APPROACHES AND STRATEGIES IN APPLIED GENOMICS FOR SOYBEAN

(AND OTHER CROPS AND SPECIES)

Mária Škrabišová, Ph.D.

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

Plant and Animal Genome Conference PAG32, San Diego, CA, USA
SoyHUB: A Platform For Soybean Applied Genomics workshop, January 13, 2025

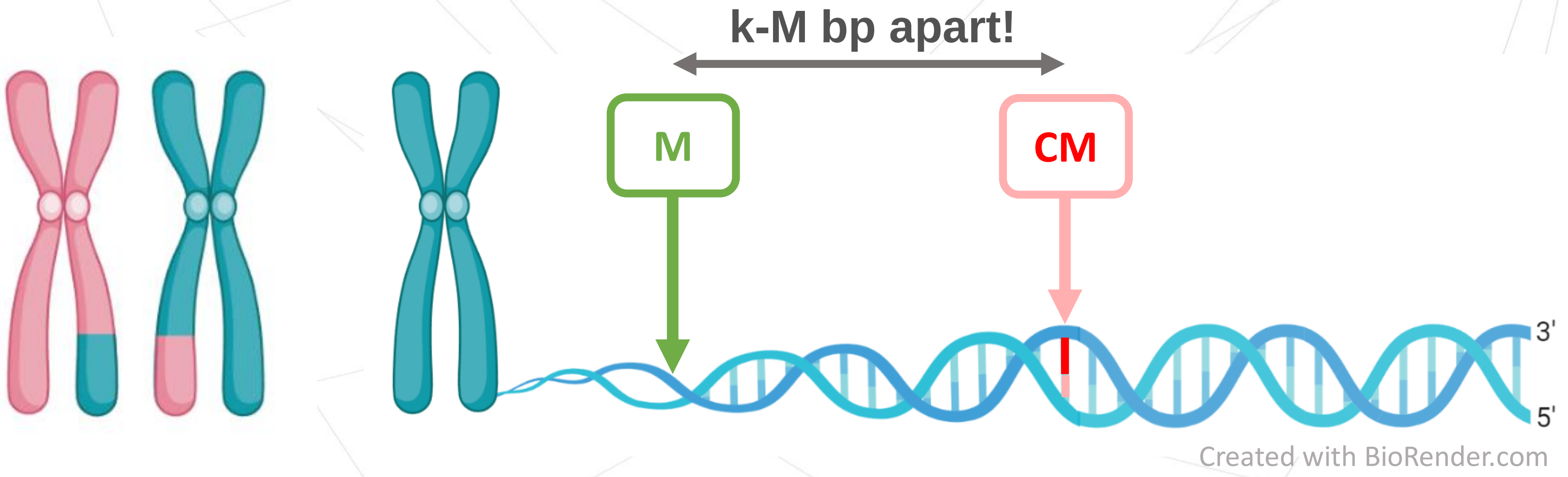
BREEDING IS NOT PRECISE ENOUGH YET!

Marker-assisted breeding

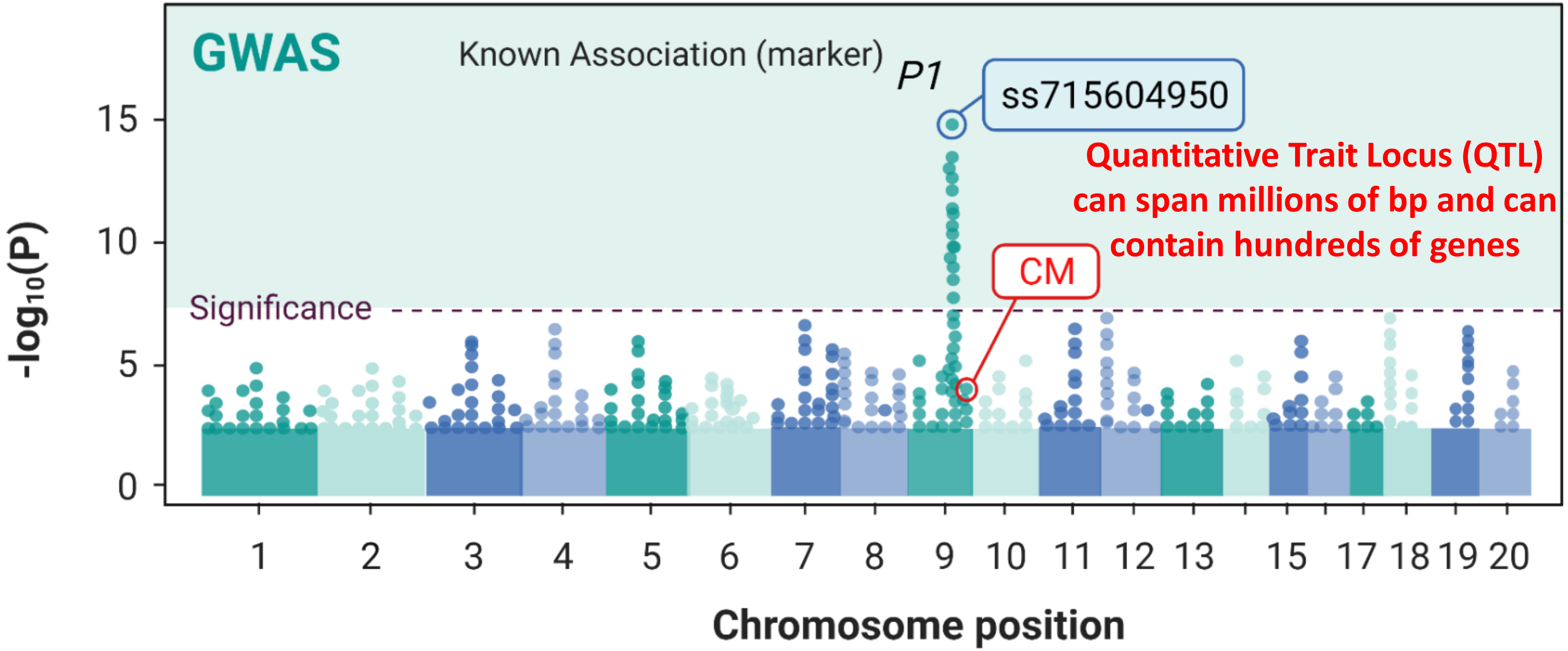
- **A marker (M) can be located far** from a gene that underlies a phenotype of interest

How to proceed? Identification of CAUSAL GENES with causative mutations (CMs)

- **Era of sequencing**
- **Association methods**



GWAS FOR MORE PRECISE BREEDING: LIMITED.

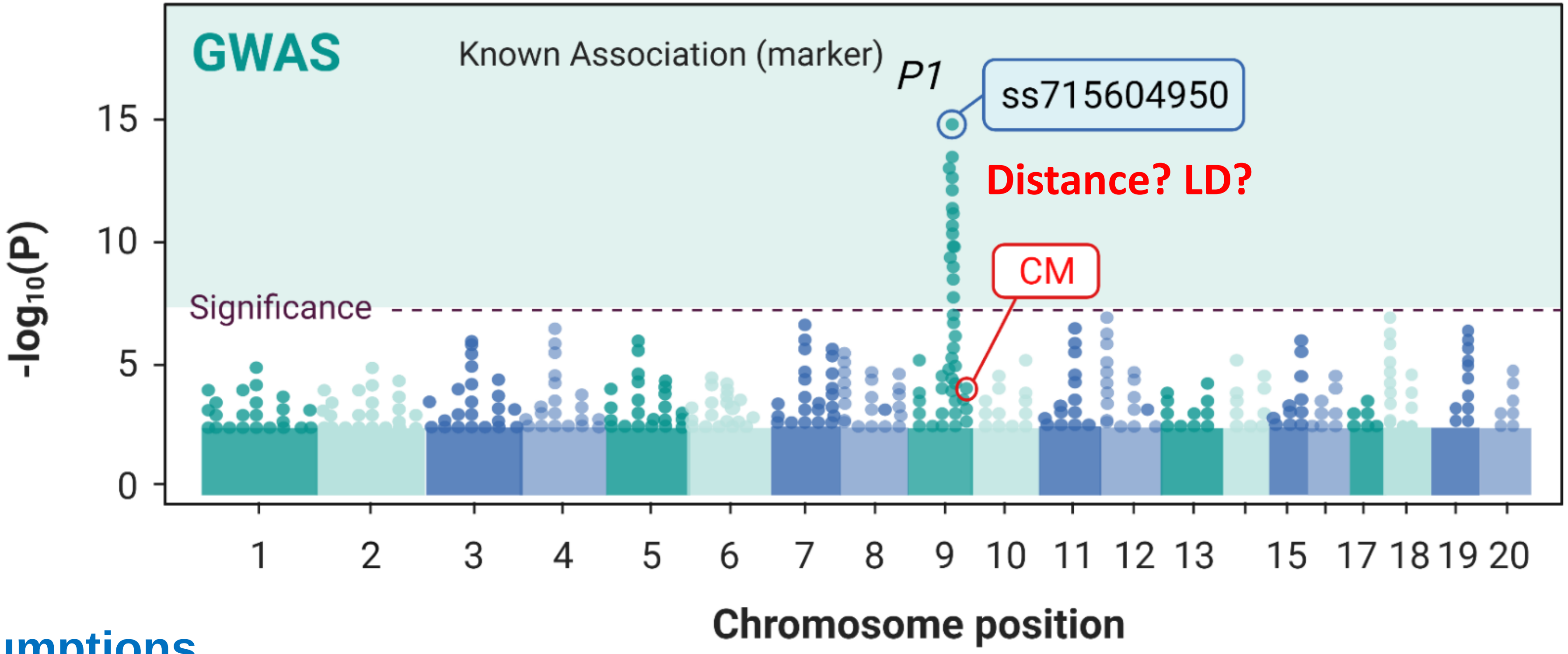


Quantitative Trait Locus (QTL) can span millions of bp and can contain hundreds of genes

Varying GWAS power caused by:

- Genotype quality: density, Indels, population structure and size
- Phenotype issues: precision, data type nature (distribution of values, frequency, etc.)
- Fitting model applied
- Other factors

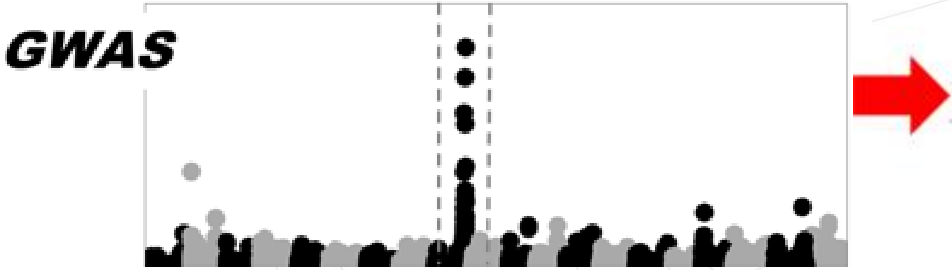
GWAS FOR MORE PRECISE BREEDING



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

HUNTING GENES FOR MORE PRECISE CROP BREEDING



Sorting the candidates

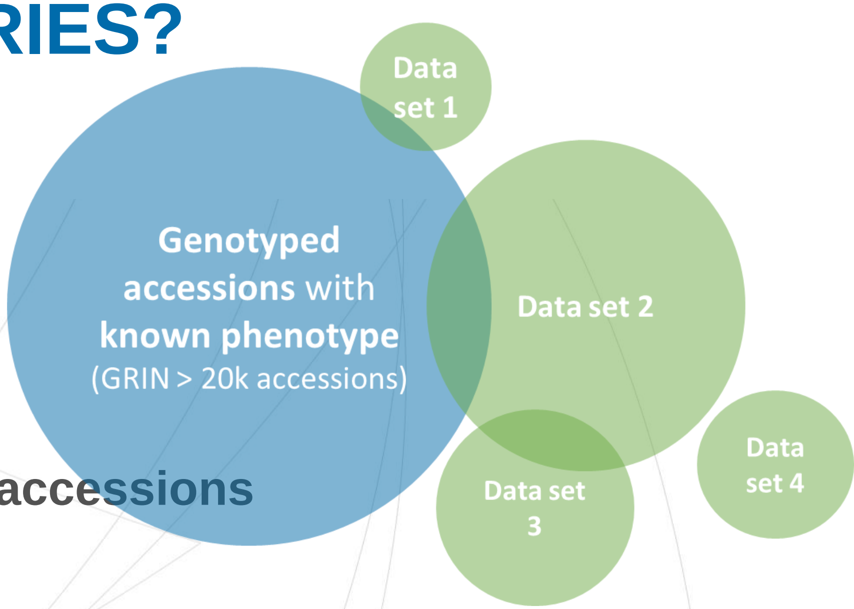
Order	Variant position	
1	CM	Highest
2	Candidate 2	
3	Candidate 3	
4	Candidate 4	
5	Candidate 5	
6	Candidate 6	
7	Candidate 7	
8	Candidate 8	
9	Candidate 9	
10	Candidate 10	Lowest

Guessing genes…
…by proximity
…by ontologies

1. How can we suppress false positives?
2. How can we derive true positives from false negatives?

LET’S CLONE GENES!

HOW TO IMPROVE GWAS-DRIVEN DISCOVERIES?



By adding power to GWAS!

