frontiers | Frontiers in Genetics

TYPE Technology and Code
PUBLISHED 09 October 2023
DOI 10.3389/fgene.2023.1251382

OPEN ACCESS

EDITED BY
Rui Yin,
University of Florida, United States

REVIEWED BY
Dailu Guan,
University of California, Davis,
United States
Xuewen Wang,
University of North Texas Health Science Center, United States

*CORRESPONDENCE
Mária Škrabišová,
maria.skrabisova@upol.cz
Trupti Joshi,
joshitr@missouri.edu

RECEIVED 01 July 2023
ACCEPTED 27 September 2023
PUBLISHED 09 October 2023

CITATION
Chan YO, Biová J, Mahmood A, Dietz N, Bilyeu K, Škrabišová M and Joshi T (2023), Genomic Variations Explorer (GenVarX): a toolset for annotating promoter and CNV regions using genotypic and phenotypic differences.
Front. Genet. 14:1251382.
doi: 10.3389/fgene.2023.1251382

Genomic Variations Explorer (GenVarX): a toolset for annotating promoter and CNV regions using genotypic and phenotypic differences

Yen On Chan[1], Jana Biová[2], Anser Mahmood[3], Nicholas Dietz[3], Kristin Bilyeu[3,4], Mária Škrabišová[2]* and Trupti Joshi[1,5,6,7]*

[1]MU Institute for Data Science and Informatics, University of Missouri-Columbia, Columbia, MO, United States, [2]Department of Biochemistry, Faculty of Science, Palacky University in Olomouc, Olomouc, Czechia, [3]Division of Plant Science and Technology, University of Missouri-Columbia, Columbia, MO, United States, [4]Plant Genetics Research Unit, United States Department of Agriculture-Agricultural Research Service, Columbia, MO, United States, [5]Christopher S. Bond Life Sciences Center, University of Missouri-Columbia, Columbia, MO, United States, [6]Department of Electrical Engineering and Computer Science, University of Missouri-Columbia, Columbia, MO, United States, [7]Department of Biomedical Informatics, Biostatistics and Medical Epidemiology, University of Missouri-Columbia, Columbia, MO, United States

The rapid growth of sequencing technology and its increasing popularity in biology-related research over the years has made whole genome re-sequencing (WGRS) data become widely available. A large amount of WGRS data can unlock the knowledge gap between genomics and phenomics through gaining an understanding of the genomic variations that can lead to phenotype changes. These genomic variations are usually comprised of allele and structural changes in DNA, and these changes can affect the regulatory mechanisms causing changes in gene expression and altering the phenotypes of organisms. In this research work, we created the GenVarX toolset, that is backed by transcription factor binding sequence data in promoter regions, the copy number variations data, SNPs and Indels data, and phenotypes data which can potentially provide insights about phenotypic differences and solve compelling questions in plant research. Analytics-wise, we have developed strategies to better utilize the WGRS data and mine the data using efficient data processing scripts, libraries, tools, and frameworks to create the interactive and visualization-enhanced GenVarX toolset that encompasses both promoter regions and copy number variation analysis components. The main capabilities of the GenVarX toolset are to provide easy-to-use interfaces for users to perform queries, visualize data, and interact with the data. Based on different input windows on the user interface, users can provide inputs corresponding to each field and submit the information as a query. The data returned on the results page is usually displayed in a tabular fashion. In addition, interactive figures are also included in the toolset to facilitate the visualization of statistical results or tool outputs. Currently, the

Abbreviations: BAM, binary alignment map; BWA, burrows-wheeler aligner; ENA, european nucleotide archive; GATK, genome analysis toolkit; GSA, genome sequence archive; GRIN, resources information network; Indels, insertions and deletions; NCBI, national center for biotechnology information; SAM, sequence alignment map; SNP, single nucleotide polymorphism; VCF, variant call format; WGRS, whole genome re-sequencing.

GenVarX toolset supports soybean, rice, and *Arabidopsis*. The researchers can access the soybean GenVarX toolset from SoyKB via https://soykb.org/SoybeanGenVarX/, rice GenVarX toolset, and *Arabidopsis* GenVarX toolset from KBCommons web portal with links https://kbcommons.org/system/tools/GenVarX/Osativa and https://kbcommons.org/system/tools/GenVarX/Athaliana, respectively.

