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

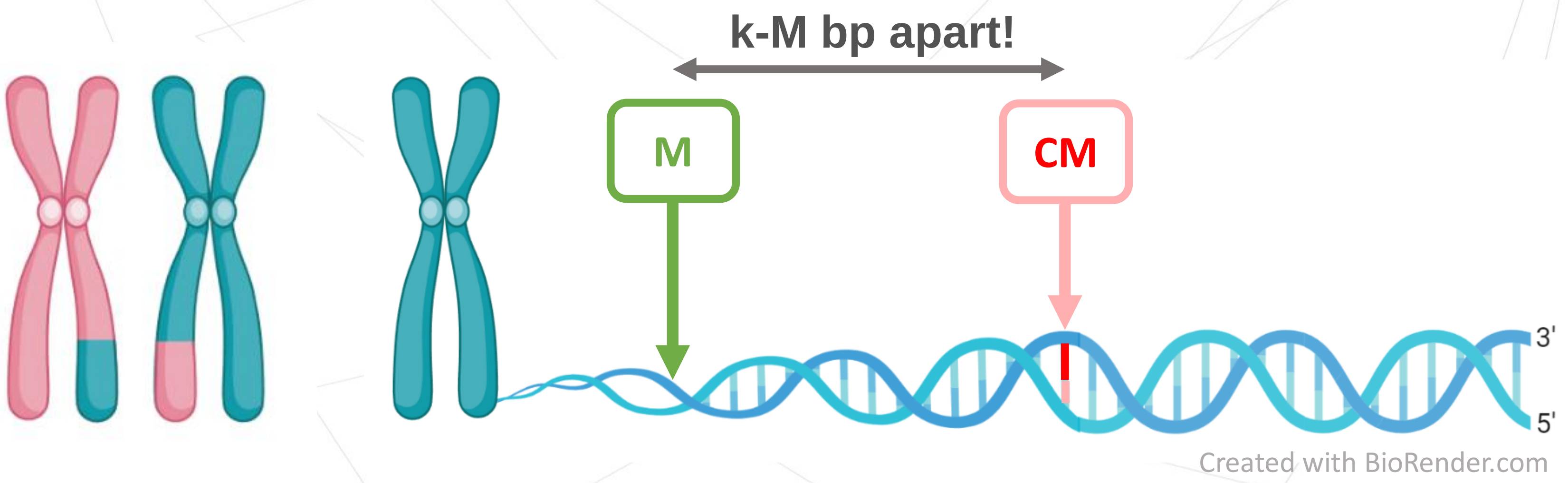
BREEDING IS NOT PRECISE ENOUGH YET!

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**



GWAS FOR MORE PRECISE BREEDING

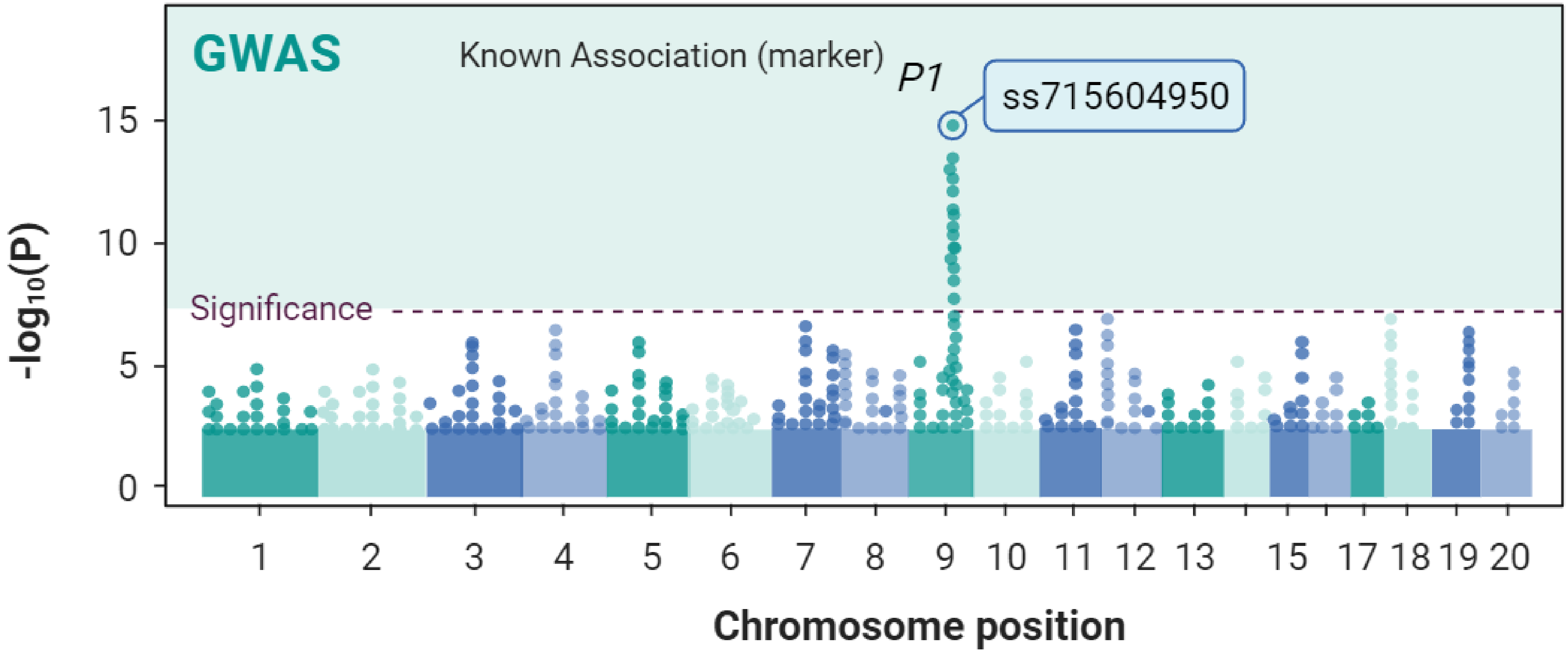
GWAS (Genome-wide association study)

- **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 ($$$)**

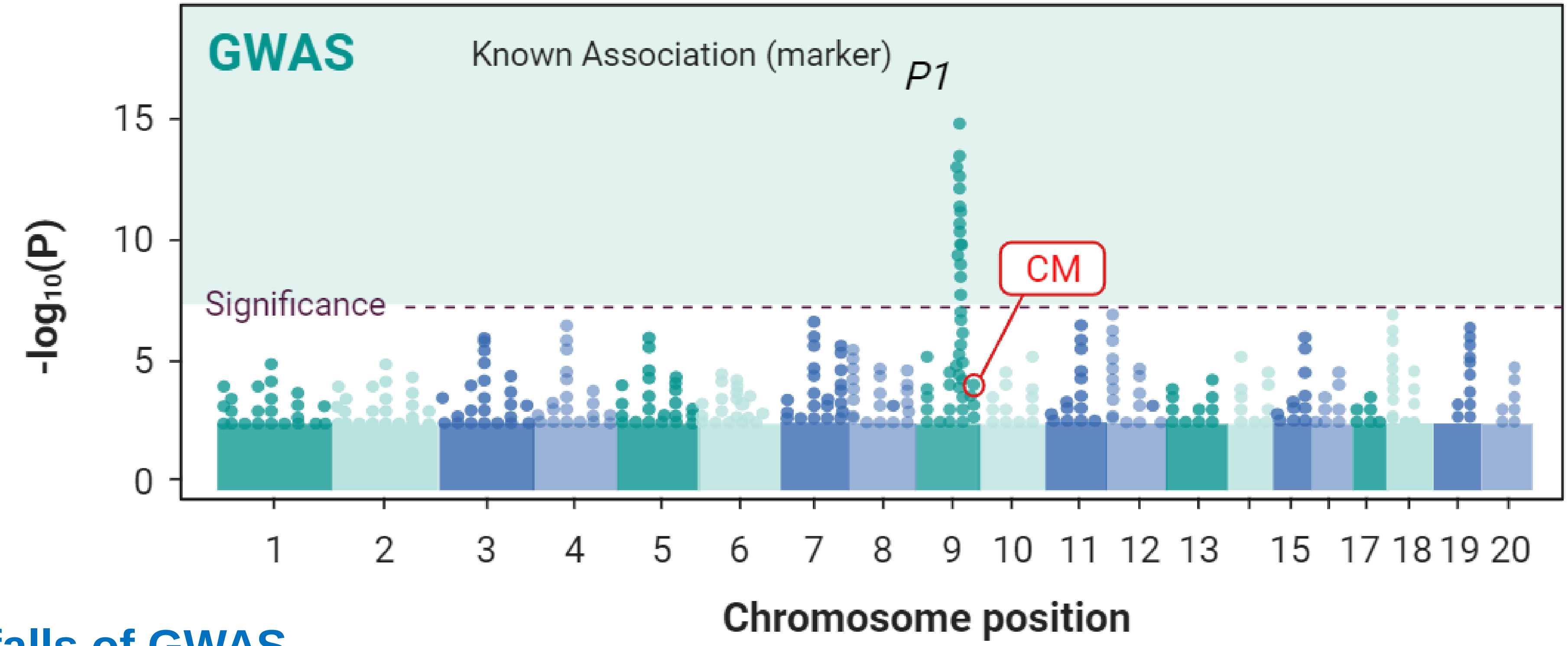
GWAS WITH GENOTYPING DATA



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, …)

