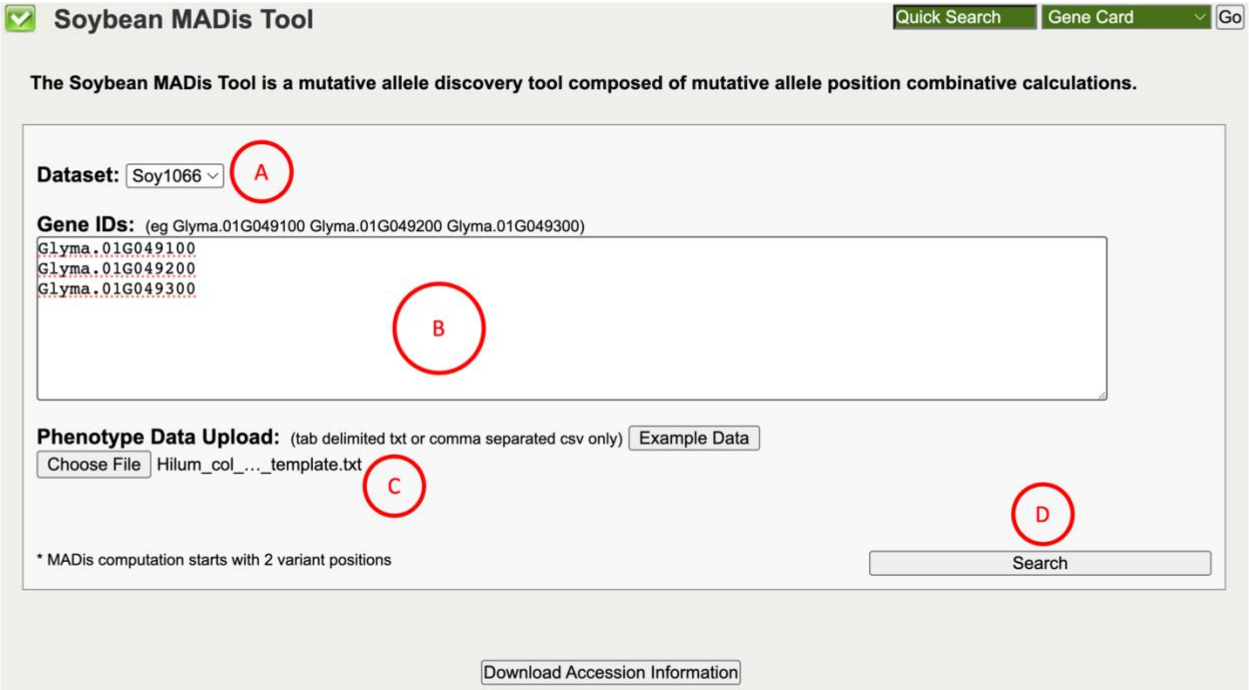
Soybean MADis Tool Guidelines

The architecture of the Soybean MADis Tool includes the MySQL database that links to the SoyKB web portals (Joshi et al. 2017; Joshi et al. 2013; Joshi et al. 2012), back-end processing code in PHP, and front-end user interfaces with interactive components and visualizations developed with HTML, CSS, and JavaScript. The purpose of the Soybean MADis Tool is for users to provide genes and phenotypes to perform calculations using the MADis algorithm and visualize the results in interactive visualizations to help researchers select the best explainable variant position combinations and advance their research. The components of this tool comprise a data input page, result pages, and an Allele Catalog visualization page (Chan et al. 2023).

On the data input page, there is only one window that consists of a dataset dropdown menu, a gene IDs input box, a phenotype data upload button, and a search button (Figure 1). When users would like to initiate a query, they can select a dataset from the dropdown menu, provide gene IDs into the Gene IDs input box with one gene ID in a new line, and upload a phenotype file in tab-delimited format before pressing the search button. The format of the phenotype file contains an accession column and a binarized phenotype data column annotating wild-type phenotype as 0 and mutative phenotype as 1. Upon filling all the required fields, a query to request MADis calculations with combinations of 2 variant positions of those inputted genes can be executed and redirect users to a result page.



