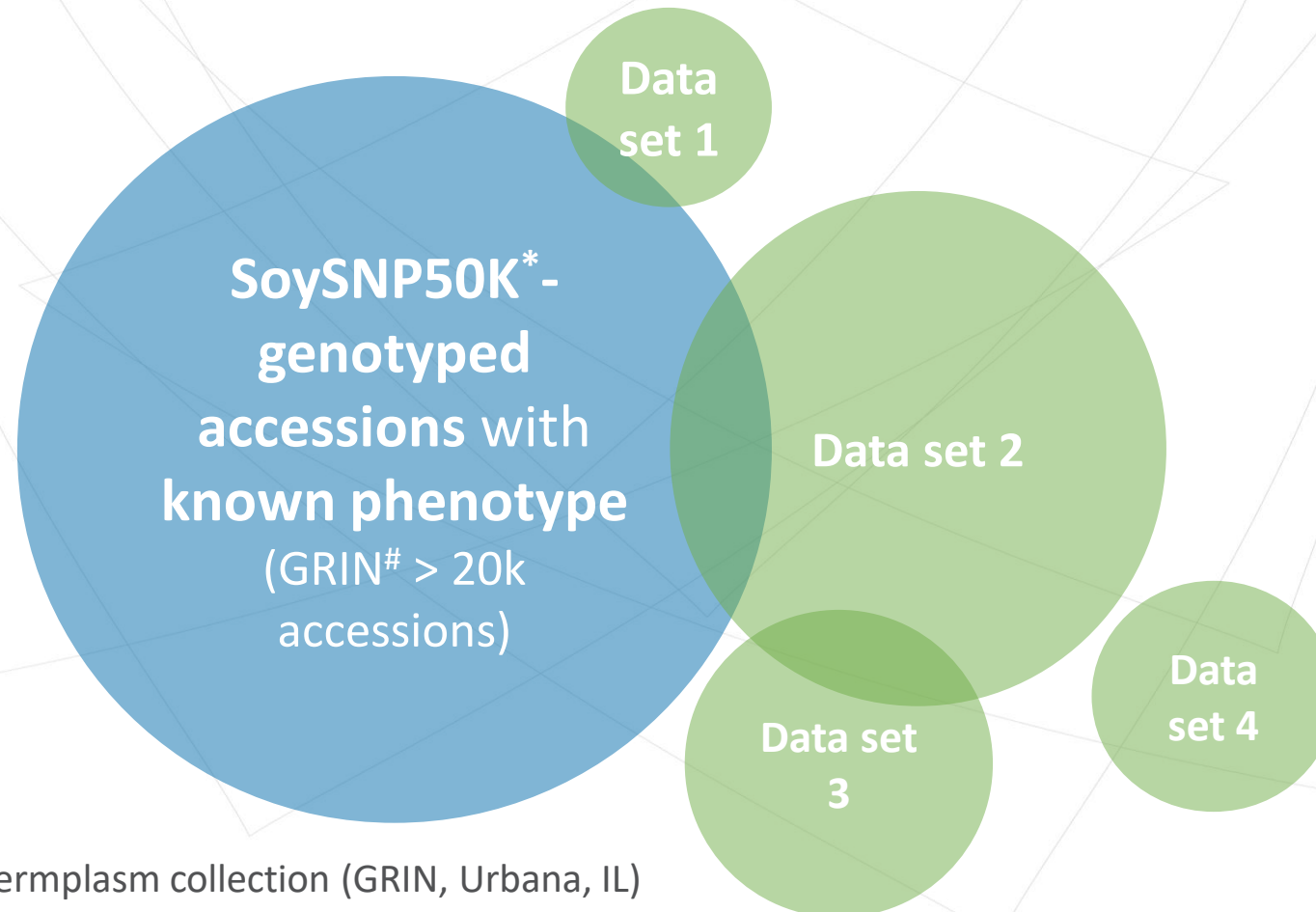


Assumptions

- There are **isolated data sets** that can be reused to boost GWAS power
- **Additional evaluation criterion** (LD independent) is required to improve post-GWAS
- There are not enough **user-friendly tools** to support the exploration of genetic diversity

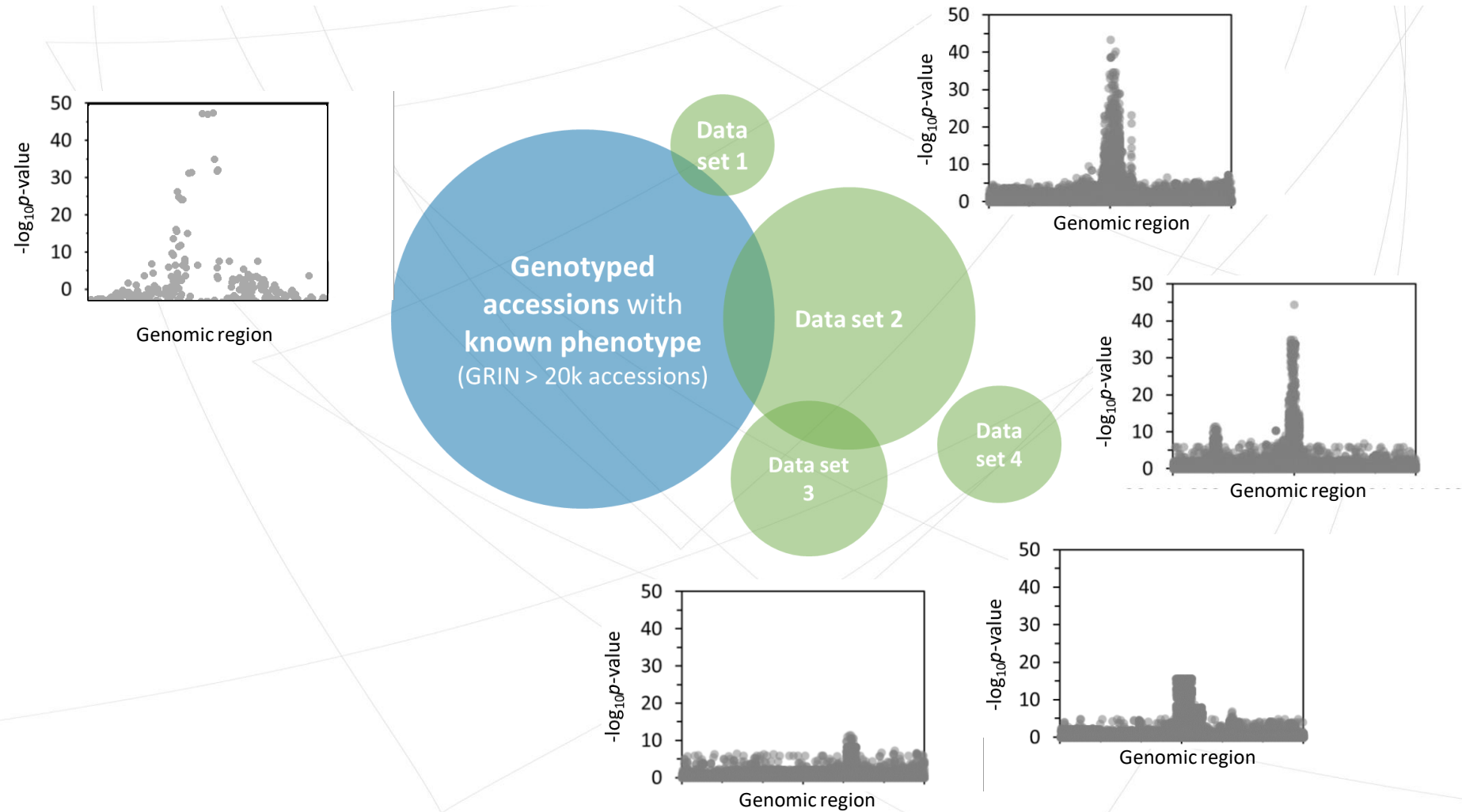
PUBLICLY AVAILABLE DATA FOR GWAS

- **Genotyped (SoySNP50K, etc.)** accessions with **known phenotype**, “n” is large
- **Resequenced** data sets with **limited phenotype** information (or unavailable), “n” is small



UNDERPOWERED GWAS FAILS IN CM PREDICTION

- GWAS is critically dependent on the data set size, the genotype quality and the phenotype frequency



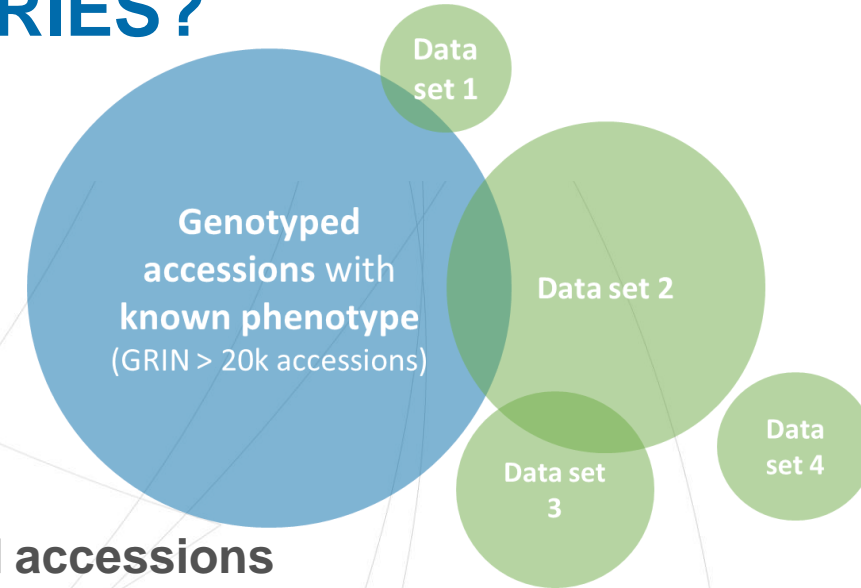
- GWAS does not work for very rare phenotypes

HOW TO IMPROVE GWAS-DRIVEN DISCOVERIES?