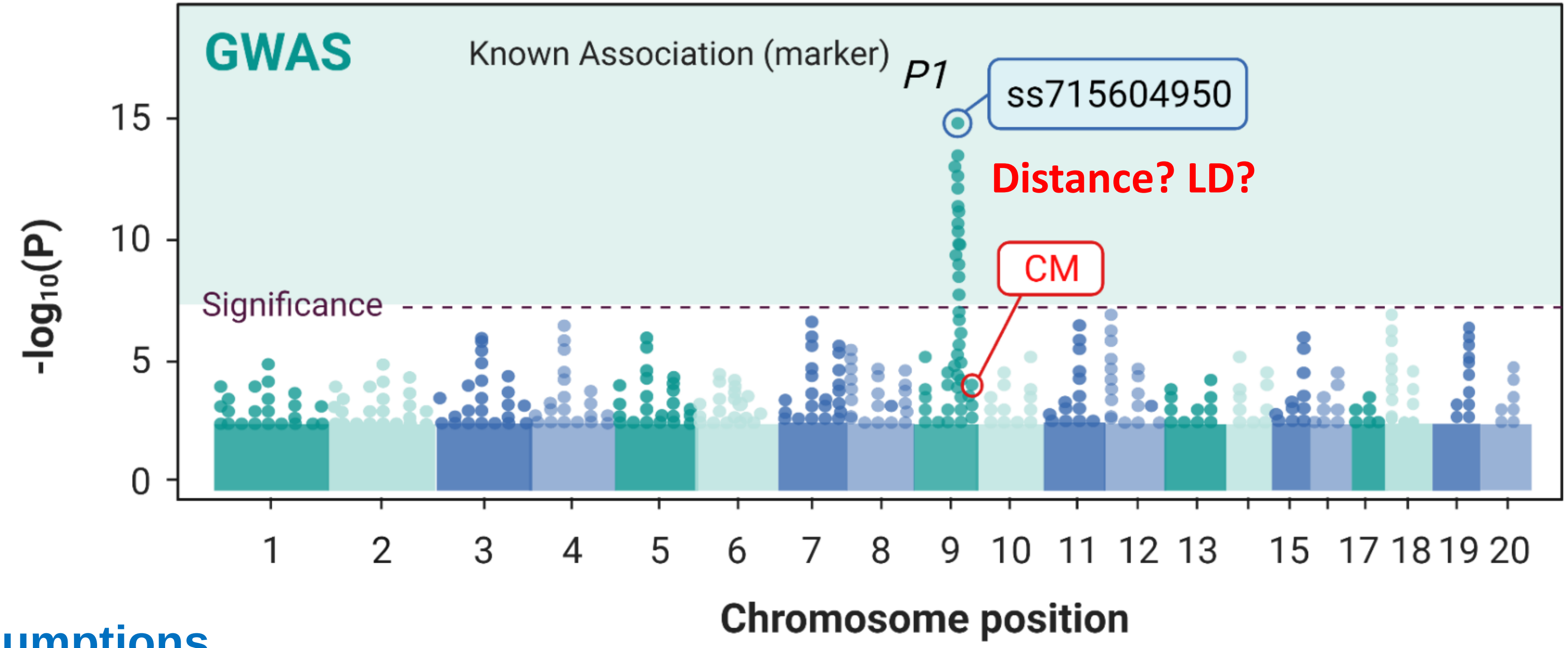
GWAS WITH RESEQUENCED DATA



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



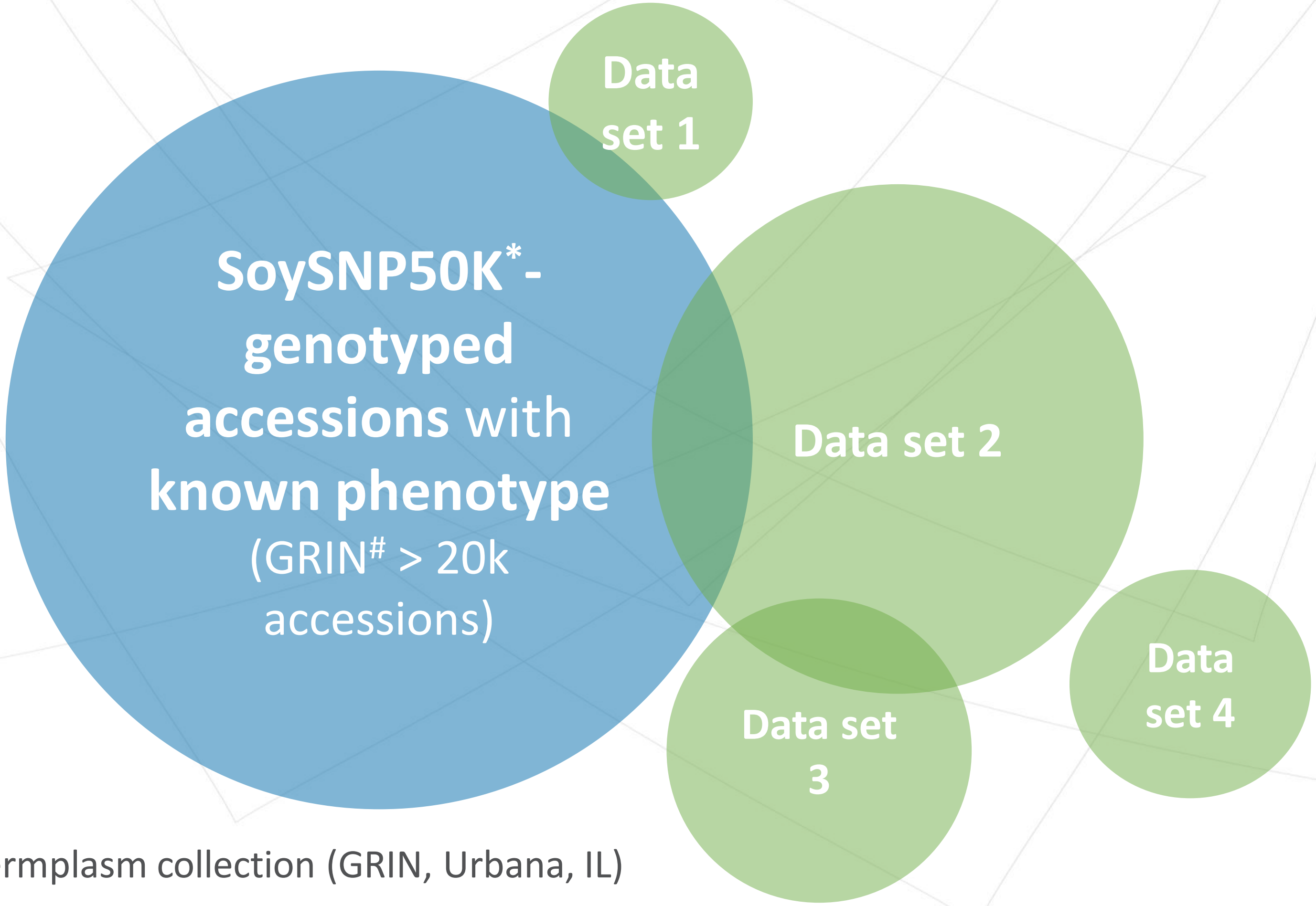
Assumptions

- 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

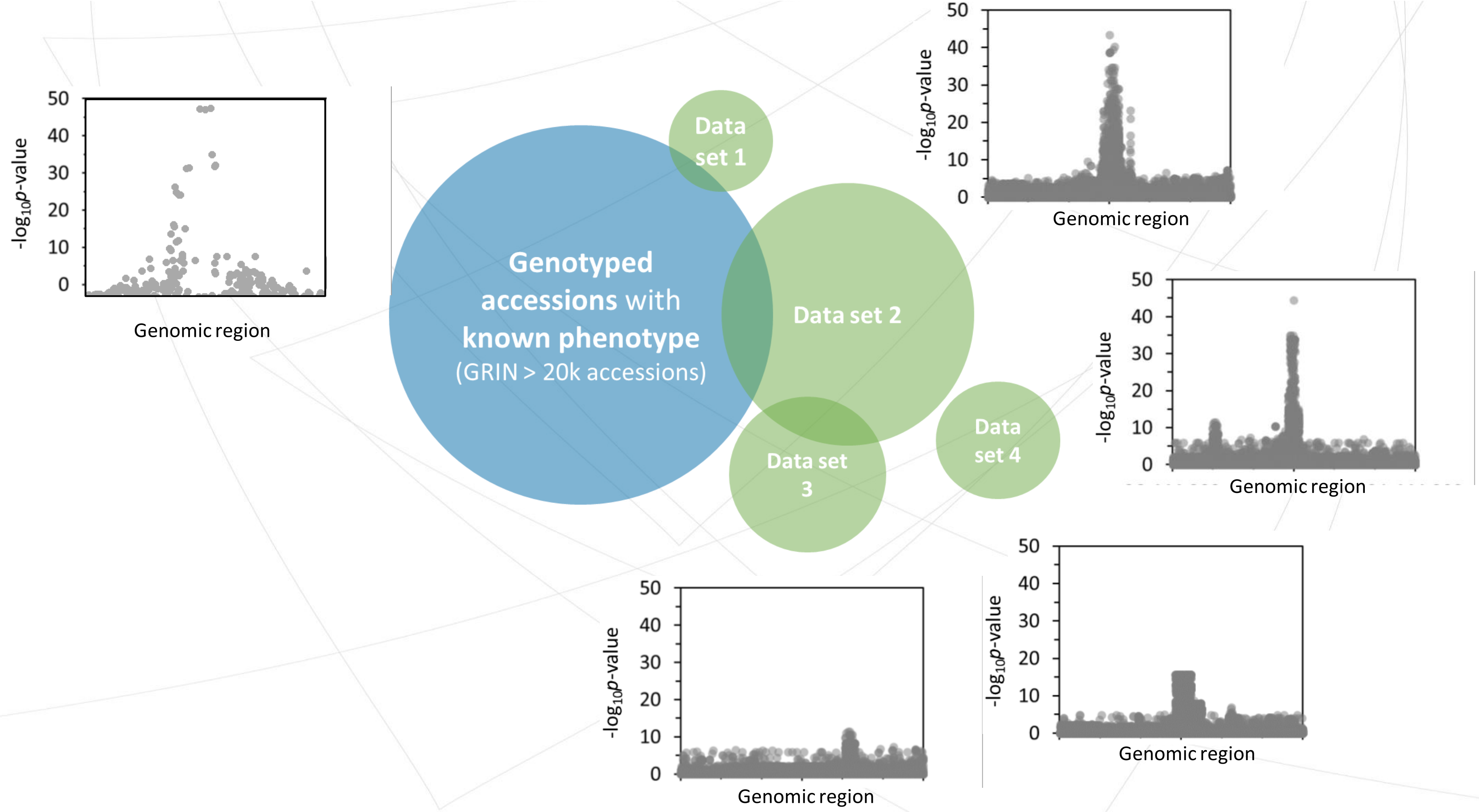
Resequenced data sets with **limited phenotype** information (or unavailable), “n” is small



*Song et al., 2015 #USDA germplasm collection (GRIN, Urbana, IL)