What is needed? Three novel concepts:

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

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

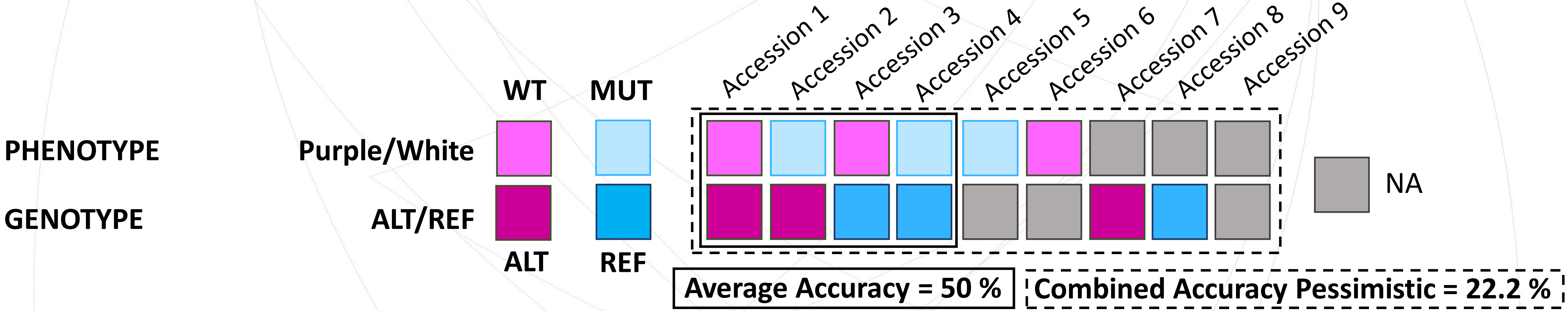
A novel Synthetic phenotype association study approach reveals the landscape of association for genomic variants and phenotypes

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

ACCURACY

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



Average Accuracy = 50 %

Combined Accuracy Pessimistic = 22.2 %

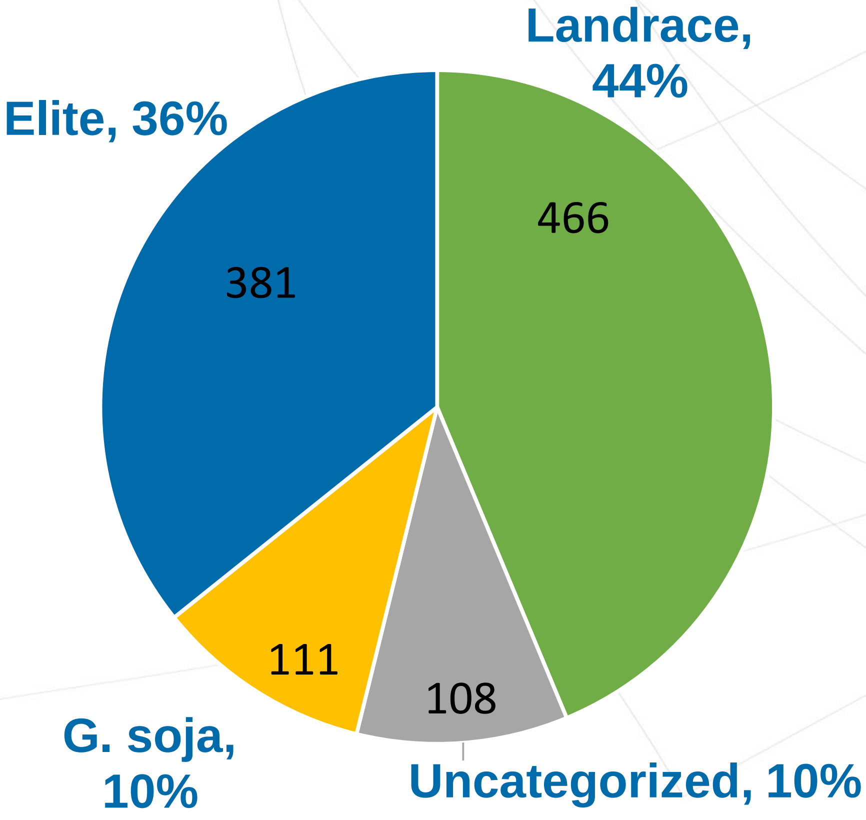
WT_Accuracy = 50 %

MUT_Accuracy = 50 %

REF = Williams 82 – white flower color – MUT phenotype

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** (Škrabišová et al., 2022)
- **SnakyVC pipeline: - Soy1066 38.3 M** (Chan et al., 2023) > **Soy2939 44.3 M** > **Soy4000+**




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

BMC Genomics

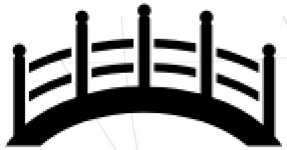
SOFTWARE **Open Access**

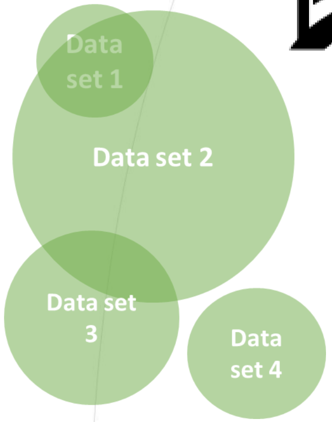
Check for updates

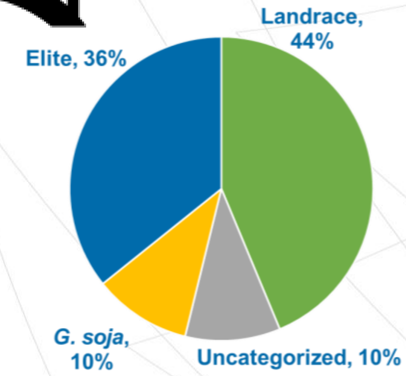
The Allele Catalog Tool: a web-based interactive tool for allele discovery and analysis

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*]

FILLING THE PHENOTYPE/GENOTYPE GAP







Real phenotype (Observed) **Flower color**		**Synthetic phenotype** CM in Glyma.13g072100 G>Indel (Zabala & Vodkin 2007) **Chr13:17,316,758 genotype**	
WT	Purple	**REF**	G
MUT	White	**ALT**	Indel

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

- Sorting the variant positions based on Accuracy calculated for the Soybean diversity panel
- 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)

New perspectives of post-GWAS analyses:
From markers to causal genes for more precise crop breeding

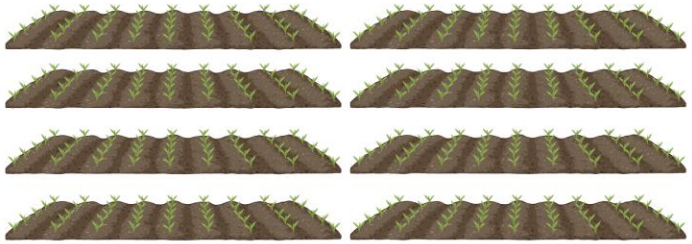
Kanovska et al., 2024

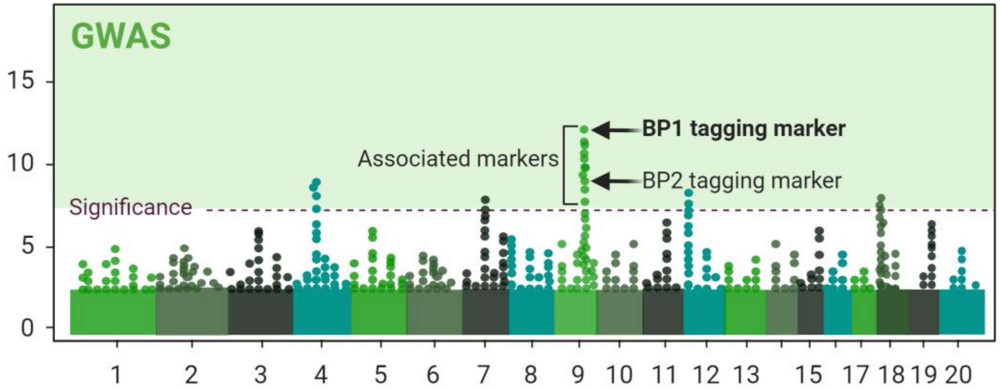
Breeding programs

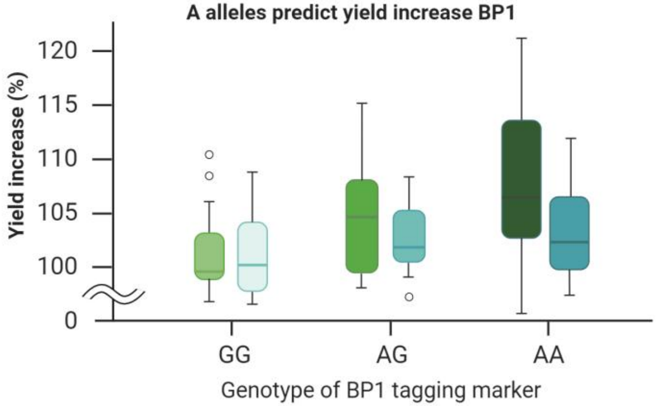
Marker identification

Marker efficiency

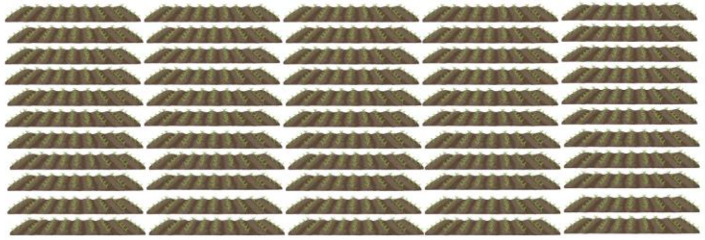
BP1

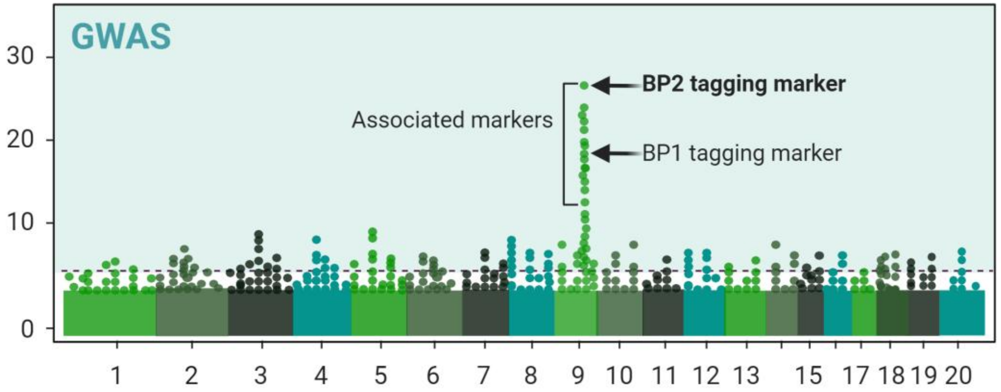


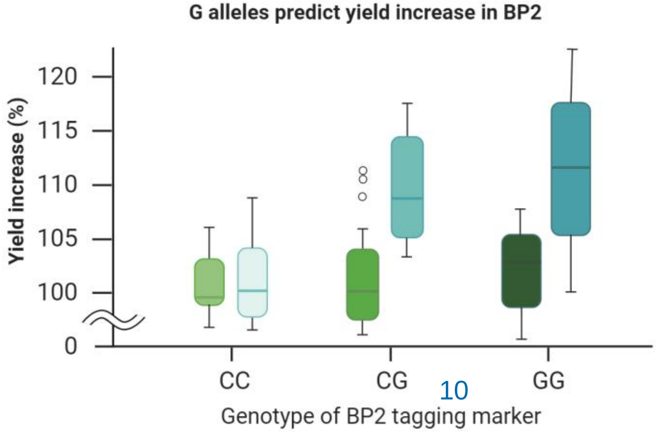




BP2

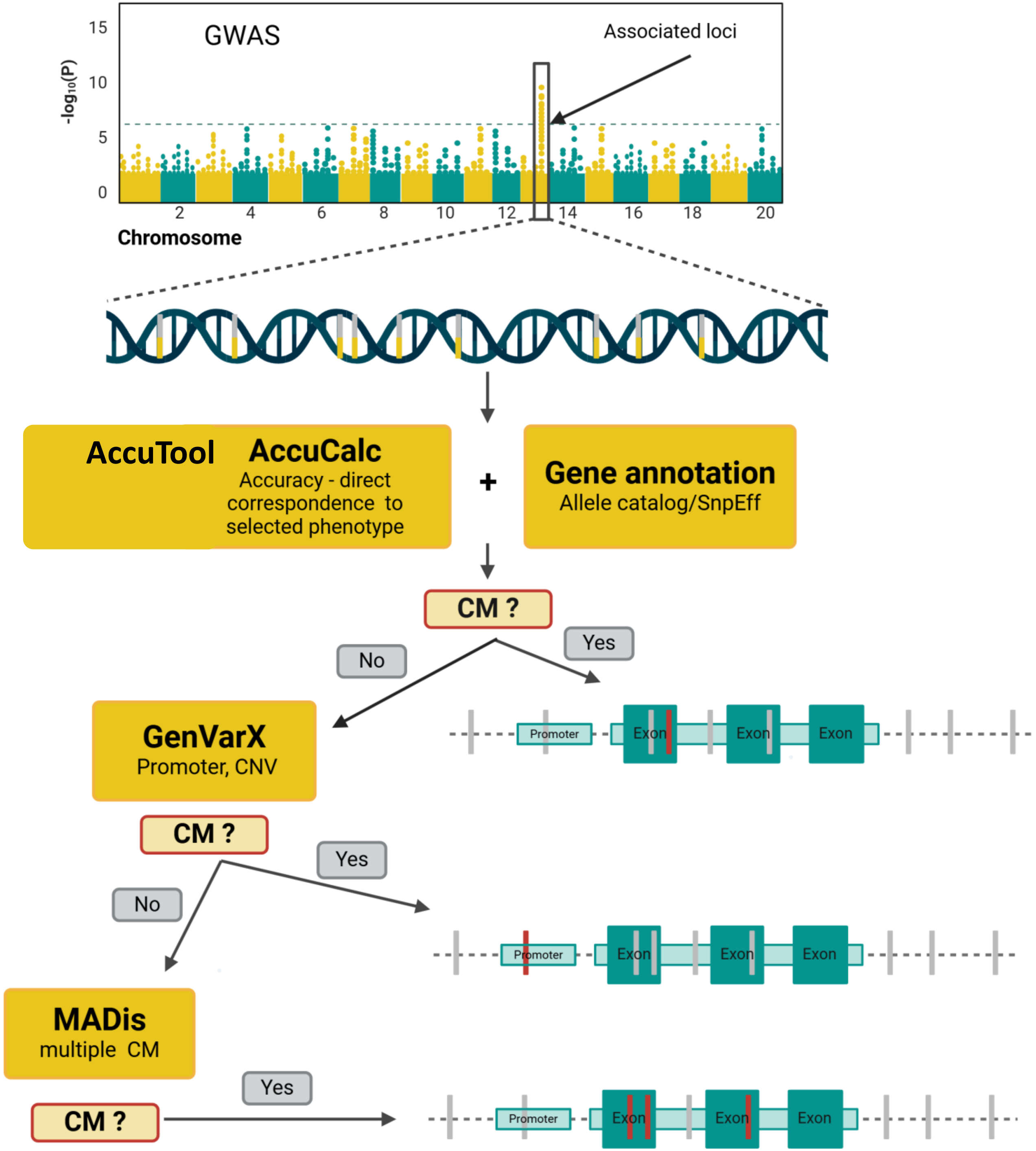






SOYHUB

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



AccuTool **AccuCalc**
Accuracy - direct correspondence to selected phenotype

+

Gene annotation
Allele catalog/SnpEff

CM ?

No

Yes

GenVarX
Promoter, CNV

CM ?