By adding power to GWAS!

What is needed? Three novel concepts:

- A junction between the missing information > **Synthetic phenotype**
- Additional GWAS evaluation criterion > **Accuracy**
- Concatenated data sets > **Curated panel of soybean resequenced accessions**



Journal of Advanced Research 42 (2022) 117–133



Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Journal of Advanced Research

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



Original Article

A novel Synthetic phenotype association study approach reveals the landscape of association for genomic variants and phenotypes



Mária Škrabišová^a, Nicholas Dietz^b, Shuai Zeng^{c,d}, Yen On Chan^{d,e}, Juexin Wang^{c,d}, Yang Liu^{d,e}, Jana Biová^a, Trupti Joshi^{c,d,e,f,*}, Kristin D. Bilyeu^{g,*}

^a Department of Biochemistry, Faculty of Science, Palacký University Olomouc, Olomouc 78371, Czech Republic

^b Division of Plant Sciences, University of Missouri, Columbia, MO 65201, USA

^c Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO 65212, USA

^d Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65212, USA

^e MU Data Science and Informatics Institute, University of Missouri, Columbia, MO 65212, USA

^f Department of Health Management and Informatics, School of Medicine, University of Missouri, Columbia, MO 65212, USA

^g Plant Genetics Research Unit, United States Department of Agriculture-Agricultural Research Service, University of Missouri, Columbia, MO 65211, USA

SYNTHETIC PHENOTYPE

- For GWAS, qualitative phenotypes are transformed into a numerical format; therefore, a genotype can be transformed the same way

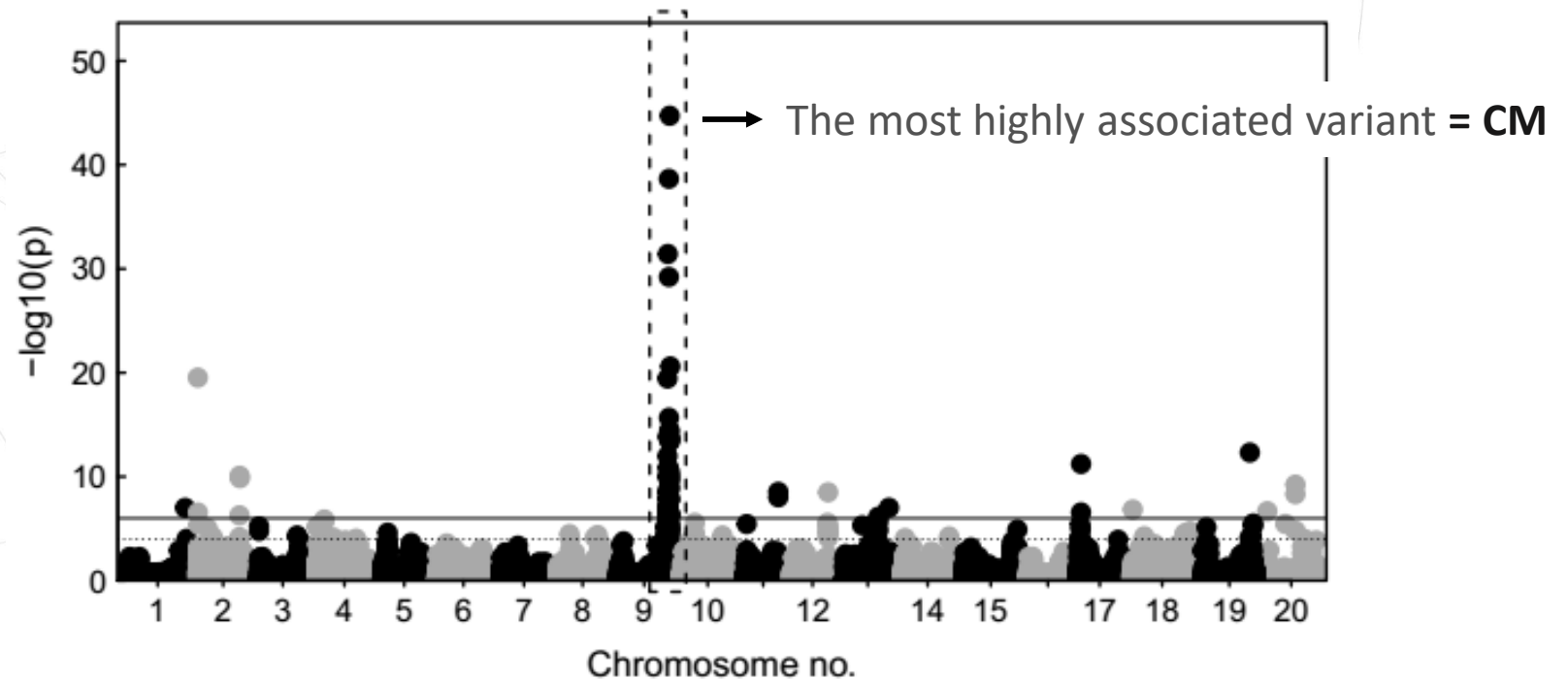
Real phenotype (Observed) Pubescence presence/absence		Synthetic phenotype CM in <i>Glyma.09g278000</i> A25T (Liu <i>et al.</i> 2020) Chr09:45,057,956 genotype	
WT	Normal	REF	T
MUT	Glabrous	ALT	A

- Every variant position can be used as a Synthetic phenotype (CM as well as TM)
- Since there is a single gene with a bi-allelic CM behind every Manhattan peak then every phenotype can be binarized (even for quantitative phenotypes)

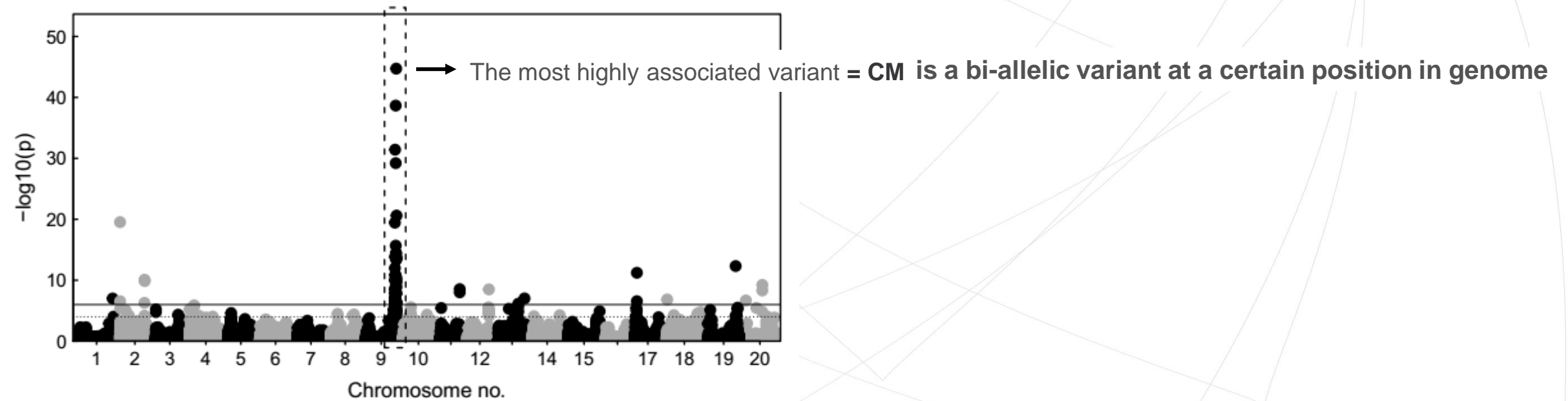
LOGIC OF SYNTHETIC PHENOTYPE: PERFECT GWAS

Prerequisites

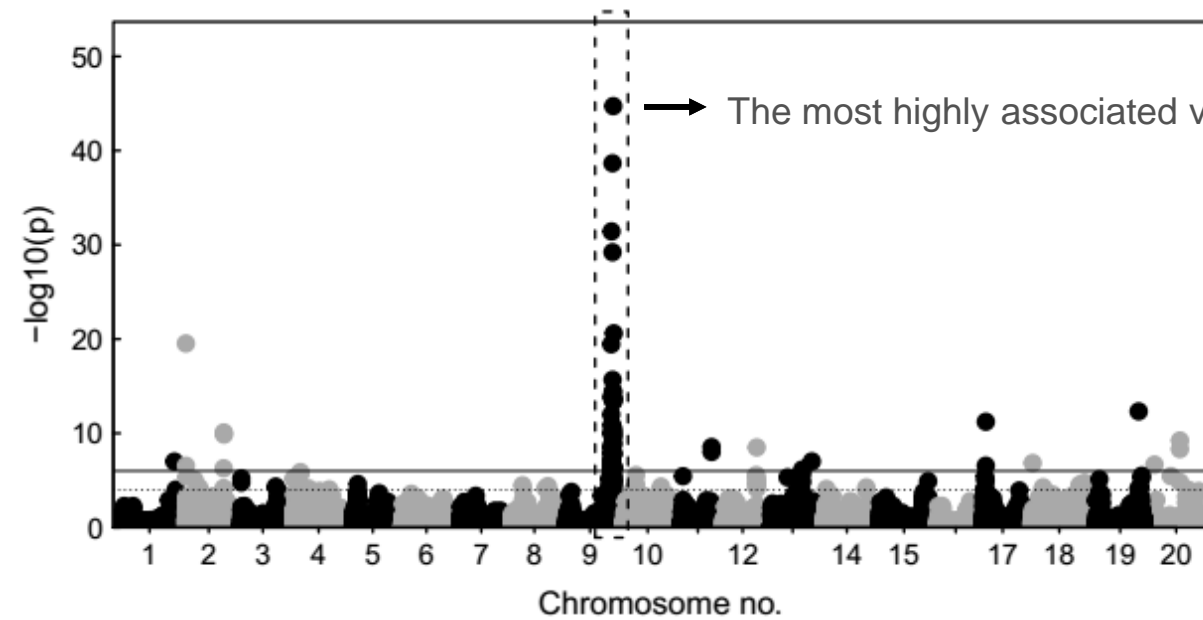
- Large “n”
- A high-quality **genotype**
- Good distribution of a **binary phenotype** (Pubescence presence/absence)
- A single **CM** in only one gene