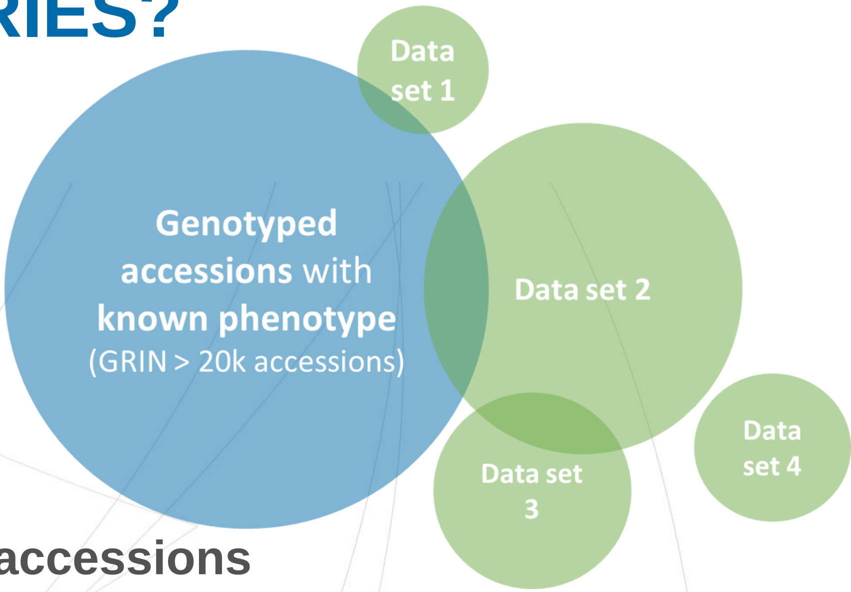
UNDERPOWERED GWAS FAILS IN CM PREDICTION

- **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

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, Palacky 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

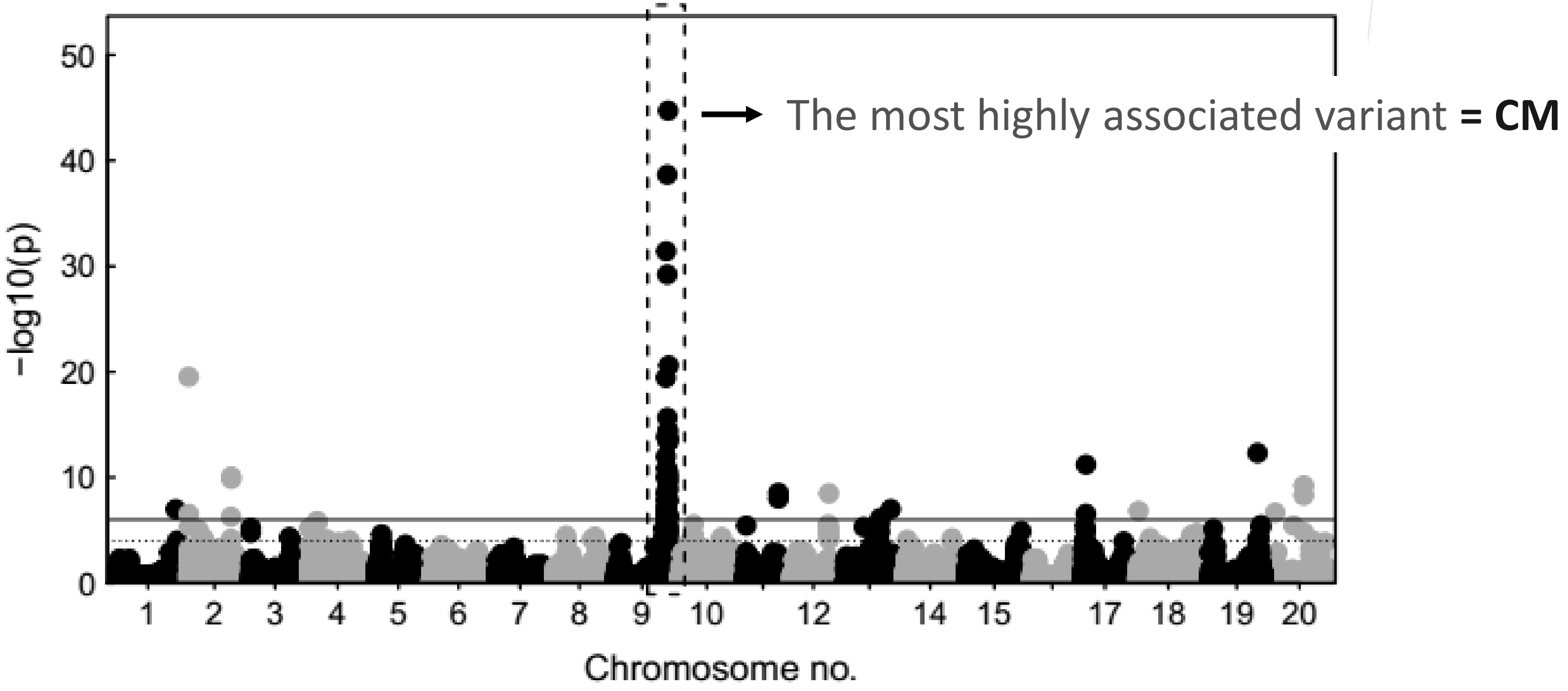
Real phenotype (Observed) **Pubescence presence/absence**		**Synthetic phenotype** CM in *Glyma.09g278000* A25T (Liu *et al.* 2020) **Chr09:45,057,956 genotype**	
WT	Normal	**REF**	T
MUT	Glabrous	**ALT**	A

- 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)

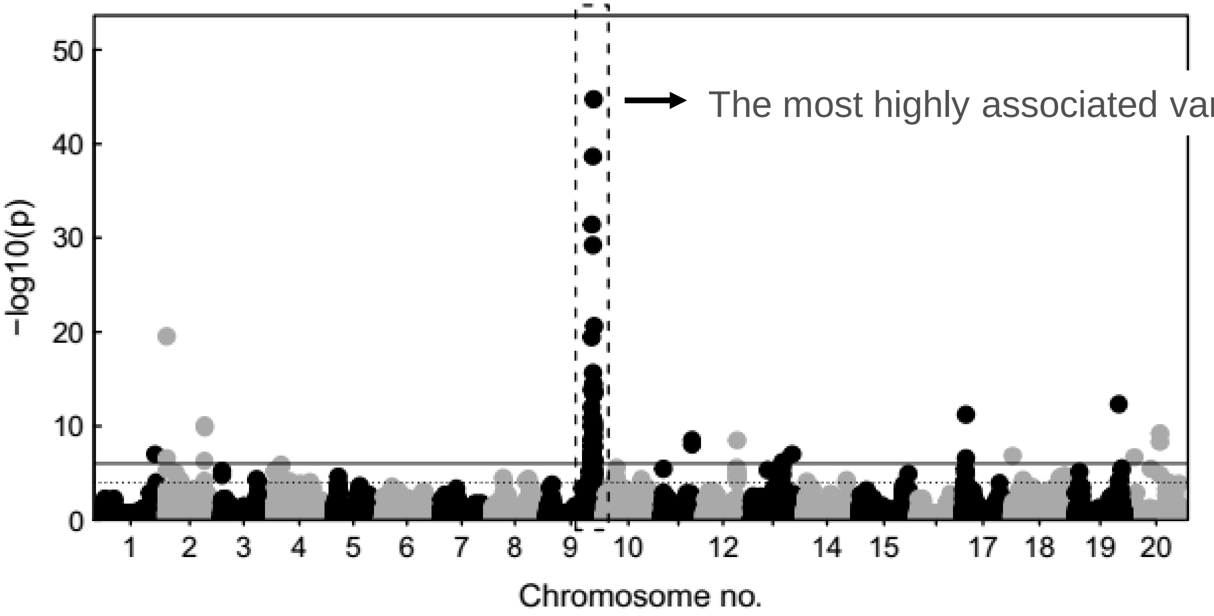
LOGIC OF SYNTHETIC PHENOTYPE: PERFECT GWAS

