User Manual for

GenVarX

Genomic **Var**iation E**x**plorer

Last updated in August 2023

Table of Contents

1 Introduction 3

2 Promoter Regions Component 3

2.1 Search By Gene IDs 4

2.2 Search By Binding TFs 7

2.3 Phenotype Viewer 10

3 Copy Number Variation Component 16

3.1 Search by Gene IDs 16

3.2 Search by Accession and Copy Numbers 19

3.3 Search by Chromosome and Region 21

3.4 Phenotype Viewer 24

4 Multi-organisms Support 30

5 References 30

1 Introduction

The exponential advancement of sequencing technology has made more whole-genome re-sequencing (WGRS) data becomes available in recent years. This increase in WGRS data can bridge and address the knowledge gap that has traditionally separated genomics from phenomics. Through complex analyses of genomic variations, WGRS data holds the key to uncovering the interplay between genetic and phenotypic deviations. The genomic variations encompassing allelic and structural alterations within DNA can potentially affect regulatory mechanisms to cause gene expression changes and phenotype alterations in organisms. In order to gain more understanding of the transcription factor binding in promoter regions and copy number variations and also assist the research community at the same time, the GenVarX toolset has been developed to serve these purposes. This toolset seamlessly integrated with transcription factor binding sequence data in promoter regions, copy number variation data, SNPs data, Indels data, and phenotypic data for users to gain insights related to phenotypic differences and solve intriguing plant research questions. The GenVarX toolset is a web-based toolset developed with programming languages like HTML, CSS, JavaScript, and PHP. The toolset consists of promoter regions and copy number variation components. In each component, there are different windows for users to bring in their data to perform queries and visualize the results in both tables and figures. There are also a lot of interactive capabilities within this toolset to assist users in gaining more understanding of the analysis and statistical results. The GenVarX toolset supports soybean, rice, and *Arabidopsis* at the moment. In order to ensure seamless navigation through the GenVarX toolset, this manual has been crafted to explain the details of each component. By offering detailed insights and guidance, this manual aims to empower users to harness the full potential of the GenVarX toolset in their pursuit of groundbreaking discoveries and advancements in plant research.

2 Promoter Regions Component

The promoter regions component provides two methods for users to perform queries (Figure 1). Users who have genes of interest can use the Search By Gene IDs window to perform queries, while those in possession of binding TFs can use the Search By Binding TFs window to perform queries.

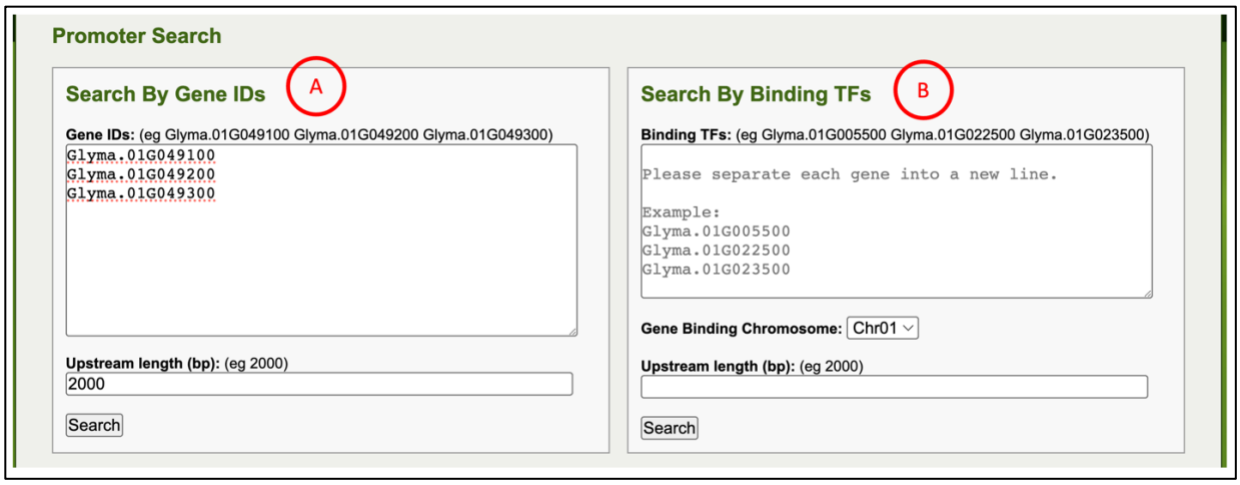


Figure 1: The promoter region component has two search windows which are (A) the Search By Gene IDs window and (B) the Search By Binding TFs window. The Search By Gene IDs window will be discussed in section 2.1, whereas the Search By Binding TFs window will be discussed in section 2.2.

2.1 Search By Gene IDs

The Search By Gene IDs window in the promoter regions component has two input boxes and a search button as shown in Figure 2.

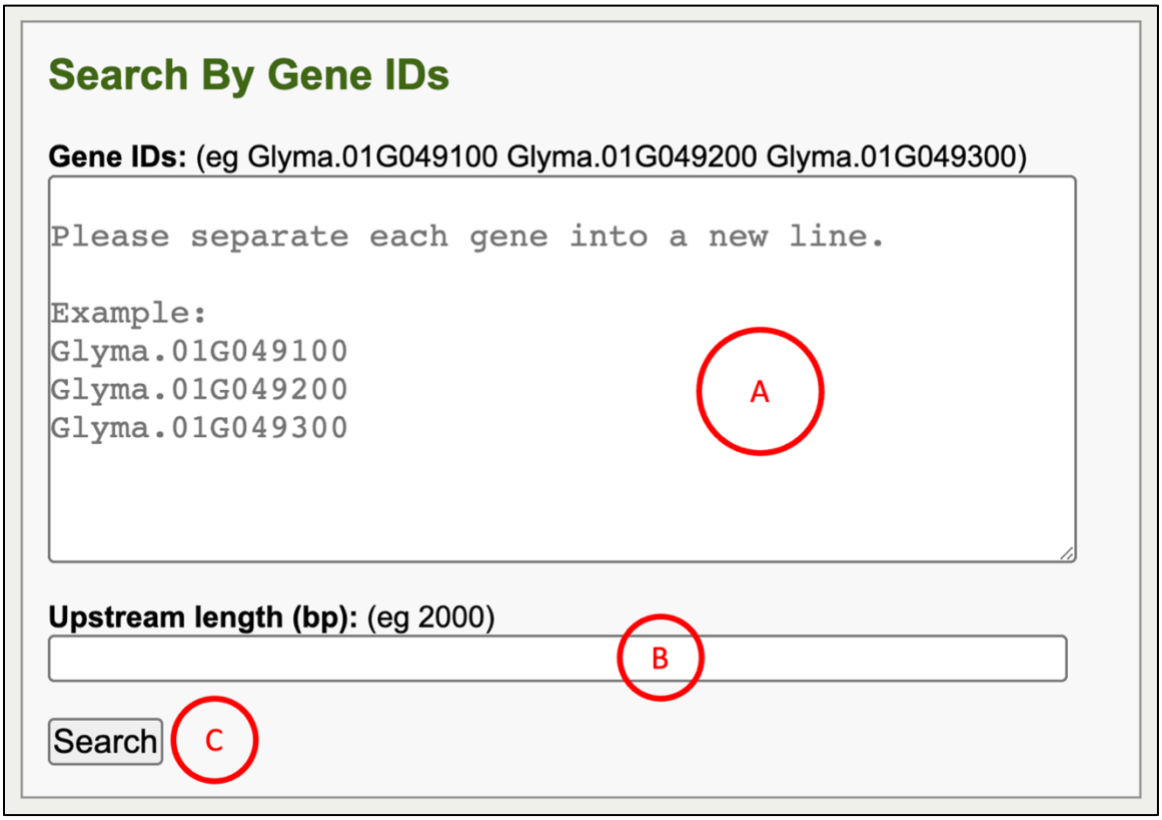


Figure 2: The Search By Gene IDs window. (A) The gene IDs input box allows users to input multiple genes, with each gene in a new line. (B) The upstream length input box can take an integer value for promoter region calculation with base-pair (bp) as the unit. (C) The search button for users to submit the query.

When users use the Search By Gene IDs window, they need to provide at least one gene (separate each gene in a new line if multiple genes are provided) and an integer value for upstream length so that they can perform a query with the search button. Users will be redirected to the result page after the query is completed. The results of each gene are shown in a table, similar to Figure 3, on the result page.

Queried Gene: Glyma.01G049300 (Chr01: 5740729 - 5741566) (+) (1)

Promoter Region: 5738728 - 5740728 (2)

(3) Gene	(4) Chromosome	(5) Start	(6) End	(7) Strand	(8) Binding_TF	(9) TF_Family	(10) Gene_Binding_Sequence	(11) Variant_Position
Glyma.01G049300	Chr01	5738941	5738961	-	Glyma.18G224500	MIKC_MADS	TTTTTCTTTTTCTTTCCCCTA	
Glyma.01G049300	Chr01	5738942	5738962	-	Glyma.18G224500	MIKC_MADS	CTTTTTCTTTTTCTTTCCCCT	
Glyma.01G049300	Chr01	5738943	5738963	-	Glyma.18G224500	MIKC_MADS	TCTTTTTCTTTTTCTTTCCCC	
Glyma.01G049300	Chr01	5738945	5738965	-	Glyma.18G224500	MIKC_MADS	TTTCTTTTTCTTTTTCTTTCC	
Glyma.01G049300	Chr01	5738947	5738967	-	Glyma.18G224500	MIKC_MADS	TTTTTCTTTTTCTTTTTCTTT	
Glyma.01G049300	Chr01	5739042	5739060	-	Glyma.04G170100	MYB	TAGGAAGTGGTTGCCAAAA	
Glyma.01G049300	Chr01	5739367	5739387	-	Glyma.18G224500	MIKC_MADS	TTTTTCTTTCTCTCTATATCT	
Glyma.01G049300	Chr01	5739371	5739391	-	Glyma.18G224500	MIKC_MADS	TTCTTTTTTCTTTCTCTCTAT	
Glyma.01G049300	Chr01	5739373	5739393	-	Glyma.18G224500	MIKC_MADS	GTTTCTTTTTTCTTTCTCTCT	
Glyma.01G049300	Chr01	5739374	5739394	-	Glyma.18G224500	MIKC_MADS	TGTTTCTTTTTTCTTTCTCTC	
Glyma.01G049300	Chr01	5739378	5739398	-	Glyma.18G224500	MIKC_MADS	TCTCTGTTTCTTTTTTCTTTC	
Glyma.01G049300	Chr01	5739380	5739400	-	Glyma.18G224500	MIKC_MADS	TTTCTCTGTTTCTTTTTTCTT	
Glyma.01G049300	Chr01	5739648	5739668	-	Glyma.18G224500	MIKC_MADS	CCTTTTTCTCCTTCTCTCCAC	
Glyma.01G049300	Chr01	5739650	5739668	-	Glyma.02G293300	C2H2	CCTTTTTCTCCTTCTCTCC	
Glyma.01G049300	Chr01	5740348	5740366	+	Glyma.02G293300	C2H2	CCTTGTCCTTCTCTTCACC	5740348
Glyma.01G049300	Chr01	5740416	5740436	+	Glyma.18G224500	MIKC_MADS	TTTTTTTTTGGTTCTTTCTTG	5740416, 5740425
Glyma.01G049300	Chr01	5740431	5740445	-	Glyma.10G142200	MYB	CCCAACCACCAAGAA	
Glyma.01G049300	Chr01	5740431	5740449	+	Glyma.04G170100	MYB	TTCTTGGTGGTTGGGCACT	
Glyma.01G049300	Chr01	5740431	5740449	-	Glyma.19G119300	MYB	AGTGCCCAACCACCAAGAA	
Glyma.01G049300	Chr01	5740433	5740446	-	Glyma.03G225200	MYB	GCCCAACCACCAAG	
Glyma.01G049300	Chr01	5740433	5740446	-	Glyma.19G222200	MYB	GCCCAACCACCAAG	
Glyma.01G049300	Chr01	5740433	5740447	-	Glyma.09G238800	MYB	TGCCCAACCACCAAG	
Glyma.01G049300	Chr01	5740701	5740720	+	Glyma.14G091200	TALE	CCCTTTGGTCTTGCTCCTTT	

Figure 3: The displayed table is generated from the outcomes obtained by querying the soybean gene. The table in the figure illustrates the information of the transcription factor binding sites for the soybean gene. Each piece of data in the table is marked with a red number to use as an index in Table 1 to explain the data in the table.

Table 1: A table to explain each piece of data and each column in Figure 3.

Index	Data/Column	Description
1	Queried Gene	The queried gene is the gene inputted by users for performing a query. The queried gene in the result page

		not only has gene ID but also has chromosome, gene start, gene end, and gene strand information showing.
2	Promoter Region	The promoter region is a region calculated based on the queried gene region and the user input upstream length.
3	Gene	This column has the queried gene.
4	Chromosome	This column has the chromosome of the queried gene.
5	Start	The starting position of the transcription factor binding sites.
6	End	The ending position of the transcription factor binding sites.
7	Strand	This column has the strand information of the binding TFs.
8	Binding TF	This column has the binding TFs that binds to the transcription factor binding sites. Each binding TF in this column is clickable.
9	TF Family	This column contains the TF families of the binding TFs.
10	Gene Binding Sequence	The reference-based gene binding sequences are shown in this column.
11	Variant Position	The variant position column shows the SNP and/or indel positions that are within the gene binding sequences.

Individuals who are interested in a specific transcription factor binding site have the option to select the corresponding binding TF. This action will lead to the display of a sequence logo figure, similar to Figure 4, on the results page. The sequence logo illustrates the potential nucleotides at each position within the gene binding sequence. Directly beneath this sequence logo, a table that provides comprehensive information including positions, a detailed breakdown of the gene binding sequence, as well as the presence of SNPs and indels, alongside accession counts from the Soy1066 data panel is presented.