PERFECT GWAS



PERFECT GWAS

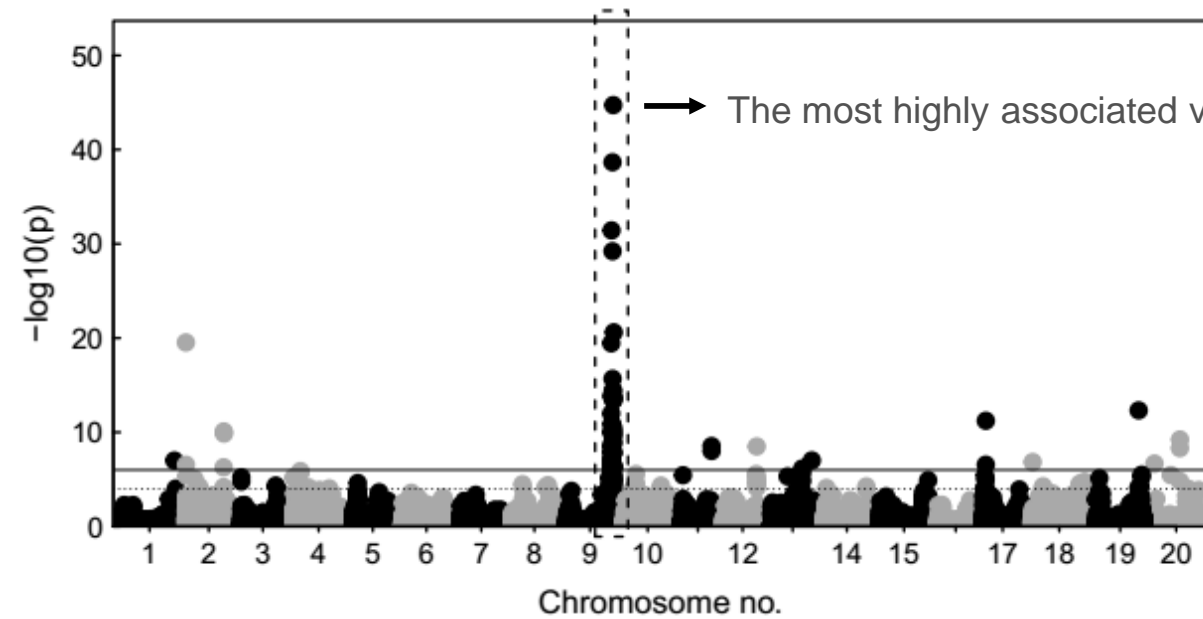


→ The most highly associated variant = **CM** is a bi-allelic variant at a certain position in genome

REF Allele

ALT Allele

PERFECT GWAS

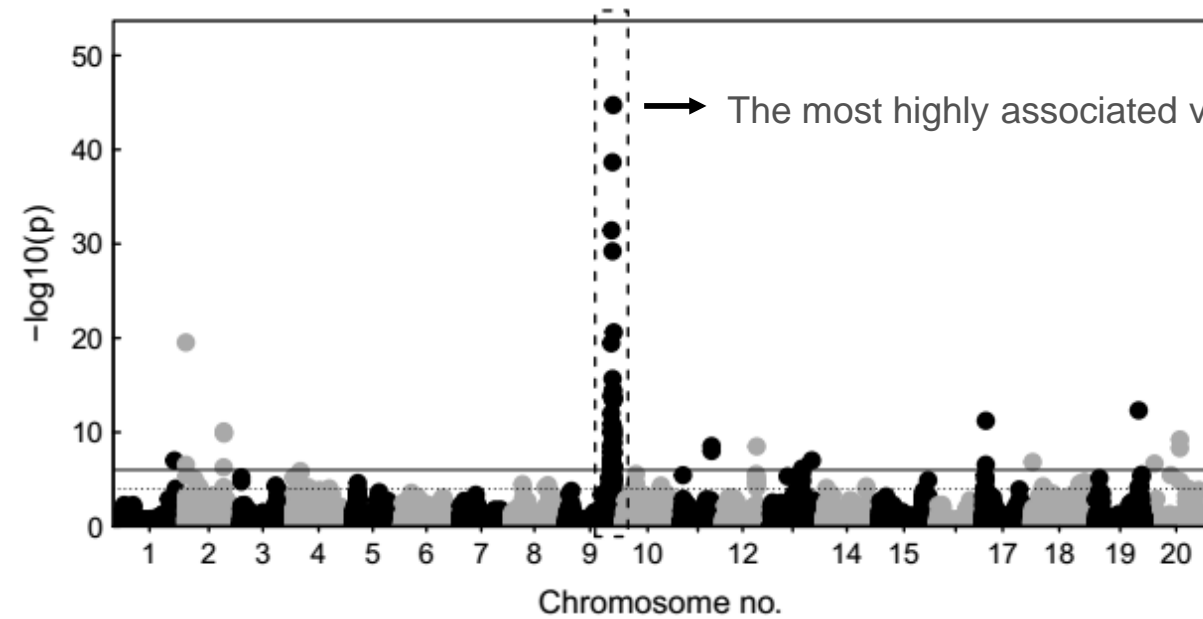


→ The most highly associated variant = **CM** is a bi-allelic variant at a certain position in genome

REF Allele = WT Phenotype

ALT Allele = MUT Phenotype

PERFECT GWAS



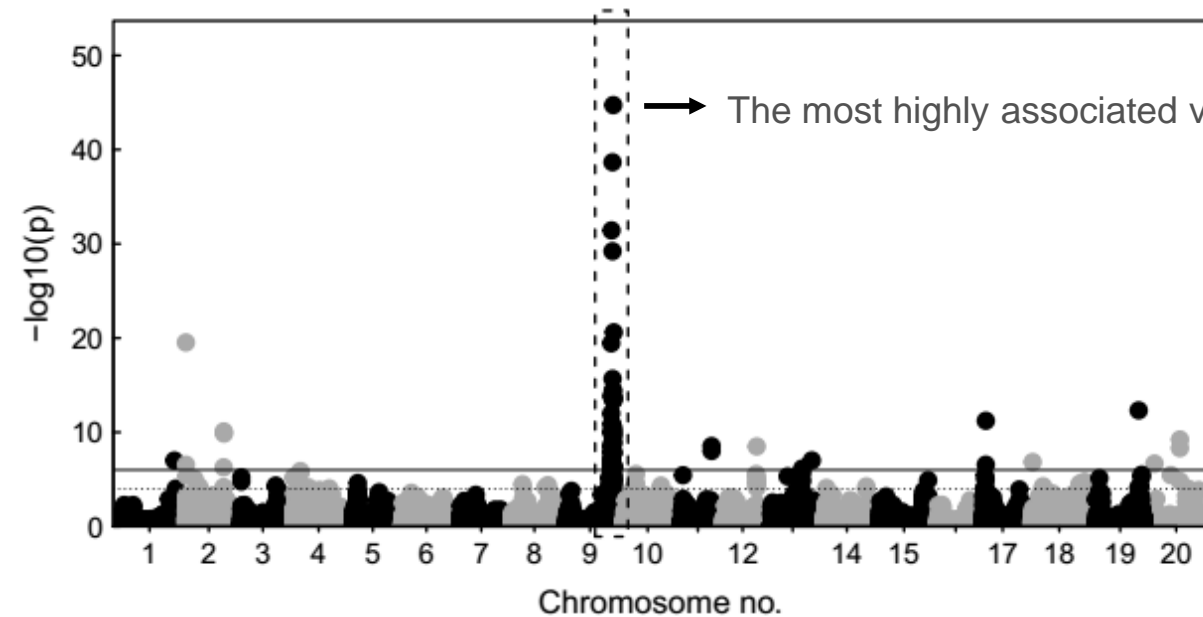
→ The most highly associated variant = **CM** is a bi-allelic variant at a certain position in genome

REF Allele = WT Phenotype

ALT Allele = MUT Phenotype

Perfect match between
phenotype and genotype

PERFECT GWAS



→ The most highly associated variant = **CM** is a bi-allelic variant at a certain position in genome

REF Allele = WT Phenotype

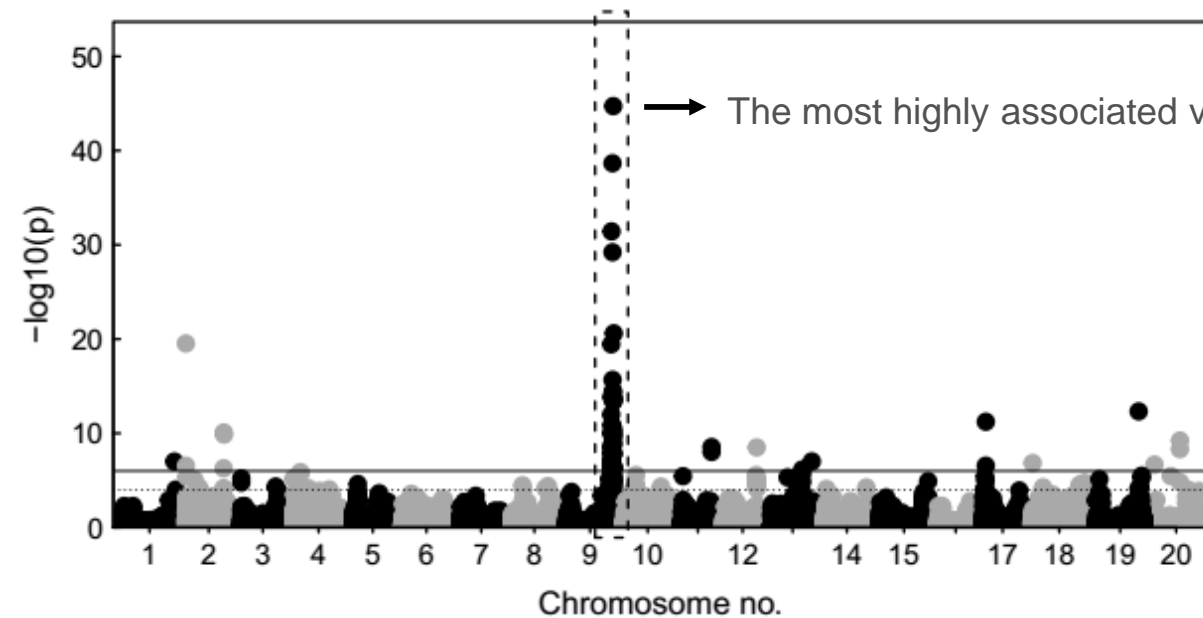
ALT Allele = MUT Phenotype

**Perfect match between
phenotype and genotype**

100% Accessions
with **WT phenotype**
have **REF Allele**

100% Accessions
with **MUT phenotype**
have **ALT Allele**

PERFECT GWAS



→ The most highly associated variant = **CM** is a bi-allelic variant at a certain position in genome