Prerequisites

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

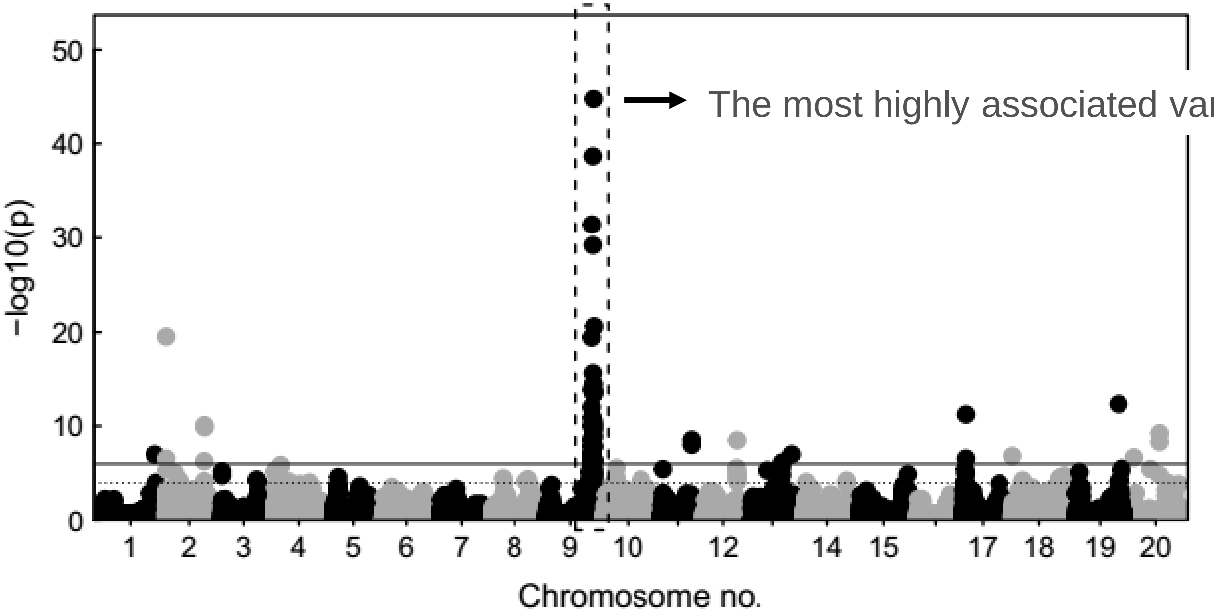


PERFECT GWAS



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

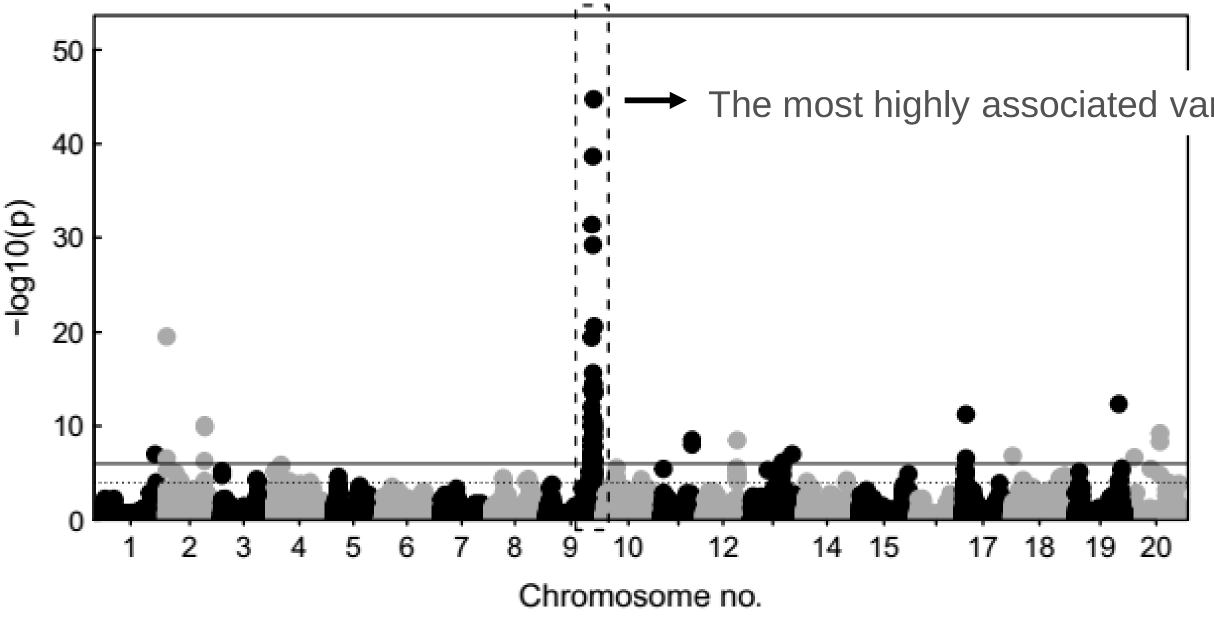
PERFECT GWAS



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

REF Allele **ALT Allele**

PERFECT GWAS

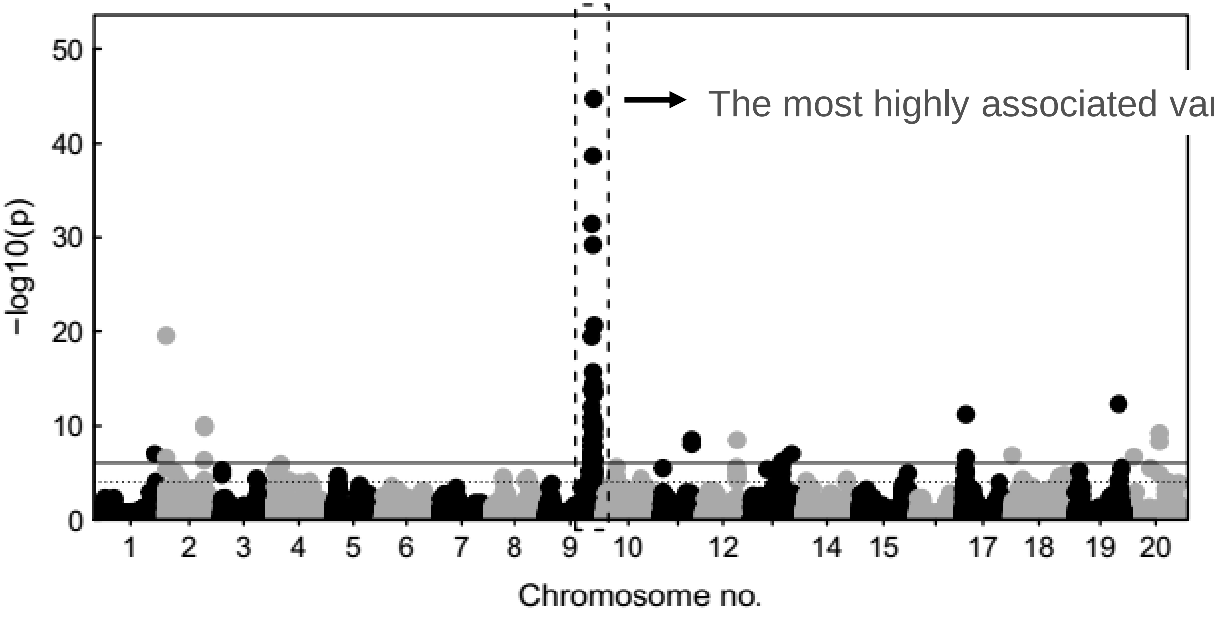


The most highly associated variant = **CM** **is a bi-allelic variant at a certain position in genome**

REF Allele = WT Phenotype

ALT Allele = MUT Phenotype

PERFECT GWAS

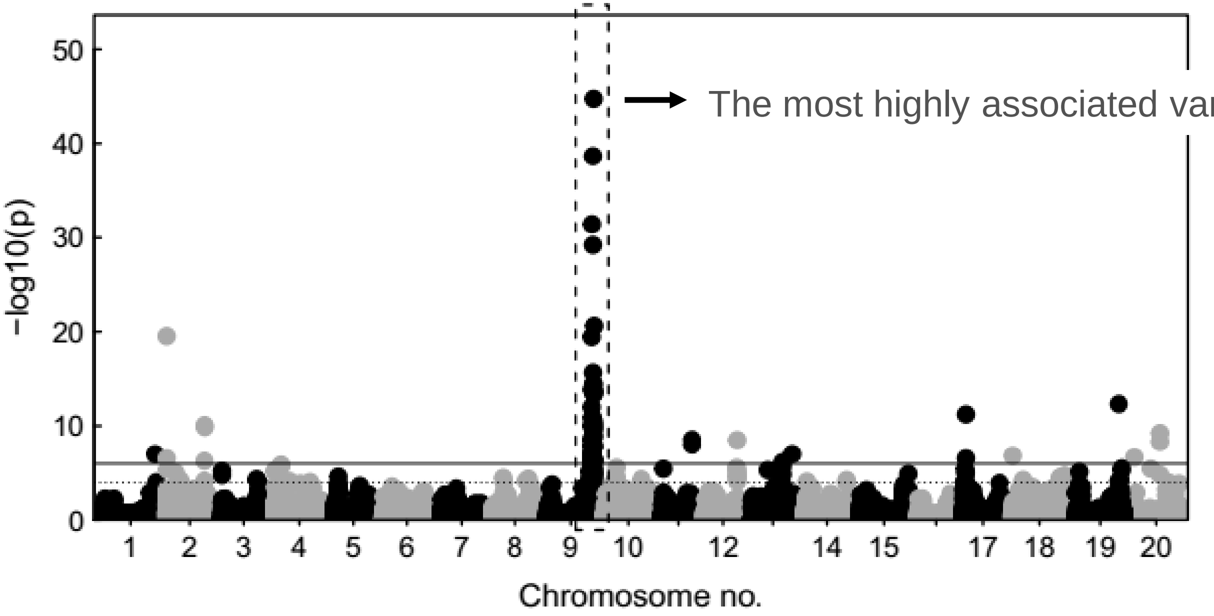


The most highly associated variant = **CM** **is a bi-allelic variant at a certain position in genome**