KEYWORDS

transcription factor, promoter, copy number variation, whole genome re-sequencing data, genomic variations, SNPs, Indels, phenotypes

Introduction

As the use of sequencing technology becomes popular in both industrial and academic sectors, a large amount of whole genome re-sequencing (WGRS) data has become publicly available for users to utilize for research activities or commercial purposes. From the WGRS data, researchers can understand the genomic variations and the potential effects on phenotypes of organisms (Bolger et al., 2017). The differences in phenotypes compared among accessions are the reflections of genomic variations and structural variations (Li et al., 2022). The genomic variations include changes of alleles in gene regions, upstream promoter regions, and intergenic regions, while the structural changes comprise insertion, deletion, and duplication of DNA segments. These allele changes can occur in the transcription factor binding domain which leads to regulatory mechanism dysfunction for some target genes. Ultimately, the issues cause alterations in gene expression patterns and lead to phenotypic diversity in organisms. Similarly, the larger structural variations seen in the form of copy number variations (CNVs) can cause gains or losses in DNA segments. These modifications in DNA can alter the copies of expressed genes and ultimately results in changes in phenotypes in organisms (Żmieńko et al., 2014). Thus, gaining an in-depth understanding of the transcription factor (TF) binding sites and the CNVs through performing analysis with extensive data and open-source tools and packages that are available online is critical in plant research and development.

Currently, there are more than 3000 soybean, 1000 *Arabidopsis*, and 3000 rice WGRS accession datasets along with phenotype datasets, annotation datasets, and transcription factors datasets scattered across different resources (The 3000 rice genomes project, 2014; Alonso-Blanco et al., 2016; Liu et al., 2020). The datasets are usually available for researchers to download in the form of static files and are not pre-integrated with other omics datasets. The publicly accessible web portals and platforms such as Plant Transcription Factor Database (PlantTFDB) (Jin et al., 2016), Plant Transcriptional Regulatory Map (PlantRegMap) (Tian et al., 2019), Gene Transcription Regulation Database (GTRD) (Yevshin et al., 2017), and JASPAR (Castro-Mondragon et al., 2021) are the main providers of transcription factors related datasets. Moreover, the National Center for Biotechnology Information (NCBI), European Nucleotide Archive (ENA), the Genome Sequence Archive (GSA) of the National Genomics Data Center (NGDC), and the CyVerse data store (Goff et al., 2011; Merchant et al., 2016) are the main resources for publicly available WGRS datasets. Likewise, there are also many open-source tools and packages available that can be applied to the WGRS data and used to perform CNV analysis like cn.MOPS (Klambauer et al., 2012), CNV-seq (Xie and Tammi, 2009), cnvScan (Samarakoon et al., 2016), and more. Nevertheless, there is a lack of interactive and visualizable web applications to integrate and query TF binding sites in promoter regions and CNVs data with the genomic variability observed from large-scale studies involving thousands of accessions to gain insight about phenotypes, perform validations, and eventually roll out new discoveries. To solve this problem, we have dedicated effort towards building an interactive and visualization-enhanced toolset for supporting this analysis.

Materials and methods

In the materials and methods section, we provide details about the datasets collected and utilized for this research work including the tools used to obtain the datasets, process the data, and the respective required inputs and outputs for these tools. The processed datasets are uploaded and stored in the MySQL database integrated into both SoyKB (Joshi et al., 2012; Joshi et al., 2013; Joshi et al., 2017) and KBCommons (Zeng et al., 2018; Zeng et al., 2019) web portals. The package used for uploading the datasets, indexing methods for the tables in the database, and technology used in the web development for the GenVarX toolset are also described.

Datasets

The GenVarX toolset currently supports the soybean, rice, and *Arabidopsis* organisms. For each organism, the toolset is divided into two components mainly for the promoter regions and the CNV analysis. In order to facilitate the functionalities of each component in each organism, many datasets are required to be collected and processed together to power the GenVarX toolset.

For the promoter regions component, the transcription factors (TFs) datasets and binding TF position probability matrices from the Plant Transcription Factor Database (PlantTFDB v5.0) (Jin et al., 2016) were acquired for each organism. In addition, the predicted transcription factor binding sites datasets and motif-gene regulation datasets from the Plant Transcriptional Regulatory Map (PlantRegMap) website (Tian et al., 2019) were required for each organism.