REF Allele = WT Phenotype

ALT Allele = MUT Phenotype

**Perfect match between
phenotype and genotype**

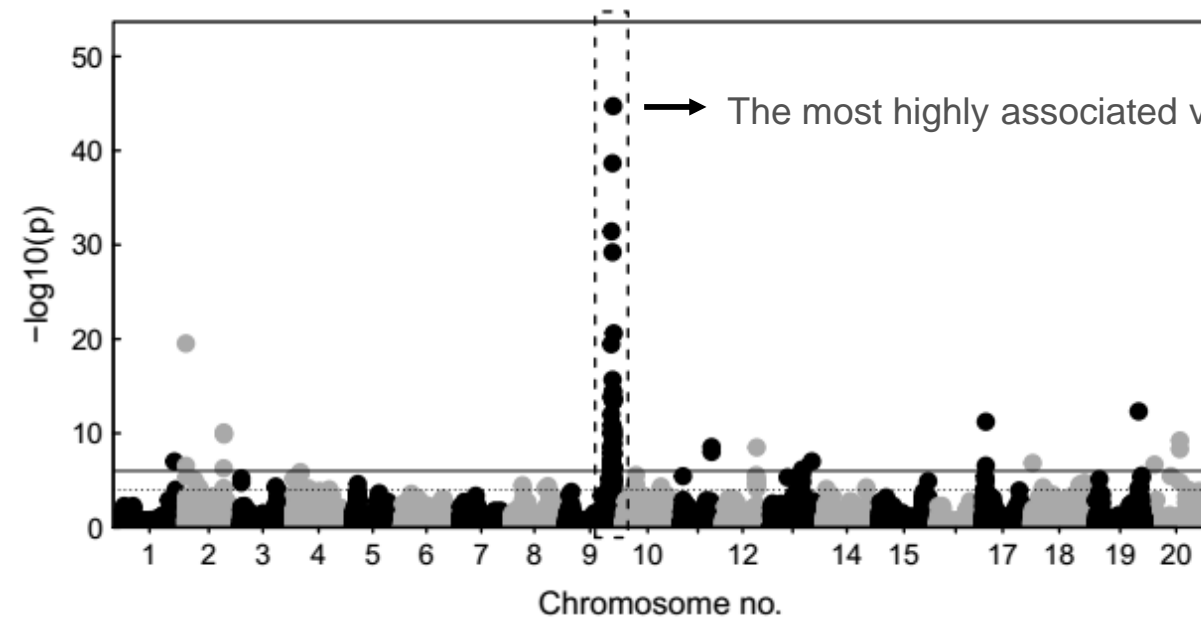
100% Accessions
with **WT** phenotype
have **REF Allele**

100% Accessions
with **MUT** phenotype
have **ALT Allele**

Phenotype and genotype categories are interchangeable

PERFECT GWAS

What if the genotype does not correlate perfectly with the phenotype?



The most highly associated variant = **CM** is a bi-allelic variant at a certain position in genome

REF Allele = WT Phenotype

ALT Allele = MUT Phenotype

Perfect match between
phenotype and genotype

100% Accessions
with **WT phenotype**
have **REF Allele**

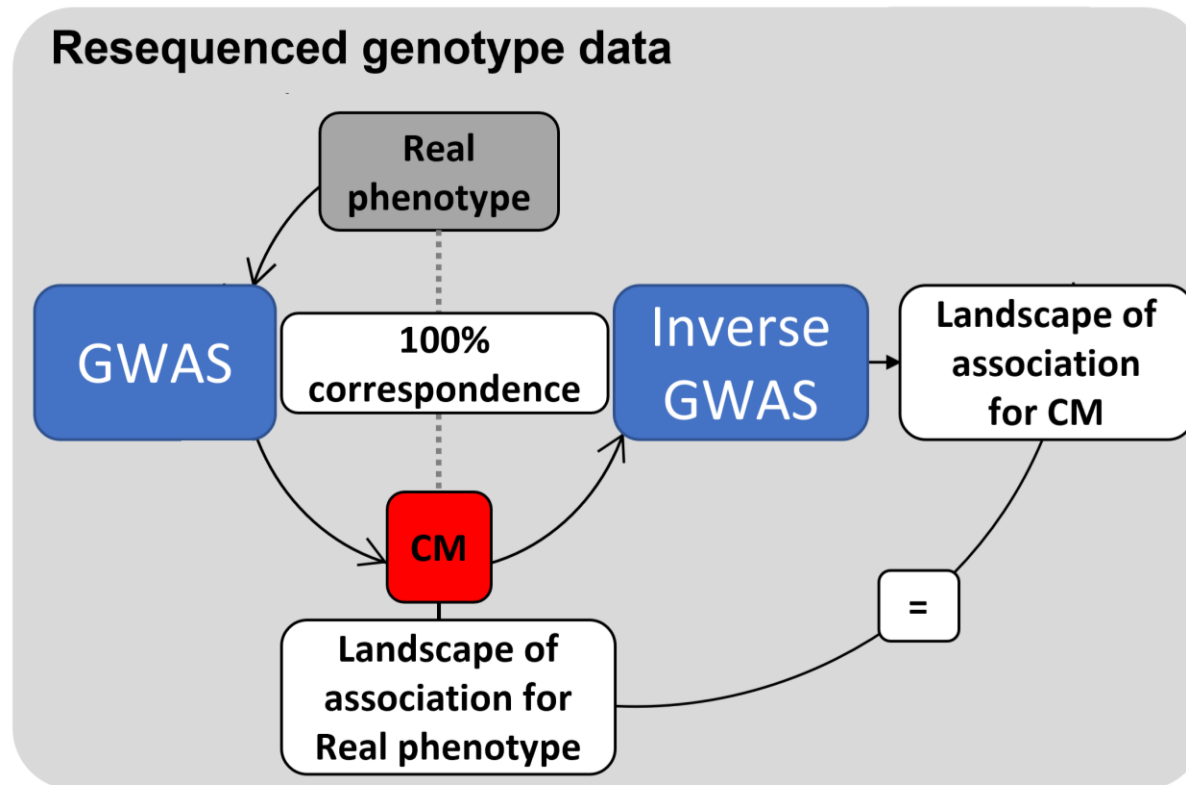
100% Accessions
with **MUT phenotype**
have **ALT Allele**

Phenotype and genotype categories are interchangeable

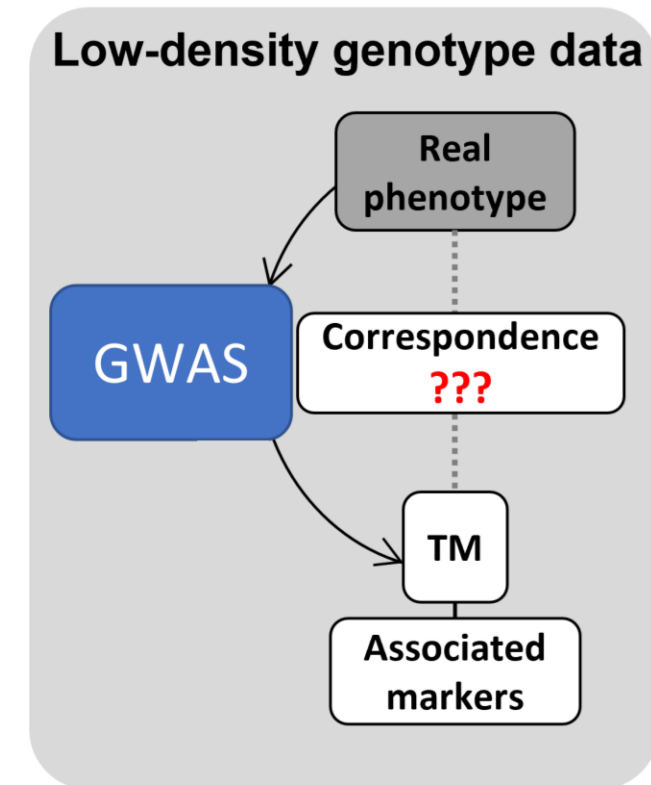
LOGIC OF DIRECT CORRESPONDENCE AS A MEASURE OF HOW WELL A VARIANT POSITION CORRELATES WITH A PHENOTYPE