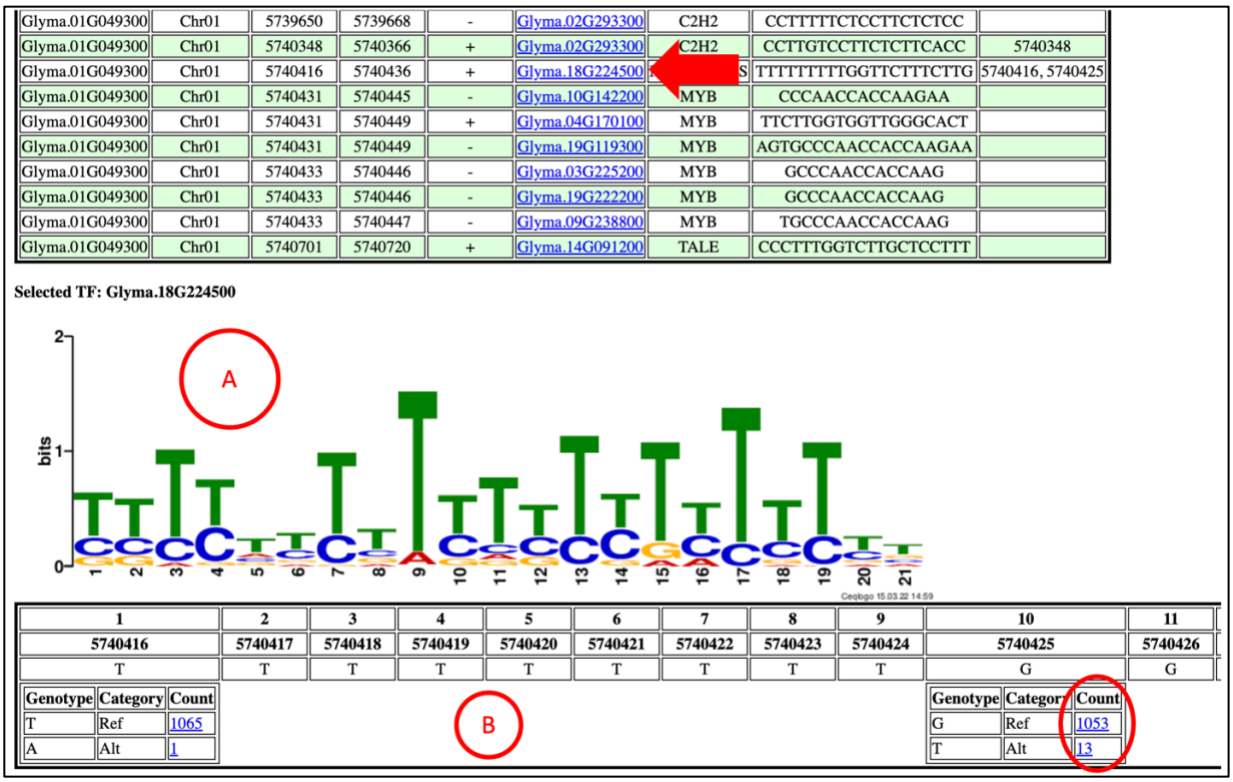


Figure 4: The sequence logo figure and the detailed gene binding sequence breakdown table will appear when users click on a gene binding TF (e.g. the gene binding TF the red arrow points to) in the table. (A) The sequence logo figure that presents the possible nucleotides of the gene binding sequence. (B) The detailed gene binding sequence breakdown table shows the positions, breakdown nucleotides, and SNPs and indels along with accession counts calculated from the Soy1066 data panel. The accession counts in the table are clickable (circled in red).

If users are interested in a SNP or indel that exists in the gene binding sequence and also in the curated Soy1066 data panel, users can click on the count to redirect to the phenotype viewer to link this SNP or indel with phenotype data. The phenotype viewer will be discussed in section 2.3 of this manual.

2.2 Search By Binding TFs

The Search By Binding TFs window in the promoter regions component has two input boxes, a dropdown menu, and a search button as shown in Figure 5.

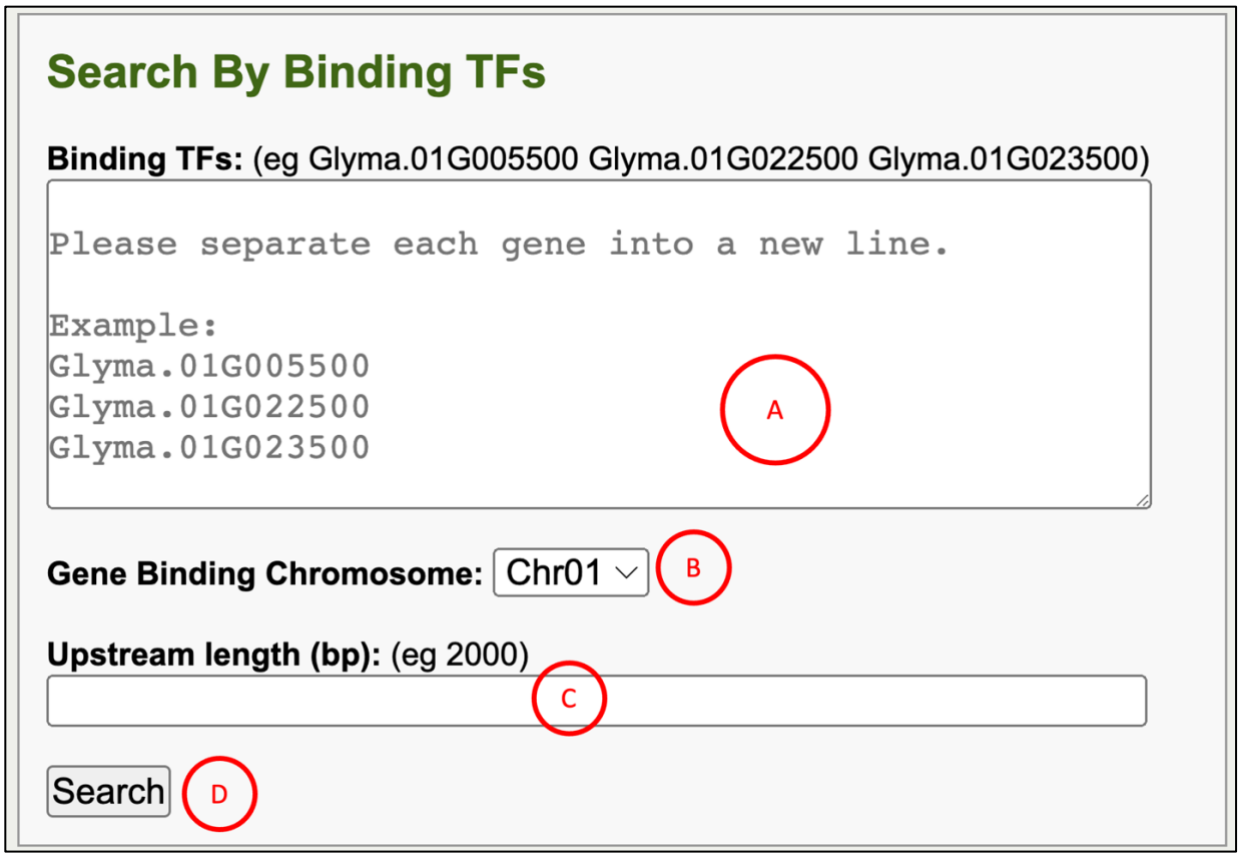
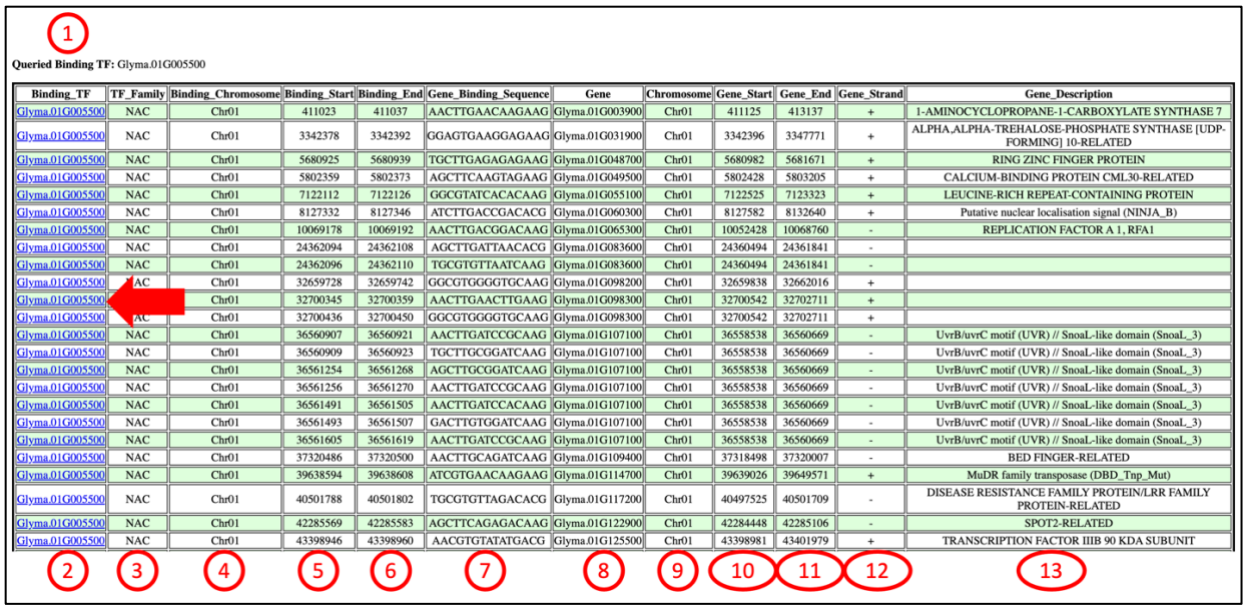


Figure 5: The Search By Binding TFs window. (A) The binding TFs input box allows users to input multiple binding TFs. Each binding TF has to be separated in a new line. (B) A dropdown menu for users to select the gene-binding chromosome, which is the chromosome that the binding TFs bind to. (C) The upstream length input box can receive an integer value as input from users to calculate and check all the promoter regions of genes. (D) The search button for users to submit the query.

Within the Search By Binding TFs window, users can input one or more binding TFs with each distinctly segmented into separate lines to ensure clarity. Moreover, users can also select the gene binding chromosome where the binding TFs bind to from the gene binding chromosome dropdown menu. Users also need to provide an integer value representing the desired upstream length, measured in base pairs, in order to check the promoter regions of genes. Upon fulfilling all the query criteria, users may initiate the search procedure by using the search button. Successful execution of the query will bring users to a results page which is as shown in Figure 6. The purpose of Figure 6 is to show all the possible genes that a binding TF can bind to.