No

Yes

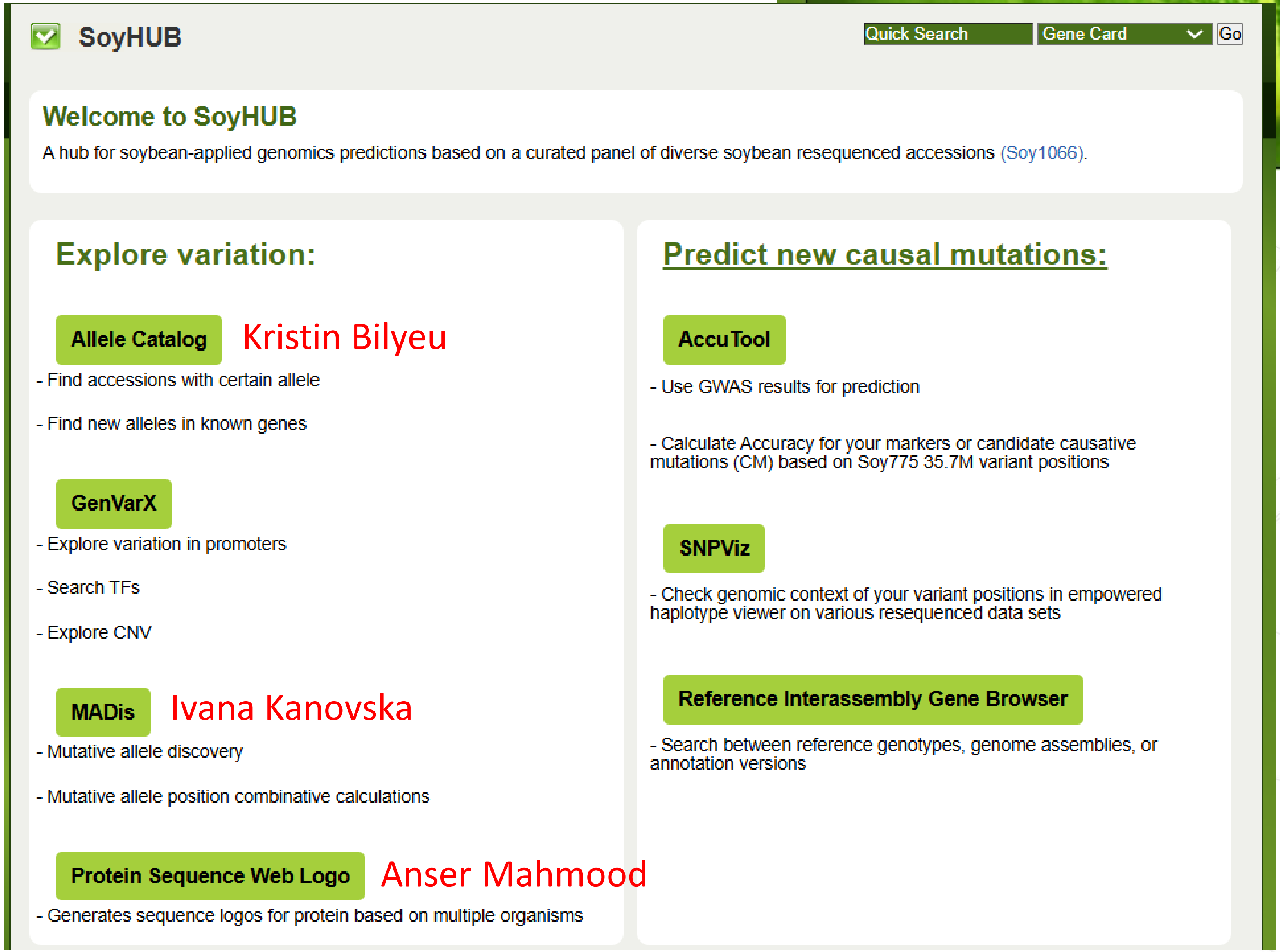
MADis
multiple CM

CM ?

Yes

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

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



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 Kristin Bilyeu

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

GenVarX

- Explore variation in promoters
- Search TFs
- Explore CNV

MADis Ivana Kanovska

- Mutative allele discovery
- Mutative allele position combinative calculations

Protein Sequence Web Logo Anser Mahmood

- 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

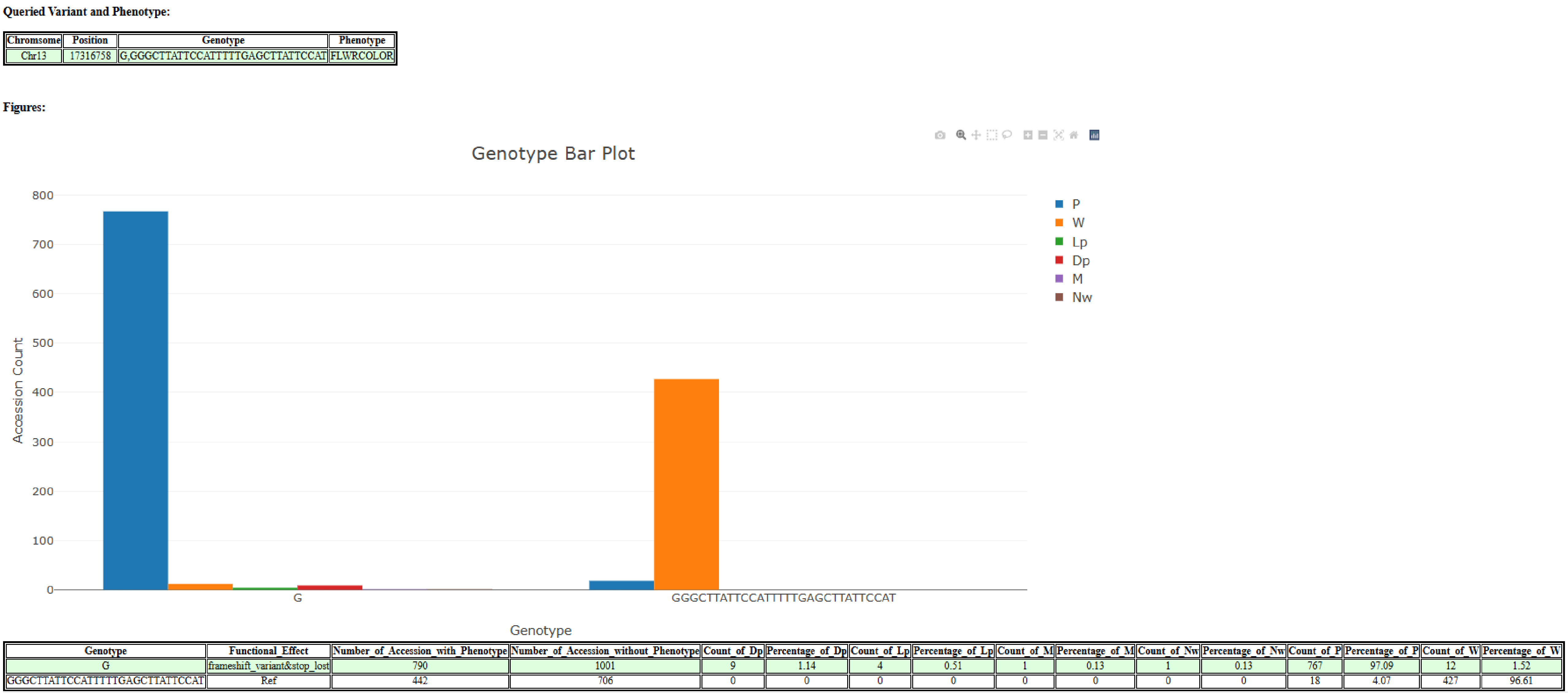
- Check genomic context of your variant positions in empowered haplotype viewer on various resequenced data sets

Reference Interassembly Gene Browser

- Search between reference genotypes, genome assemblies, or annotation versions

- **A curated panel of soybean resequenced accessions (Soy2939)**
- **Phenotypes (GRIN)**
- **Tools**

LINKING GENOTYPES WITH QUALITATIVE PHENOTYPES: *W1* FLOWER COLOR



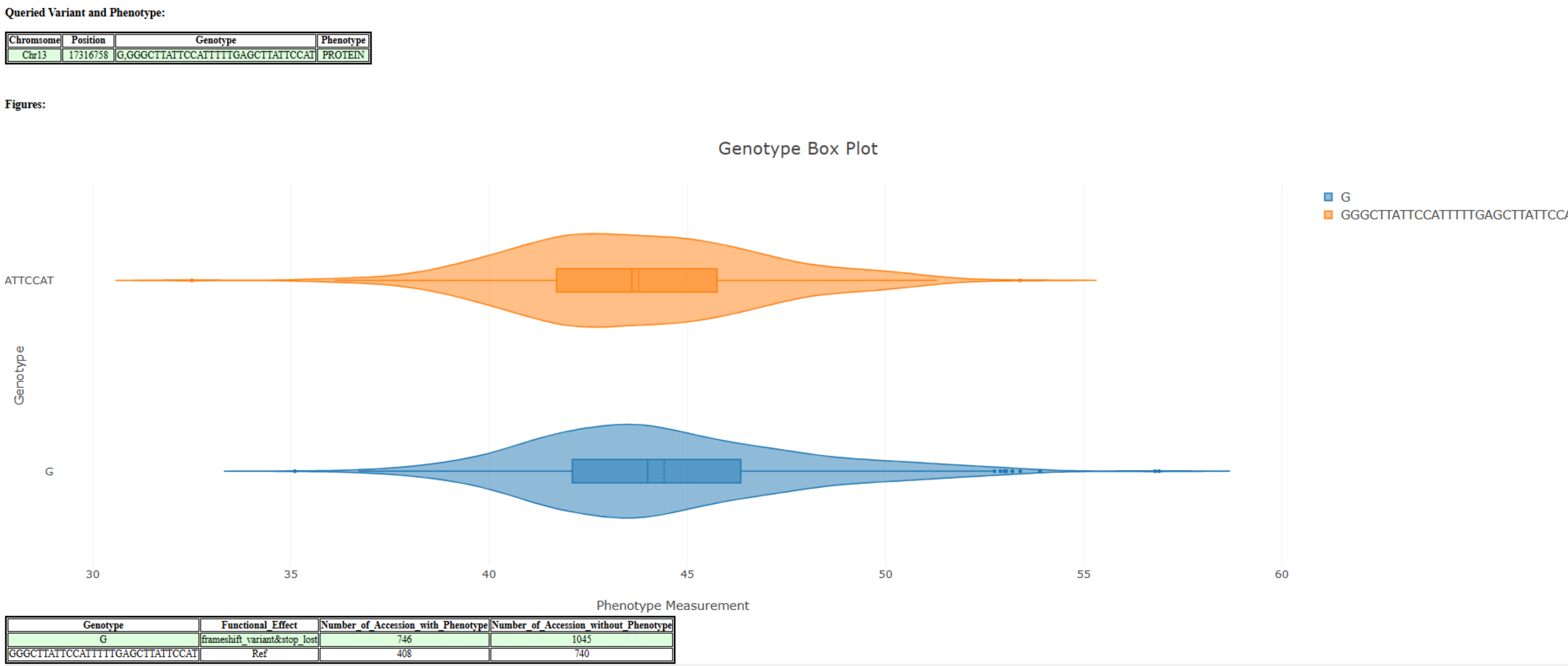
Queried Variant and Phenotype:

Chromsome	Position	Genotype	Phenotype
Chr13	17316758	G,GGGCTTATTCCATTTTTGAGCTTATTCCAT	FLWRCOLOR

Figures:

Genotype	Functional_Effect	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype	Count_of_Dp	Percentage_of_Dp	Count_of_Lp	Percentage_of_Lp	Count_of_M	Percentage_of_M	Count_of_Nw	Percentage_of_Nw	Count_of_P	Percentage_of_P	Count_of_W	Percentage_of_W
G	frameshift_variant&stop_lost	790	1001	9	1.14	4	0.51	1	0.13	1	0.13	767	97.09	12	1.52
GGGCTTATTCCATTTTTGAGCTTATTCCAT	Ref	442	706	0	0	0	0	0	0	0	0	18	4.07	427	96.61

LINKING GENOTYPES WITH QUANTITATIVE PHENOTYPES: PROTEIN



Queried Variant and Phenotype:

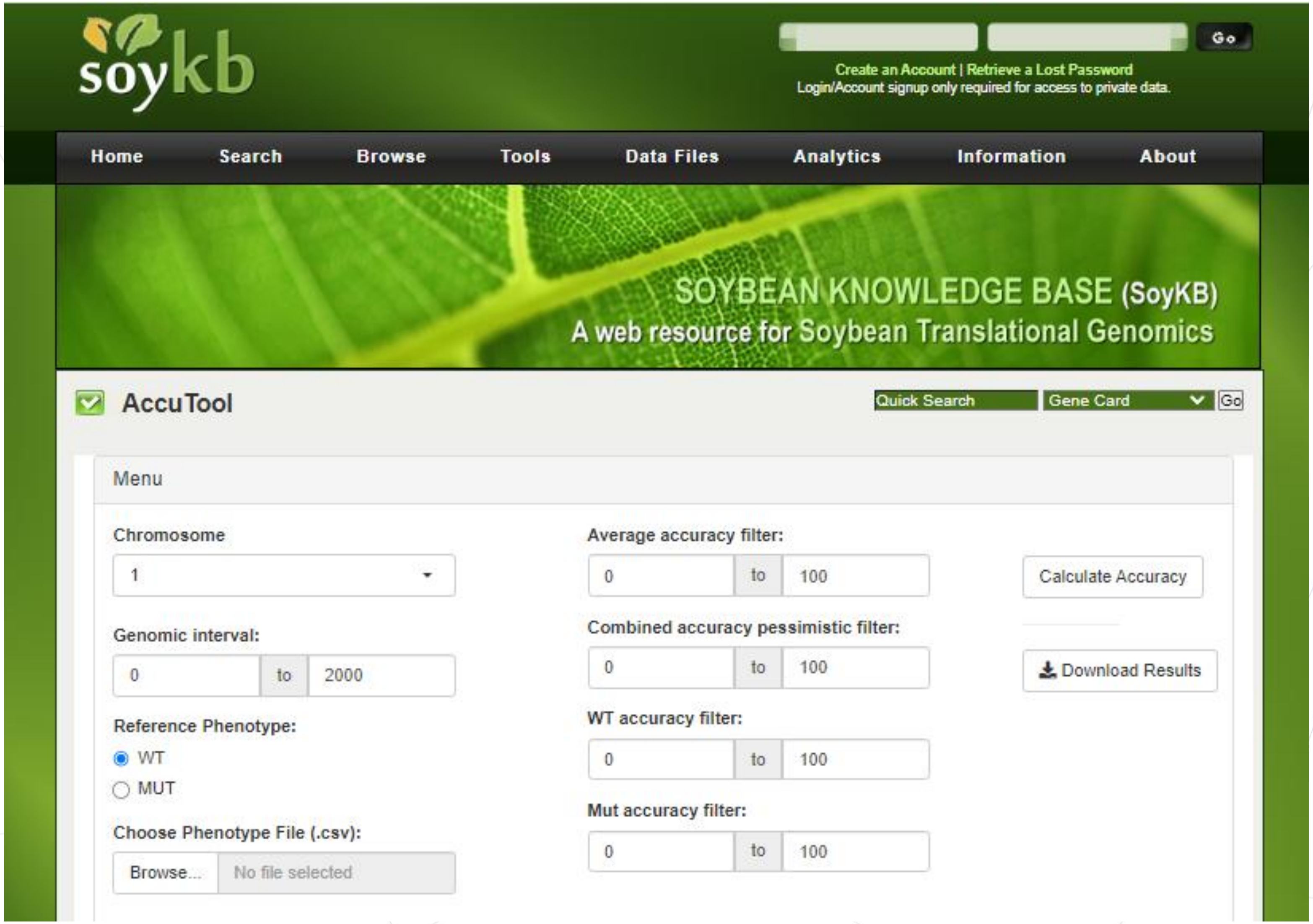
Chromsome	Position	Genotype	Phenotype
Chr13	17316758	G,GGGCTTATTCCATTTTTGAGCTTATTCCAT	PROTEIN