- Between markers, CMs and phenotypes

a



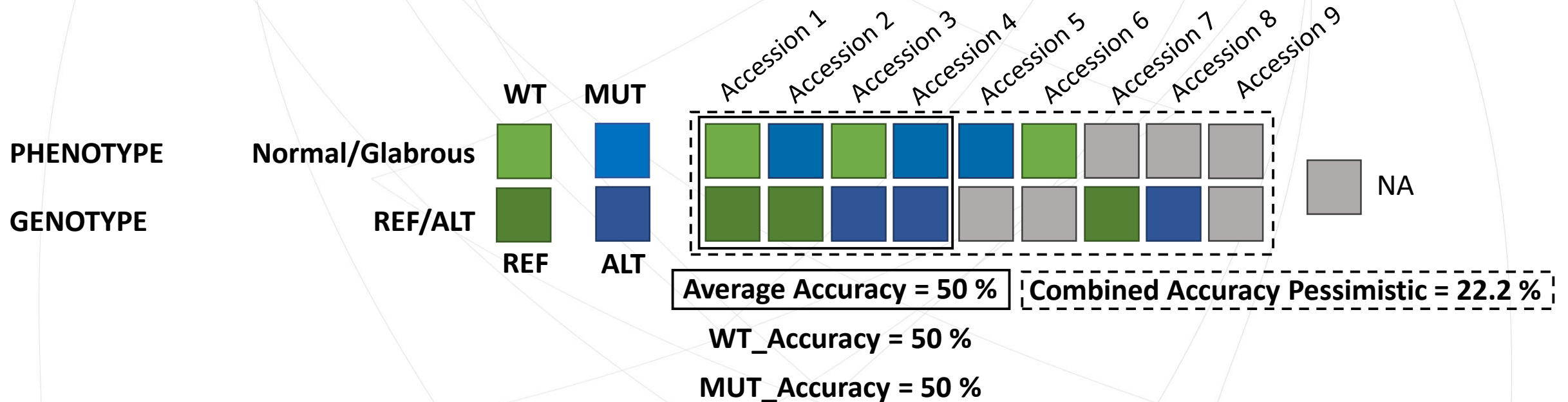
b



- 100% correspondence between *P1-CM* and the presence/absence of trichomes
- Correspondence between *P1-CM* and its TM is not perfect

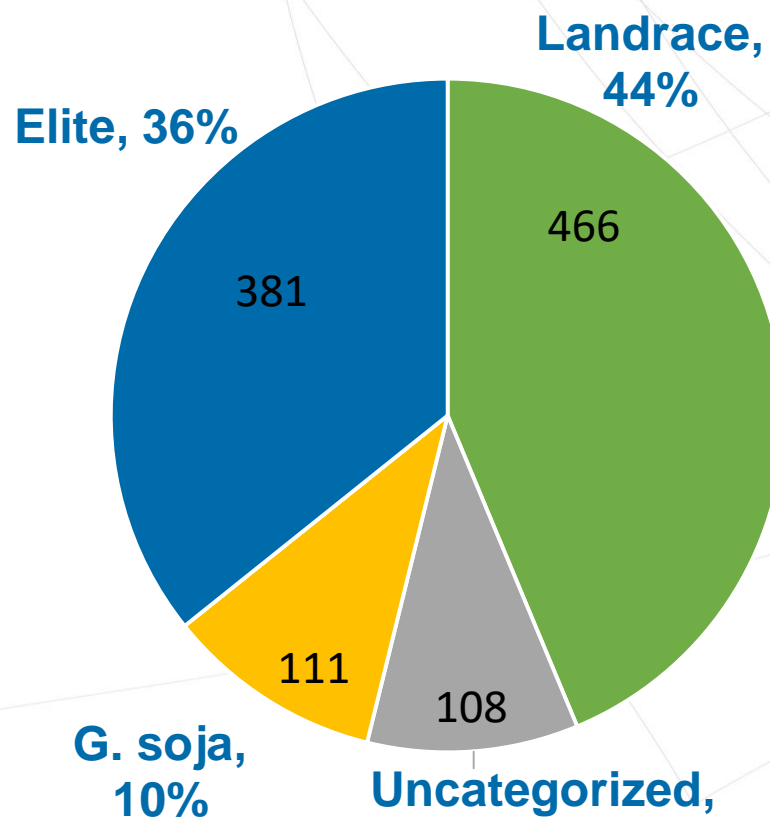
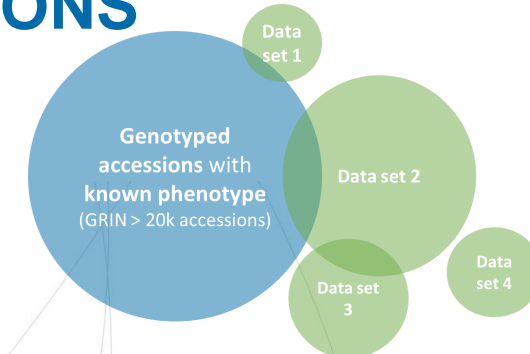
ACCURACY

- Combination of sensitivity and specificity
- A measure of direct correspondence between variant positions and phenotypes



A CURATED PANEL OF SOYBEAN RESEQUENCED ACCESSIONS

- All publicly available resequenced data sets with accessions from over 35 countries
- **Soy775: 35.7 M variant positions** - **1** glabrous accession (Škrabišová et al., 2022)
- **SnakyVC pipeline:** - Soy1066 38.3 M (**7**, Chan et al., 2023) > **Soy2939 44.3 M (18)**



Chan et al. *BMC Genomics* (2023) 24:107
<https://doi.org/10.1186/s12864-023-09161-3>

BMC Genomics

SOFTWARE

Open Access

The Allele Catalog Tool: a web-based interactive tool for allele discovery and analysis

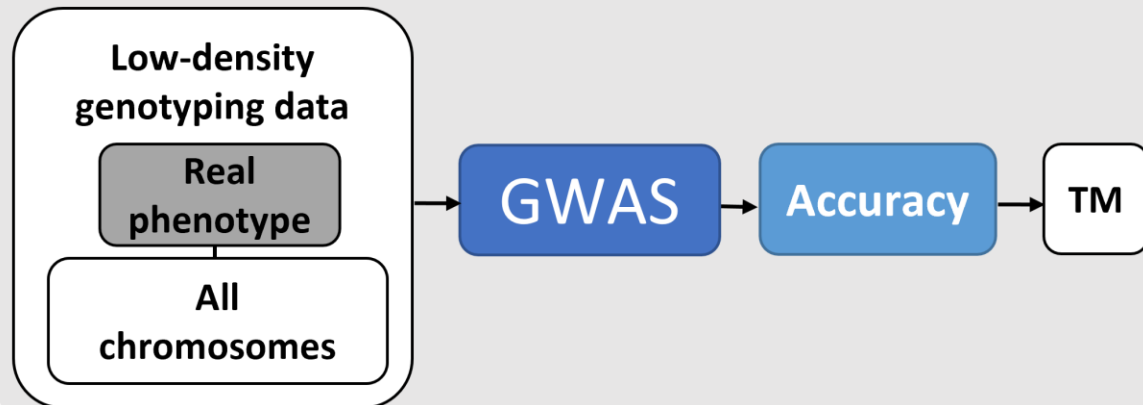
Yen On Chan^{1,2}, Nicholas Dietz³, Shuai Zeng⁴, Juexin Wang^{2,4}, Sherry Flint-Garcia⁵, M. Nancy Salazar-Vidal^{3,6}, Mária Škrabišová⁷, Kristin Bilyeu^{5*} and Trupti Joshi^{1,2,4,8*} 



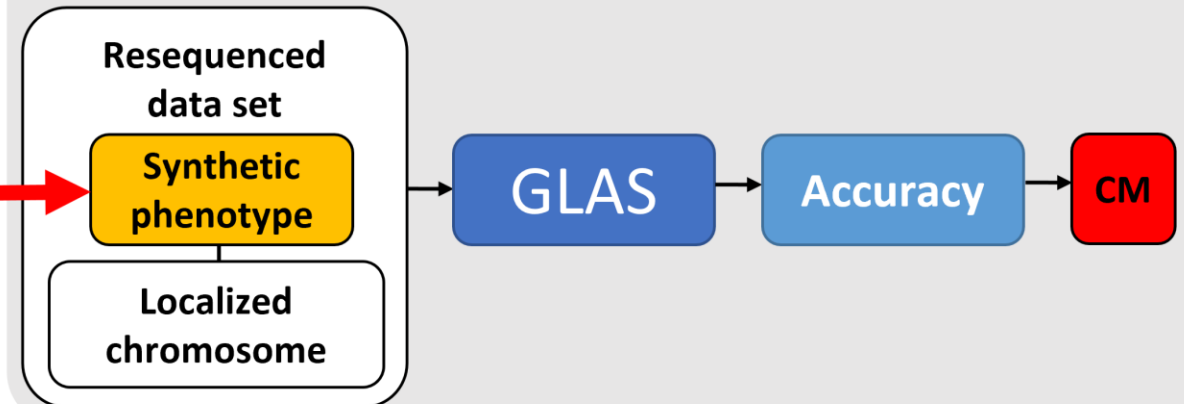
SYNTHETIC PHENOTYPE TO CAUSATIVE MUTATION STRATEGY (SP2CM)

SP2CM: Synthetic phenotype to CM strategy

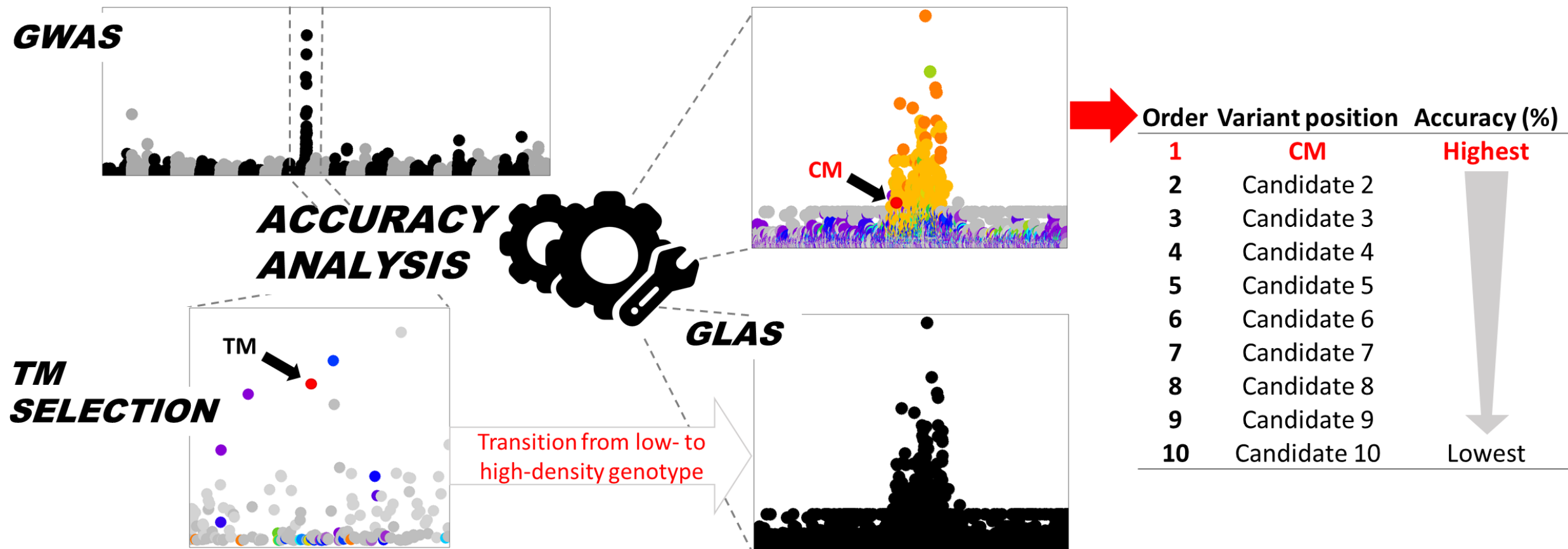
PART 1: Identification of the most accurate tagging marker