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

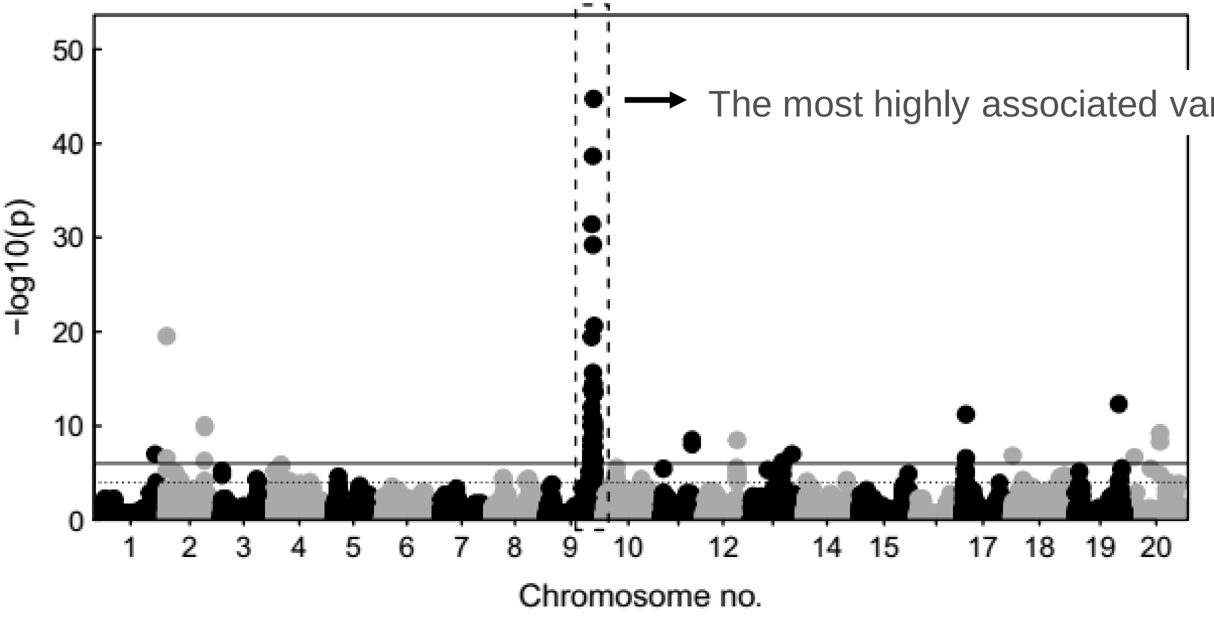
ALT Allele = MUT Phenotype

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

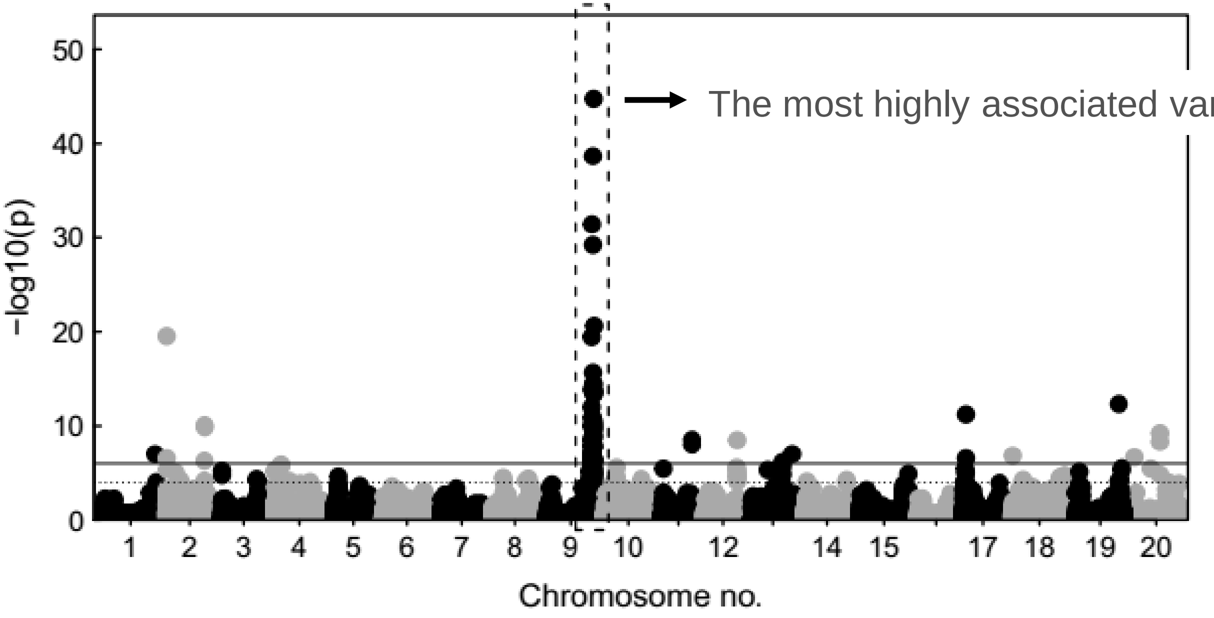
100% Accessions with WT phenotype have REF Allele

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

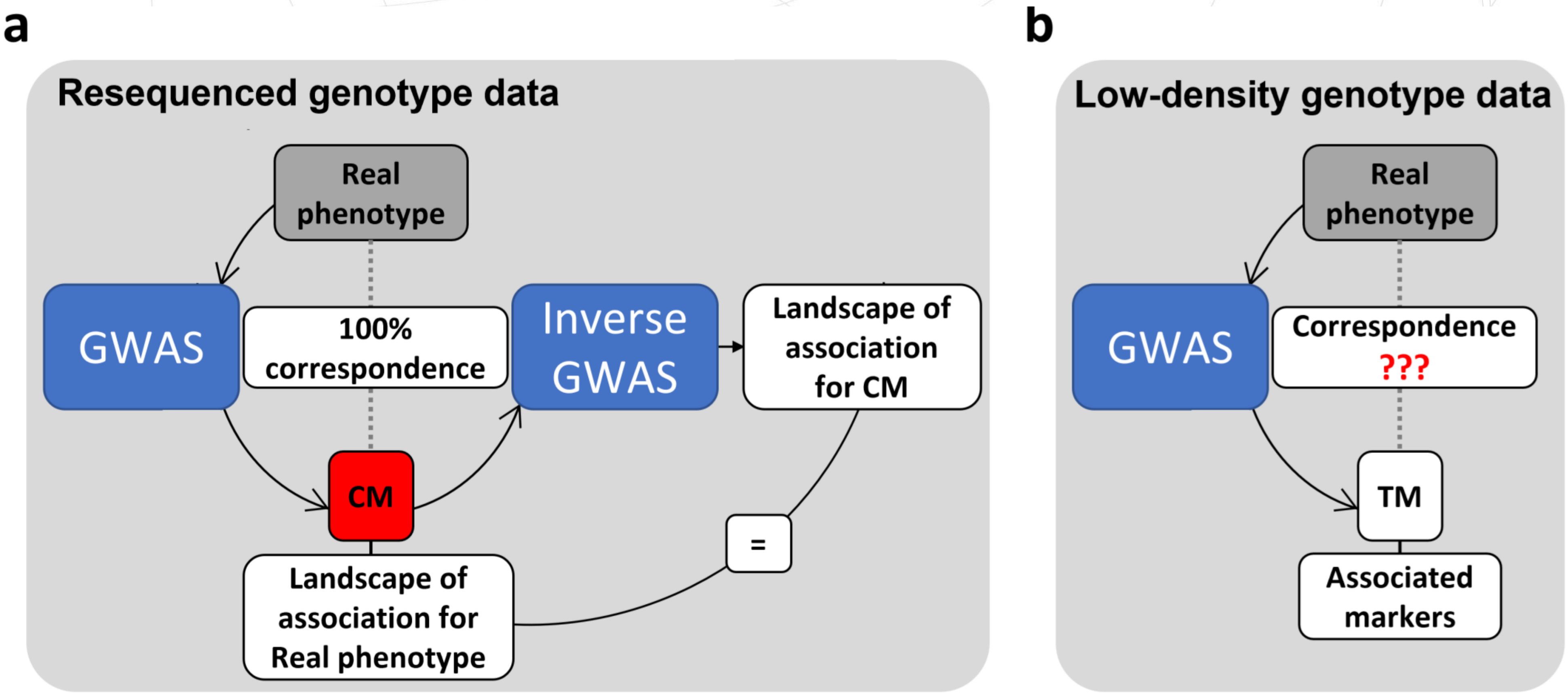
100% Accessions with WT phenotype have REF Allele

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

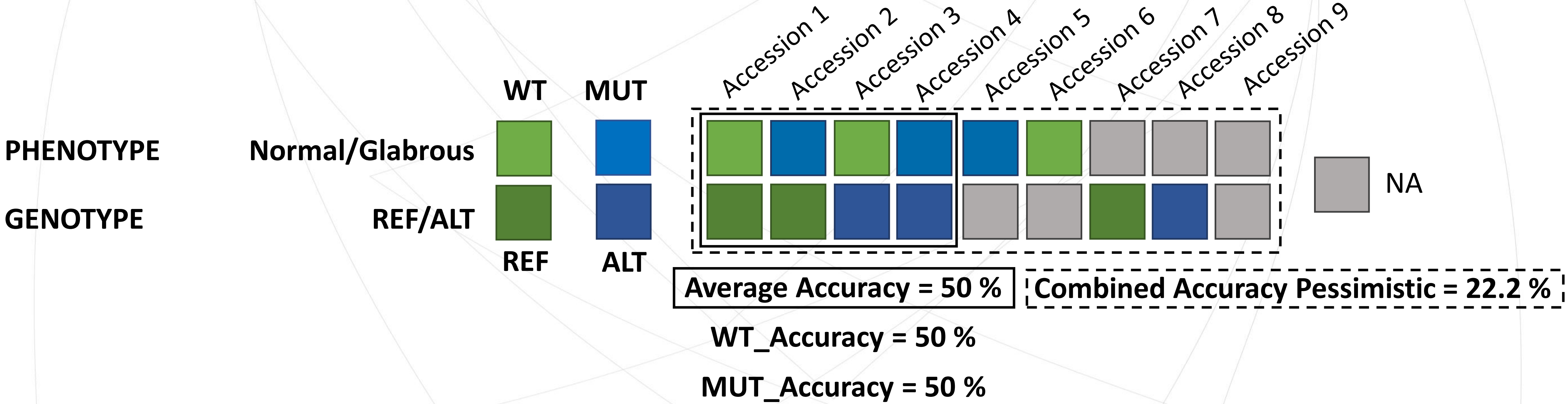
- Between markers, CMs and phenotypes



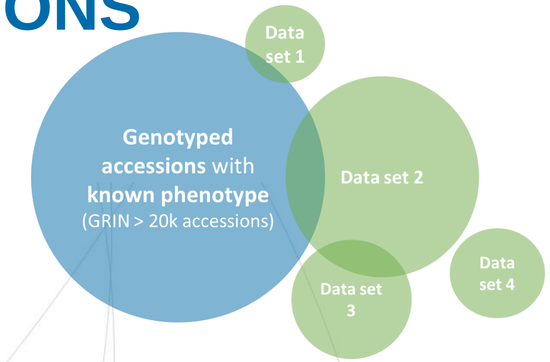
- **100% correspondence between *P1-CM* and the presence/absence of trichomes**
- Correspondence between *P1*-CM and its TM is not perfect

ACCURACY

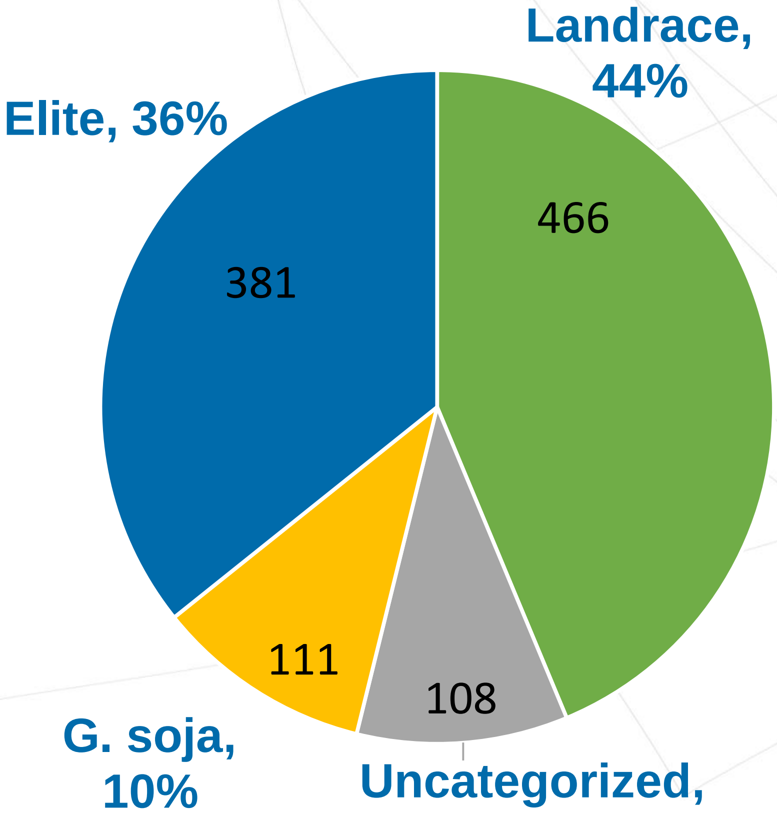
- Combination of sensitivity and specificity
- A measure of direct correspondence between variant positions and phenotypes