For the CNV analysis component, we acquired WGRS datasets for each organism (in different formats depending on their availability in public resources) and generated the CNV results. The WGRS files mainly include sequencing data in FASTQ format and the reads mapped to genome sequences in Binary Alignment

Map (BAM) format. The FASTQ format datasets are further processed and converted into BAM files as the BAM files are the required input format for our selected tool for generating CNV results.

For the CNV analysis component in soybean, we have acquired WGRS datasets of 1066 distinct soybean accessions from publicly available datasets including Zhou302v2 (Zhou et al., 2015), Liu304 (Liu et al., 2020), USB-15x (Valliyodan et al., 2021), USB-40x (Valliyodan et al., 2021), Soja (Kim et al., 2010), and MSMC (Valliyodan and Nguyen, 2006). The PGen pipeline (Liu et al., 2016) was used for mapping these reads to Williams 82 version 2 (Wm82.a2.v1) reference genome and for SNP and Indel calling to generate both BAM and Variant Call Format (VCF) files. The mapped soybean sequencing reads in BAM format for all datasets were collectively stored on the Cyverse Data Store. The soybean Wm82.a2.v1 reference genome was downloaded from the Phytozome website (Goodstein et al., 2011). We have also collected soybean phenotype datasets from the Germplasm Resources Information Network (GRIN) database.

Similarly, we have acquired 3000 BAM files and VCF files of rice from the 3000 Rice Genome on Amazon Web Services (AWS) (2014). The Nipponbare reference genome that was used to generate these BAM files was downloaded from the Rice Annotation Project Database (RAP-DB) (Sakai et al., 2013). Regarding the rice phenotypic datasets, our research group acquired the datasets from the International Rice Research Institute's official web portal (https://snp-seek.irri.org/_download.zul).

Furthermore, a selected set of 1043 *Arabidopsis* FASTQ files have also been acquired from the SRA Run Selector with project number PRJNA273563 using the SRA Toolkit 3.0.0 (https://github.com/ncbi/sra-tools). These *Arabidopsis* accessions are part of the 1135 *Arabidopsis* accessions presented in the 1001 Genomes Consortium (Alonso-Blanco et al., 2016). Since the collected files are in FASTQ format, we have also collected the *Arabidopsis thaliana* TAIR10 reference genome from the Phytozome website for mapping the FASTQ files to the reference genome and generating the BAM files. We also directly downloaded the *Arabidopsis* VCF files from the 1001 Genomes Consortium (https://1001genomes.org/data/GMI-MPI/releases/v3.1/).

Data processing

In data processing, the datasets for the promoter regions component and the CNV analysis component of the GenVarX toolset were processed separately using a different set of open-source and publicly available tools. There are four types of datasets that are utilized in the promoter regions component. The datasets are TF datasets, predicted transcription factor binding sites datasets, motif-gene regulation datasets, and binding TF position probability matrices. The first three types of datasets only need some additional processing for extracting the mandatory columns that are used in the promoter regions component. The binding TF position probability matrices, on the other hand, are processed further using the Ceqlogo tool in the Meme Suite (Bailey et al., 2015). The details about the binding TF position probability matrices and the sequence logo figures will be utilized in the promoter regions component to overlap with genomic variations (SNP and Indels) datasets and show the potential impacts of nucleotide variations on conserved positions in motifs.

For the CNV analysis component, there were two types of datasets involved, the FASTQ and BAM files. The FASTQ files were required to be processed and converted into BAM files so that they can be used as inputs to the cn. MOPS package. In the data processing of the FASTQ files, the files were aligned with their respective reference genome using the Burrows-Wheeler Aligner tool (BWA) version 0.7.17 (Li and Durbin, 2009) to create outputs in Sequence Alignment Map (SAM) format. Moreover, the SortSam, MarkDuplicates, and AddOrReplaceReadGroups commands in the Genome Analysis Toolkit (GATK) version 4.2.6.1 (McKenna et al., 2010) were used to sort the SAM files, mark duplications, assign reads to read groups, and output the final BAM files required for the next step.