Figures:

Genotype	Functional_Effect	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype
G	frameshift_variant&stop_lost	746	1045
GGGCTTATTCCATTTTTGAGCTTATTCCAT	Ref	408	740

ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS

Dr. Nicholas Dietz



SOYBEAN KNOWLEDGE BASE (SoyKB)

A web resource for Soybean Translational Genomics

AccuTool

Menu

Chromosome: 1

Genomic interval: 0 to 2000

Reference Phenotype:
- ● WT
- ○ MUT

Choose Phenotype File (.csv): No file selected

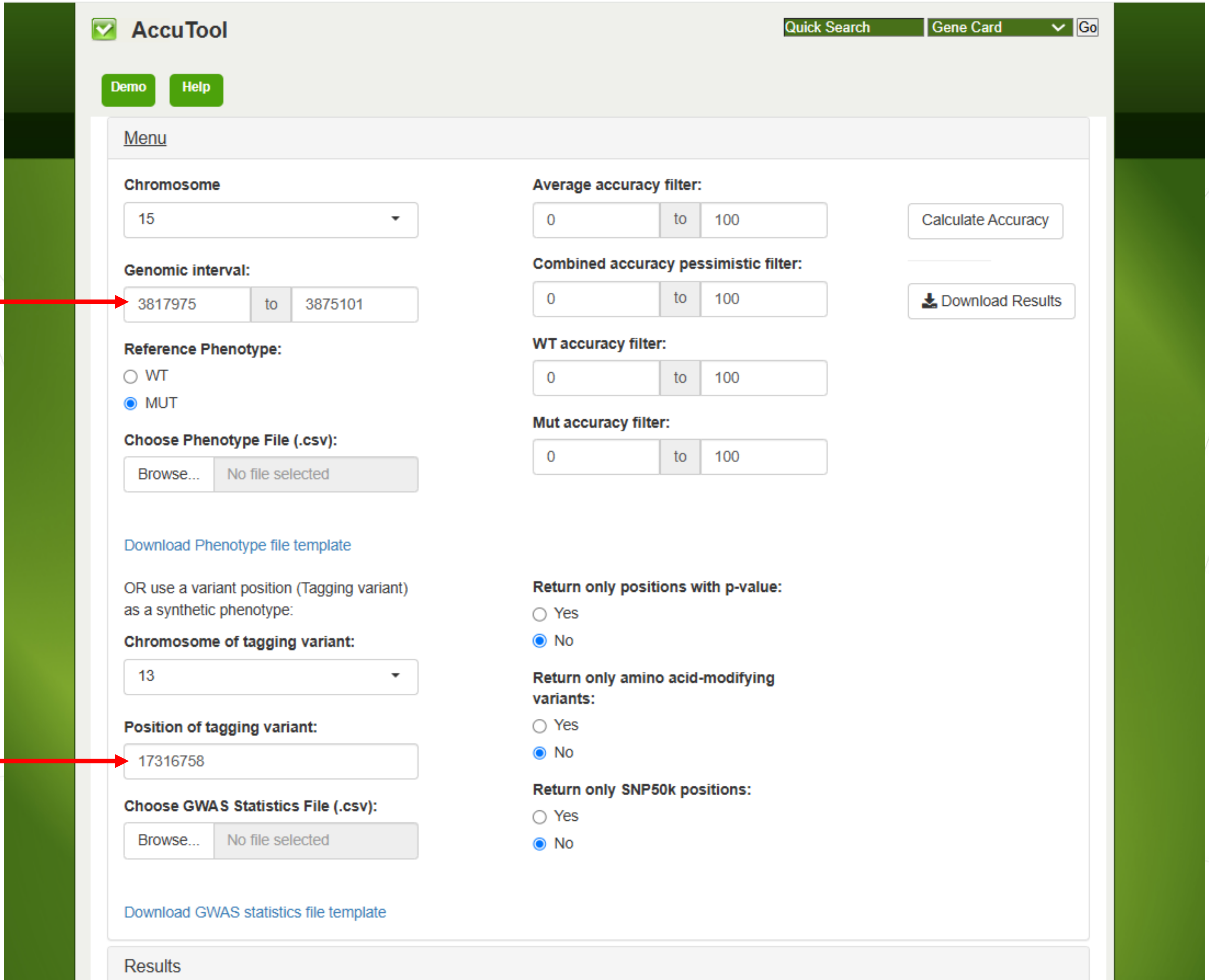
Average accuracy filter: 0 to 100

Combined accuracy pessimistic filter: 0 to 100

WT accuracy filter: 0 to 100

Mut accuracy filter: 0 to 100

ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS



Chromosome: 15

Genomic interval: 3817975 to 3875101

Reference Phenotype:
- ○ WT
- ● MUT

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

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

Chromosome of tagging variant: 13

Position of tagging variant: 17316758

Choose GWAS Statistics 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

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

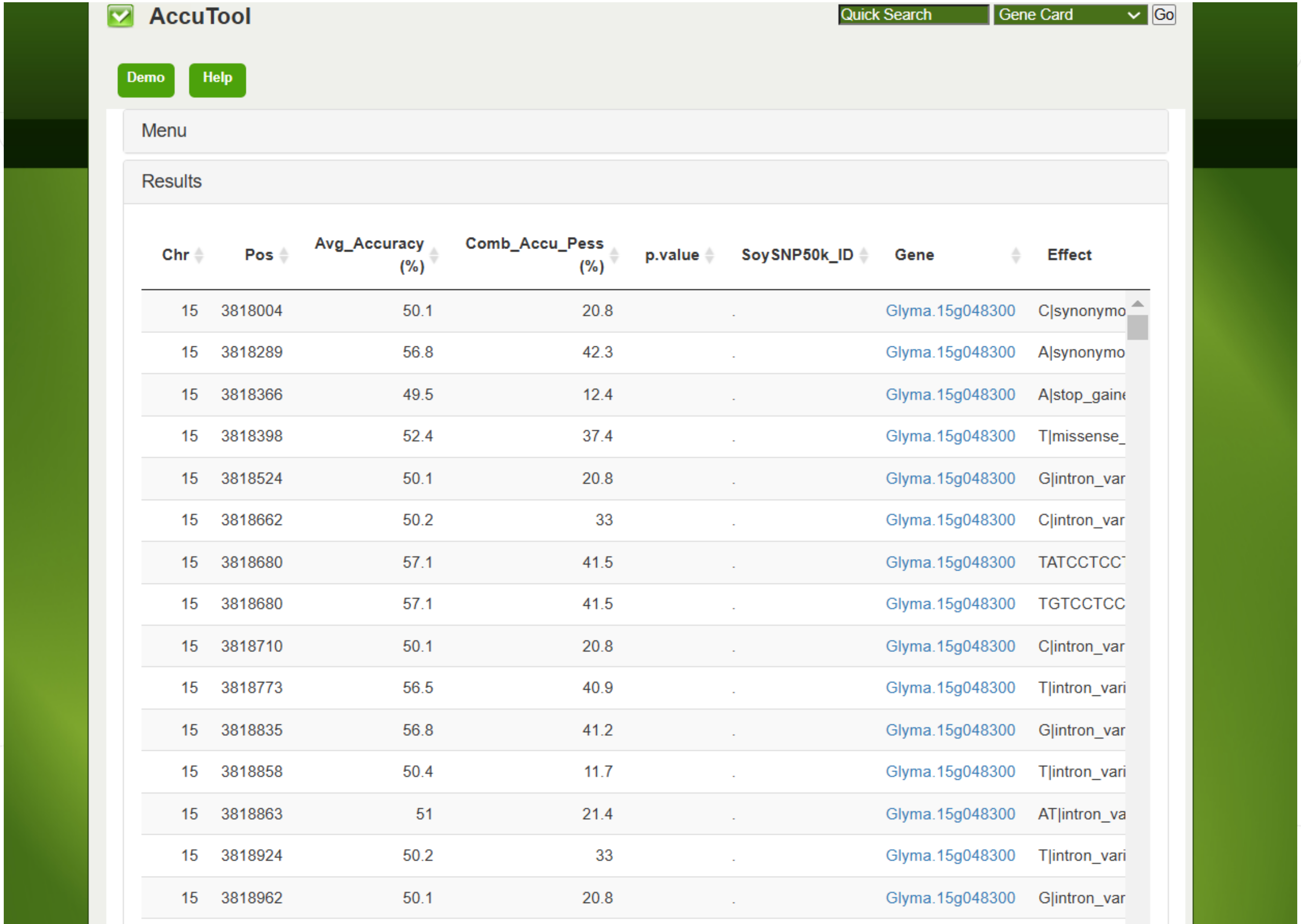
Return only amino acid-modifying variants:
- ○ Yes
- ● No

Return only SNP50k positions:
- ○ Yes
- ● No

Protein QTL

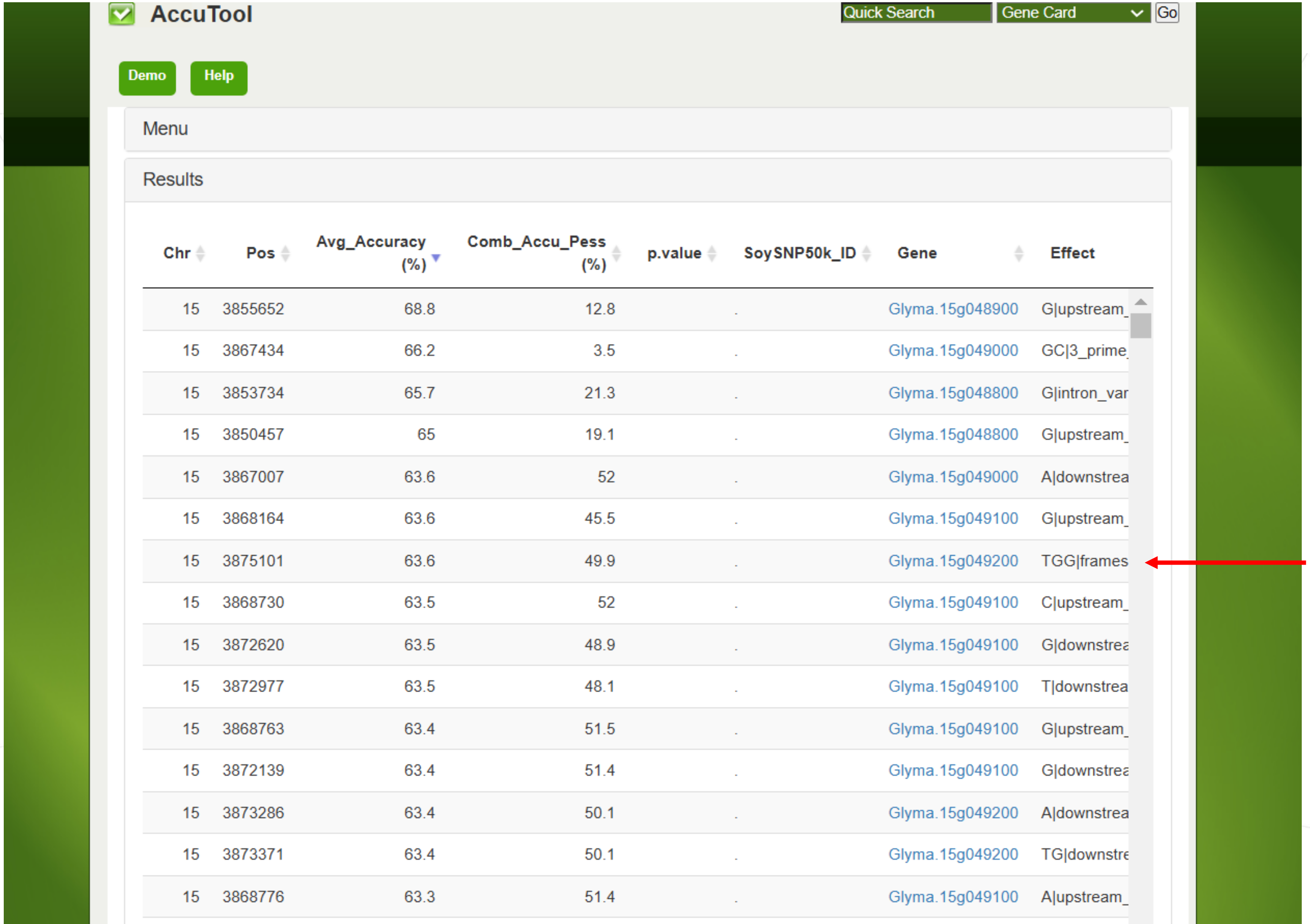
W1 FLOWER COLOR
Zabala & Vodkin 2007

ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS



Chr	Pos	Avg_Accuracy (%)	Comb_Accu_Pess (%)	p.value	SoySNP50k_ID	Gene	Effect
15	3818004	50.1	20.8		.	Glyma.15g048300	C|synonymo
15	3818289	56.8	42.3		.	Glyma.15g048300	A|synonymo
15	3818366	49.5	12.4		.	Glyma.15g048300	A|stop_gaine
15	3818398	52.4	37.4		.	Glyma.15g048300	T|missense_
15	3818524	50.1	20.8		.	Glyma.15g048300	G|intron_var
15	3818662	50.2	33		.	Glyma.15g048300	C|intron_var
15	3818680	57.1	41.5		.	Glyma.15g048300	TATCCTCC
15	3818680	57.1	41.5		.	Glyma.15g048300	TGTCCTCC
15	3818710	50.1	20.8		.	Glyma.15g048300	C|intron_var
15	3818773	56.5	40.9		.	Glyma.15g048300	T|intron_vari
15	3818835	56.8	41.2		.	Glyma.15g048300	G|intron_var
15	3818858	50.4	11.7		.	Glyma.15g048300	T|intron_vari
15	3818863	51	21.4		.	Glyma.15g048300	AT|intron_va
15	3818924	50.2	33		.	Glyma.15g048300	T|intron_vari
15	3818962	50.1	20.8		.	Glyma.15g048300	G|intron_var

ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS

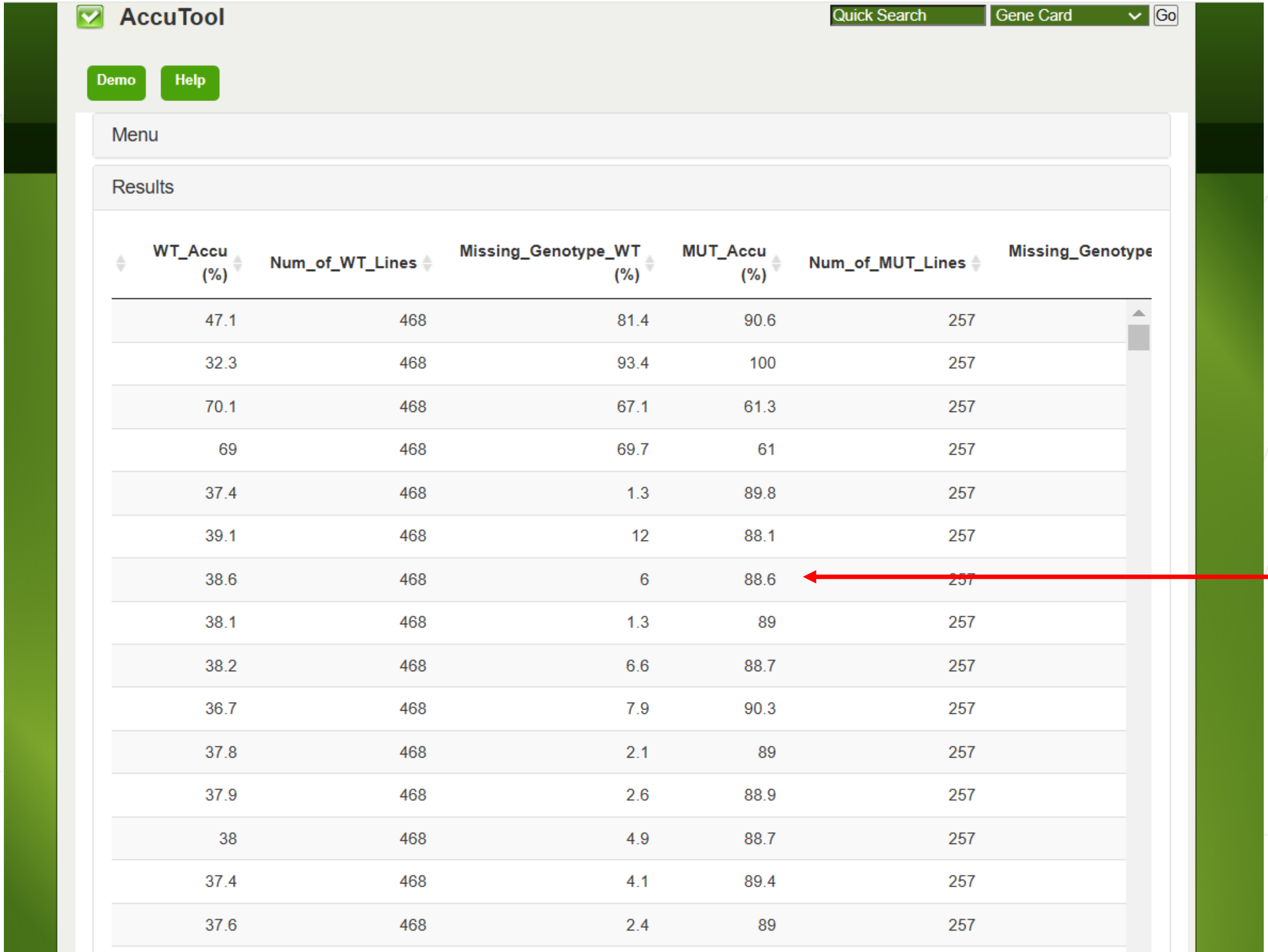


Chr	Pos	Avg_Accuracy (%)	Comb_Accu_Pess (%)	p.value	SoySNP50k_ID	Gene	Effect
15	3855652	68.8	12.8		.	Glyma.15g048900	G|upstream_
15	3867434	66.2	3.5		.	Glyma.15g049000	GC|3_prime_
15	3853734	65.7	21.3		.	Glyma.15g048800	G|intron_var
15	3850457	65	19.1		.	Glyma.15g048800	G|upstream_
15	3867007	63.6	52		.	Glyma.15g049000	A|downstrea
15	3868164	63.6	45.5		.	Glyma.15g049100	G|upstream_
15	3875101	63.6	49.9		.	Glyma.15g049200	TGG|frames
15	3868730	63.5	52		.	Glyma.15g049100	C|upstream_
15	3872620	63.5	48.9		.	Glyma.15g049100	G|downstrea
15	3872977	63.5	48.1		.	Glyma.15g049100	T|downstrea
15	3868763	63.4	51.5		.	Glyma.15g049100	G|upstream_
15	3872139	63.4	51.4		.	Glyma.15g049100	G|downstrea
15	3873286	63.4	50.1		.	Glyma.15g049200	A|downstrea
15	3873371	63.4	50.1		.	Glyma.15g049200	TG|downstre
15	3868776	63.3	51.4		.	Glyma.15g049100	A|upstream_

SWEET39

Miao et al., 2020;
Wang et al., 2020

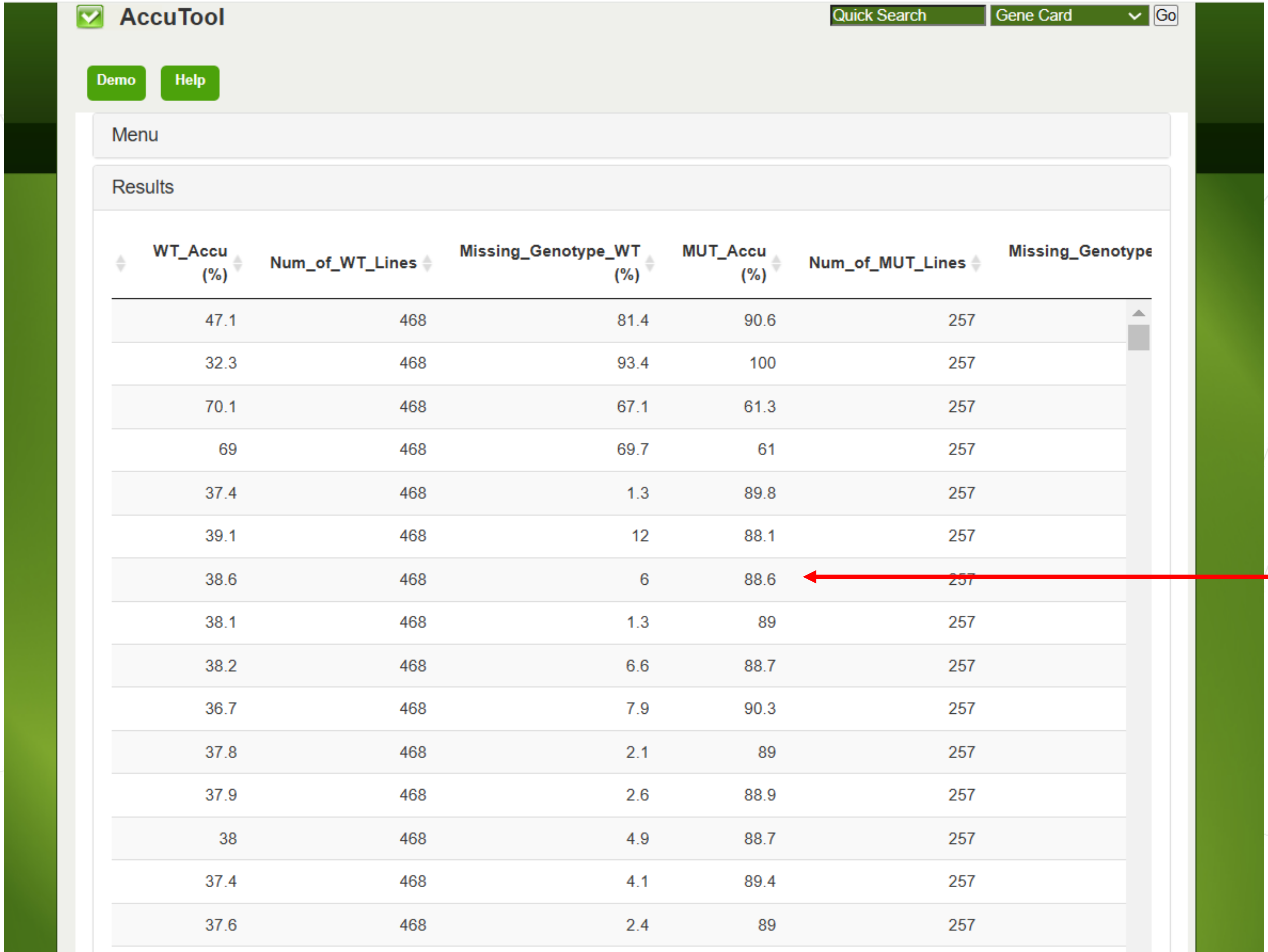
ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS



WT_Accu (%)	Num_of_WT_Lines	Missing_Genotype_WT (%)	MUT_Accu (%)	Num_of_MUT_Lines	Missing_Genotype
47.1	468	81.4	90.6	257	
32.3	468	93.4	100	257	
70.1	468	67.1	61.3	257	
69	468	69.7	61	257	
37.4	468	1.3	89.8	257	
39.1	468	12	88.1	257	
38.6	468	6	88.6	257	
38.1	468	1.3	89	257	
38.2	468	6.6	88.7	257	
36.7	468	7.9	90.3	257	
37.8	468	2.1	89	257	
37.9	468	2.6	88.9	257	
38	468	4.9	88.7	257	
37.4	468	4.1	89.4	257	
37.6	468	2.4	89	257	

Glyma.15g049200
SWEET39 T/TGG

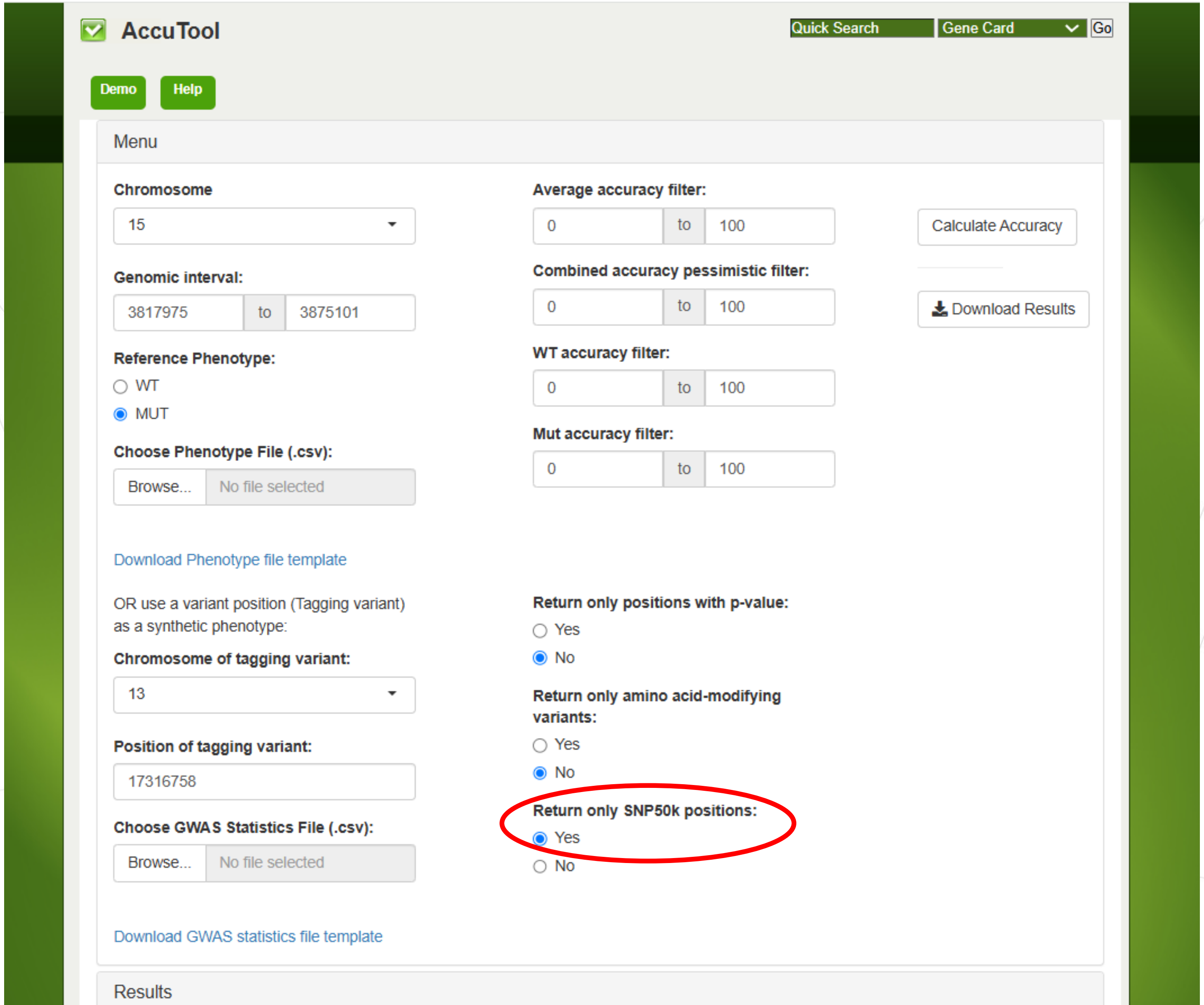
ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS



WT_Accu (%)	Num_of_WT_Lines	Missing_Genotype_WT (%)	MUT_Accu (%)	Num_of_MUT_Lines	Missing_Genotype
47.1	468	81.4	90.6	257	
32.3	468	93.4	100	257	
70.1	468	67.1	61.3	257	
69	468	69.7	61	257	
37.4	468	1.3	89.8	257	
39.1	468	12	88.1	257	
38.6	468	6	88.6	257	
38.1	468	1.3	89	257	
38.2	468	6.6	88.7	257	
36.7	468	7.9	90.3	257	
37.8	468	2.1	89	257	
37.9	468	2.6	88.9	257	
38	468	4.9	88.7	257	
37.4	468	4.1	89.4	257	
37.6	468	2.4	89	257	

Glyma.15g049200
SWEET39 T/TGG

ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS



Menu

Chromosome
15

Genomic interval:
3817975 to 3875101

Reference Phenotype:
○ WT
● MUT

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

Download Phenotype file template

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

Chromosome of tagging variant:
13

Position of tagging variant:
17316758

Choose GWAS Statistics File (.csv):
Browse... No file selected

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

Results

ACCUTOOL: DIRECT CORRESPONDENCE ANALYSIS

Chr	Pos	Avg_Accuracy (%)	Comb_Accu_Pess (%)	p.value	SoySNP50k_ID	Gene	Effect	WT_Accu (%)	Num_of_WT_Lines	Missing_Genotype_WT (%)	MUT_Accu (%)
15	3846538	57.2	44		ss715621777	Glyma.15g048700	G|missense_va	24.7	468	3	89.8
15	3851525	47.4	38.2		ss715621779	Glyma.15g048800	G|upstream_ge	30	468	3.2	64.8
15	3853359	60.1	56.1		ss715621780	Glyma.15g048800	A|intron_variant	66.5	468	3	53.8
15	3863922	58.8	54.3		ss715621781	.	C|intergenic_re	63.7	468	2.8	53.9

ACCUCALC: DIRECT CORRESPONDENCE ANALYSIS FOR OTHER SPECIES

Dr. Jana Biová

Poster no.: **P062**

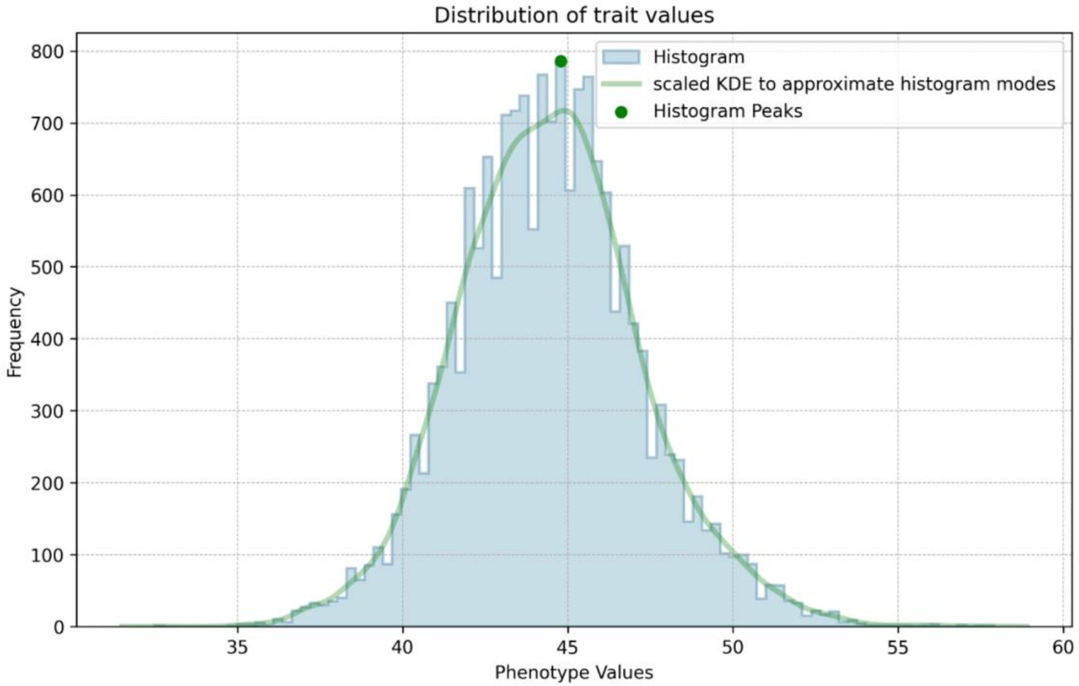
AccuCalc for other species

https://github.com/Biovja/AccuCalc

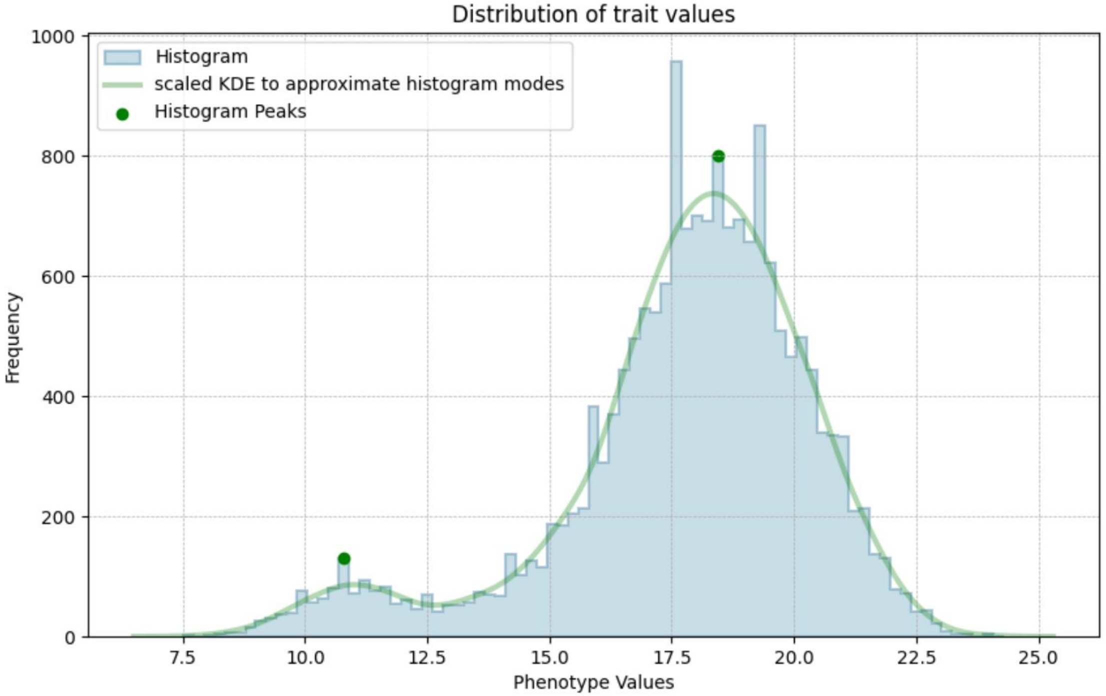
Biova et al., 2024

LINKING GENOTYPES WITH PHENOTYPES:
QUALITATIVE, DISPROPORTIONAL PHENOTYPES OF COMPLEX TRAITS

- Multimodality



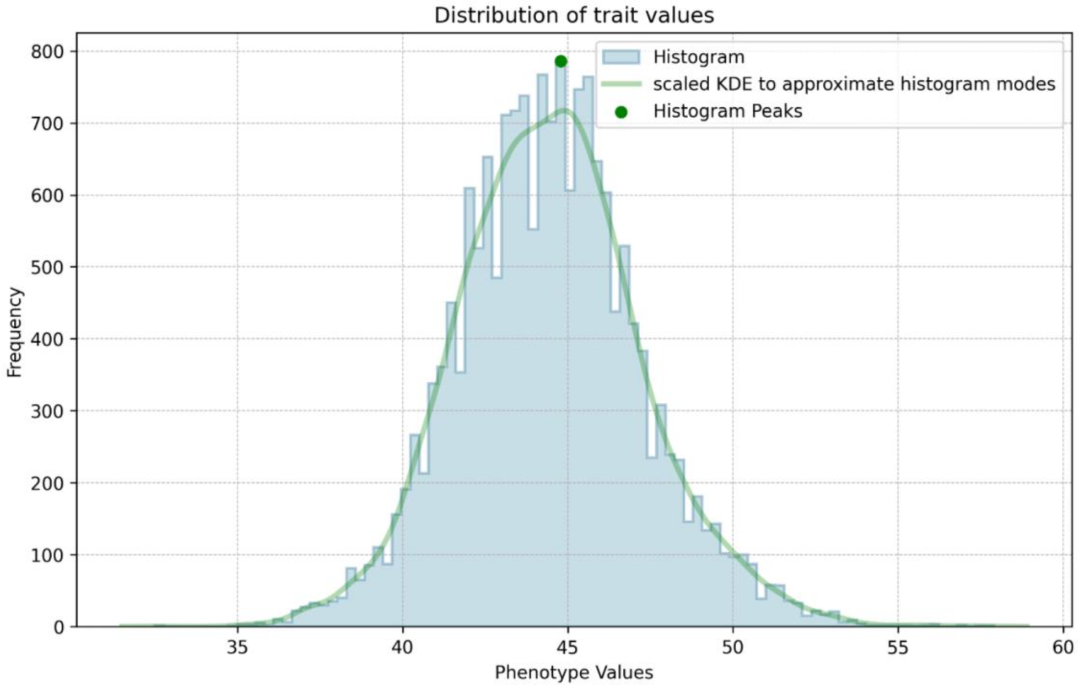
distribution of protein content among soybean samples in the GRIN dataset



distribution of oil content among soybean samples in the GRIN dataset

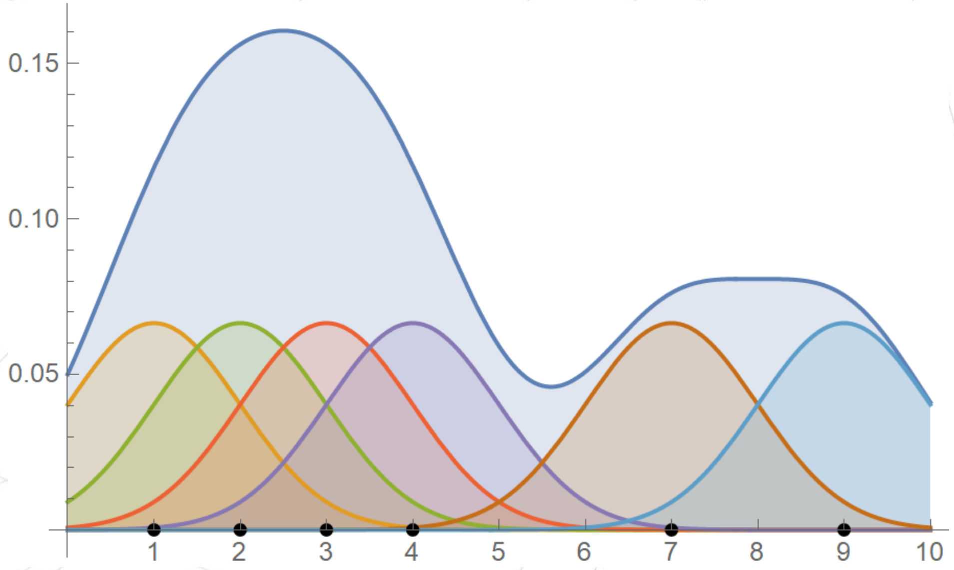
LINKING GENOTYPES WITH PHENOTYPES:
QUALITATIVE, DISPROPORTIONAL PHENOTYPES OF COMPLEX TRAITS

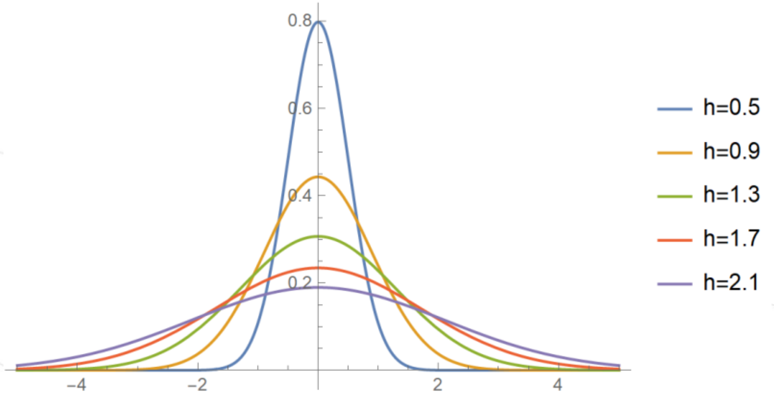
- Multimodality



distribution of protein content among soybean samples in the GRIN dataset

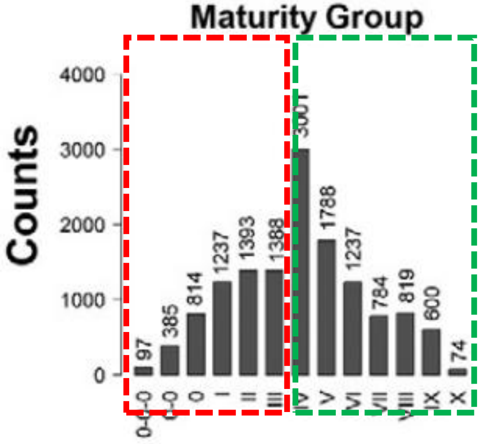
- Kernel density estimate

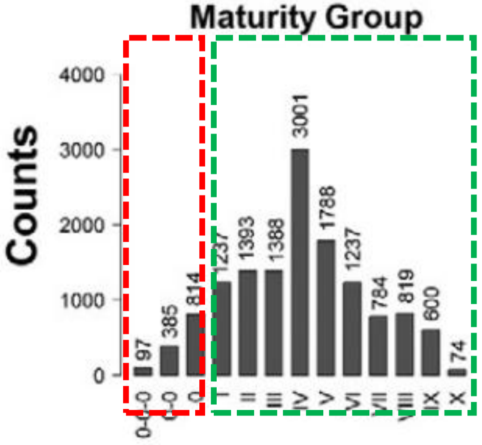


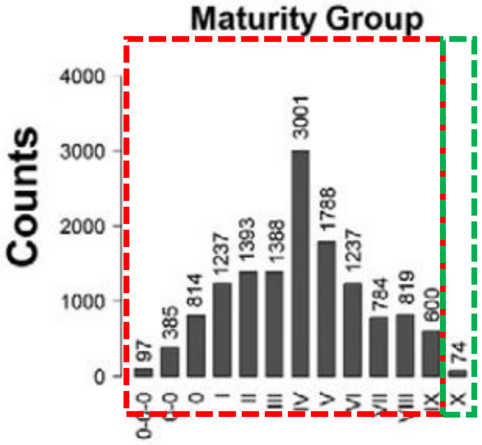


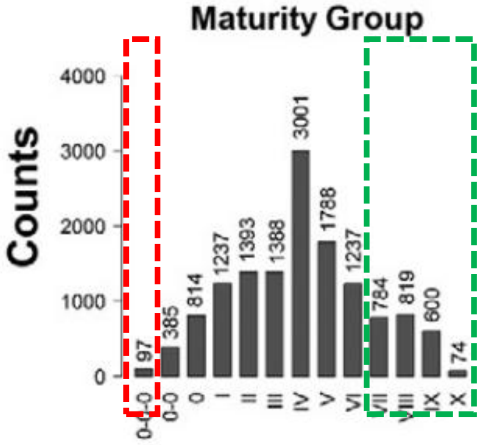
LINKING GENOTYPES WITH PHENOTYPES:
QUALITATIVE, DISPROPORTIONAL PHENOTYPES OF COMPLEX TRAITS