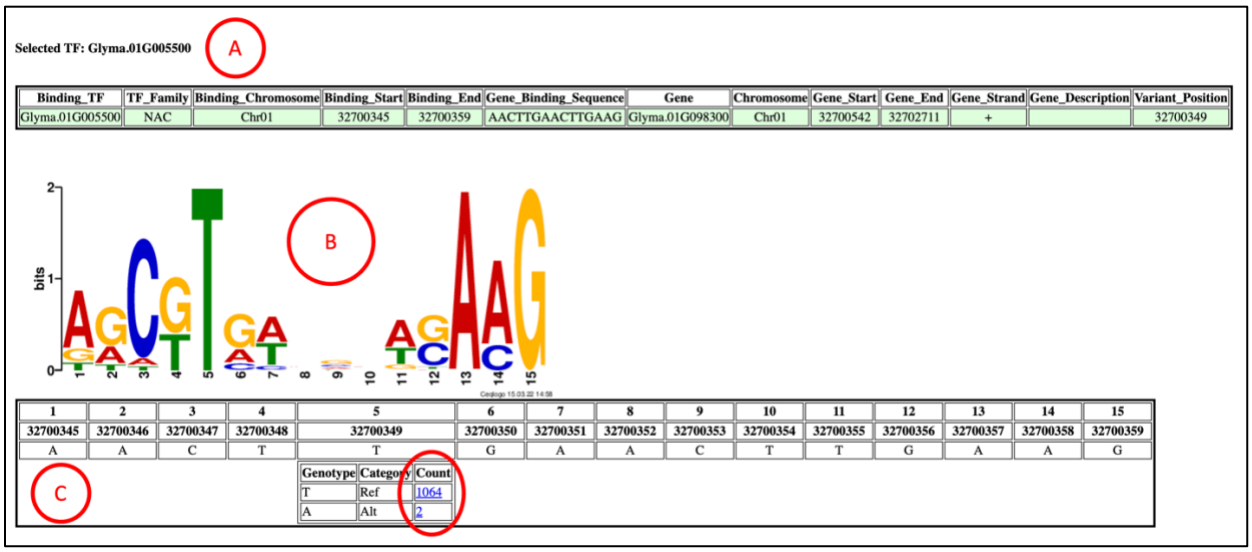


Binding_TF	TF_Family	Binding_Chromosome	Binding_Start	Binding_End	Gene_Binding_Sequence	Gene	Chromosome	Gene_Start	Gene_End	Gene_Strand	Gene_Description
Glyma.01G005500	NAC	Chr01	411023	411037	AACTTGAACAAGAAG	Glyma.01G003900	Chr01	411125	413137	+	1-AMINOCYCLOPROPANE-1-CARBOXYLATE SYNTHASE 7
Glyma.01G005500	NAC	Chr01	3342378	3342392	GGAGTGAAGGAGAAG	Glyma.01G031900	Chr01	3342396	3347771	+	ALPHA,ALPHA-TREHALOSE-PHOSPHATE SYNTHASE [UDP-FORMING] 10-RELATED
Glyma.01G005500	NAC	Chr01	5680925	5680939	TGCTTGAGAGAGAAG	Glyma.01G048700	Chr01	5680982	5681671	+	RING ZINC FINGER PROTEIN
Glyma.01G005500	NAC	Chr01	5802359	5802373	AGCTTCAAGTAGAAG	Glyma.01G049500	Chr01	5802428	5803205	+	CALCIUM-BINDING PROTEIN CML30-RELATED
Glyma.01G005500	NAC	Chr01	7122112	7122126	GGCGTATCACACAAG	Glyma.01G055100	Chr01	7122525	7123323	+	LEUCINE-RICH REPEAT-CONTAINING PROTEIN
Glyma.01G005500	NAC	Chr01	8127332	8127346	ATCTTGACCGACACG	Glyma.01G060300	Chr01	8127582	8132640	+	Putative nuclear localisation signal (NINJA_B)
Glyma.01G005500	NAC	Chr01	10069178	10069192	AACTTGACGGACAAG	Glyma.01G065300	Chr01	10052428	10068760	-	REPLICATION FACTOR A 1, RFA1
Glyma.01G005500	NAC	Chr01	24362094	24362108	AGCTTGATTAACACG	Glyma.01G083600	Chr01	24360494	24361841	-	
Glyma.01G005500	NAC	Chr01	24362096	24362110	TGCGTGTTAATCAAG	Glyma.01G083600	Chr01	24360494	24361841	-	
Glyma.01G005500	NAC	Chr01	32659728	32659742	GGCGTGGGGTGCAAG	Glyma.01G098200	Chr01	32659838	32662016	+	
Glyma.01G005500	NAC	Chr01	32700345	32700359	AACTTGAACTTGAAG	Glyma.01G098300	Chr01	32700542	32702711	+	
Glyma.01G005500	NAC	Chr01	32700436	32700450	GGCGTGGGGTGCAAG	Glyma.01G098300	Chr01	32700542	32702711	+	
Glyma.01G005500	NAC	Chr01	36560907	36560921	AACTTGATCCGCAAG	Glyma.01G107100	Chr01	36558538	36560669	-	UvrB/uvrC motif (UVR) // SnoaL-like domain (SnoaL_3)
Glyma.01G005500	NAC	Chr01	36560909	36560923	TGCTTGCGGATCAAG	Glyma.01G107100	Chr01	36558538	36560669	-	UvrB/uvrC motif (UVR) // SnoaL-like domain (SnoaL_3)
Glyma.01G005500	NAC	Chr01	36561254	36561268	AGCTTGCGGATCAAG	Glyma.01G107100	Chr01	36558538	36560669	-	UvrB/uvrC motif (UVR) // SnoaL-like domain (SnoaL_3)
Glyma.01G005500	NAC	Chr01	36561256	36561270	AACTTGATCCGCAAG	Glyma.01G107100	Chr01	36558538	36560669	-	UvrB/uvrC motif (UVR) // SnoaL-like domain (SnoaL_3)
Glyma.01G005500	NAC	Chr01	36561491	36561505	AACTTGATCCACAAG	Glyma.01G107100	Chr01	36558538	36560669	-	UvrB/uvrC motif (UVR) // SnoaL-like domain (SnoaL_3)
Glyma.01G005500	NAC	Chr01	36561493	36561507	GACTTGTGGATCAAG	Glyma.01G107100	Chr01	36558538	36560669	-	UvrB/uvrC motif (UVR) // SnoaL-like domain (SnoaL_3)
Glyma.01G005500	NAC	Chr01	36561605	36561619	AACTTGATCCGCAAG	Glyma.01G107100	Chr01	36558538	36560669	-	UvrB/uvrC motif (UVR) // SnoaL-like domain (SnoaL_3)
Glyma.01G005500	NAC	Chr01	37320486	37320500	AACTTGCAGATCAAG	Glyma.01G109400	Chr01	37318498	37320007	-	BED FINGER-RELATED
Glyma.01G005500	NAC	Chr01	39638594	39638608	ATCGTGAACAAGAAG	Glyma.01G114700	Chr01	39639026	39649571	+	MuDR family transposase (DBD_Tnp_Mut)
Glyma.01G005500	NAC	Chr01	40501788	40501802	TGCGTGTTAGACACG	Glyma.01G117200	Chr01	40497525	40501709	-	DISEASE RESISTANCE FAMILY PROTEIN/LRR FAMILY PROTEIN-RELATED
Glyma.01G005500	NAC	Chr01	42285569	42285583	AGCTTCAGAGACAAG	Glyma.01G122900	Chr01	42284448	42285106	-	SPOT2-RELATED
Glyma.01G005500	NAC	Chr01	43398946	43398960	AACGTGTATATGACG	Glyma.01G125500	Chr01	43398981	43401979	+	TRANSCRIPTION FACTOR IIIB 90 KDA SUBUNIT

Figure 6: The result page rendered as the query from the Search By Binding TFs window is successfully executed. The information in each column will be further explained in Table 2. The binding TFs in the table are clickable. Clicking on the binding TF to which the red arrow points will be redirected to another result page as presented in Figure 7.

Table 2: A table to explain each piece of data and each column in Figure 6.

Index	Data/Column	Description
1	Queried Binding TF	The binding TF of which the query and the result table are based on.
2	Binding TF	The binding TFs in the query.
3	TF Family	The TF family column shows the families of the binding TFs .
4	Binding Chromosome	The chromosomes that the binding TFs bind to.
5	Binding Start	The start positions where the binding TFs bind to.
6	Binding End	The end positions where the binding TFs bind to.
7	Gene Binding Sequence	The gene binding sequences of the reference.
8	Gene	The genes to which the binding TFs attach upstream.
9	Chromosome	The chromosomes of the genes.
10	Gene Start	The start positions of the genes.
11	Gene End	The end positions of the genes.
12	Gene Strand	The strands of the genes.
13	Gene Description	The gene descriptions of the genes.



Binding_TF	TF_Family	Binding_Chromosome	Binding_Start	Binding_End	Gene_Binding_Sequence	Gene	Chromosome	Gene_Start	Gene_End	Gene_Strand	Gene_Description	Variant_Position
Glyma.01G005500	NAC	Chr01	32700345	32700359	AACTTGAACTTGAAG	Glyma.01G098300	Chr01	32700542	32702711	+		32700349

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
32700345	32700346	32700347	32700348	32700349	32700350	32700351	32700352	32700353	32700354	32700355	32700356	32700357	32700358	32700359
A	A	C	T	T	G	A	A	C	T	T	G	A	A	G

Genotype	Category	Count
T	Ref	1064
A	Alt	2

Figure 7: The result page of the selected binding TF. (A) The information of the selected binding TF. (B) The sequence logo figure that presents the possible nucleotides of the gene binding sequence. (C) The detailed gene binding sequence breakdown table shows the positions, breakdown nucleotides, and SNPs and indels along with accession counts calculated from the Soy1066 data panel. The accession counts in the table are clickable (circled in red).

If users are interested in a SNP or indel that exists in the gene binding sequence and also in the curated Soy1066 data panel, users can click on the count to redirect to the phenotype viewer to link this SNP or indel with phenotype data. The phenotype viewer will be discussed in section 2.3 of the manual.

2.3 Phenotype Viewer

The purpose of the phenotype viewer in the promoter regions component is to provide capabilities for users to connect genotypes with phenotypes and visualize the phenotype data in the form of tables and figures. When users click on an accession count of a genotype, which is circled in red in Figure 4 or 7, users will be redirected to the phenotype viewer. The phenotype viewer (Figure 8) has an accordion with different tabs to include coordinate information, genotype checkboxes, and phenotype checkboxes that are separated into different categories under different tabs. Under the accordion, there are several buttons with different functionalities for users to interact with the phenotype viewer to get their desired data. The functionalities of the buttons are displayed in Table 3.

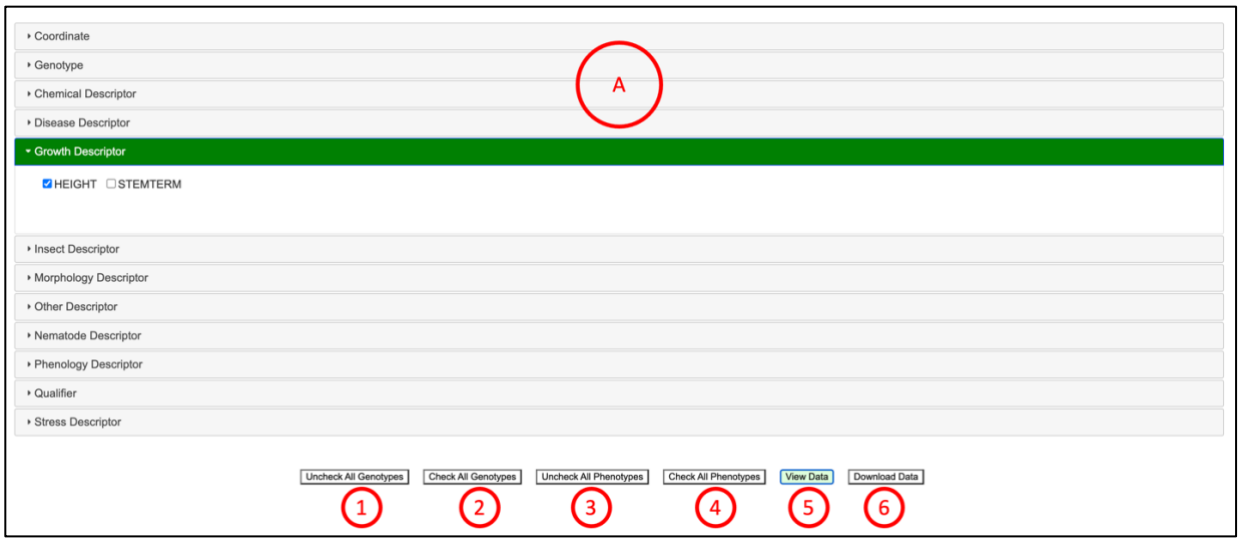


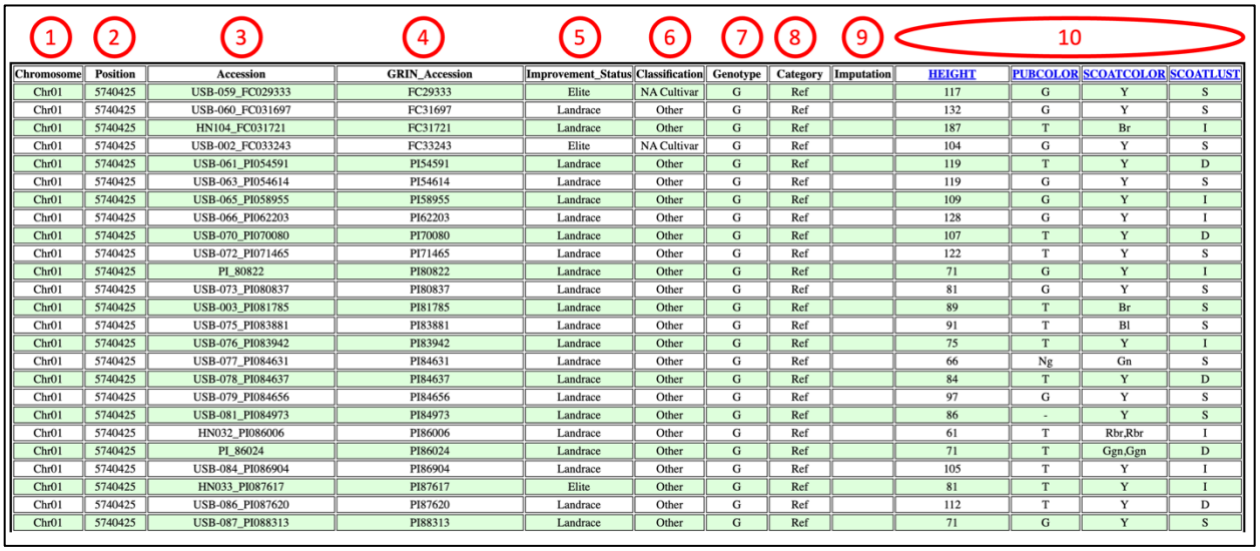
Figure 8: The accordion and buttons in the phenotype viewer. (A) The accordion section has different tabs, and each tab is clickable to expand. The coordinate tab has the chromosome and position information of the selected genotype. The genotype tab has all the possible genotypes in the form of checkboxes for users to select. The rest of the tabs correspond to different phenotype categories, and each tab has phenotype trait checkboxes for users to select phenotypes of interest. The buttons under the accordion are illustrated in Table 3.

Table 3: A table to explain the functionalities of the buttons under the accordion in Figure 8.

Index	Button	Description
1	Uncheck All Genotypes	This button unchecks all the genotype checkboxes in the genotype tab of the accordion.
2	Check All Genotypes	This button checks all the genotype checkboxes in the genotype tab of the accordion.
3	Uncheck All Phenotypes	This button unchecks all the phenotype checkboxes in all the phenotype tabs of the accordion.
4	Check All Phenotypes	This button checks all the phenotype checkboxes in all the phenotype tabs of the accordion.
5	View Data	This button can be clicked to query genotype and phenotype information and show the results in a table. An example of the table is shown in Figure 9. This button will not work, and an error message is prompted to users if none of the genotypes in the genotype tab of the accordion is selected.
6	Download data	This button can be clicked to query genotype and phenotype information and download it as a static file.