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)

```
Glyma.01G049100
Glyma.01G049200
Glyma.01G049300
```

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

Choose File Hilum_col_…_template.txt

* MADis computation starts with 2 variant positions

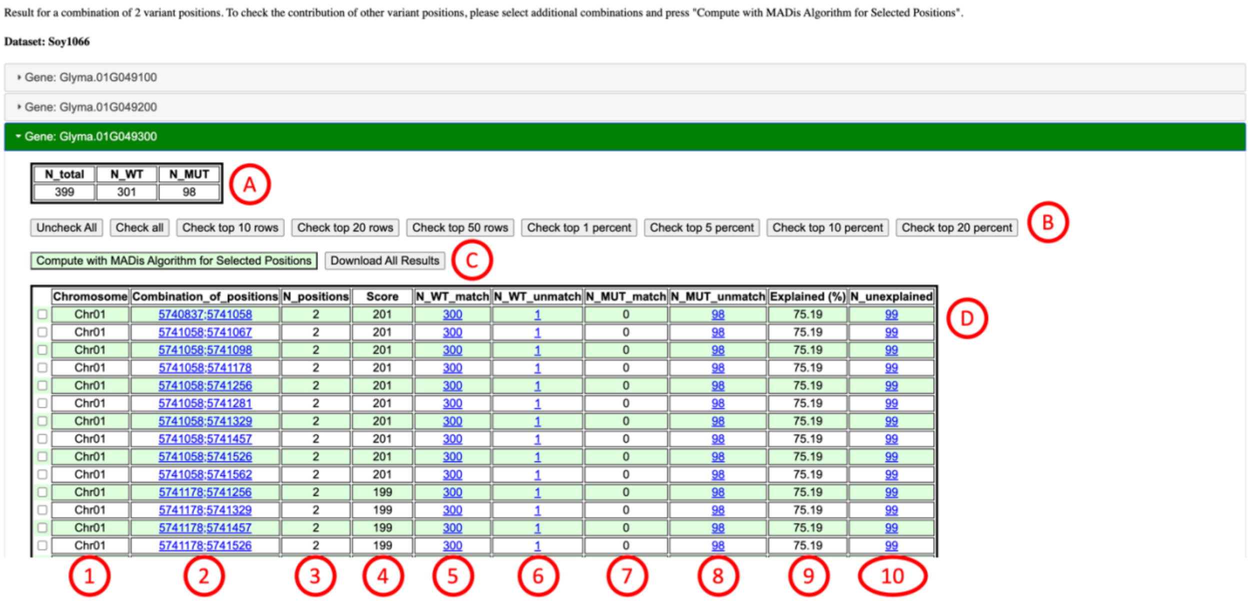
Figure 1: The data input page of the Soybean MADis Tool that consists of only one window. Within the window, there are (A) a dataset selection dropdown menu, (B) a multi-line input box for at least one gene IDs with one in a new line, (C) a phenotype data upload button (example data is also provided), and (D) a search button for initiating a query.

The MADis initial round of calculations starts with combinations of 2 variant positions. The total data size for computing with combinations of 2 variant positions can be calculated using the combination formula as follows:

$$\text{Total data size} = {}_{n}C_{2} \text{ ---------- } (1)$$

where n denotes the total number of variant positions in a gene

The data size is feasible for the computing environment even when the total number of variant positions in a gene is huge. The results of the calculations are returned to the result page and rendered as tables for user visualizations once the calculation is completed (Figure 2). In Figure 2, the details of the columns marked from 1 to 10 are illustrated in Table 1 so that users can better understand the meaning of each column.



Result for a combination of 2 variant positions. To check the contribution of other variant positions, please select additional combinations and press "Compute with MADis Algorithm for Selected Positions".

Dataset: Soy1066

N_total	N_WT	N_MUT
399	301	98

	Chromosome	Combination_of_positions	N_positions	Score	N_WT_match	N_WT_unmatch	N_MUT_match	N_MUT_unmatch	Explained (%)	N_unexplained
☐	Chr01	5740837;5741058	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741067	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741098	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741178	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741256	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741281	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741329	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741457	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741526	2	201	300	1	0	98	75.19	99
☐	Chr01	5741058;5741562	2	201	300	1	0	98	75.19	99
☐	Chr01	5741178;5741256	2	199	300	1	0	98	75.19	99
☐	Chr01	5741178;5741329	2	199	300	1	0	98	75.19	99
☐	Chr01	5741178;5741457	2	199	300	1	0	98	75.19	99
☐	Chr01	5741178;5741526	2	199	300	1	0	98	75.19	99