- Categorical phenotype









JOINT EFFORTS FOR SOYBEAN APPLIED GENOMICS

Legume Genomics


Dr. Mária Škrabišová


Dr. Jana Biová


Ivana Kaňovská,
Ph.D. student


Alžbeta A. Rástocká,
graduate student


Jana Slivková

Funding

UNITED SOYBEAN BOARD™

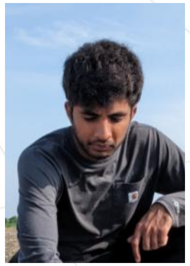
MISSOURI SOYBEANS

Applied Genomics

USDA


Dr. Kristin Bilyeu


Dr. Nicholas Dietz

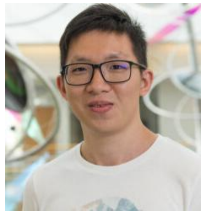

Anser Mahmood,
Ph.D. student


Dr. Nathan Grant

Bioinformatics


Dr. Trupti Joshi

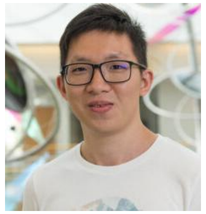

Dr. Shuai Zeng


Dr. Yen On Chan


Manish Sridhar
Immadi

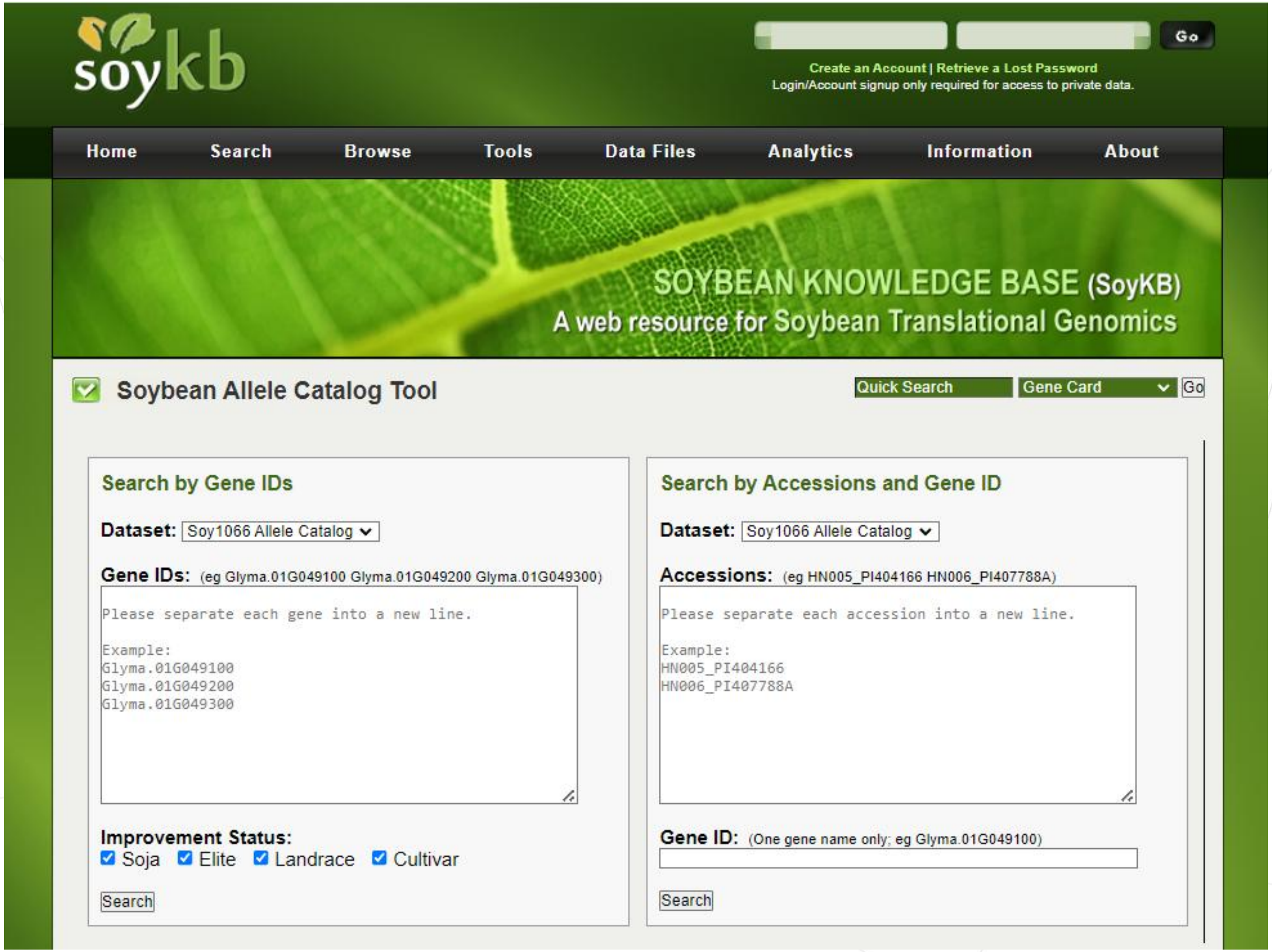
THANK YOU FOR YOUR ATTENTION!

SOYBEAN ALLELE CATALOG



Dr. Yen On Chan

Chan et al., 2023a



Login/Account signup only required for access to private data.

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

Soybean Allele Catalog Tool

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 by Accessions and Gene ID

Dataset: Soy1066 Allele Catalog

Accessions: (eg HN005_PI404166 HN006_PI407788A)

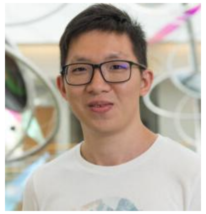
```
Please separate each accession into a new line.

Example:
HN005_PI404166
HN006_PI407788A
```

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

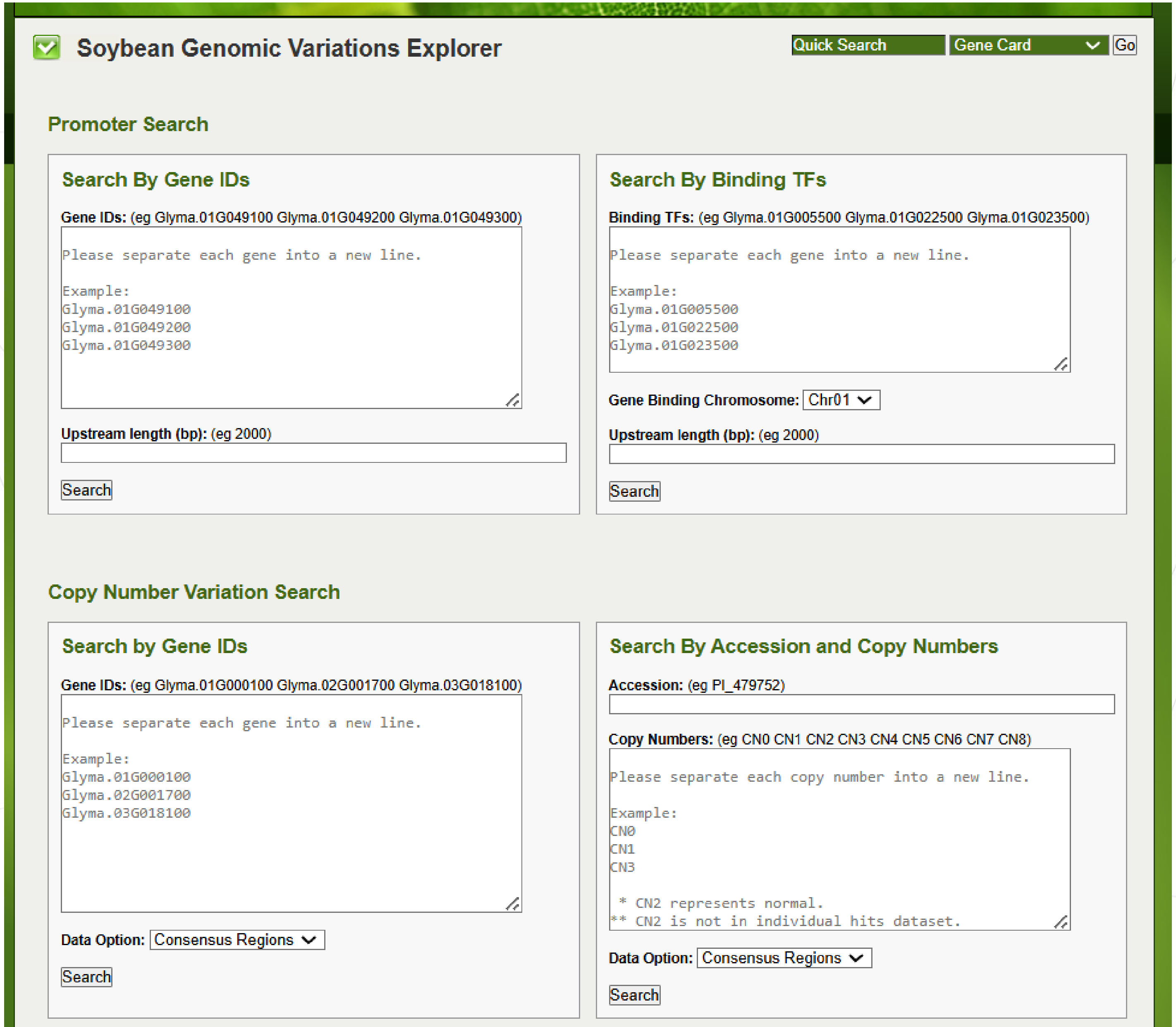
Dr. Kristin Bilyeu

GENVARX: DIRECT CORRESPONDENCE ANALYSIS



Dr. Yen On Chan

Chan et al., 2023b



Soybean Genomic Variations Explorer

Promoter Search

Search By Gene IDs

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

```
Please separate each gene into a new line.

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

Upstream length (bp): (eg 2000)

Search By Binding TFs

Binding TFs: (eg Glyma.01G005500 Glyma.01G022500 Glyma.01G023500)

```
Please separate each gene into a new line.

Example:
Glyma.01G005500
Glyma.01G022500
Glyma.01G023500
```

Gene Binding Chromosome: Chr01

Upstream length (bp): (eg 2000)

Copy Number Variation Search

Search by Gene IDs

Gene IDs: (eg Glyma.01G000100 Glyma.02G001700 Glyma.03G018100)

```
Please separate each gene into a new line.

Example:
Glyma.01G000100
Glyma.02G001700
Glyma.03G018100
```

Data Option: Consensus Regions

Search By Accession and Copy Numbers

Accession: (eg PI_479752)

Copy Numbers: (eg CN0 CN1 CN2 CN3 CN4 CN5 CN6 CN7 CN8)

```
Please separate each copy number into a new line.