When users land on the phenotype viewer (Figure 8), they can select the genotype of interest by checking the genotype checkboxes in the genotype tab of the accordion. Besides that, users

can also use the “Uncheck All Genotypes” and “Check All Genotypes” buttons to uncheck or check all genotypes. Apart from that, users can also select the phenotypes of interest by using the checkboxes in the phenotype tabs of the accordion. Another option to select phenotypes is to use the “Uncheck All Phenotypes” and “Check All Phenotypes” to uncheck or check all phenotypes. Users must choose at least one genotype and phenotype in order for the “View Data” button and “Download Data” button to function properly. When users click on the “View Data” button, a query will be submitted to the backend code to get the genotype and phenotype data and present the data in a table (Figure 9). On the other hand, if the “Download Data” button is pressed, a query will be submitted to the backend code to get the genotype and phenotype data and then transfer the data in the form of a static file to the user end. The genotype and phenotype data in the table or static file consists of at least 10 columns in the case of soybean data (Figure 9). The details of these columns are demonstrated in Table 4.



Chromosome	Position	Accession	GRIN_Accession	Improvement_Status	Classification	Genotype	Category	Imputation	HEIGHT	PUBCOLOR	SCOATCOLOR	SCOATLUST
Chr01	5740425	USB-059_FC029333	FC29333	Elite	NA Cultivar	G	Ref		117	G	Y	S
Chr01	5740425	USB-060_FC031697	FC31697	Landrace	Other	G	Ref		132	G	Y	S
Chr01	5740425	HN104_FC031721	FC31721	Landrace	Other	G	Ref		187	T	Br	I
Chr01	5740425	USB-002_FC033243	FC33243	Elite	NA Cultivar	G	Ref		104	G	Y	S
Chr01	5740425	USB-061_PI054591	PI54591	Landrace	Other	G	Ref		119	T	Y	D
Chr01	5740425	USB-063_PI054614	PI54614	Landrace	Other	G	Ref		119	G	Y	S
Chr01	5740425	USB-065_PI058955	PI58955	Landrace	Other	G	Ref		109	G	Y	I
Chr01	5740425	USB-066_PI062203	PI62203	Landrace	Other	G	Ref		128	G	Y	I
Chr01	5740425	USB-070_PI070080	PI70080	Landrace	Other	G	Ref		107	T	Y	D
Chr01	5740425	USB-072_PI071465	PI71465	Landrace	Other	G	Ref		122	T	Y	S
Chr01	5740425	PI_80822	PI80822	Landrace	Other	G	Ref		71	G	Y	I
Chr01	5740425	USB-073_PI080837	PI80837	Landrace	Other	G	Ref		81	G	Y	S
Chr01	5740425	USB-003_PI081785	PI81785	Landrace	Other	G	Ref		89	T	Br	S
Chr01	5740425	USB-075_PI083881	PI83881	Landrace	Other	G	Ref		91	T	Bl	S
Chr01	5740425	USB-076_PI083942	PI83942	Landrace	Other	G	Ref		75	T	Y	I
Chr01	5740425	USB-077_PI084631	PI84631	Landrace	Other	G	Ref		66	Ng	Gn	S
Chr01	5740425	USB-078_PI084637	PI84637	Landrace	Other	G	Ref		84	T	Y	D
Chr01	5740425	USB-079_PI084656	PI84656	Landrace	Other	G	Ref		97	G	Y	S
Chr01	5740425	USB-081_PI084973	PI84973	Landrace	Other	G	Ref		86	-	Y	S
Chr01	5740425	HN032_PI086006	PI86006	Landrace	Other	G	Ref		61	T	Rbr,Rbr	I
Chr01	5740425	PI_86024	PI86024	Landrace	Other	G	Ref		71	T	Ggn,Ggn	D
Chr01	5740425	USB-084_PI086904	PI86904	Landrace	Other	G	Ref		105	T	Y	I
Chr01	5740425	HN033_PI087617	PI87617	Elite	Other	G	Ref		81	T	Y	I
Chr01	5740425	USB-086_PI087620	PI87620	Landrace	Other	G	Ref		112	T	Y	D
Chr01	5740425	USB-087_PI088313	PI88313	Landrace	Other	G	Ref		71	G	Y	S

Figure 9: The genotype and phenotype data rendered in a tabular form after the “View Data” button is clicked. There are at least 10 columns in the table. The details of each column are listed in Table 4.

Table 4: A table explains the columns in the table that hold the genotype and phenotype data.

Index	Column	Description
1	Chromosome	The chromosome of the genotypes.
2	Position	The position of the genotypes.
3	Accession	The accession names of the accessions that are in the Soy1066 data panel.
4	GRIN_Accession	The alternative accession names that collected from the GRIN database.
5	Improvement_Status	The improvement status of the soybean accessions. There are a few categories such as Soja, Elite, and Landrace.

6	Classification	This column has the classifications of the soybean accessions, which include North America (NA) Ancestor, NA Cultivar, and Other.
7	Genotype	The genotype selected by users in the genotype tab of the accordion.
8	Category	This column indicates whether a genotype as a reference allele (Ref) or an alternate allele (Alt).
9	Imputation	This column shows whether a genotype is an imputed genotype. A “+” sign indicates that the genotype is an imputed genotype.
10	Phenotype columns selected by users	These columns contain the phenotypes selected by users and also the phenotypic measurements correspond to the accessions. Each phenotype header is clickable to redirect to the phenotype data visualization page to visualize the distributions of the phenotype data (Figure 10 and 11).

Each phenotype column header in the genotype and phenotype data table is clickable. When users click on a phenotype column header, a new query is initiated to fetch and process the phenotype data. Consequently, users are redirected to the phenotype data visualization page. Figure 10 and Figure 11 serve as the phenotype data visualization pages, with the appropriate one shown based on the data type (quantitative or qualitative). During the processing step, the algorithm first determines whether the phenotype data is quantitative or qualitative. Quantitative data is presented in a violin plot (Figure 10B), while qualitative data is displayed in a bar plot (Figure 11B). In Figure 10C and Figure 11C, the methods for summarizing the phenotype data differ based on the data type (quantitative or qualitative). For quantitative data, the summary is derived solely from genotypes and accessions with or without phenotype data, summarized in the form of accession counts (Figure 10C). Conversely, qualitative data includes not only counts of accessions with or without phenotype data, but also provides a breakdown of each phenotype category with corresponding counts and percentages (Figure 11C). Furthermore, irrespective of whether the phenotype data is quantitative or qualitative, the improvement status summary bar plots are accessible. These plots serve to summarize the distribution of accessions across genotypes and improvement status categories (Figure 10D and 11D).

Figure 10: The phenotype data visualization page is displayed when the phenotype data is quantitative data. (A) A table to show the selected genotypes and phenotype. (B) A violin plot for users to visualize the distribution of the quantitative phenotype data. (C) A summary table that summarizes the accessions based on genotypes and with or without phenotype data. (D) An improvement status summary bar plot summarizes the accessions based on genotypes and improvement status.

Figure 11: The phenotype data visualization page is displayed when the phenotype data is qualitative data. (A) A table to show the selected genotypes and phenotype. (B) A bar plot for users to visualize the distribution of the qualitative phenotype data. (C) A summary table not only summarizes the accessions based on genotypes and with or without phenotype data but also contains the breakdown of counts and percentages for each qualitative category. (D) An improvement status summary bar plot summarizes the accessions based on genotypes and improvement status.

3 Copy Number Variation Component

The copy number variation component offers three methods for users to conduct queries (Figure 12). Users with specific genes of interest can utilize the Search By Gene IDs window to perform queries. Meanwhile, users intrigued by accessions and their corresponding copy numbers can employ the Search By Accession and Copy Numbers window for their queries. For those interested in a particular chromosome and region, the Search By Chromosome and Region window is available to facilitate queries.

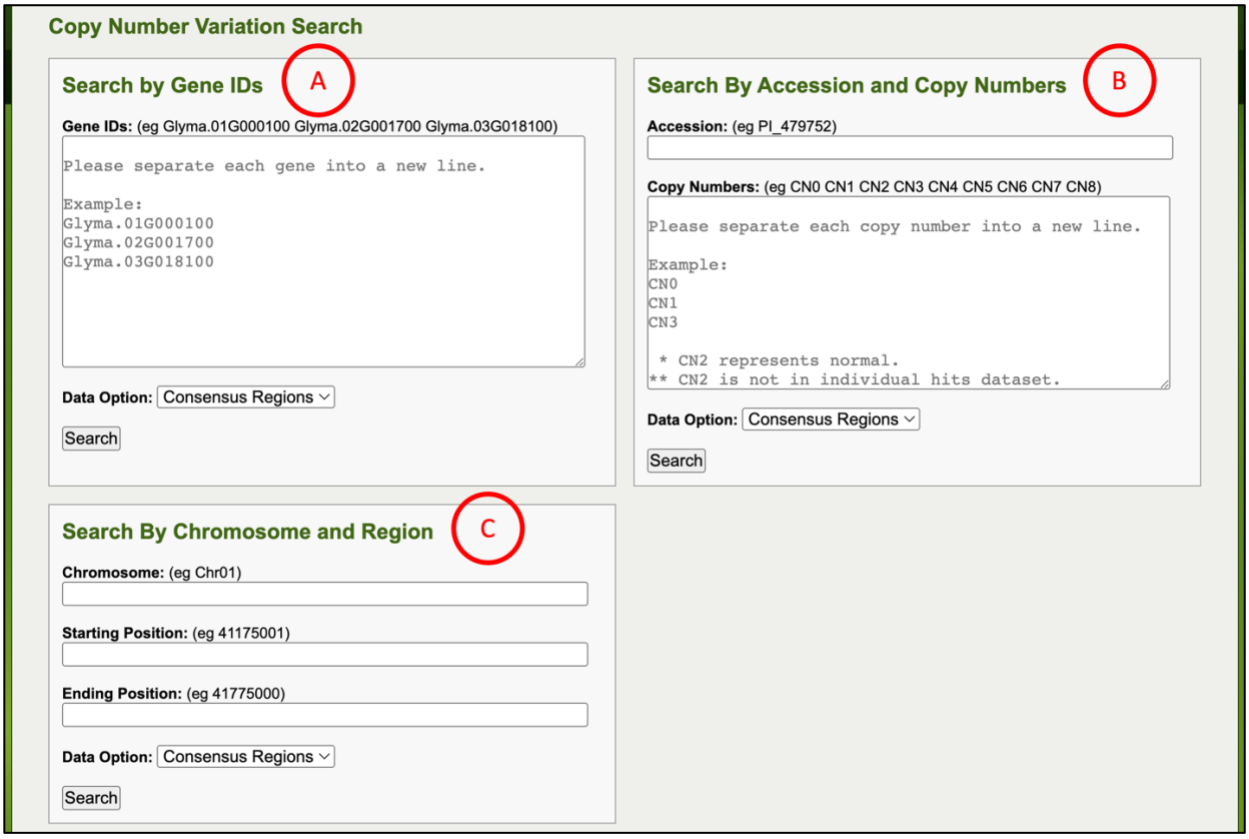


Figure 12: The copy number variation component comprises three search windows: (A) the Search By Gene IDs window, (B) the Search By Accession and Copy Numbers window, and (C) the Search By Chromosome and Region window. We will dive into the details of each window in the following sections: section 3.1 will cover the Search By Gene IDs window, section 3.2 will explore the Search By Accession and Copy Numbers window, and section 3.3 will provide insights into the Search By Chromosome and Region window.

3.1 Search by Gene IDs

The Search By Gene IDs window in the copy number variation component has an input box, a dropdown menu, and a search button as shown in Figure 13.