Figure 2: The result page of the MADis calculation with combinations of 2 variant positions. (A) A table that shows the total number of accessions (N_total), number of wild-type accessions (N_WT), and number of mutative accessions (N_MUT). (B) A group of buttons for users to easily select the checkboxes in the table in D. (C) The “Compute with MADis Algorithm for Selected Positions” allows users to take the variant positions of which the checkboxes are check and perform further MADis calculations with more combinations. Besides that, a “Download All Results” button is also provided for users to download the MADis results. (D) A table that shows

the results of MADis calculations with combinations of 2 variant positions, and the details of each column is described in Table 1.

Table 1: A table that explains the meaning of all the columns marked from 1 to 10 in Figure 2.

Index	Column	Description
1	Chromosome	The chromosome of the gene and those variant positions.
2	Combination of Positions	The possible combination of positions in the MADis calculations.
3	N_Positions	The numbers of positions in the combinations.
4	Score	The calculated scores for each position combination.
5	N_WT_match	The numbers of wild-type accessions that match the wild-type phenotypes.
6	N_WT_unmatch	The numbers of wild-type accessions that do not match the wild-type phenotypes.
7	N_MUT_match	The numbers of mutative accessions that match the mutative phenotypes.
8	N_MUT_unmatch	The numbers of mutative accessions that do not match the mutative phenotypes.
9	Explained (%)	The percentages that explain the mutations. The formula for this calculation is as follow: $$percentage = 100 * \frac{(N_{MUT_{match}}) + (N_{WT_{match}})}{total\ number\ of\ accessions}$$

10	N_unexplained	The numbers of wild-type and mutative accessions that do not match with their corresponding wild-type or mutative phenotypes.

If users check the checkboxes in the initial round of MADis results (Figure 2D) and select “Compute with MADis Algorithm for Selected Positions” button (Figure 2C), a second round of MADis calculations will be performed. As we are allowing users to perform with a maximum of combinations of 7 variant positions in the second round of analysis, the total data size to compute can be calculated using the formula below.

$$\text{Total data size} = {}_{m}C_{7} + {}_{m}C_{6} + {}_{m}C_{5} + {}_{m}C_{4} + {}_{m}C_{3} + {}_{m}C_{2} \text{ ---------- } (2)$$

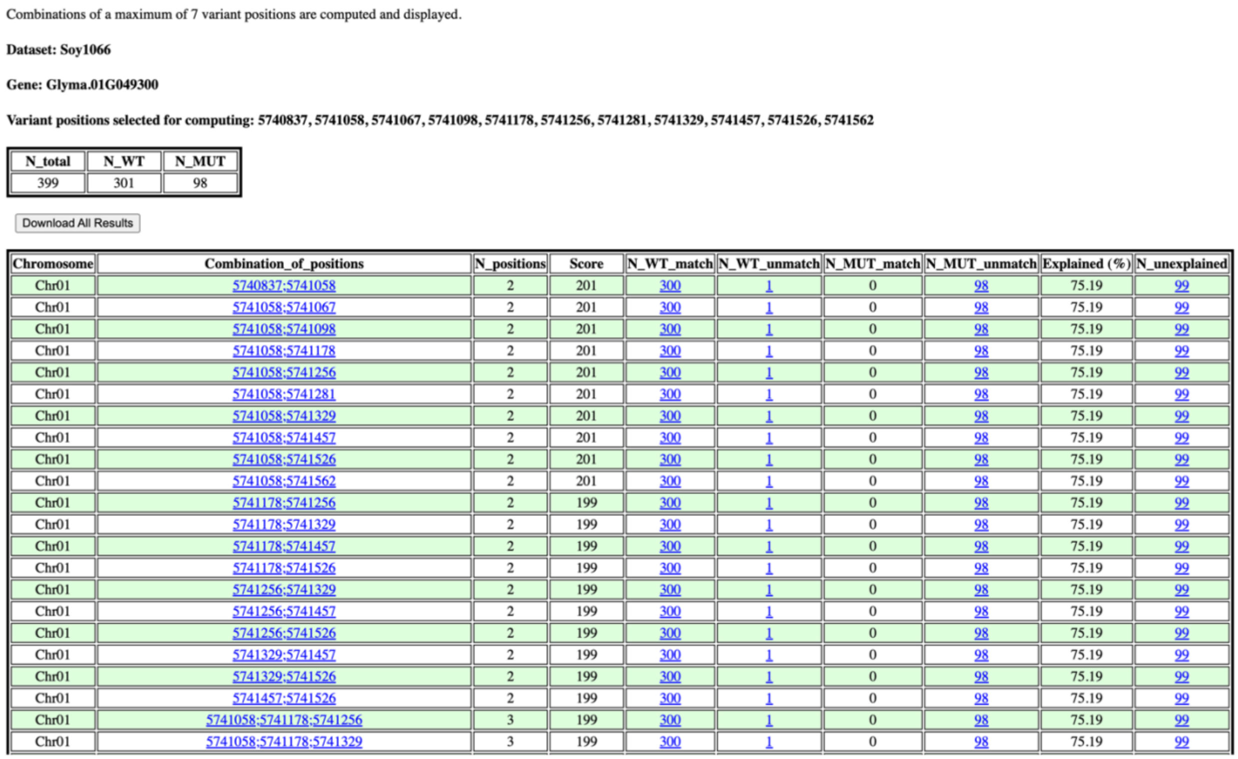
where m is the total number of variant positions selected for computing in this round

