AccuCalc and MADis, tools for improving causative mutation discoveries in associated loci

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

1 Department of Biochemistry, Faculty of Science, Palacký University in Olomouc, Olomouc, Czech Republic
2 Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO, 65212, USA
3 MU Data Science and Informatics Institute, University of Missouri, Columbia, MO, 65212, USA
4 Department of Electrical Engineering and Computer Science
5 Department of Health Management and Informatics, School of Medicine, University of Missouri, Columbia, MO, 65212, USA
6 Plant Genetics Research Unit, United States Department of Agriculture-Agricultural Research Service, University of Missouri, Columbia MO, USA

INTRODUCTION

Identifying causative mutations (CMs) is essential for advancing functional genomic research. However, current methods that associate phenotypes with genotypes often designate large genomic regions as associated loci. Consequently, further analysis is required to pinpoint candidate genes and their corresponding CMs. To address this need, we developed two new tools for CM identification: Python-based package for Accuracy calculation and Synthetic phenotype creation (AccuCalc) [1] and Multiple Alleles Discovery (MADis) tool [2]. The tools are together with Genomic Variations Explorer (GenVarX) [3] part of a set for CM discoveries.

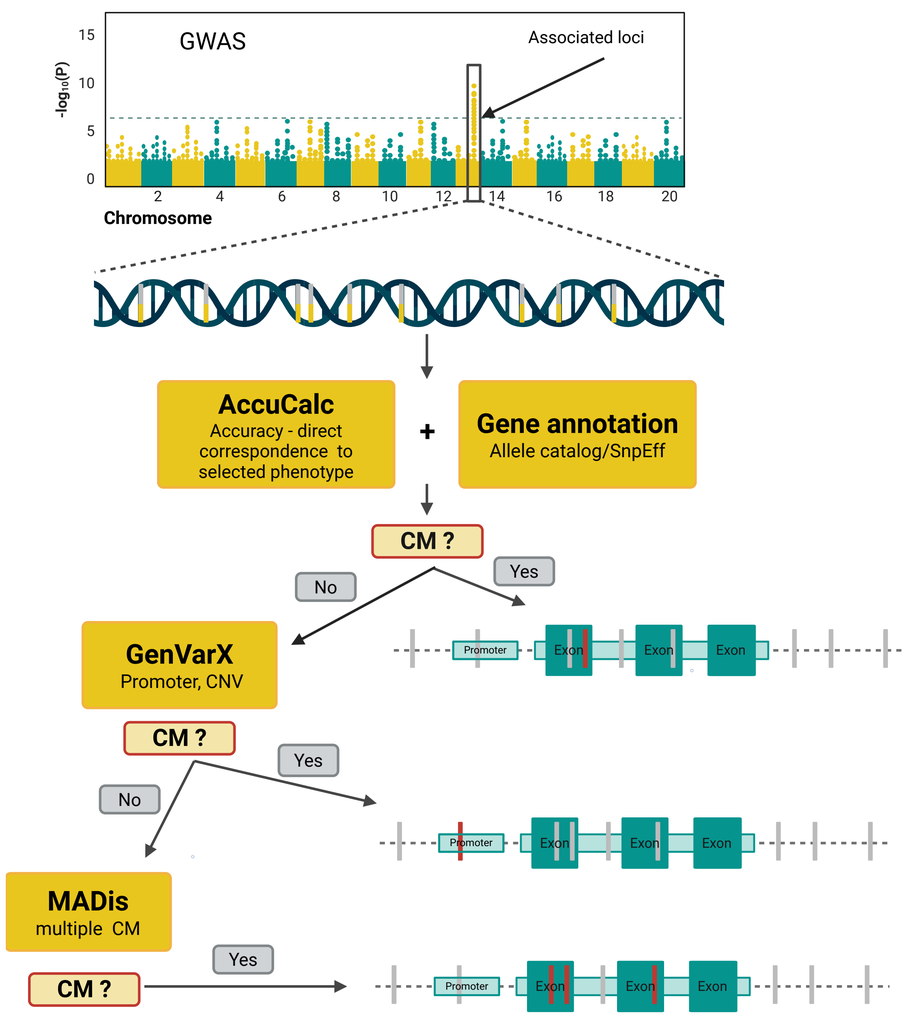
Tools

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

The GitHub for Python package AccuCal. Allow processing of user-provided data regardless of the organism.

The GitHub for Python package MADis. Allow processing of user-provided data regardless of the organism.

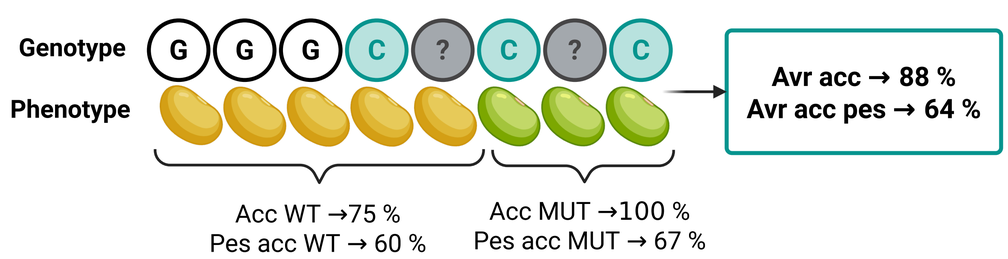
RESULTS



← The diagram shows the tools created for post-GWAS analysis and identification of CM with example usage. For simplicity, a random region of the genome associated in GWAS with a theoretical phenotype is used here, and only a single candidate gene from the associated region is shown. Yellow boxes represent the tools and their main usage in identifying CM. Light blue shows examples of candidate genes and their CM types identified by appropriate tools. The gray dashed line in the background represents the surroundings of the gene, which is divided into the promoter (light rectangle with the word promoter), darkly depicted exons, and lightly depicted introns and UTRs. The gray vertical lines for the example candidate genes represent observed mutations in the gene and its surroundings, while the red lines represent CM identified by the tool.