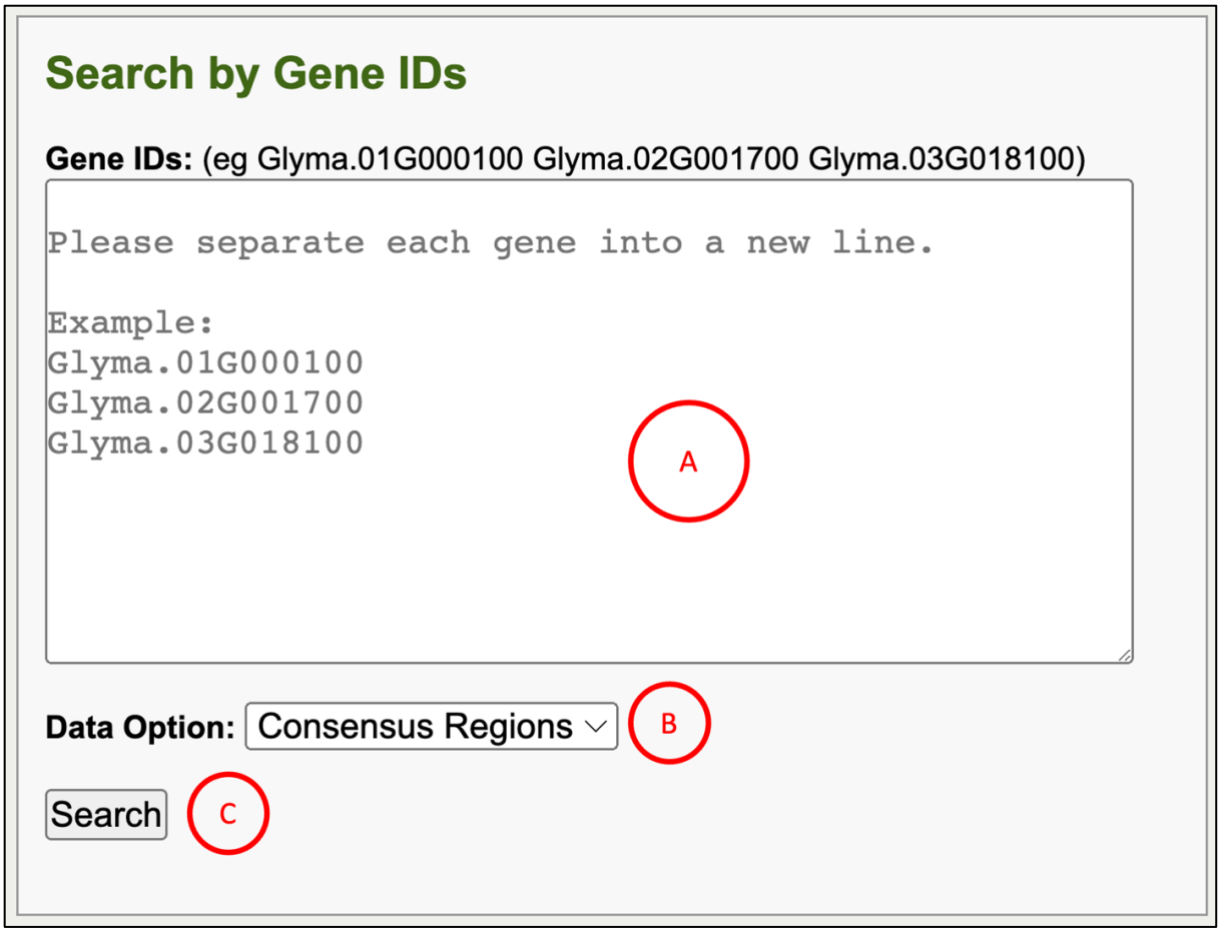
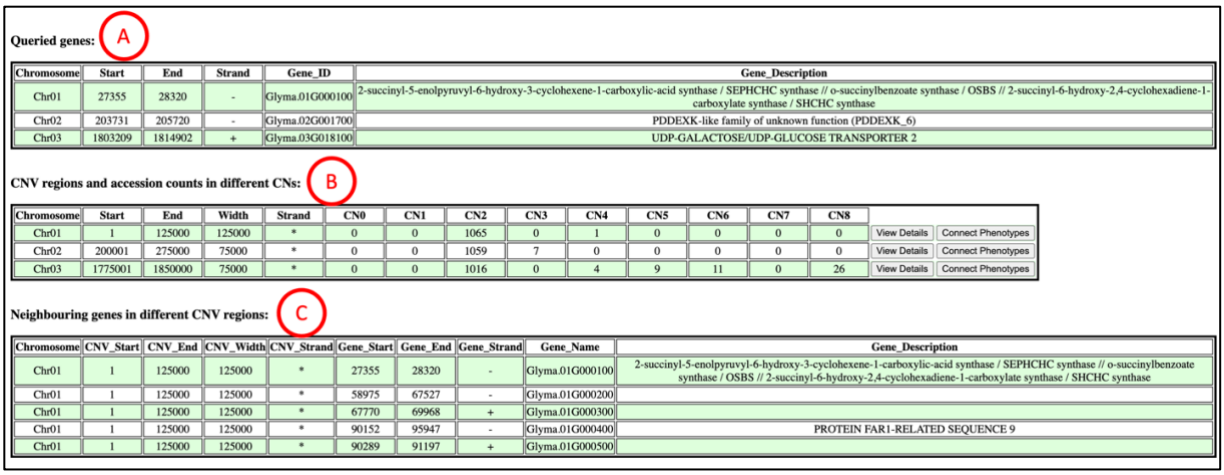


Figure 13: The Search By Gene IDs window. (A) The gene IDs input box allows users to input multiple genes, with each gene in a new line. (B) The data option is for users to select the copy number variation data type of interest for the query. (C) The search button for users to submit the query.

Upon using the Search By Gene IDs window, users are required to input a minimum of one gene. If multiple genes are provided, they should be separated by new lines. The data option dropdown menu contains two options which are the individual hits and consensus regions. Both individual hits and consensus regions are the copy number variation results analyzed using the Soy1066 data panel with the cn.MOPS package [1]. The individual hits are the outputs analyzed for each accession, and the consensus regions are copy number variation regions summarized across all accessions [1]. From these two options, users are required to select one before they can click on the submit button to initiate a query and be redirected to the results page (Figure 14). There are three sections on the result page which are the queried genes section (Figure 14A), the copy number variation regions and accessions counts section (Figure 14B), and the neighboring genes section (Figure 14C).



Queried genes: A

Chromosome	Start	End	Strand	Gene_ID	Gene_Description
Chr01	27355	28320	-	Glyma.01G000100	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid synthase / SEPHCHC synthase // o-succinylbenzoate synthase / OSBS // 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase / SHCHC synthase
Chr02	203731	205720	-	Glyma.02G001700	PDDEXK-like family of unknown function (PDDEXK_6)
Chr03	1803209	1814902	+	Glyma.03G018100	UDP-GALACTOSE/UDP-GLUCOSE TRANSPORTER 2

CNV regions and accession counts in different CNs: B

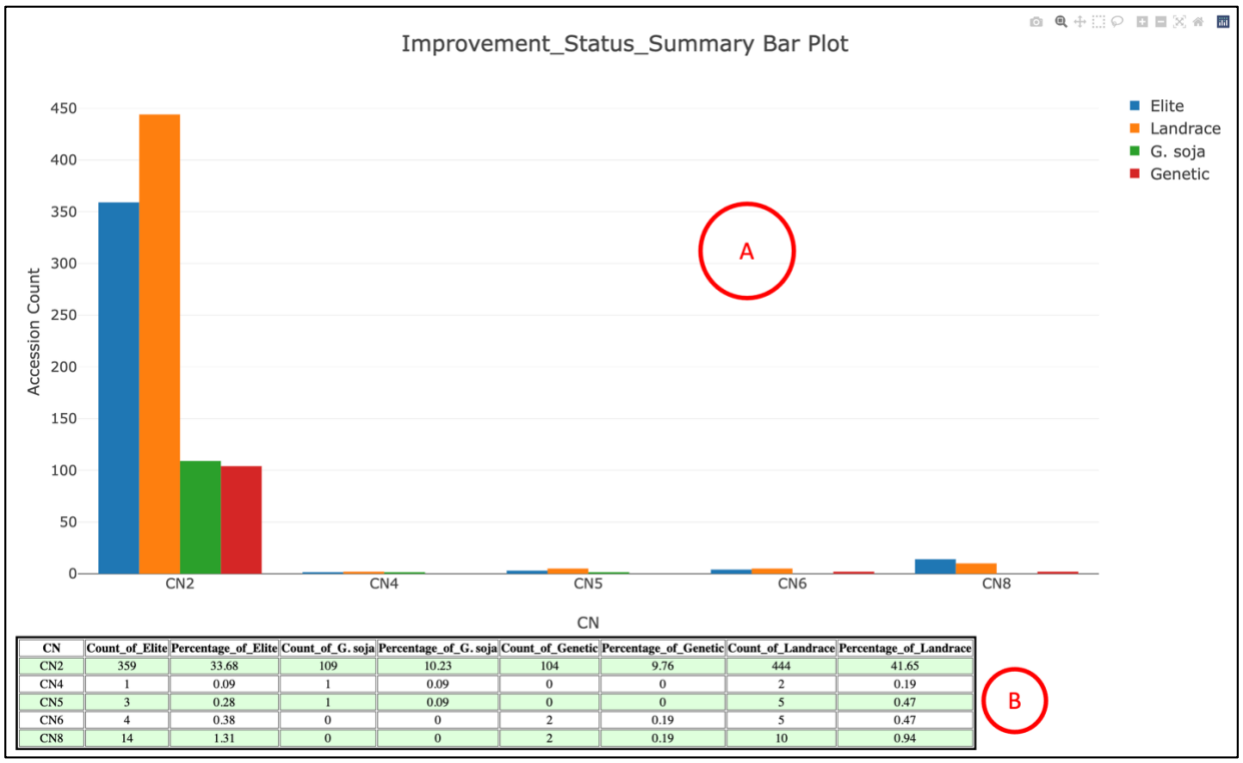
Chromosome	Start	End	Width	Strand	CN0	CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8		
Chr01	1	125000	125000	*	0	0	1065	0	1	0	0	0	0	View Details	Connect Phenotypes
Chr02	200001	275000	75000	*	0	0	1059	7	0	0	0	0	0	View Details	Connect Phenotypes
Chr03	1775001	1850000	75000	*	0	0	1016	0	4	9	11	0	26	View Details	Connect Phenotypes

Neighbouring genes in different CNV regions: C

Chromosome	CNV_Start	CNV_End	CNV_Width	CNV_Strand	Gene_Start	Gene_End	Gene_Strand	Gene_Name	Gene_Description
Chr01	1	125000	125000	*	27355	28320	-	Glyma.01G000100	2-succinyl-5-enolpyruvyl-6-hydroxy-3-cyclohexene-1-carboxylic-acid synthase / SEPHCHC synthase // o-succinylbenzoate synthase / OSBS // 2-succinyl-6-hydroxy-2,4-cyclohexadiene-1-carboxylate synthase / SHCHC synthase
Chr01	1	125000	125000	*	58975	67527	-	Glyma.01G000200	
Chr01	1	125000	125000	*	67770	69968	+	Glyma.01G000300	
Chr01	1	125000	125000	*	90152	95947	-	Glyma.01G000400	PROTEIN FAR1-RELATED SEQUENCE 9
Chr01	1	125000	125000	*	90289	91197	+	Glyma.01G000500	

Figure 14: The result page rendered when users make a query using the Search By Gene IDs window in the copy number variation component. There are three sections on this result page which are (A) the queried genes section, (B) the copy number variation regions and accessions counts section, and (C) the neighboring genes section.

The queried genes section shows information on the genes inputted by the users. The information includes chromosomes, start positions, end positions, strands, gene IDs, and gene descriptions. The copy number variation regions and accessions counts section contains a table with chromosomes, start positions, end positions, and strand information of the copy number variation regions. Each copy number variation region has accession counts of the Soy1066 data panel that are categorized based on copy numbers ranging from CN0 to CN8 by the copy number variation analysis using the cn.MOPS package. Among the copy numbers, CN0 and CN1 represent loss, CN2 represent normal, and CN3 to CN8 represent gain. Furthermore, each copy number variation region in this section also has two buttons: the “View Details” button is for improvement status distribution visualization, and the “Connect Phenotypes” button is to connect copy numbers with phenotypes in the phenotype viewer. The improvement status distribution visualization is a new web page to show the improvement status distribution similar to Figure 15. The phenotype viewer will be discussed in section 3.4 of this manual. The neighboring genes section demonstrates the information of all the genes that are enclosed in each copy number variation region. For each copy number variation region, there is a table to show the chromosome, copy number variation region start position, end position, and strand. Within that table, the information of each gene, such as gene start position, gene end position, gene name, and gene description, that is enclosed within the copy number variation region are also listed in each row.



CN	Count_of_Elite	Percentage_of_Elite	Count_of_G. soja	Percentage_of_G. soja	Count_of_Genetic	Percentage_of_Genetic	Count_of_Landrace	Percentage_of_Landrace
CN2	359	33.68	109	10.23	104	9.76	444	41.65
CN4	1	0.09	1	0.09	0	0	2	0.19
CN5	3	0.28	1	0.09	0	0	5	0.47
CN6	4	0.38	0	0	2	0.19	5	0.47
CN8	14	1.31	0	0	2	0.19	10	0.94

B

Figure 15: The improvement status distribution visualization generated based on the copy numbers, improvement status, and accessions in the Soy1066 data panel. (A) An improvement status summary bar plot summarizes the accessions based on copy numbers and improvement status. (B) A table that summarizes the accession counts and percentages according to copy numbers and improvement status.

3.2 Search by Accession and Copy Numbers

The Search By Accession and Copy Numbers window in the copy number variation component has two input boxes, a dropdown menu, and a search button as shown in Figure 16.

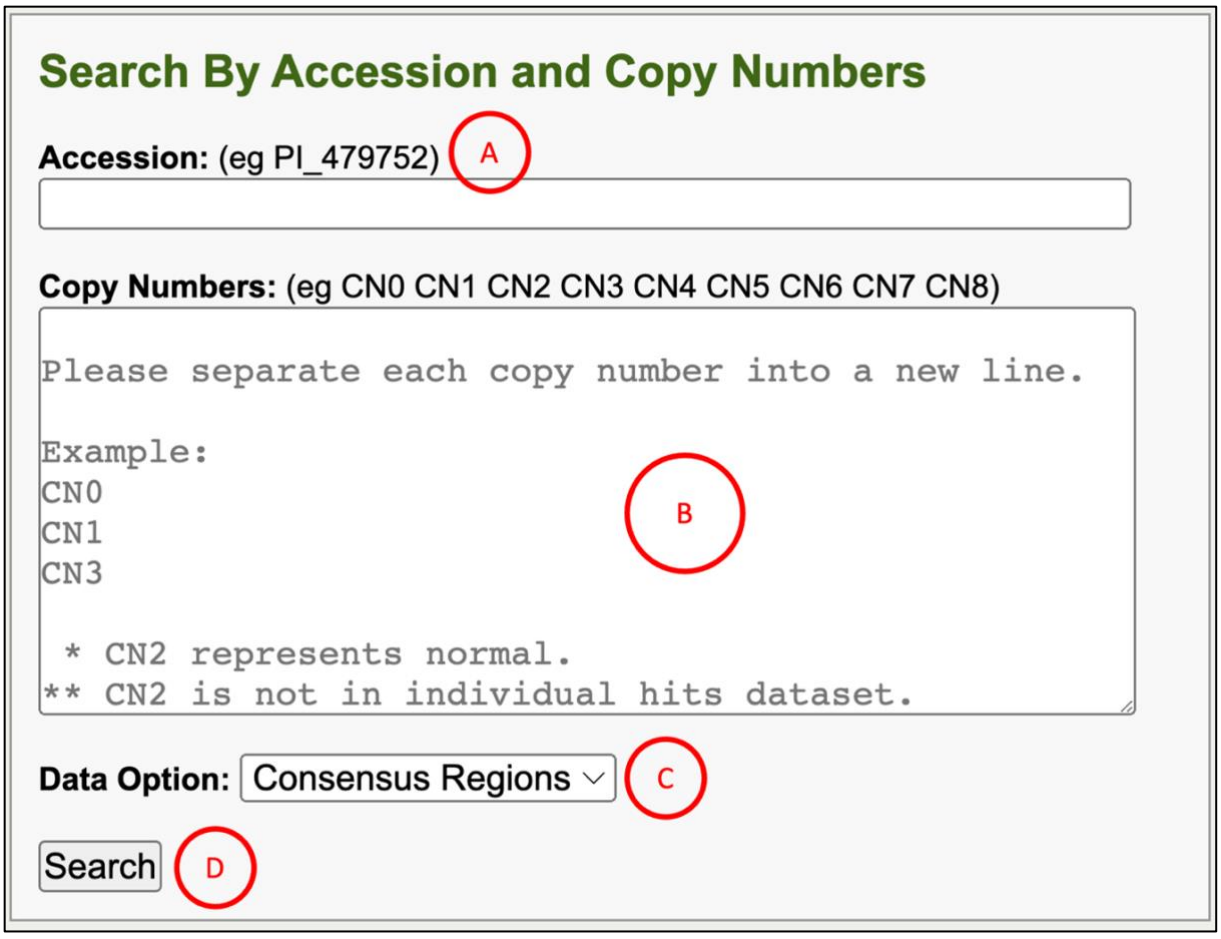


Figure 16: The Search By Accession and Copy Numbers window. (A) The accession input box that only allows one accession as input. (B) The copy numbers input box that allows users to input multiple copy numbers. Each copy number has to be separated into a new line. (C) The data option is for users to select the copy number variation data type of interest for the query. (D) The search button for users to submit the query.