The reason to have two rounds of MADis computations is because running MADis algorithm on combinations of 7 variant positions in the first round can have huge computational costs if the total number of variant positions in a gene is high. The formula to calculate the total data size running combinations of 7 variant positions in the first round is as follows.

$$\text{Total data size} = {}_{n}C_{7} + {}_{n}C_{6} + {}_{n}C_{5} + {}_{n}C_{4} + {}_{n}C_{3} + {}_{n}C_{2} \text{ ---------- } (3)$$

where n denotes the total number of variant positions in a gene in the first round

Hence, the total data size can be very huge if the n in formula 3 is huge. In fact, the total data size calculated from formula 1 is always less than formula 3 as n can only be a positive integer. To avoid the heavy computational costs, the second round of MADis calculation approach is developed for users to select important variant positions in the first round results and then do a second round of MADis calculation. In that case, the total data size in the second round is always lower or equal as $m \leq n$ stands in any case. This approach can not only reduce computational time but also reduce resource allocation to increase computational efficiency. Figure 3 is the result page that demonstrates the second round of calculations. The interface in Figure 3 is similar to Figure 2.



Combinations of a maximum of 7 variant positions are computed and displayed.

Dataset: Soy1066

Gene: Glyma.01G049300

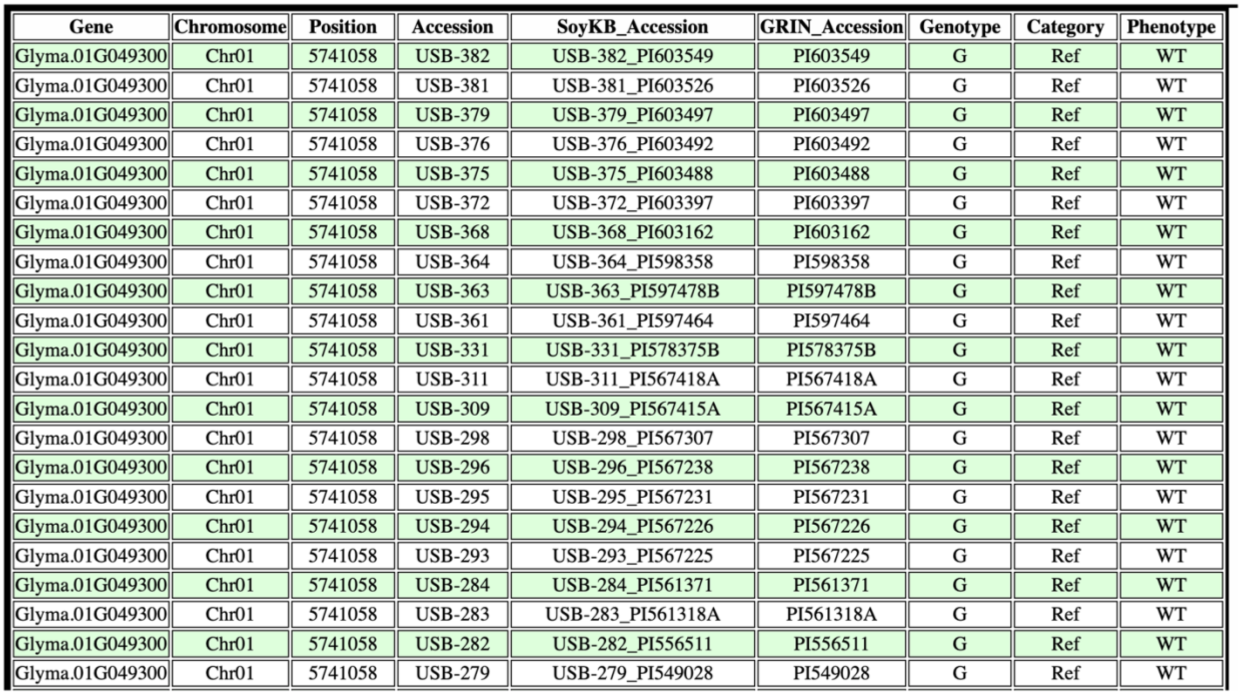
Variant positions selected for computing: 5740837, 5741058, 5741067, 5741098, 5741178, 5741256, 5741281, 5741329, 5741457, 5741526, 5741562