PART 2: Landscape of association for a genomic variant



SP2CM – NAVIGATION SCHEME



CASE STUDIES

PHENOTYPES

Proportional

Disproportional

Rare

MG: *E1*

Protein: *SWEET39*

MG: *E2*

Protein: *CCT*

MG: *E3*

Pod shattering: *Pdh1*

Quantitative

Qualitative

PUB_C: *T*

Leaf shape: *Ln*

SCN_R: *SNAP18*

FLWRC: *W1*

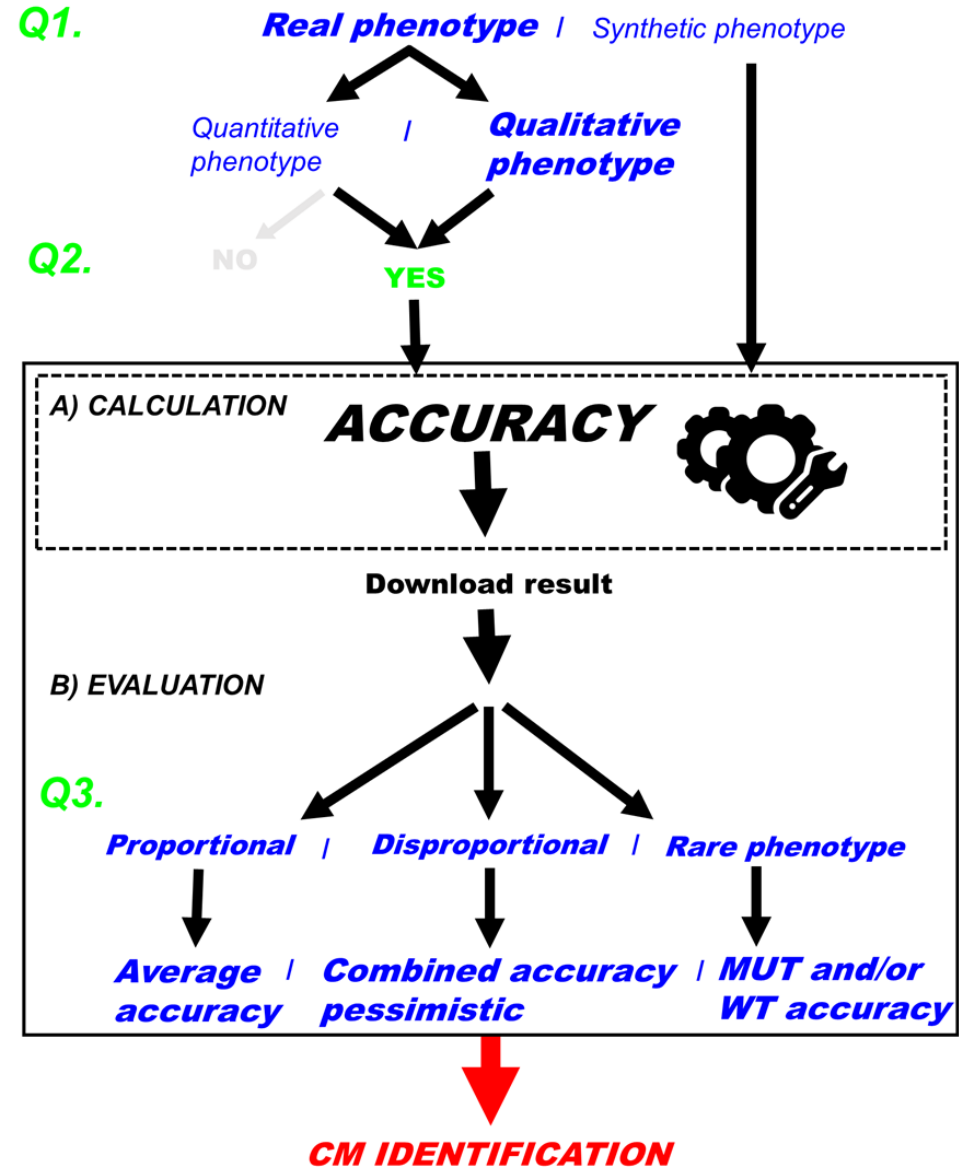
PUB_C: *Td*

Stay green: *D1*

SCN_R: *SHMT08*

STEMTERM: *Dt1*

SP2CM DECISION TREE



SP2CM IMPROVES GWAS-DRIVEN DISCOVERIES

- **Leverages** both resequenced and genotyping **data**
- Helps to decide whether to invest in additional resequencing or phenotyping
- **Narrows down the number of candidate genes**
- **Assists in identifying CM**

SP2CM IMPROVES GWAS-DRIVEN DISCOVERIES

- **Leverages** both resequenced and genotyping **data**
- Helps to decide whether to invest in additional resequencing or phenotyping
- **Narrows down the number of candidate genes**
- **Assists in identifying CM**

Identified genes:

Pod shattering: *NST1A*

Pod color: *L1*

Seed coat color: *O*

Soybean cyst nematode resistance: *SNAP11*

Pubescence density: *P1*

Pod color: *L2*

NEW GENES IDENTIFIED

- Pod color *L2*



OPEN ACCESS

EDITED BY

Li Ma,
University of Maryland, College Park,
United States

REVIEWED BY

Shoaib Ur Rehman,
Muhammad Nawaz Shareef University of
Agriculture, Pakistan
Sridhar Malkaram,
West Virginia State University,
United States

*CORRESPONDENCE

Kristin Bilyeu,
✉ kristin.bilyeu@usda.gov
Mária Škrabišová,
✉ maria.skrabisova@upol.cz

[†]These authors have contributed equally
to this work



TYPE Original Research
PUBLISHED 08 January 2024
DOI 10.3389/fgene.2023.1320652

Natural and artificial selection of multiple alleles revealed through genomic analyses

Jana Biová^{1†}, Ivana Kaňovská^{1†}, Yen On Chan^{2,3},
Manish Sridhar Immadi⁴, Trupti Joshi^{2,3,4,5}, Kristin Bilyeu^{6*} and
Mária Škrabišová^{1*}

¹Department of Biochemistry, Faculty of Science, Palacký University in Olomouc, Olomouc, Czechia, ²MU Institute for Data Science and Informatics, University of Missouri-Columbia, Columbia, MO, United States, ³Christopher S. Bond Life Sciences Center, University of Missouri-Columbia, Columbia, MO, United States, ⁴Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, Columbia, MO, United States, ⁵Department of Biomedical Informatics, Biostatistics and Medical Epidemiology, University of Missouri-Columbia, Columbia, MO, United States, ⁶United States Department of Agriculture-Agricultural Research Service, Plant Genetics Research Unit, Columbia, MO, United States

SOYBEAN APPLIED GENOMICS HUB

The screenshot displays the Soybean Knowledge Base (SoyKB) website. At the top left is the 'soykb' logo. To the right is a login section with 'Username' and 'Password' input fields, a 'Go' button, and links for 'Create an Account', 'Retrieve a Lost Password', and 'Login/Account signup only required for access to private data.' Below this is a dark navigation bar with links: Home, Search, Browse, Tools, Soy Hub, Data Files, Analytics, Information, and About. A large green banner features a soybean leaf image and the text 'SOYBEAN KNOWLEDGE BASE (SoyKB) A web resource for Soybean Translational Genomics'. Below the banner is a 'SoyHub' section with a 'Quick Search' bar, a 'Gene Card' dropdown, and a 'Go' button. The main content area is divided into two columns. The left column, titled 'Explore variation:', contains 'Allele Catalog' (with sub-points: Find accessions with certain allele, Find new alleles in known genes) and 'GenVarX' (with sub-points: Explore variation in promoters, Search TFs). The right column, titled 'Predict new causal mutations:', contains 'AccuTool' (with sub-points: Use GWAS results for prediction, Calculate Accuracy for your markers or candidate causative mutations (CM) based on Soy775 35.7M variant positions) and 'SNPViz' (with sub-point: Check genomic context of your variant positions in empowered haplotype viewer on various resequenced data sets).

soykb

Username Password Go

Create an Account | Retrieve a Lost Password
Login/Account signup only required for access to private data.

Home Search Browse Tools Soy Hub Data Files Analytics Information About

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

☒ **SoyHub** Quick Search Gene Card Go

Welcome to Soy Hub
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

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

SOYBEAN ALLELE CATALOG



Create an Account | Retrieve a Lost Password
Login/Account signup only required for access to private data.