Before initiating a query, users are required to input an accession into the designated accession input box and provide at least one copy number. Additionally, users have the option to select either individual hits or consensus regions from the data option dropdown menu. Once all mandatory fields are completed, users can proceed by clicking the search button to execute their query. Users will be redirected to the result page to visualize the results in tabular form (Figure 17).

CNV regions and CNs of accession PI_479752:

Chromosome	Start	End	Width	Strand	Accession	CN
Chr01	19375001	19500000	125000	*	PI_479752	CN0
Chr01	37825001	38400000	575000	*	PI_479752	CN0
Chr03	8775001	9025000	250000	*	PI_479752	CN0
Chr03	13300001	13525000	225000	*	PI_479752	CN0
Chr03	14850001	16200000	1350000	*	PI_479752	CN0
Chr05	24350001	24725000	375000	*	PI_479752	CN0
Chr06	30350001	30775000	425000	*	PI_479752	CN0
Chr07	8800001	8900000	100000	*	PI_479752	CN0
Chr08	24175001	24575000	400000	*	PI_479752	CN0
Chr08	33000001	33075000	75000	*	PI_479752	CN0
Chr08	34975001	35100000	125000	*	PI_479752	CN0
Chr09	14775001	15025000	250000	*	PI_479752	CN0
Chr09	17000001	17300000	300000	*	PI_479752	CN0
Chr09	22650001	26900000	4250000	*	PI_479752	CN0
Chr09	28450001	29025000	575000	*	PI_479752	CN0
Chr09	30150001	30250000	100000	*	PI_479752	CN0
Chr10	13950001	14600000	650000	*	PI_479752	CN0
Chr10	34275001	35300000	1025000	*	PI_479752	CN0
①	②	③	④	⑤	⑥	⑦

Figure 17: A table on the result page that shows all the copy number variation regions based on the accession and copy numbers inputted by users. Each column in this table is further described in Table 5.

Table 5: A table explains the columns in the table that hold the copy number variation regions of an accession with different copy numbers.

Index	Column	Description
1	Chromosome	The chromosomes of the copy number variation regions.
2	Start	The start position of the copy number variation regions.
3	End	The end position of the copy number variation regions.
4	Width	The width of the copy number variation regions.
5	Strand	The strand of the copy number variation regions.
6	Accession	The accessions that have these copy number variation regions.
7	CN	The copy numbers of these copy number variation regions. CN0 and CN1 represent loss, CN2 represents normal, and CN3 to CN8 represent gain.

3.3 Search by Chromosome and Region

The Search By Chromosome and Region window within the copy number variation component features input fields for specifying a chromosome, start position, and end position of a region. Alongside these, there is a dropdown menu and a search button, as depicted in Figure 18.

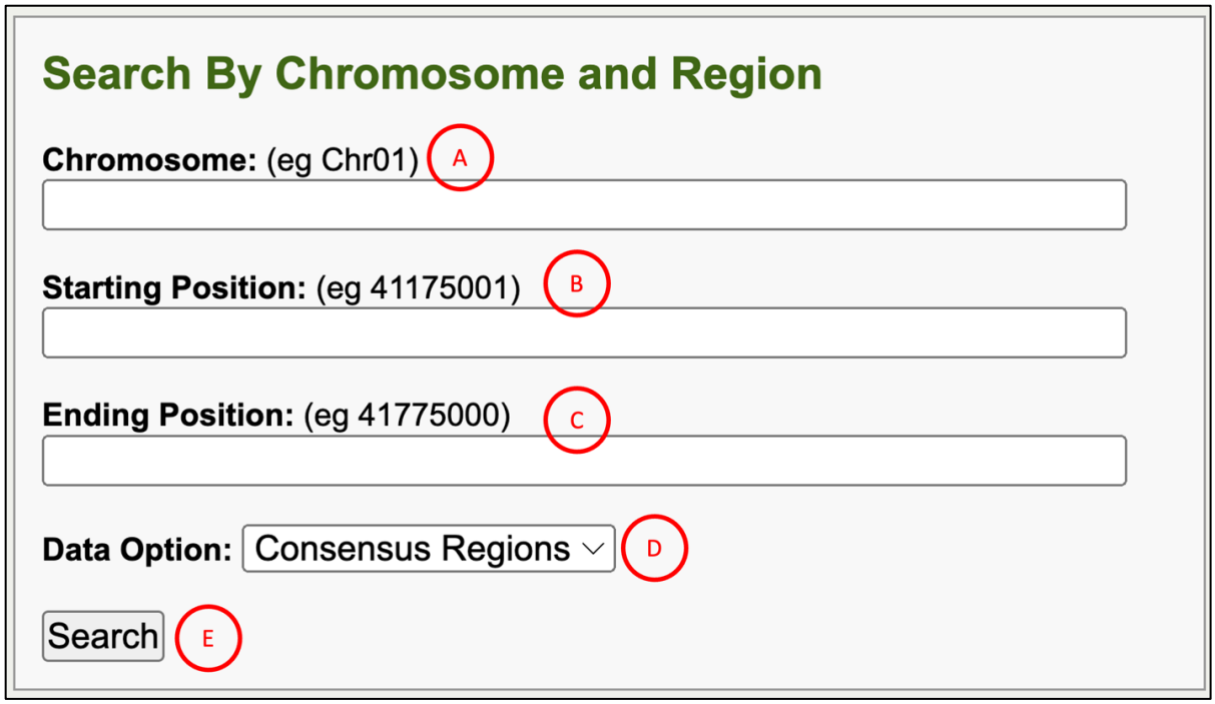
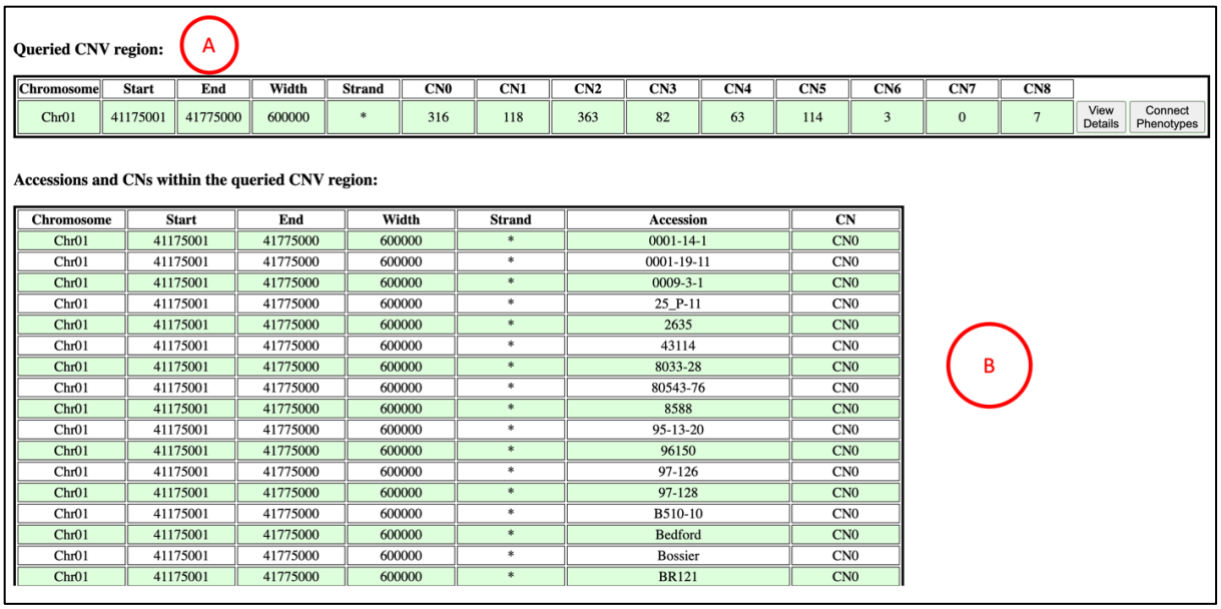


Figure 18: The Search By Chromosome and Region window. (A) The chromosome input box for users to provide a chromosome. (B) The start position input box that allows users to input a start position. (C) The end position input box that allows users to input an end position. (C) The data option is for users to select the copy number variation data type of interest for the query. (D) The search button for users to submit the query.

When users use the Search By Chromosome and Region window in the copy number variation component, they are required to input a single chromosome, a distinct start position, and a separate end position. Furthermore, they can choose between the individual hits or consensus region options from the dropdown menu. Upon completion of these steps, users should click the search button to initiate their query. Upon successful execution, they will be redirected to the results page (Figure 19).



Queried CNV region: A

Chromosome	Start	End	Width	Strand	CN0	CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8		
Chr01	41175001	41775000	600000	*	316	118	363	82	63	114	3	0	7	View Details	Connect Phenotypes

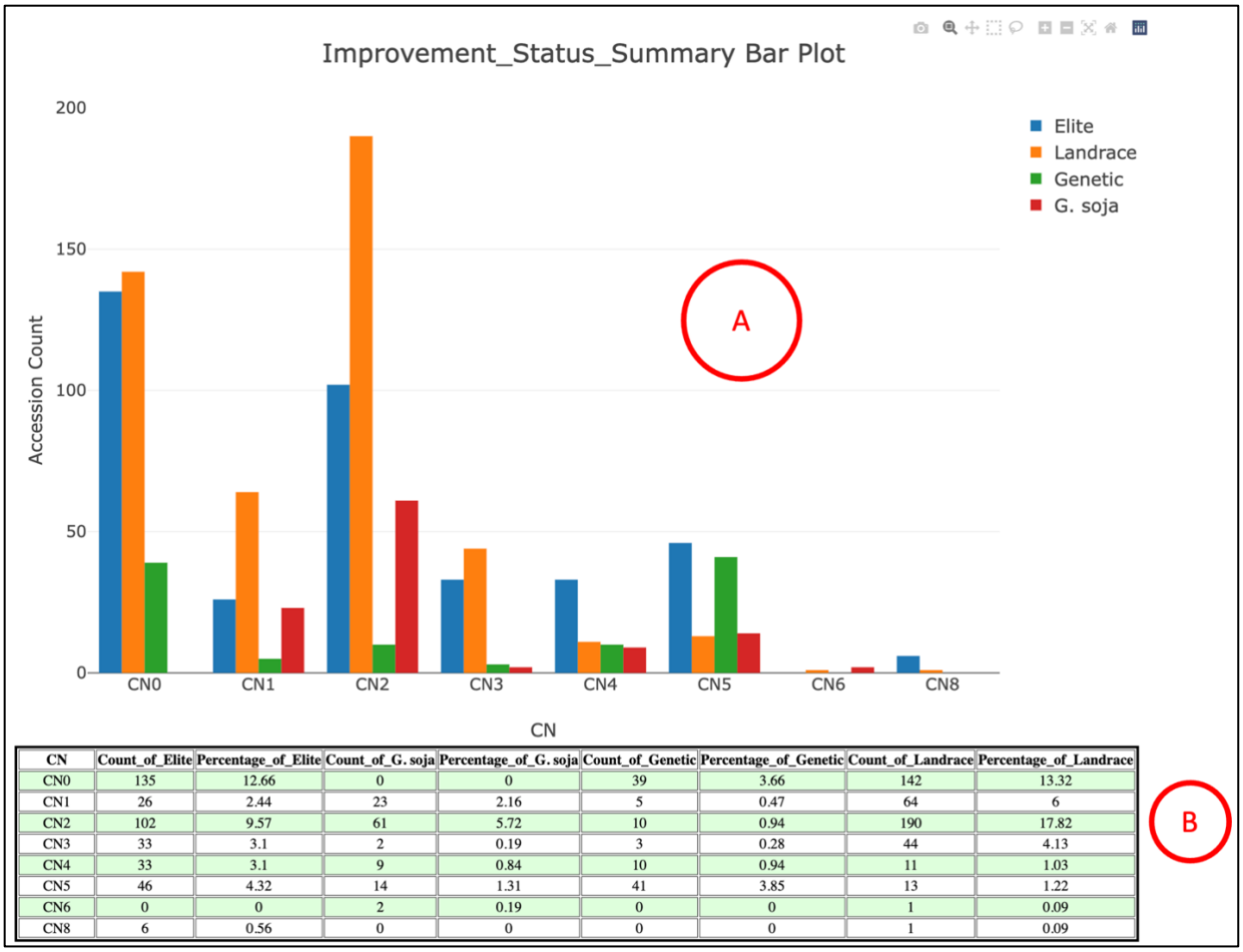
Accessions and CNs within the queried CNV region:

Chromosome	Start	End	Width	Strand	Accession	CN
Chr01	41175001	41775000	600000	*	0001-14-1	CN0
Chr01	41175001	41775000	600000	*	0001-19-11	CN0
Chr01	41175001	41775000	600000	*	0009-3-1	CN0
Chr01	41175001	41775000	600000	*	25_P-11	CN0
Chr01	41175001	41775000	600000	*	2635	CN0
Chr01	41175001	41775000	600000	*	43114	CN0
Chr01	41175001	41775000	600000	*	8033-28	CN0
Chr01	41175001	41775000	600000	*	80543-76	CN0
Chr01	41175001	41775000	600000	*	8588	CN0
Chr01	41175001	41775000	600000	*	95-13-20	CN0
Chr01	41175001	41775000	600000	*	96150	CN0
Chr01	41175001	41775000	600000	*	97-126	CN0
Chr01	41175001	41775000	600000	*	97-128	CN0
Chr01	41175001	41775000	600000	*	B510-10	CN0
Chr01	41175001	41775000	600000	*	Bedford	CN0
Chr01	41175001	41775000	600000	*	Bossier	CN0
Chr01	41175001	41775000	600000	*	BR121	CN0

B

Figure 19: The result page rendered when users make a query using the Search By Chromosome and Region window in the copy number variation component. There are two sections on this result page: (A) a table that shows the copy number variation regions within the user defined queried region along with copy numbers and accession counts, and (B) each copy number variation region has a table to list out the accessions and the corresponding copy numbers of those accessions.