N_total	N_WT	N_MUT
399	301	98

Chromosome	Combination_of_positions	N_positions	Score	N_WT_match	N_WT_unmatch	N_MUT_match	N_MUT_unmatch	Explained (%)	N_unexplained
Chr01	5740837;5741058	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741067	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741098	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741178	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741256	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741281	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741329	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741457	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741526	2	201	300	1	0	98	75.19	99
Chr01	5741058;5741562	2	201	300	1	0	98	75.19	99
Chr01	5741178;5741256	2	199	300	1	0	98	75.19	99
Chr01	5741178;5741329	2	199	300	1	0	98	75.19	99
Chr01	5741178;5741457	2	199	300	1	0	98	75.19	99
Chr01	5741178;5741526	2	199	300	1	0	98	75.19	99
Chr01	5741256;5741329	2	199	300	1	0	98	75.19	99
Chr01	5741256;5741457	2	199	300	1	0	98	75.19	99
Chr01	5741256;5741526	2	199	300	1	0	98	75.19	99
Chr01	5741329;5741457	2	199	300	1	0	98	75.19	99
Chr01	5741329;5741526	2	199	300	1	0	98	75.19	99
Chr01	5741457;5741526	2	199	300	1	0	98	75.19	99
Chr01	5741058;5741178;5741256	3	199	300	1	0	98	75.19	99
Chr01	5741058;5741178;5741329	3	199	300	1	0	98	75.19	99