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

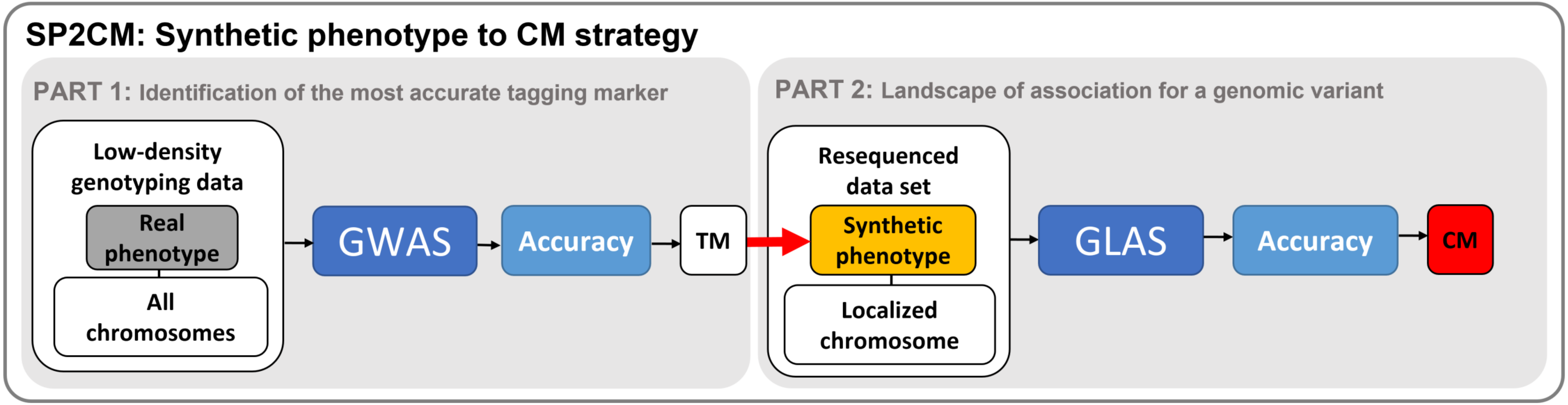
SOFTWARE Open Access

Check for updates

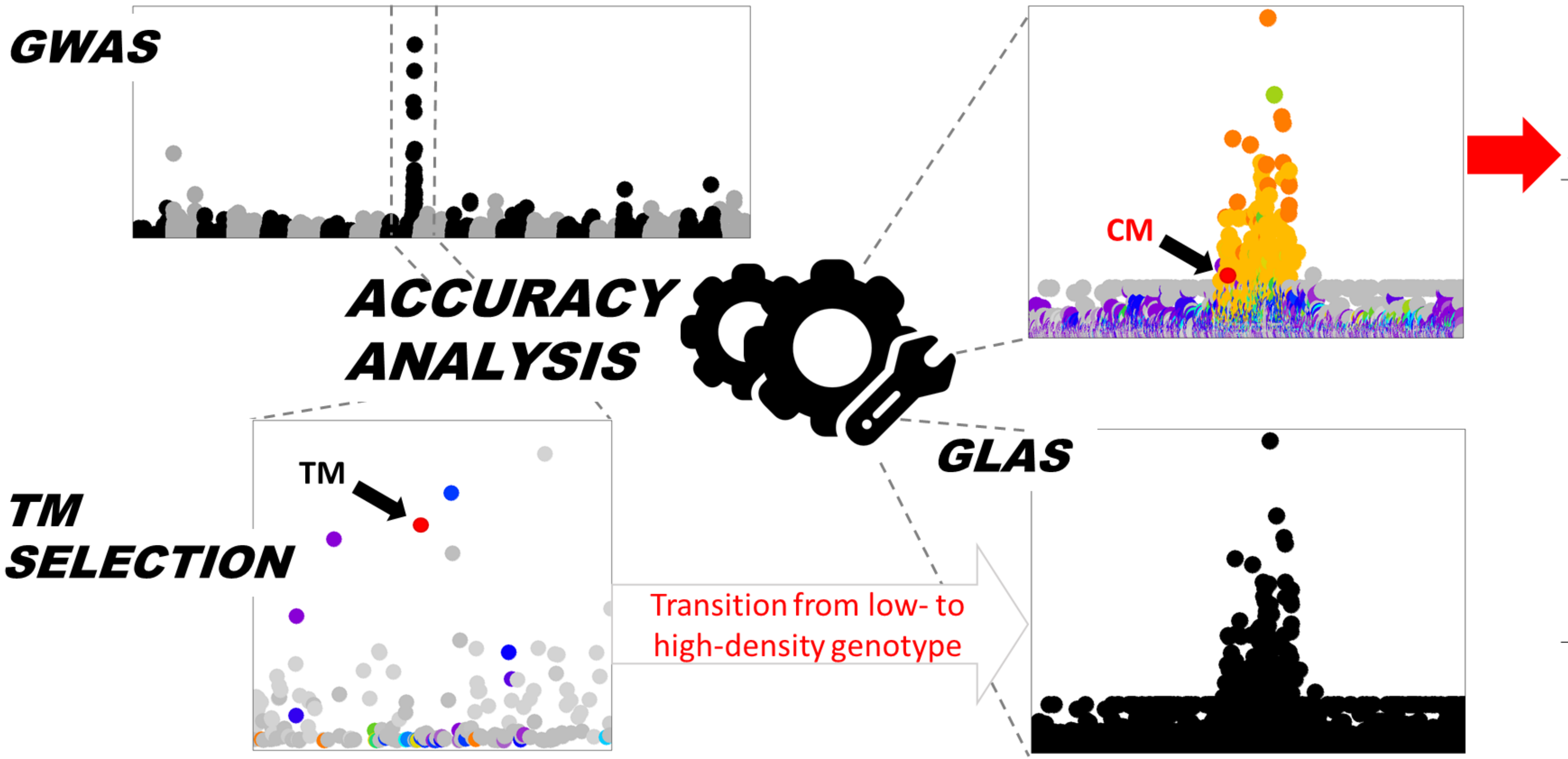
The Allele Catalog Tool: a web-based interactive tool for allele discovery and analysis

Yen On Chan[1,2], Nicholas Dietz[3], Shuai Zeng[4], Juexin Wang[2,4], Sherry Flint-Garcia[5], M. Nancy Salazar-Vidal[3,6], Mária Škrabišová[7], Kristin Bilyeu[5*] and Trupti Joshi[1,2,4,8*]

SYNTHETIC PHENOTYPE TO CAUSATIVE MUTATION STRATEGY (SP2CM)



SP2CM – NAVIGATION SCHEME



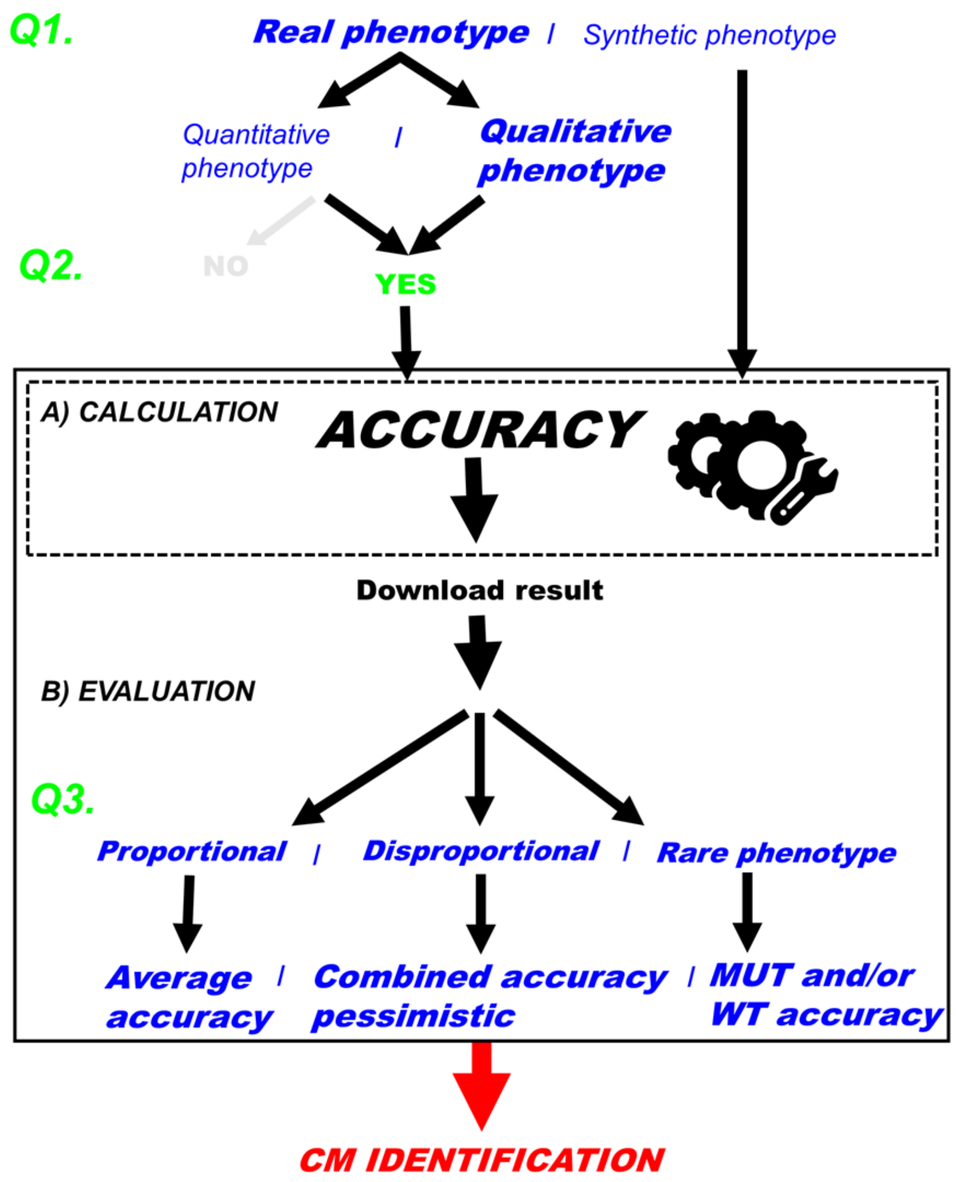
Order	Variant position	Accuracy (%)
1	CM	Highest
2	Candidate 2	
3	Candidate 3	
4	Candidate 4	
5	Candidate 5	
6	Candidate 6	
7	Candidate 7	
8	Candidate 8	
9	Candidate 9	
10	Candidate 10	Lowest