AccuCalc

↓ The Accuracy shows the correspondence between the genotype and phenotype. The average accuracy (Avr acc) counts only with the available data, and pessimistic (pes) accuracy adds information about result reliability by lowering the value in proportion to the missing data.



↓The top 3 best accuracy results for *Glyma.18G273600* (and gene surrounding ± 100 kbp). The result shows no genetic variant with high accuracy that can be considered a single or majority CM candidate, pointing to the different CM types.

#CHROM	POS	REF/ALT	Avr acc(%)	Acc WT(%)	Acc MUT(%)	Avr acc pes(%)	Acc pes WT(%)	Acc pes MUT(%)	Count ref	Count alt	Ref phen for Acc	Count WT	Count MUT
18	55626903	C/A	67.80	46.54	89.06	67.80	46.54	89.06	649	417	MUT	419	64
18	55615040	G/T	67.22	56.32	78.12	67.22	56.32	78.12	511	555	MUT	419	64
18	55720759	C/T	67.20	62.53	71.88	67.20	62.53	71.88	414	652	MUT	419	64

MADis

→ The best-scored MADis output for *Glyma.18G273600* and *Glyma.19G120400*. The MADis results for the analysis of potential multiple CMs in *Glyma.19G120400* affecting the pod color trait. In the analysis, the 127 blackpod-colored accessions and 605 brown and tanpod-colored accessions were used for MADis prediction. The MADis result was a perfect correspondence for more than 98% of analyzed accessions, implying the discovered multiple CMs in *Glyma.19G120400* as a strong candidate. MADis result for *Glyma.19G120400* shows the example of a CMs discovery by MADis. The MADis promotor (new addition to MADis) results for the analysis of potential multiple CMs in *Glyma.18G273600* affecting the plant architecture. In the analysis, the 483 accessions with non-determinated stem termination accessions and 64 accessions with semi-determinated stem termination were used for MADis prediction in promoter. The number of explained mutant accessions, 17 from 64, suggests that the phenotype can be affected by in the analysis undetected genomic variation, for example transposon located upstream of *Glyma.18G273600* [4].

L1 Glyma.19G120400

Genomic variants	N combination	Score	Max score	N WT	N WT explained	N MUT	N MUT explained	Explained phen (%)
19-37,806,091 19-37,806,119 19-37,806,160 19-37,819,253 19-37,819,382 19-37,819,390	6	704	732	127	124	605	594	98.10

Dt2 Glyma.18G273600

Genomic variants	N combination	Score	Max score	N WT	N WT explained	N MUT	N MUT explained	Explained phen (%)
18-55,635,871 18-55,637,359	2	371	483	419	410	64	17	88.41

REFERENCES

[1] Biová, J., Dietz, N., Chan, Y. O., Joshi, T., Bilyeu, K., & Škrabišová, M. (2023). AccuCalc: A Python Package for Accuracy Calculation in GWAS. Genes 2023, Vol. 14, Page 123, 14(1), 123. https://doi.org/10.3390/GENES14010123

[2] Biová, J., Kaňovská, I., Chan, Y. O., Immadi, M. S., Joshi, T., Bilyeu, K., & Škrabišová, M. (2024). Natural and artificial selection of multiple alleles revealed through genomic analyses. Frontiers in Gene tics, 14, 1320652. https://doi.org/10.3389/FGENE.2023.1320652

[3] Chan Y.O., Biova J., Mahmood A., Dietz N., Škrabišová M., Bilyeu K., Joshi T. (2023) Genomic Variations Explorer (GenVarX): A Toolset for Annotating Promoter and CNV Regions Using Genotypic and Phenotypic Differences. Regular paper at APBC 2023. Accepted. Invited for full paper in Frontiers

[4] Ping J, Liu Y, Sun L, Zhao M, Li Y, She M, Sui Y, Lin F, Liu X, Tang Z, Nguyen H, Tian Z, Qiu L, Nelson RL, Clemente TE, Specht JE, Ma J. Dt2 is a gain-of-function MADS-domain factor gene that specifies semideterminacy in soybean. Plant Cell. 2014 Jul;26(7):2831-42. doi: 10.1105/tpc.114.126938. Epub 2014 Jul 8. PMID: 25005919; PMCID: PMC4145117.

ACKNOWLEDGEMENTS

Applied Genomics to Improve Soybean Seed protein, #1920-152-0131-C; Enhancing Soybean Applied Genomics Tools for Improving Soybean, #2220–152-0202; Leveraging Genomics to Enhance the US Soybean Quality Reputation, #2332-201-0101; USA Soybean Quality Reputation Enhanced with Genomics Integration, #24-201 S-B-1-A/2432-201-0101; Palacký University Internal Grant Agency, #IGA_2020_013; IGA_PrF_2021_015; #IGA_PrF_2022_025; IGA_PrF_2023_022.

Legume Generation (Boosting innovation in breeding for the next generation of legume crops for Europe) has received funding from the European Union through Horizon Europe under grant agreement No 101081329 and co-funding from UK Research and Innovation (UKRI) from the UK government's Horizon Europe funding guarantee. It also receives support from the governments of Switzerland and New Zealand.

www.legumegeneration.eu @LegumeHubEU linkedIn.com/company/legume-hub-eu