HomeSearchBrowseToolsData FilesAnalyticsInformationAbout



SOYBEAN KNOWLEDGE BASE (SoyKB)

A web resource for Soybean Translational Genomics

☒ Soybean Allele Catalog Tool

Quick SearchGene CardGo

Search by Gene IDs

Dataset: Soy1066 Allele Catalog

Gene IDs: (eg Glyma.01G049100 Glyma.01G049200 Glyma.01G049300)

Please separate each gene into a new line.
Example:
Glyma.01G049100
Glyma.01G049200
Glyma.01G049300

Improvement Status:
☒ Soja ☒ Elite ☒ Landrace ☒ Cultivar

Search

Search by Accessions and Gene ID

Dataset: Soy1066 Allele Catalog

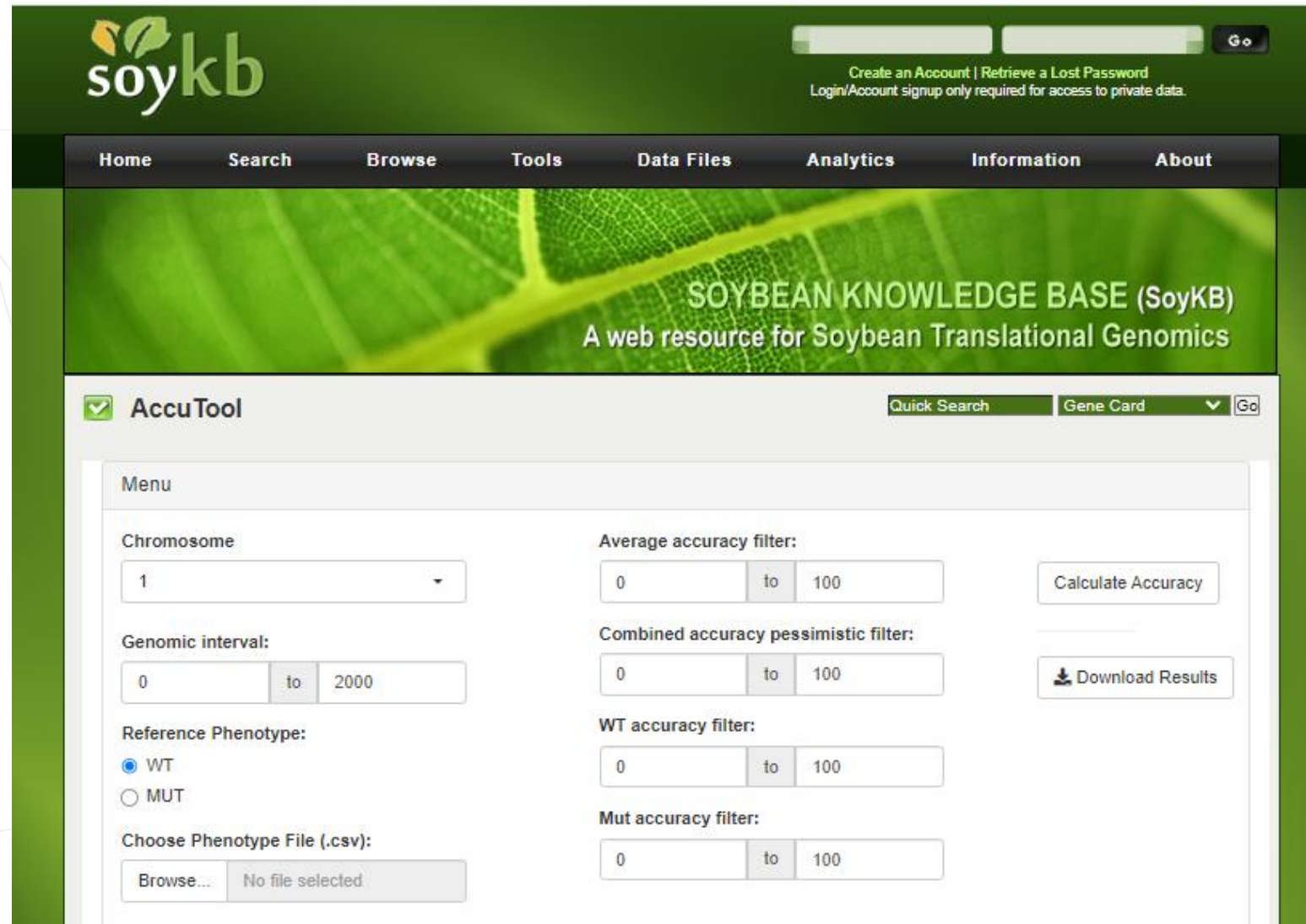
Accessions: (eg HN005_P1404166 HN006_P1407788A)

Please separate each accession into a new line.
Example:
HN005_P1404166
HN006_P1407788A

Gene ID: (One gene name only; eg Glyma.01G049100)

Search

ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS



The screenshot displays the SoyKB website interface. At the top, the 'soykb' logo is on the left, and a navigation bar contains links for Home, Search, Browse, Tools, Data Files, Analytics, Information, and About. Below the navigation bar is a banner for 'SOYBEAN KNOWLEDGE BASE (SoyKB)' with the tagline 'A web resource for Soybean Translational Genomics'. The main content area features the 'AccuTool' section, which includes a 'Menu' tab and several input fields for data analysis. The 'Chromosome' field is set to '1'. The 'Genomic interval' is set from '0' to '2000'. The 'Reference Phenotype' section has 'WT' selected. The 'Choose Phenotype File (.csv)' section shows a 'Browse...' button and 'No file selected'. On the right, there are four accuracy filter sections: 'Average accuracy filter', 'Combined accuracy pessimistic filter', 'WT accuracy filter', and 'Mut accuracy filter', each with input fields for '0' and '100'. A 'Calculate Accuracy' button is located below the first two filter sections, and a 'Download Results' button is below the last two. A 'Go' button is in the top right corner of the page.

soykb

Create an Account | Retrieve a Lost Password
Login/Account signup only required for access to private data.