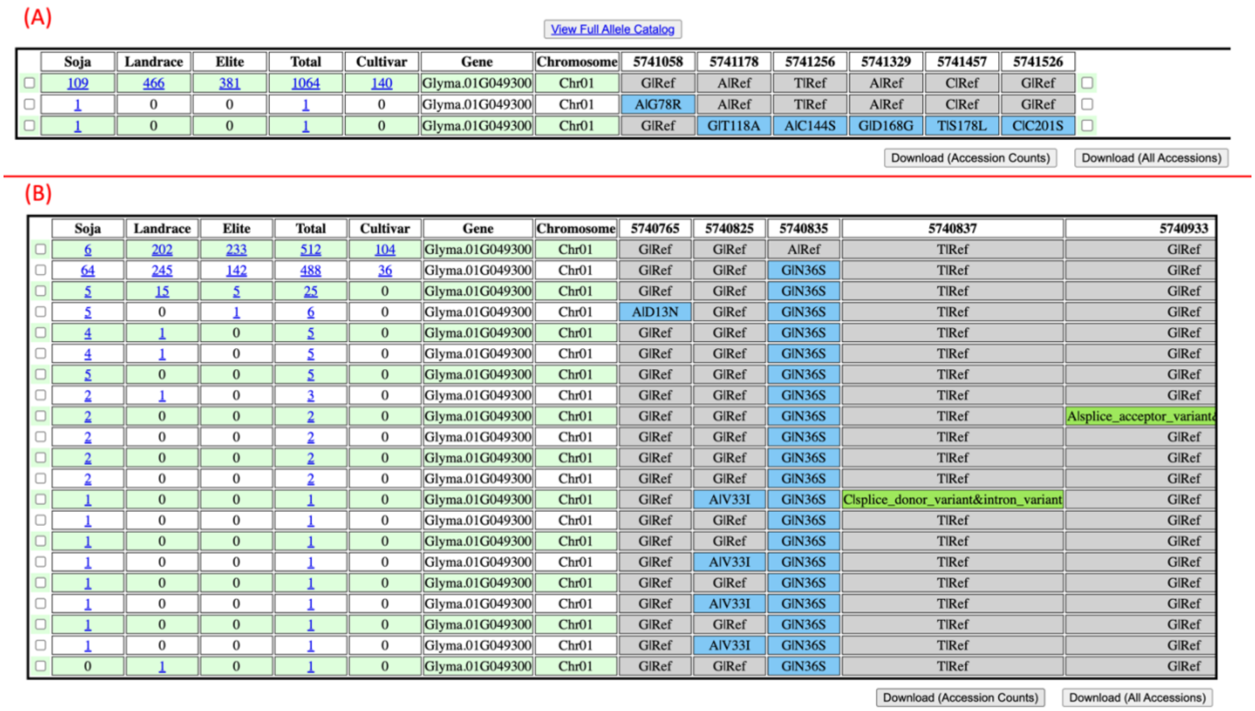
Figure 3: The result page of the second round of the MADis calculation with a maximum of combinations of 7 variant positions. The structure of this result page is similar to the structure of the result page for the first round of calculations. Please refer to Table 1 for the details of each column of the table.

In order to increase interactivity with users and assist them in gaining more information about the variant positions and accession counts, links are added to the table on the result page for users to visualize the accessions and phenotypes loaded into a table that installed in the modal pop-up box (Figure 4), or to click on the combination of positions to redirect to the Allele Catalog visualization (Figure 5). The modal pop-up box appears when users click on the accession counts that are larger than zero in the N_WT_match, N_WT_unmatch, N_MUT_match, and N_MUT_unmatch columns. As shown in Figure 4, the table in the modal pop-up box contains information such as gene, chromosome, positions, accession names, alternative SoyKB accession names, alternative Germplasm Resources Information Network (GRIN) accession names from the GRIN database, genotypes, genotype categories (reference (Ref) or alternate (Alt)), and phenotypes (wild-type phenotype (WT) or mutated phenotype (MUT)) of the variant position combinations. Additionally, the position combinations in the combination of positions column are also clickable to open a new page with the Allele Catalog of the corresponding variant positions (Figure 5). On that new page, users can also visualize the full Allele Catalog of the gene by clicking the "View Full Allele Catalog" button. The structures and functions of the Allele Catalog visualization are illustrated in detail in (Chan et al. 2023).



Gene	Chromosome	Position	Accession	SoyKB_Accession	GRIN_Accession	Genotype	Category	Phenotype
Glyma.01G049300	Chr01	5741058	USB-382	USB-382_PI603549	PI603549	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-381	USB-381_PI603526	PI603526	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-379	USB-379_PI603497	PI603497	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-376	USB-376_PI603492	PI603492	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-375	USB-375_PI603488	PI603488	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-372	USB-372_PI603397	PI603397	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-368	USB-368_PI603162	PI603162	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-364	USB-364_PI598358	PI598358	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-363	USB-363_PI597478B	PI597478B	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-361	USB-361_PI597464	PI597464	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-331	USB-331_PI578375B	PI578375B	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-311	USB-311_PI567418A	PI567418A	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-309	USB-309_PI567415A	PI567415A	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-298	USB-298_PI567307	PI567307	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-296	USB-296_PI567238	PI567238	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-295	USB-295_PI567231	PI567231	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-294	USB-294_PI567226	PI567226	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-293	USB-293_PI567225	PI567225	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-284	USB-284_PI561371	PI561371	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-283	USB-283_PI561318A	PI561318A	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-282	USB-282_PI556511	PI556511	G	Ref	WT
Glyma.01G049300	Chr01	5741058	USB-279	USB-279_PI549028	PI549028	G	Ref	WT

Figure 4: The table in the modal pop-up box to demonstrate the genotypes and phenotypes of a combination of variant positions.