Example:
CN0
CN1
CN3

 * CN2 represents normal.
** CN2 is not in individual hits dataset.
```

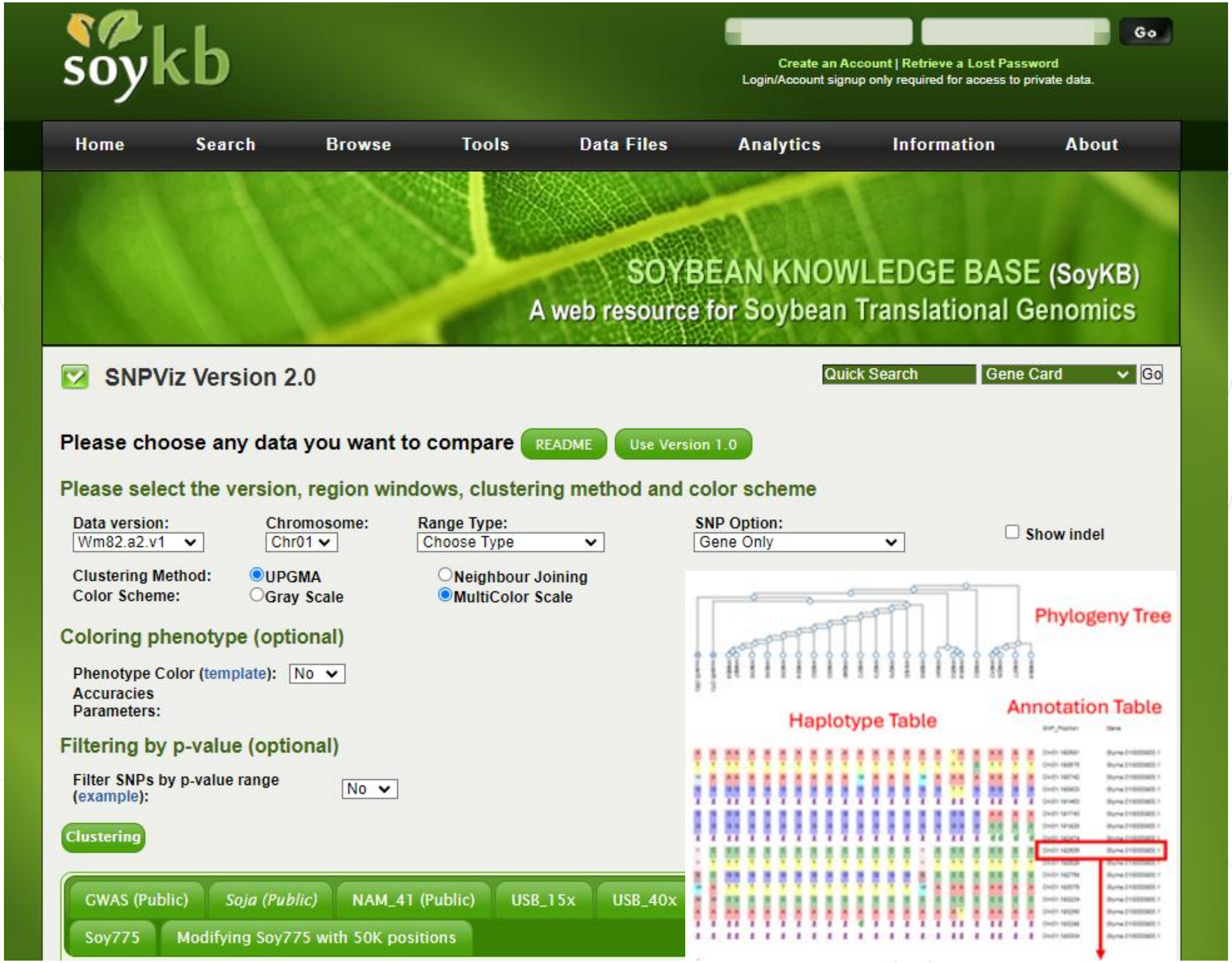
Data Option: Consensus Regions

SNPVIZ 2.0: EMPOWERED HAPLOTYPE VIEWER

Dr. Shuai Zeng

Zeng et al., 2021

Zeng et al., 2020



soykb

Go

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 | Go

Please choose any data you want to compare README | Use Version 1.0

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

Data version: Wm82.a2.v1

Chromosome: Chr01

Range Type: Choose Type

SNP Option: Gene Only

☐ Show indel

Clustering Method: ● UPGMA ○ Neighbour Joining

Color Scheme: ○ Gray Scale ● MultiColor Scale

Coloring phenotype (optional)

Phenotype Color (template): No

Accuracies

Parameters:

Filtering by p-value (optional)

Filter SNPs by p-value range (example): No

Clustering

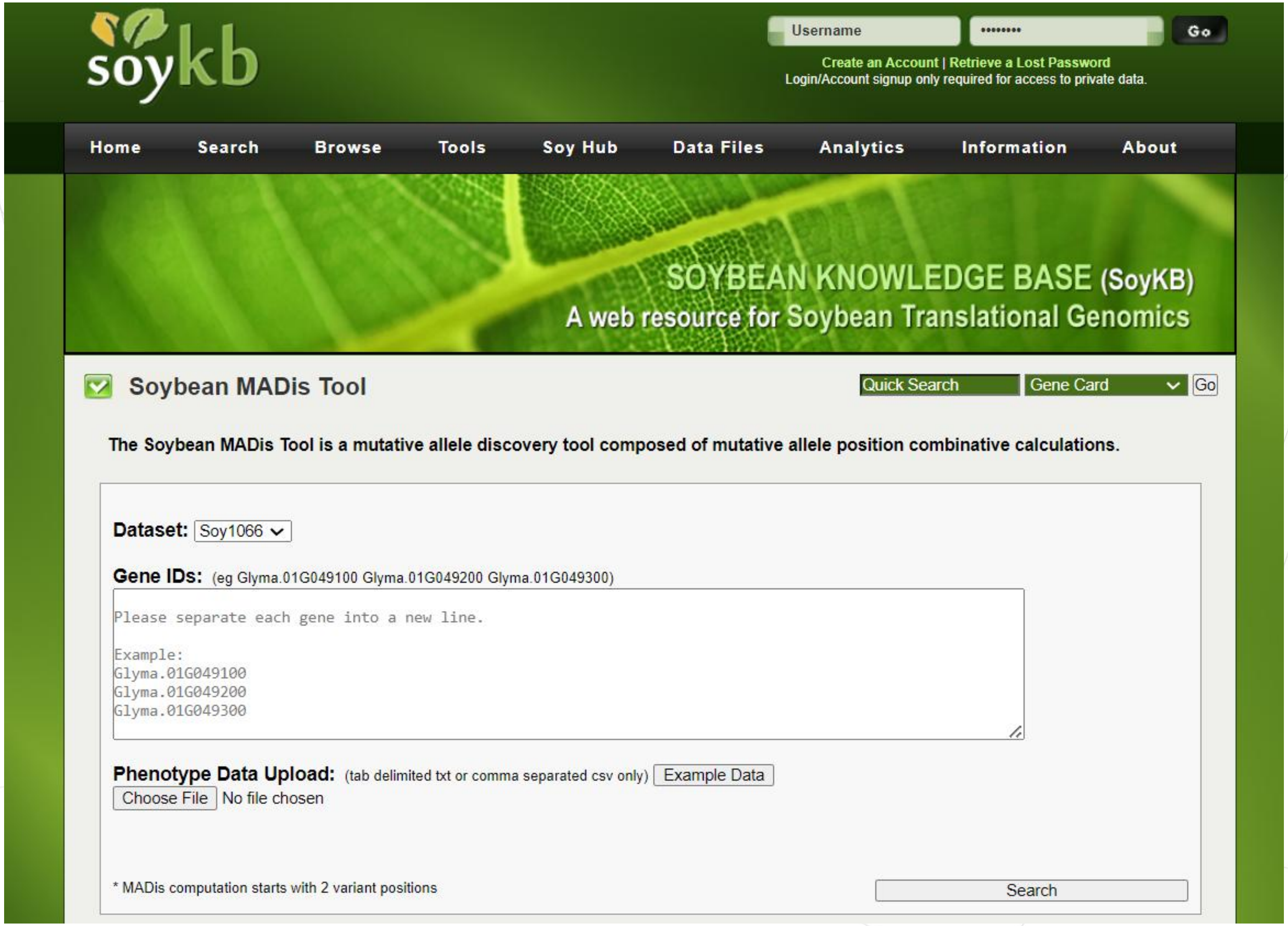
GWAS (Public) | *Soja (Public)* | NAM_41 (Public) | USB_15x | USB_40x

Soy775 | Modifying Soy775 with 50K positions

MADIS: MUTATIVE ALLELE DISCOVERY TOOL

Biova et al., 2024

Dr. Jana Biová



Soybean MADis Tool

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

Dataset: Soy1066

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

```
Please separate each gene into a new line.

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

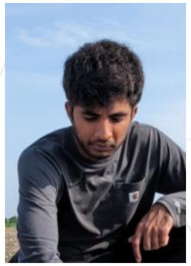
Phenotype Data Upload: (tab delimited txt or comma separated csv only)

* MADis computation starts with 2 variant positions

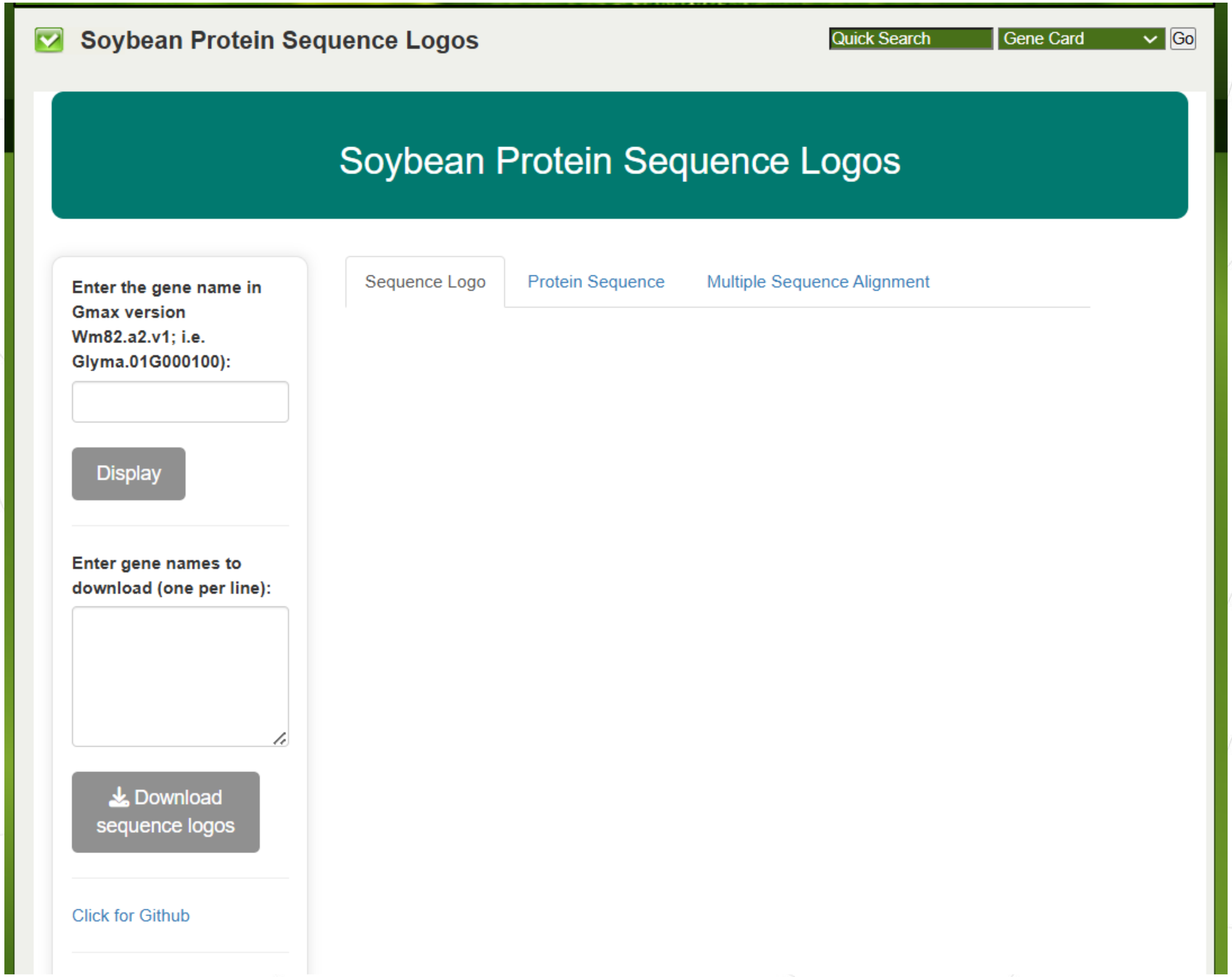
Ivana Kaňovská
Ph.D. student

Poster 559

PROTEIN SEQUENCE WEB LOGO: CONSERVED AMINO ACID SEARCH



Anser Mahmood,
Ph.D. student

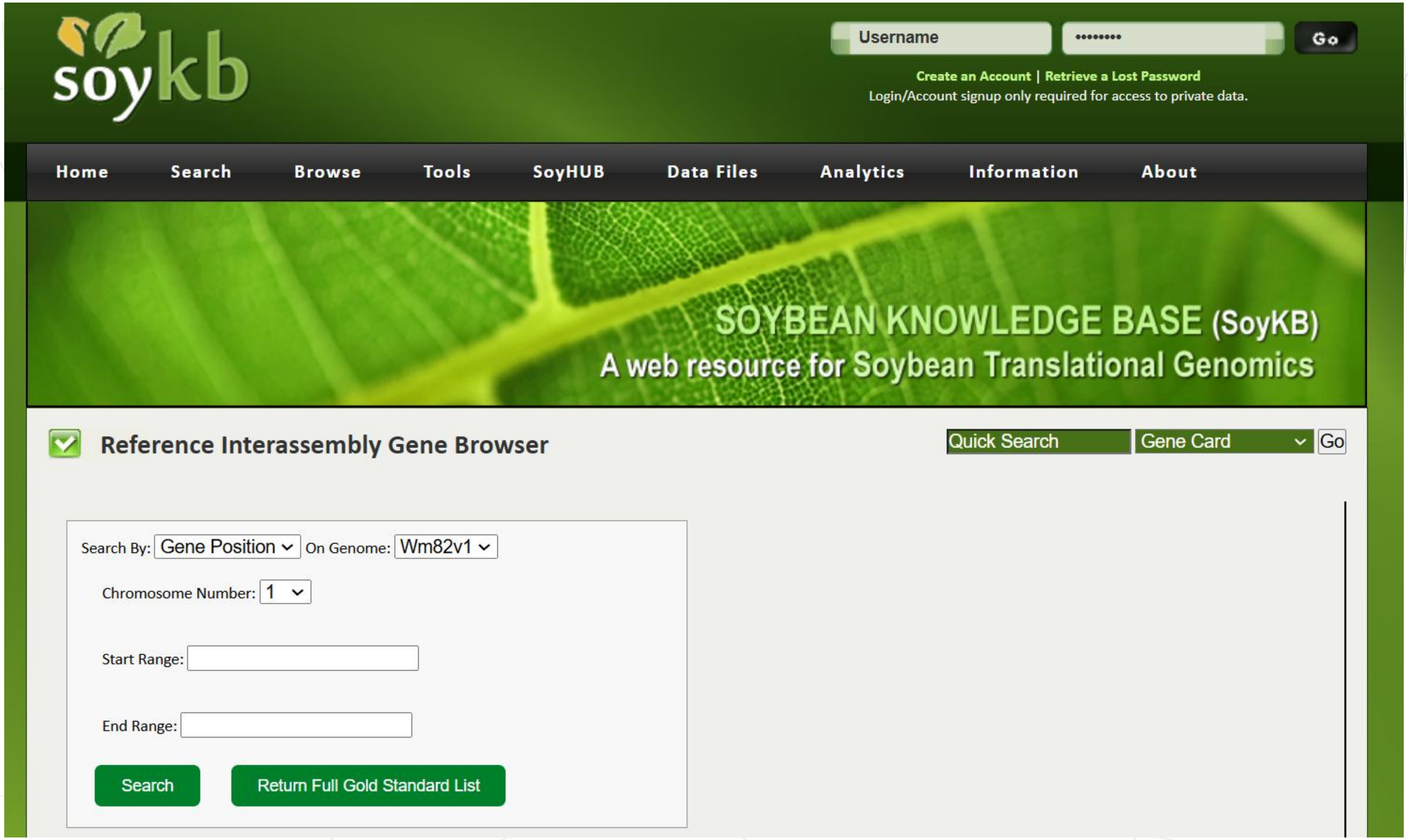


Enter the gene name in Gmax version Wm82.a2.v1; i.e. Glyma.01G000100):

Enter gene names to download (one per line):

REFERENCE INTERASSEMBLY GENE BROWSER: GENE ID SEARCH

Travers Connor,
student



Login/Account signup only required for access to private data.

Reference Interassembly Gene Browser

Search By: Gene Position On Genome: Wm82v1

Chromosome Number: 1

Start Range:

End Range:

Alžbeta A. Rástocká,
graduate student

SOYHUB HOSTS DATA SETS

Data storage & Analysis support

- The North latitude soybean data set
- European resequenced data set
- Curation of incoming data sets

KB Commons for other species

- Bean?
- Pea?

Funded by
the European Union

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!