The idea of Figure 19A is similar to Figure 14B. There are two buttons for each copy number variation region which are the “View Details” button and the “Connect Phenotypes” button. The “View Details” button is for improvement status distribution visualization, and the “Connect Phenotypes” button is to connect copy numbers with phenotypes in the phenotype viewer. The improvement status distribution visualization is a new web page to show the improvement status distribution like Figure 20, which is also similar to Figure 15. The phenotype viewer will be discussed in section 3.4 of this manual.



CN	Count_of_Elite	Percentage_of_Elite	Count_of_G. soja	Percentage_of_G. soja	Count_of_Genetic	Percentage_of_Genetic	Count_of_Landrace	Percentage_of_Landrace
CN0	135	12.66	0	0	39	3.66	142	13.32
CN1	26	2.44	23	2.16	5	0.47	64	6
CN2	102	9.57	61	5.72	10	0.94	190	17.82
CN3	33	3.1	2	0.19	3	0.28	44	4.13
CN4	33	3.1	9	0.84	10	0.94	11	1.03
CN5	46	4.32	14	1.31	41	3.85	13	1.22
CN6	0	0	2	0.19	0	0	1	0.09
CN8	6	0.56	0	0	0	0	1	0.09

Figure 20: The improvement status distribution visualization generated based on the copy numbers, improvement status, and accessions in the Soy1066 data panel. (A) An improvement status summary bar plot summarizes the accessions based on copy numbers and improvement status. (B) A table that summarizes the accession counts and percentages according to copy numbers and improvement status.

3.4 Phenotype Viewer

The purpose of the phenotype viewer in the copy number variation component is to provide capabilities for users to connect copy numbers with phenotypes and visualize the phenotype data in the form of tables and figures. When users click on the “Connect Phenotypes” button in Figure 14B and Figure 19A, users will be redirected to the phenotype viewer. The phenotype viewer (Figure 21) has an accordion with different tabs to include copy number variation region information, copy number (CN) checkboxes, and phenotype checkboxes that are separated into different categories under different tabs. Under the accordion, there are several buttons with different functionalities for users to interact with the phenotype viewer to get their desired data. The functionalities of the buttons are displayed in Table 6.

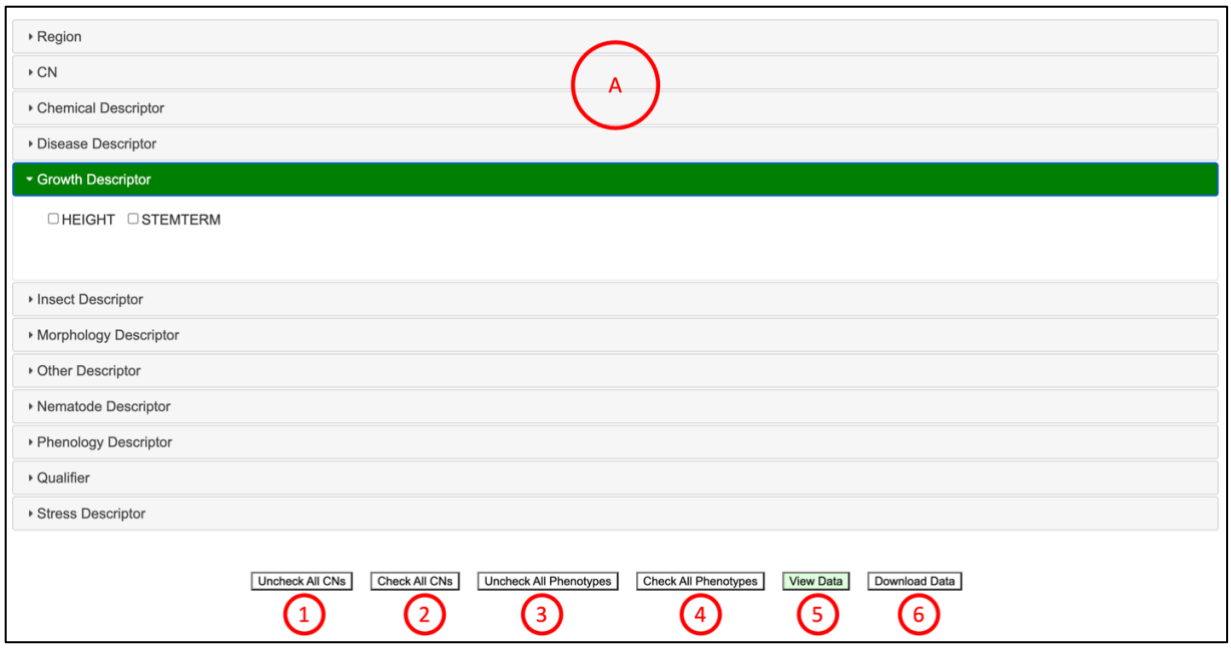
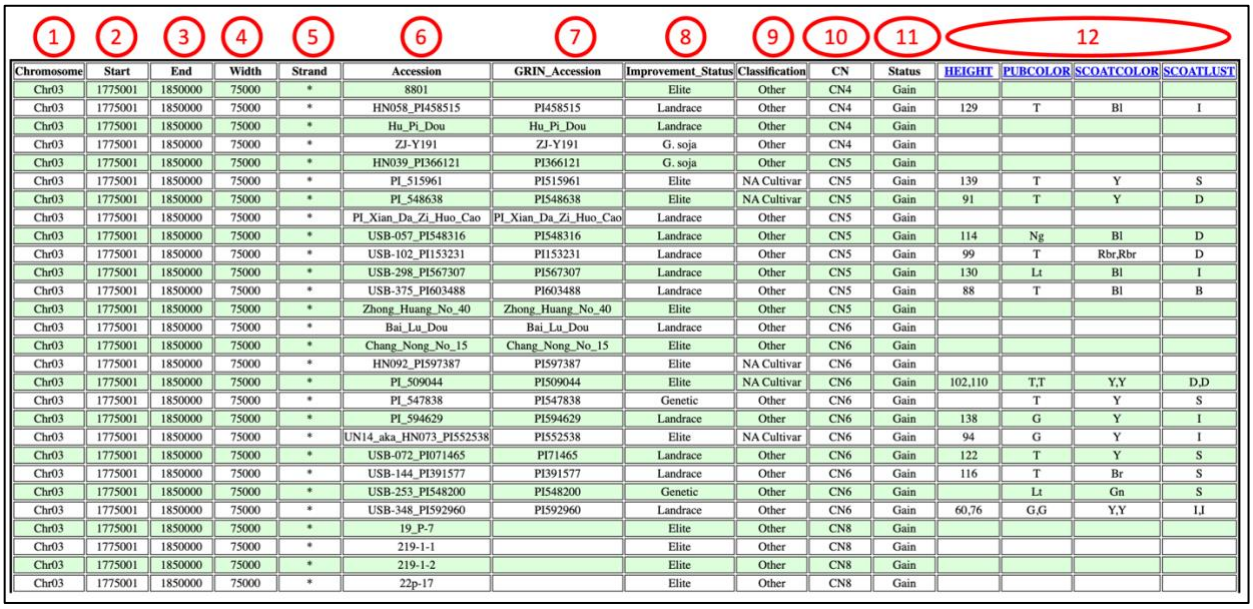


Figure 21: The accordion and buttons in the phenotype viewer. (A) The accordion section has different tabs, and each tab is clickable to expand. The copy number variation region tab has the chromosome, start position, end position, and data option information of the selected copy number variation region. The CN tab has all the possible copy numbers in the form of checkboxes for users to select. The rest of the tabs correspond to different phenotype categories, and each tab has phenotype trait checkboxes for users to select phenotypes of interest. The buttons under the accordion are illustrated in Table 6.

Table 6: A table to explain the functionalities of the buttons under the accordion in Figure 21.

Index	Button	Description
1	Uncheck All CNs	This button unchecks all the copy number checkboxes in the CN tab of the accordion.
2	Check All CNs	This button checks all the copy number checkboxes in the CN tab of the accordion.
3	Uncheck All Phenotypes	This button unchecks all the phenotype checkboxes in all the phenotype tabs of the accordion.
4	Check All Phenotypes	This button checks all the phenotype checkboxes in all the phenotype tabs of the accordion.
5	View Data	This button can be clicked to query copy number and phenotype information and show the results in a table. An example of the table is shown in Figure 22. This button will not work, and an error message is prompted to users if none of the copy number in the CN tab of the accordion is selected.
6	Download data	This button can be clicked to query copy number and phenotype information and download it as a static file.

When users land on the phenotype viewer (Figure 21), they can select the copy numbers of interest by checking the copy number checkboxes in the CN tab of the accordion. Besides that, users can also use the “Uncheck All CNs” and “Check All CNs” buttons to uncheck or check all copy numbers. Apart from that, users can also select the phenotypes of interest by using the checkboxes in the phenotype tabs of the accordion. Another option to select phenotypes is to use the “Uncheck All Phenotypes” and “Check All Phenotypes” to uncheck or check all phenotypes. Users must choose at least one copy number and phenotype in order for the “View Data” button and “Download Data” button to function properly. When users click on the “View Data” button, a query will be submitted to the backend code to get the copy number and phenotype data and present the data in a table (Figure 22). On the other hand, if the “Download Data” button is pressed, a query will be submitted to the backend code to get the copy number and phenotype data and then transfer the data in the form of a static file to the user end. The copy number and phenotype data in the table or static file consists of at least 12 columns in the case of soybean data (Figure 22). The details of these columns are demonstrated in Table 7.



1	2	3	4	5	6	7	8	9	10	11	12			
Chromosome	Start	End	Width	Strand	Accession	GRIN_Accession	Improvement_Status	Classification	CN	Status	HEIGHT	PUBCOLOR	SCOATCOLOR	SCOATLUST
Chr03	1775001	1850000	75000	*	8801		Elite	Other	CN4	Gain				
Chr03	1775001	1850000	75000	*	HN058_PI458515	PI458515	Landrace	Other	CN4	Gain	129	T	Bl	I
Chr03	1775001	1850000	75000	*	Hu_Pi_Dou	Hu_Pi_Dou	Landrace	Other	CN4	Gain				
Chr03	1775001	1850000	75000	*	ZJ-Y191	ZJ-Y191	G. soja	Other	CN4	Gain				
Chr03	1775001	1850000	75000	*	HN039_PI366121	PI366121	G. soja	Other	CN5	Gain				
Chr03	1775001	1850000	75000	*	PI_515961	PI515961	Elite	NA Cultivar	CN5	Gain	139	T	Y	S
Chr03	1775001	1850000	75000	*	PI_548638	PI548638	Elite	NA Cultivar	CN5	Gain	91	T	Y	D
Chr03	1775001	1850000	75000	*	PI_Xian_Da_Zi_Huo_Cao	PI_Xian_Da_Zi_Huo_Cao	Landrace	Other	CN5	Gain				
Chr03	1775001	1850000	75000	*	USB-057_PI548316	PI548316	Landrace	Other	CN5	Gain	114	Ng	Bl	D
Chr03	1775001	1850000	75000	*	USB-102_PI153231	PI153231	Landrace	Other	CN5	Gain	99	T	Rbr,Rbr	D
Chr03	1775001	1850000	75000	*	USB-298_PI567307	PI567307	Landrace	Other	CN5	Gain	130	Lt	Bl	I
Chr03	1775001	1850000	75000	*	USB-375_PI603488	PI603488	Landrace	Other	CN5	Gain	88	T	Bl	B
Chr03	1775001	1850000	75000	*	Zhong_Huang_No_40	Zhong_Huang_No_40	Elite	Other	CN5	Gain				
Chr03	1775001	1850000	75000	*	Bai_Lu_Dou	Bai_Lu_Dou	Landrace	Other	CN6	Gain				
Chr03	1775001	1850000	75000	*	Chang_Nong_No_15	Chang_Nong_No_15	Elite	Other	CN6	Gain				
Chr03	1775001	1850000	75000	*	HN092_PI597387	PI597387	Elite	NA Cultivar	CN6	Gain				
Chr03	1775001	1850000	75000	*	PI_509044	PI509044	Elite	NA Cultivar	CN6	Gain	102,110	T,T	Y,Y	D,D
Chr03	1775001	1850000	75000	*	PI_547838	PI547838	Genetic	Other	CN6	Gain		T	Y	S
Chr03	1775001	1850000	75000	*	PI_594629	PI594629	Landrace	Other	CN6	Gain	138	G	Y	I
Chr03	1775001	1850000	75000	*	UN14_aka_HN073_PI552538	PI552538	Elite	NA Cultivar	CN6	Gain	94	G	Y	I
Chr03	1775001	1850000	75000	*	USB-072_PI071465	PI71465	Landrace	Other	CN6	Gain	122	T	Y	S
Chr03	1775001	1850000	75000	*	USB-144_PI391577	PI391577	Landrace	Other	CN6	Gain	116	T	Br	S
Chr03	1775001	1850000	75000	*	USB-253_PI548200	PI548200	Genetic	Other	CN6	Gain		Lt	Gn	S
Chr03	1775001	1850000	75000	*	USB-348_PI592960	PI592960	Landrace	Other	CN6	Gain	60,76	G,G	Y,Y	I,I
Chr03	1775001	1850000	75000	*	19_P-7		Elite	Other	CN8	Gain				
Chr03	1775001	1850000	75000	*	219-1-1		Elite	Other	CN8	Gain				
Chr03	1775001	1850000	75000	*	219-1-2		Elite	Other	CN8	Gain				
Chr03	1775001	1850000	75000	*	22p-17		Elite	Other	CN8	Gain				

Figure 22: The copy number and phenotype data rendered in a tabular form after the “View Data” button is clicked. There are at least 12 columns in the table. The details of each column are listed in Table 7.