Home Search Browse Tools Data Files Analytics Information About

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

✓ AccuTool Quick Search Gene Card Go

Menu

Chromosome
1

Genomic interval:
0 to 2000

Reference Phenotype:
☒ WT
☐ MUT

Choose Phenotype File (.csv):
Browse... No file selected

Average accuracy filter:
0 to 100

Combined accuracy pessimistic filter:
0 to 100

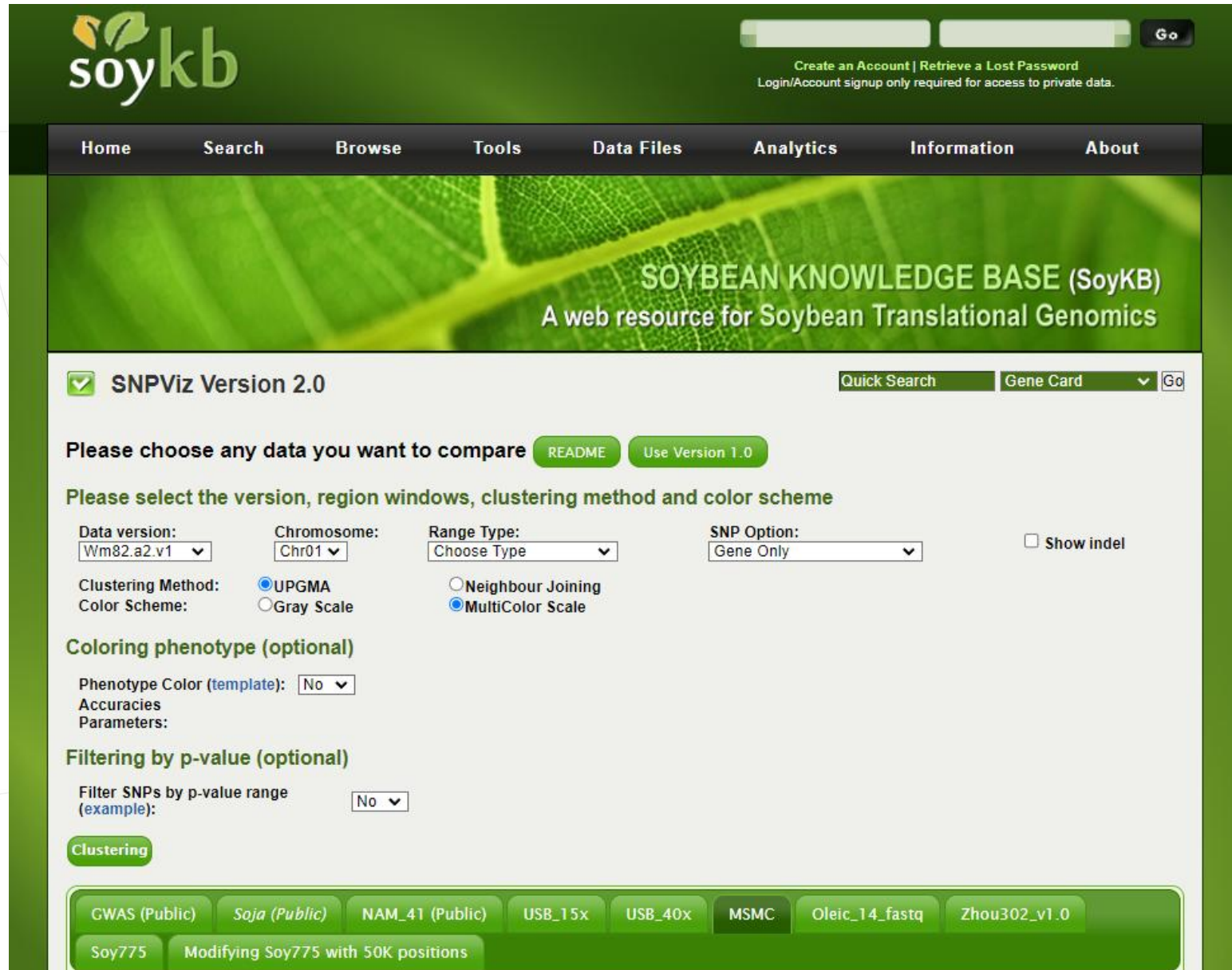
WT accuracy filter:
0 to 100

Mut accuracy filter:
0 to 100

Calculate Accuracy

Download Results

SNPVIZ 2.0: EMPOWERED HAPLOTYPE VIEWER



The screenshot displays the SoyKB website's SNPviz 2.0 interface. At the top, the 'soykb' logo is on the left, and a 'Go' button is on the right. Below the logo, a navigation bar contains links: Home, Search, Browse, Tools, Data Files, Analytics, Information, and About. A large green banner with a soybean leaf image reads 'SOYBEAN KNOWLEDGE BASE (SoyKB) A web resource for Soybean Translational Genomics'. The main content area features a 'SNPViz Version 2.0' section with a 'Quick Search' button and a 'Gene Card' dropdown. Below this, a 'Please choose any data you want to compare' section includes a 'README' button and a 'Use Version 1.0' button. The 'Please select the version, region windows, clustering method and color scheme' section contains several dropdown menus: 'Data version' (Wm82.a2.v1), 'Chromosome' (Chr01), 'Range Type' (Choose Type), and 'SNP Option' (Gene Only). There are also radio buttons for 'Clustering Method' (UPGMA selected, Neighbour Joining) and 'Color Scheme' (Gray Scale selected, MultiColor Scale). A 'Show indel' checkbox is present. The 'Coloring phenotype (optional)' section has a 'Phenotype Color (template)' dropdown (No) and links for 'Accuracies' and 'Parameters'. The 'Filtering by p-value (optional)' section has a 'Filter SNPs by p-value range (example)' dropdown (No). A 'Clustering' button is located below the filtering section. At the bottom, a row of buttons lists various datasets: GWAS (Public), Soja (Public), NAM_41 (Public), USB_15x, USB_40x, MSMC, Oleic_14_fastq, Zhou302_v1.0, Soy775, and Modifying Soy775 with 50K positions.

soykb

Create an Account | Retrieve a Lost Password
Login/Account signup only required for access to private data.

Home Search Browse Tools Data Files Analytics Information About

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

☒ **SNPViz Version 2.0** Quick Search Gene Card

Please choose any data you want to compare

Please select the version, region windows, clustering method and color scheme

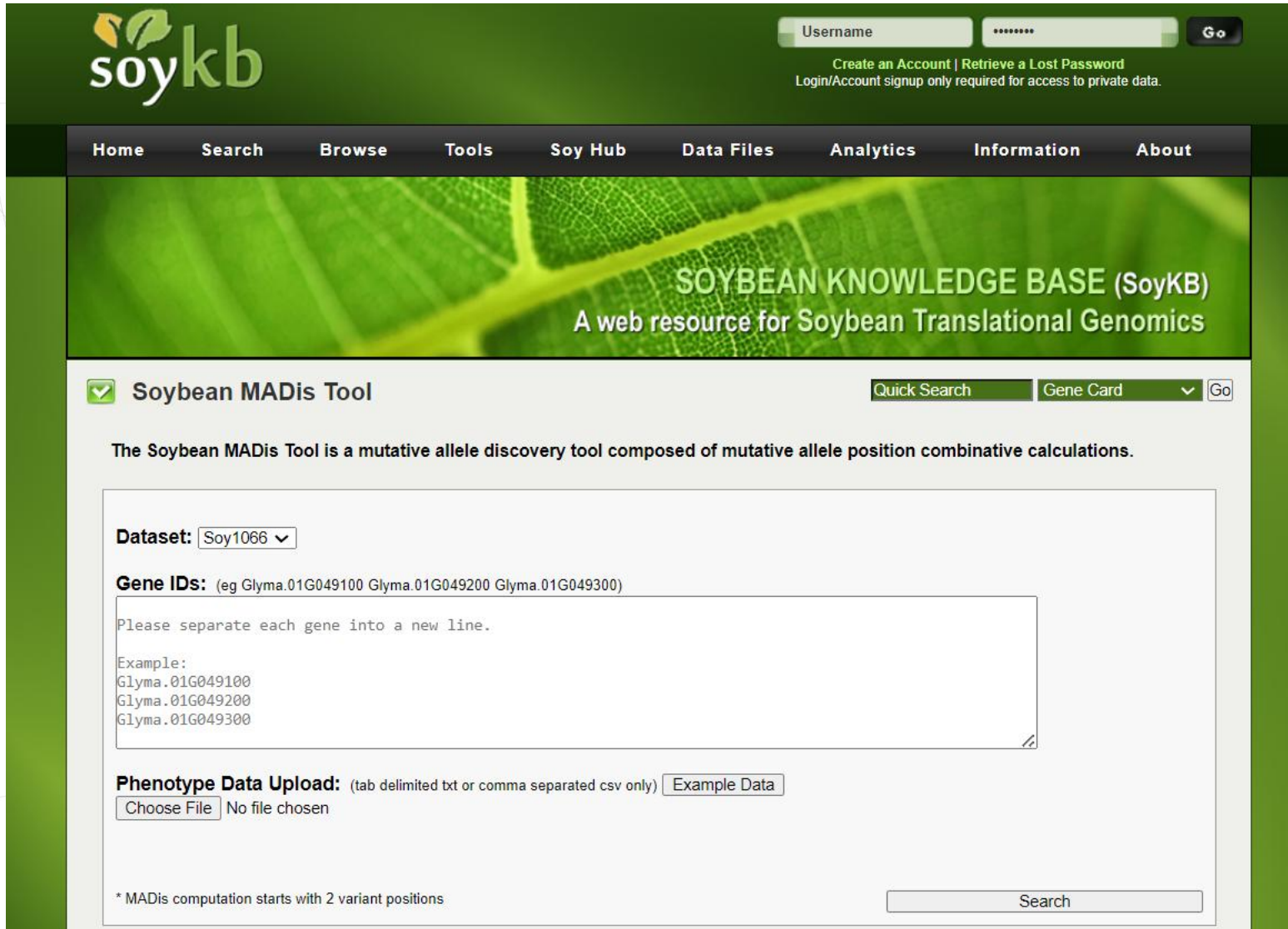
Data version: Chromosome: Range Type: SNP Option: ☐ Show indel

Clustering Method: ☒ UPGMA ☐ Neighbour Joining
Color Scheme: ☐ Gray Scale ☒ MultiColor Scale

Coloring phenotype (optional)
Phenotype Color (template): [Accuracies](#)
[Parameters](#)

Filtering by p-value (optional)
Filter SNPs by p-value range (example):

MADIS: MUTATIVE ALLELE DISCOVERY TOOL



The screenshot displays the Soybean Knowledge Base (SoyKB) website. At the top, there is a navigation bar with links: Home, Search, Browse, Tools, Soy Hub, Data Files, Analytics, Information, and About. Below this is a banner for the Soybean Knowledge Base (SoyKB) with the text "A web resource for Soybean Translational Genomics". The main content area features the "Soybean MADis Tool" section. It includes a "Quick Search" button, a "Gene Card" dropdown, and a "Go" button. A description states: "The Soybean MADis Tool is a mutative allele discovery tool composed of mutative allele position combinative calculations." Below this, there is a "Dataset" dropdown menu set to "Soy1066". The "Gene IDs" section provides an example of gene IDs: "Glyma.01G049100 Glyma.01G049200 Glyma.01G049300". A text input field is provided for entering gene IDs, with a note: "Please separate each gene into a new line." Below the input field, there is a "Phenotype Data Upload" section with a note: "(tab delimited txt or comma separated csv only)". It includes a "Choose File" button and a "No file chosen" status. An "Example Data" link is also present. At the bottom, there is a "Search" button and a note: "* MADis computation starts with 2 variant positions".

soykb

Username

[Create an Account](#) | [Retrieve a Lost Password](#)
Login/Account signup only required for access to private data.

[Home](#) [Search](#) [Browse](#) [Tools](#) [Soy Hub](#) [Data Files](#) [Analytics](#) [Information](#) [About](#)

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

☒ **Soybean MADis Tool**

The Soybean MADis Tool is a mutative allele discovery tool composed of mutative allele position combinative calculations.

Dataset:

Gene IDs: (eg Glyma.01G049100 Glyma.01G049200 Glyma.01G049300)

Please separate each gene into a new line.

Example:
Glyma.01G049100
Glyma.01G049200
Glyma.01G049300

Phenotype Data Upload: (tab delimited txt or comma separated csv only)

No file chosen

* MADis computation starts with 2 variant positions

GWAS TO GENES STRATEGY

It can be used for other species too!

- Arabidopsis
- Rice
- Cotton

We offer training!

- Assistance with analyses on our data OR on your own data
- Tools guidance

We are open to collaboration!

- Let's clone genes together
- Let's identify limitations of different genomes

UTILIZATION OF GWAS TO GENES STRATEGY FOR EUROPEAN SOYBEANS & KNOWLEDGE TRANSFER TO OTHER LEGUMES

Selection of precise markers

- Soybean maturity genes
- Food-grade traits
- Yield-related traits

Identification of new candidate genes

- Soybean maturity genes
- Food-grade traits
- Yield-related traits

Improving pre-breeding for legumes

- Exploration of natural diversity by contrasting the worldwide genetic pool



Funded by
the European Union

JOINT EFFORTS FOR SOYBEAN APPLIED GENOMICS

Legume Genomics



Dr. Mária Škrabišová



Jana Slivková, M.Sc.



Jana Biová, M.Sc.



Ivana Kaňovská, M.Sc.

Funding



Applied Genomics



Dr. Kristin Bilyeu



Dr. Nicholas Dietz



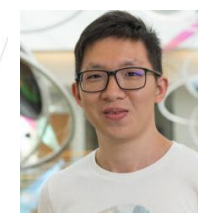
Anser Mahmood



Dr. Trupti Joshi



Dr. Shuai Zeng



Yen On Chan



Manish Sridhar
Immadi

Bioinformatics

Acknowledgement



Legume Generation (Boosting innovation in breeding for the next generation of legume crops for Europe) has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No.101081329. It also receives support from the governments of the United Kingdom, Switzerland and New Zealand.



THANK YOU FOR YOUR ATTENTION!