CASE STUDIES

PHENOTYPES

	Proportional	Disproportional	Rare
Quantitative	MG: *E1* Protein: *SWEET39* MG: *E2* Protein: *CCT* MG: *E3* Pod shattering: *Pdh1*		
Qualitative	PUB_C: *T* FLWRC: *W1* PUB_C: *Td* STEMTERM: *Dt1*	Leaf shape: *Ln* Stay green: *D1*	SCN_R: *SNAP18* SCN_R: *SHMT08*

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***

frontiers | Frontiers in Genetics

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

Check for updates

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

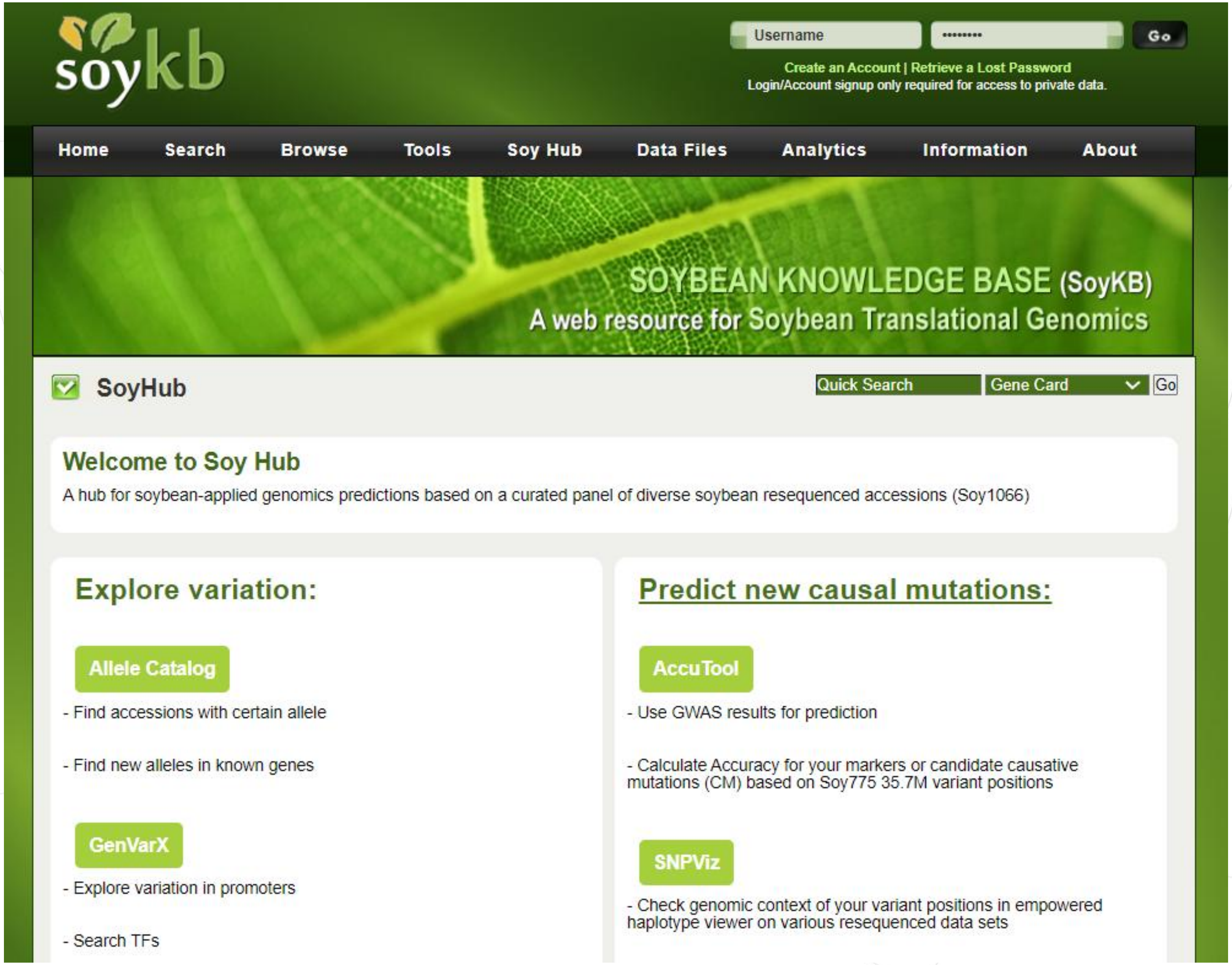
†These authors have contributed equally to this work

Natural and artificial selection of multiple alleles revealed through genomic analyses

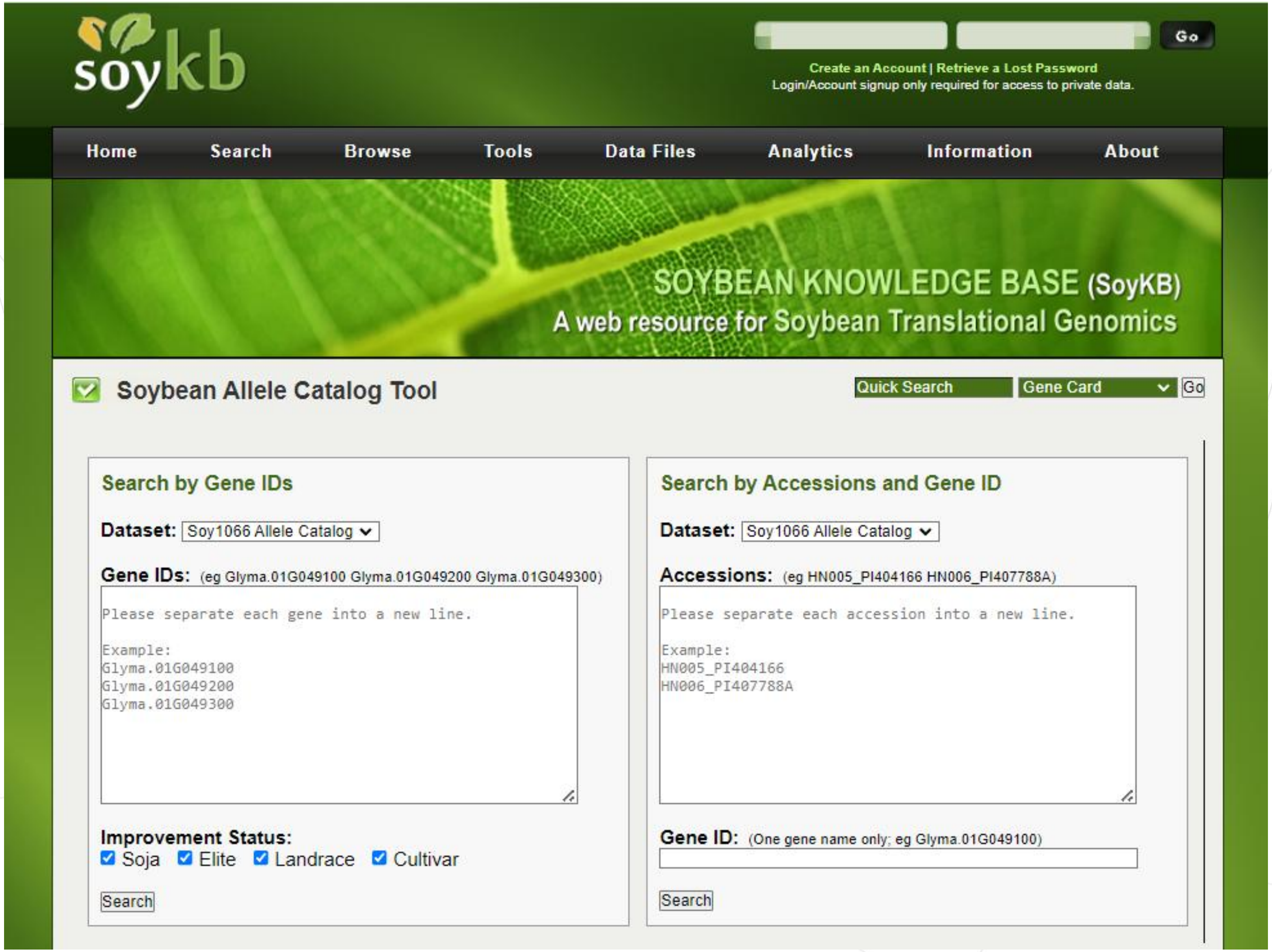
Jana Biová[1†], Ivana Kaňovská[1†], Yen On Chan[2,3], Manish Sridhar Immadi[4], Trupti Joshi[2,3,4,5], Kristin Bilyeu[6*] and Mária Škrabišová[1*]