Table 7: A table explains the columns in the table that hold the copy number and phenotype data.

Index	Column	Description
1	Chromosome	The chromosomes of the copy number variation regions.
2	Start	The start positions of the copy number variation regions.

3	End	The end positions of the copy number variation regions.
4	Width	The strands of the copy number variation regions.
5	Strand	The widths of the copy number variation regions.
6	Accession	The accession names of the accessions that are in the Soy1066 data panel.
7	GRIN_Accession	The alternative accession names that collected from the GRIN database.
8	Improvement_Status	The improvement status of the soybean accessions. There are a few categories such as Soja, Elite, and Landrace.
9	Classification	This column has the classifications of the soybean accessions, which include North America (NA) Ancestor, NA Cultivar, and Other.
10	CN	The copy numbers selected by users in the CN tab of the accordion.
11	Status	This column indicates whether a copy number represent loss, normal, or gain.
12	Phenotype columns selected by users	These columns contain the phenotypes selected by users and also the phenotypic measurements correspond to the accessions. Each phenotype header is clickable to redirect to the phenotype data visualization page to visualize the distributions of the phenotype data (Figure 23 and 24).

Each phenotype column header in the copy number and phenotype data table is clickable. When users click on a phenotype column header, a new query is initiated to fetch and process the phenotype data. Subsequently, users are redirected to the phenotype data visualization page. Figure 23 and Figure 24 serve as the phenotype data visualization pages, with the appropriate one shown based on the data type (quantitative or qualitative). During the processing step, the algorithm first determines whether the phenotype data is quantitative or qualitative. Quantitative data is presented in a violin plot (Figure 23B), while qualitative data is displayed in a bar plot (Figure 24B). In Figure 23C and Figure 24C, the methods for summarizing the phenotype data differ based on the data type (quantitative or qualitative). For quantitative data, the summary is derived solely from copy numbers and accessions with or without phenotype data, summarized in the form of accession counts (Figure 23C). Conversely, qualitative data includes not only counts of accessions with or without phenotype data, but also provides a breakdown of each phenotype category with corresponding counts and percentages (Figure 24C). Furthermore, irrespective of whether the phenotype data is quantitative or qualitative, the improvement status summary bar plots are accessible. These plots serve to summarize the distribution of accessions across copy numbers and improvement status categories (Figure 23D and 24D).

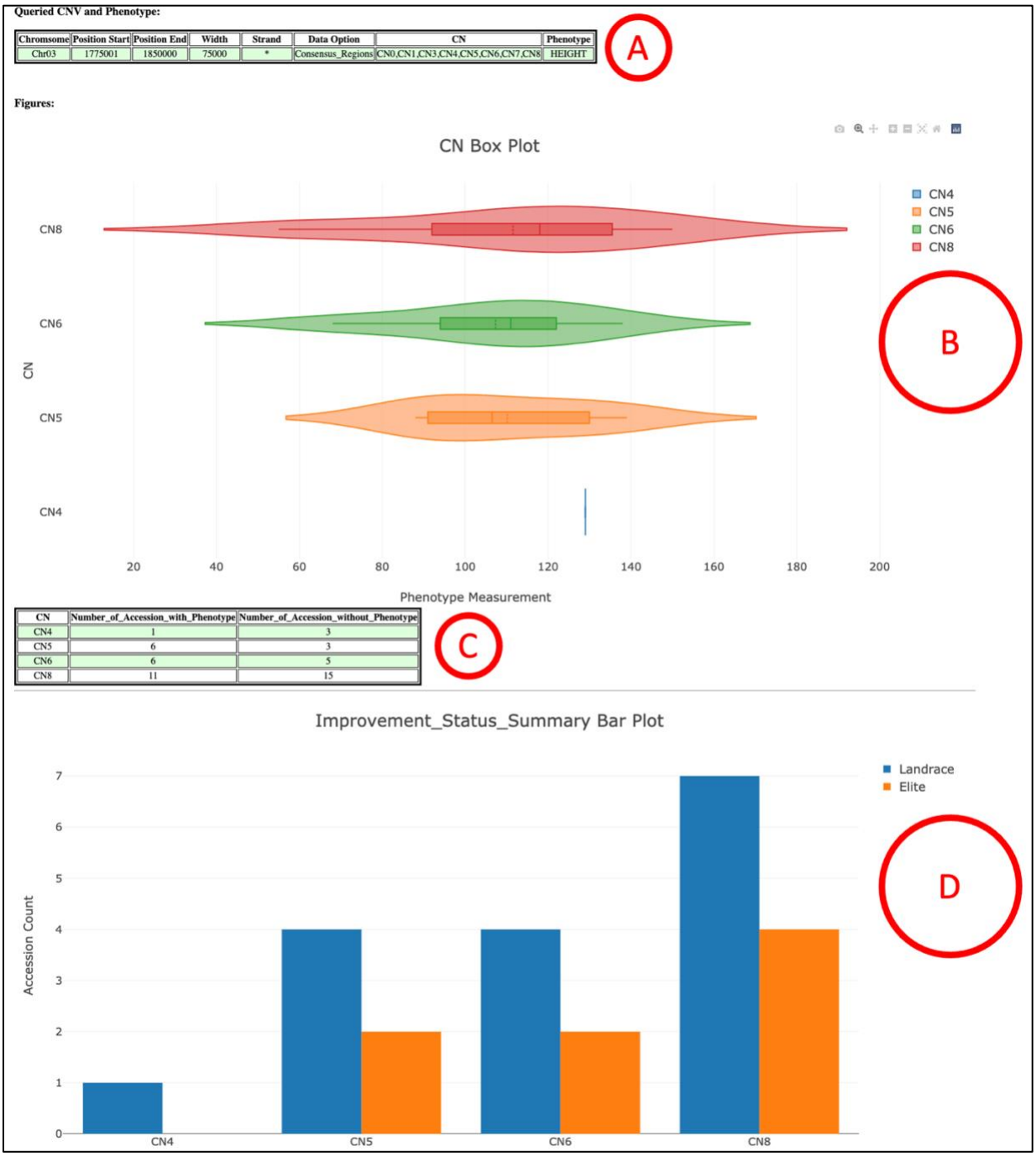


Figure 23: The phenotype data visualization page is displayed when the phenotype data is quantitative data. (A) A table to show the selected copy numbers and phenotype. (B) A violin plot for users to visualize the distribution of the quantitative phenotype data. (C) A summary table that summarizes the accessions based on copy numbers and with or without phenotype data. (D) An improvement status summary bar plot summarizes the accessions based on copy numbers and improvement status.

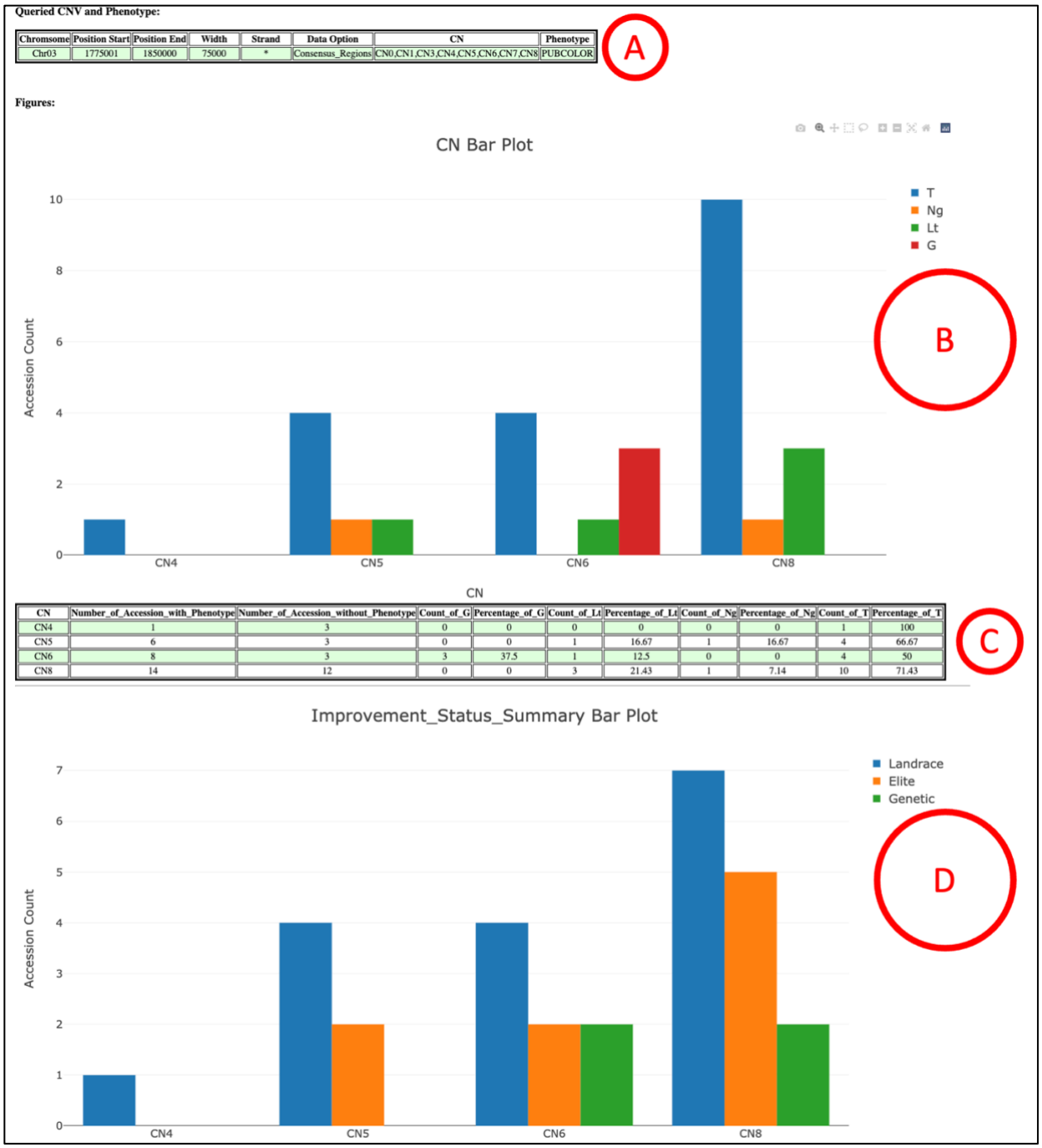


Figure 24: The phenotype data visualization page is displayed when the phenotype data is qualitative data. (A) A table to show the selected copy numbers and phenotype. (B) A bar plot for users to visualize the distribution of the qualitative phenotype data. (C) A summary table not only summarizes the accessions based on copy numbers and with or without phenotype data but also contains the breakdown of counts and percentages for each qualitative category. (D) An improvement status summary bar plot summarizes the accessions based on copy numbers and improvement status.

4 Multi-organisms Support

The GenVarX toolset currently supports soybean, rice, and *Arabidopsis*. Throughout this manual, all figures and examples are taken from the soybean GenVarX toolset, offering a comprehensive illustration of its capabilities. Importantly, the GenVarX toolsets for rice and Arabidopsis encompass identical functionalities, ensuring that researchers working with these plant species can expect consistent and reliable support. For user convenience, the soybean GenVarX toolset is conveniently accessible through the SoyKB web portals. Similarly, the rice and Arabidopsis GenVarX toolsets are readily available via the KBCommons web portals, creating a cohesive and user-friendly experience across different plant research domains. The links to access each GenVarX toolset are listed as follows:

- Soybean GenVarX Toolset: https://soykb.org/SoybeanGenVarX/
- Rice GenVarX Toolset: https://kbcommons.org/system/tools/GenVarX/Osativa
- *Arabidopsis* GenVarX Toolset: https://kbcommons.org/system/tools/GenVarX/Athaliana

5 References

1. Klambauer G, Schwarzbauer K, Mayr A, Clevert DA, Mitterecker A, Bodenhofer U, Hochreiter S: **cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate**. *Nucleic Acids Res* 2012, **40**(9):e69.