(A)

	Soja	Landrace	Elite	Total	Cultivar	Gene	Chromosome	5741058	5741178	5741256	5741329	5741457	5741526	
☐	109	466	381	1064	140	Glyma.01G049300	Chr01	G|Ref	A|Ref	T|Ref	A|Ref	C|Ref	G|Ref	☐
☐	1	0	0	1	0	Glyma.01G049300	Chr01	A|G78R	A|Ref	T|Ref	A|Ref	C|Ref	G|Ref	☐
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	G|T118A	A|C144S	G|D168G	T|S178L	C|C201S	☐

(B)

	Soja	Landrace	Elite	Total	Cultivar	Gene	Chromosome	5740765	5740825	5740835	5740837	5740933
☐	6	202	233	512	104	Glyma.01G049300	Chr01	G|Ref	G|Ref	A|Ref	T|Ref	G|Ref
☐	64	245	142	488	36	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	5	15	5	25	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	5	0	1	6	0	Glyma.01G049300	Chr01	A|D13N	G|Ref	G|N36S	T|Ref	G|Ref
☐	4	1	0	5	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	4	1	0	5	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	5	0	0	5	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	2	1	0	3	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	2	0	0	2	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	A|splice_acceptor_variant
☐	2	0	0	2	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	2	0	0	2	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	2	0	0	2	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	A|V33I	G|N36S	C|splice_donor_variant&intron_variant	G|Ref
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	A|V33I	G|N36S	T|Ref	G|Ref
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	A|V33I	G|N36S	T|Ref	G|Ref
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref
☐	1	0	0	1	0	Glyma.01G049300	Chr01	G|Ref	A|V33I	G|N36S	T|Ref	G|Ref
☐	0	1	0	1	0	Glyma.01G049300	Chr01	G|Ref	G|Ref	G|N36S	T|Ref	G|Ref

Figure 5: The Allele Catalog visualization that has been integrated into the Soybean MADis tool. (A) An Allele Catalog visualization with only the user-selected combination of variant positions. (B) The full Allele Catalog visualization contains all the variant positions in a gene, and it is also the output after clicking the “View Full Allele Catalog” button.

The Soybean MADis Tool is presently available on the SoyKB web portals at https://soykb.org/SoybeanMADisTool/. This tool is freely accessible through standard web browsers, making it a user-friendly resource. It holds significant value for the soybean research community, offering the capability to conduct data analysis using the MADis algorithm.

References

Chan, Yen On, Nicholas Dietz, Shuai Zeng, Juexin Wang, Sherry Flint-Garcia, M. Nancy Salazar-Vidal, Mária Škrabišová, Kristin Bilyeu, and Trupti Joshi. 2023. "The Allele Catalog Tool: a web-based interactive tool for allele discovery and analysis." *BMC Genomics* 24 (1):107. doi: 10.1186/s12864-023-09161-3.

Joshi, Trupti, Michael R. Fitzpatrick, Shiyuan Chen, Yang Liu, Hongxin Zhang, Ryan Z. Endacott, Eric C. Gaudiello, Gary Stacey, Henry T. Nguyen, and Dong Xu. 2013. "Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding." *Nucleic Acids Research* 42 (D1):D1245-D1252. doi: 10.1093/nar/gkt905.

Joshi, Trupti, Kapil Patil, Michael R. Fitzpatrick, Levi D. Franklin, Qiuming Yao, Jeffrey R. Cook, Zheng Wang, Marc Libault, Laurent Brechenmacher, Babu Valliyodan, Xiaolei Wu, Jianlin Cheng, Gary Stacey, Henry T. Nguyen, and Dong Xu. 2012. "Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics." *BMC Genomics* 13 (1):S15. doi: 10.1186/1471-2164-13-S1-S15.

Joshi, Trupti, Jiaojiao Wang, Hongxin Zhang, Shiyuan Chen, Shuai Zeng, Bowei Xu, and Dong Xu. 2017. "The Evolution of Soybean Knowledge Base (SoyKB)." In *Plant Genomics Databases: Methods and Protocols*, edited by Aalt D. J. van Dijk, 149-159. New York, NY: Springer New York.