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

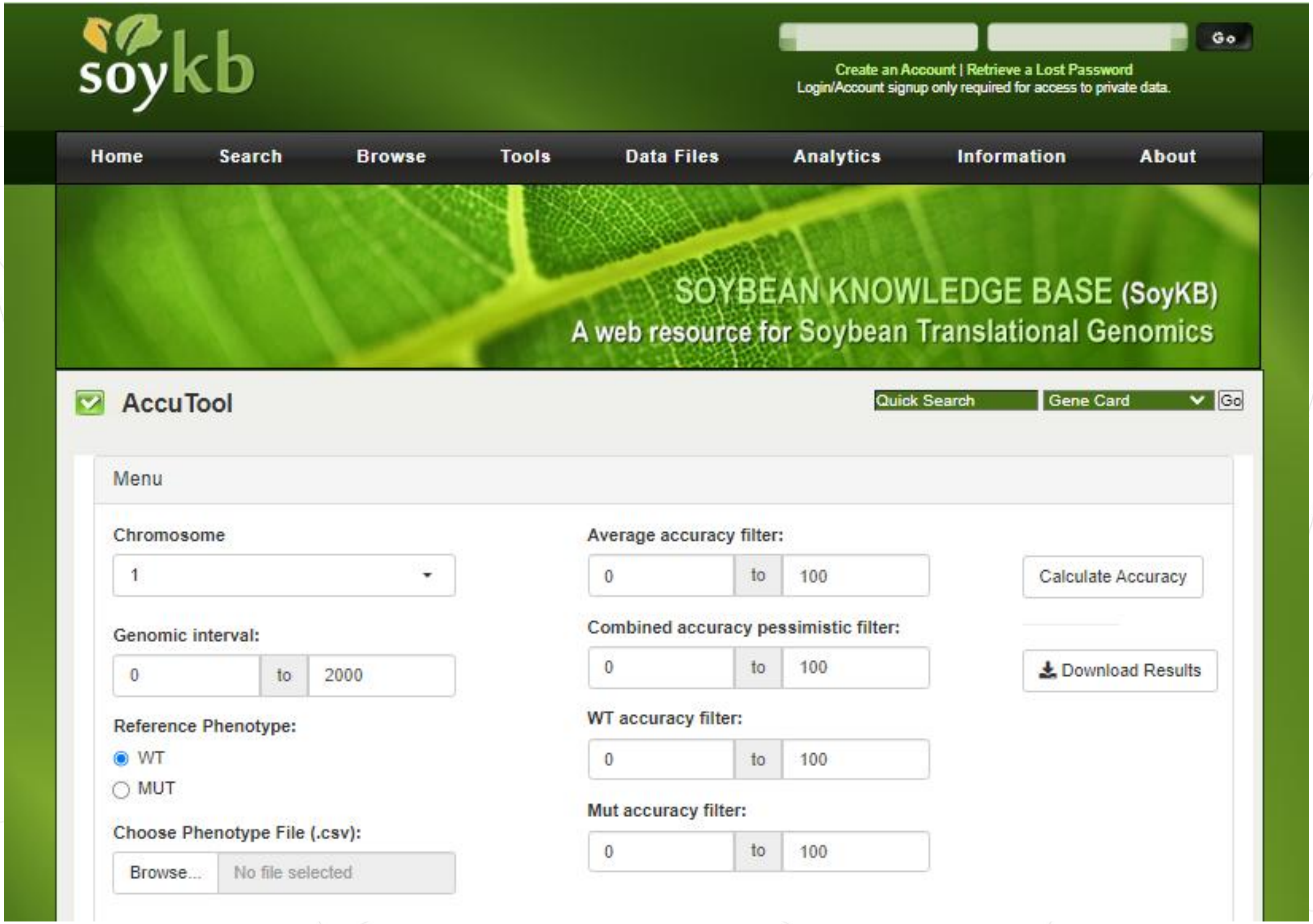
SOYBEAN APPLIED GENOMICS HUB



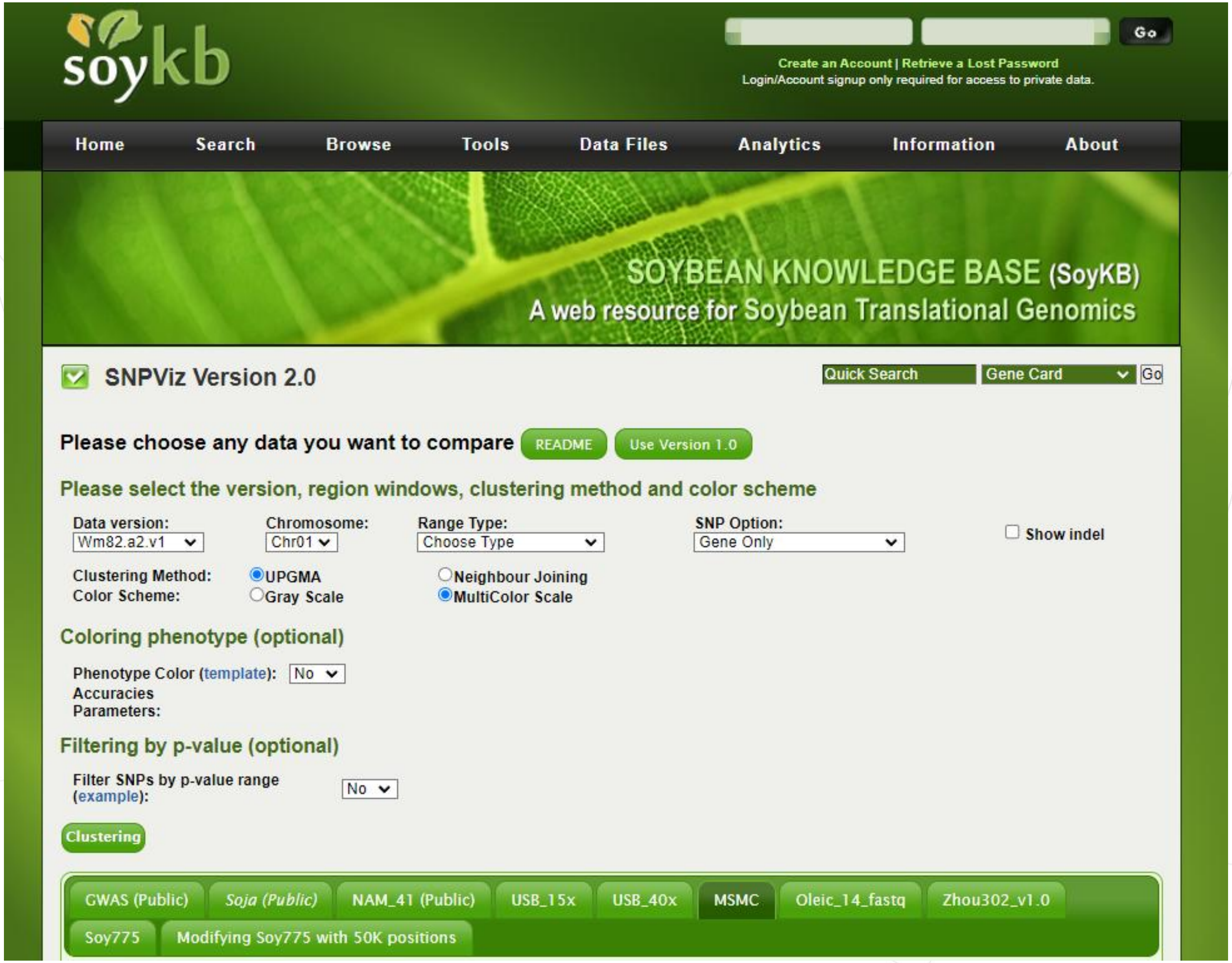
SOYBEAN ALLELE CATALOG



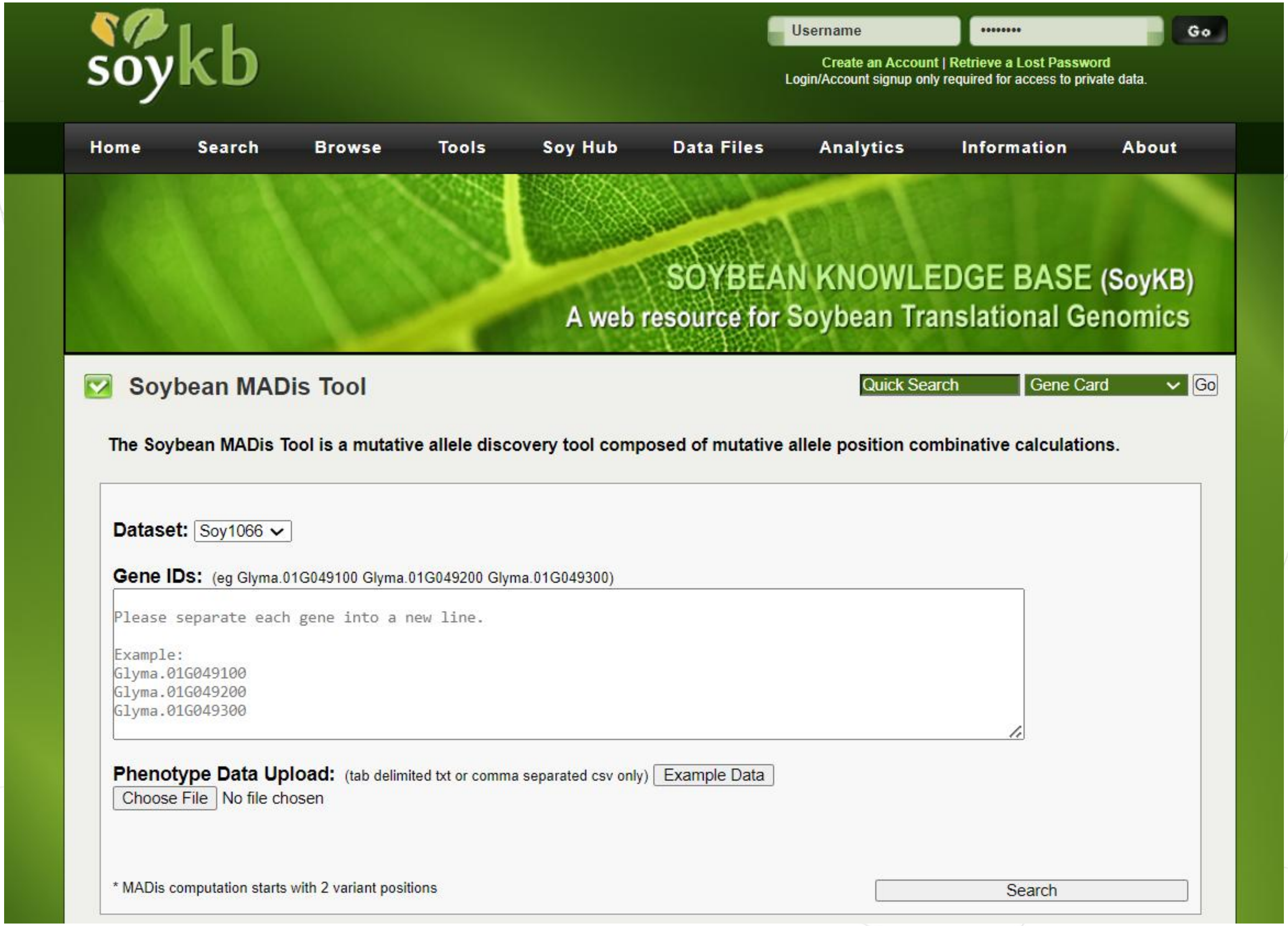
ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS



SNPVIZ 2.0: EMPOWERED HAPLOTYPE VIEWER



MADIS: MUTATIVE ALLELE DISCOVERY TOOL



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

Legume Generation

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.

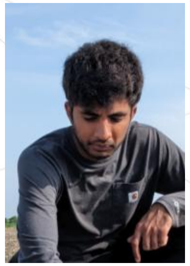
Ivana Kaňovská, M.Sc.

Funding

Applied Genomics

Dr. Kristin Bilyeu

Dr. Nicholas Dietz

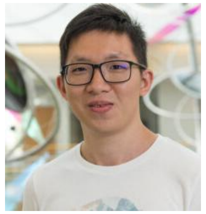


Anser Mahmood

Bioinformatics

Dr. Trupti Joshi

Dr. Shuai Zeng



Yen On Chan

Manish Sridhar Immadi

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!