In order to generate the CNV results, we utilized the cn. MOPS R package (Klambauer et al., 2012) written in C++ and R by Klambauer et al. (2012). The cn. MOPS package can take in BAM files to calculate the coverage depth of each position across accessions and perform read variations decomposition based on the mixture components and Poisson distributions across accessions using a Bayesian method (Klambauer et al., 2012). Because of its implementation, this package can achieve a low false discovery rate (FDR) as high noise data is filtered out during the calculation process. The cn. MOPS package has demonstrated good performance and significance using metrics such as the precision-recall area-under-curve (PR AUC) and recall rate by comparing itself with other methods. The outputs of this package that were useful to our CNV analysis component are the CNV individual hits of each accession and the CNV consensus regions across all samples.

For the analysis, we have written R scripts to perform CNV calculations for each organism separately. Each R script that uses the cn. MOPS package takes in the file paths of the BAM files as inputs. The R script was designed to perform CNV analysis on the chromosomal sequences of an organism. Other non-chromosomal sequences like scaffold sequences, mitochondria sequences, and more were not included in the CNV analysis. Upon the completion of CNV analysis, CNV individual hits and consensus regions were collected and ready to be uploaded to the database.

Data storing

In the data storing section, we provide details about database, data upload methods, and data indexing methods. Uploading data into databases and storing the data with proper indexing can enhance the data query by speeding up the process and preserving the data for long-term usage. In our GenVarX toolset development, we adopted this common practice in order to provide good services to users.

In the GenVarX toolset development, the MySQL database integrated into the SoyKB and KBCommons web portals was utilized to store the datasets for both the promoter regions component and the CNV analysis component of different organisms. With regard to uploading the datasets to the MySQL database, the open-source SQLAlchemy package (Bayer, 2012) written in Python was used to assist the data upload process. For datasets that are very large in size such as the CNV consensus regions datasets and genotype datasets, the B+ tree indexing method was used to index the tables in the database. Having all necessary datasets in the database, queries from the web applications can be facilitated as the data in the database can be searched and returned to the users.

Web development

The GenVarX toolset is presented as a web-based application to the users. Therefore, the user-interactive parts of the GenVarX toolset were web-focused. In web development, several programming languages, libraries, and frameworks were utilized. Because the soybean GenVarX toolset and the GenVarX toolset of other organisms are being deployed on different platforms, the used frameworks are slightly different between the two GenVarX toolsets.

In the development of the soybean GenVarX toolset, HTML, CSS, JavaScript, PHP, and SQL programming languages were utilized in coding both the promoter and the CNV analysis components. Among the four programming languages, HTML, CSS, and JavaScript were focused on the front-end of the web application, while PHP and SQL were focused on the back-end development. In the front-end of the web application, the jQuery JavaScript library was used for enhancing the JavaScript functions and sending requests to the back-end for data retrieval. The PHP back-end of the web application was focused on rendering PHP code and communicating with the database to collect data from the database using SQL queries. The entire technological structure is closer to Linux, Apache, MySQL, and PHP (LAMP) stack.

Likewise, the GenVarX toolsets for other organisms also used the same programming languages, and the programming languages were focused on the front-end and back-end, respectively, similar to the soybean GenVarX toolset. The jQuery JavaScript library was also used for the same purposes in the GenVarX toolsets for the other organisms. Nevertheless, the framework used in the GenVarX toolsets for the other organisms was the Laravel framework which encompasses the back-end of the web application in classes that extends the based controller class and manages all the routes of the web application in one place. Using the Laravel framework, the GenVarX toolset can work for many organisms in the same code. Hence, using this framework simplifies the web development processes.

The deployment of the soybean GenVarX toolset and the other universal GenVarX toolsets are on different websites. The soybean GenVarX toolset was deployed on the SoyKB website while the rest of the toolsets were deployed on the KBCommons website. The links to access the GenVarX toolset are placed under the tool section of both websites. Users can click on the links to get redirected to the GenVarX toolset. For simplicity, below are the links to access the different GenVarX toolsets hosted on the SoyKB and KBCommons websites:

- 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

Results

In the results section, the promoter regions and the CNV analysis components of the GenVarX toolset are illustrated. The results focus on the functionalities, interfaces, inputs, and outputs of each component of the GenVarX toolset.

Promoter regions component

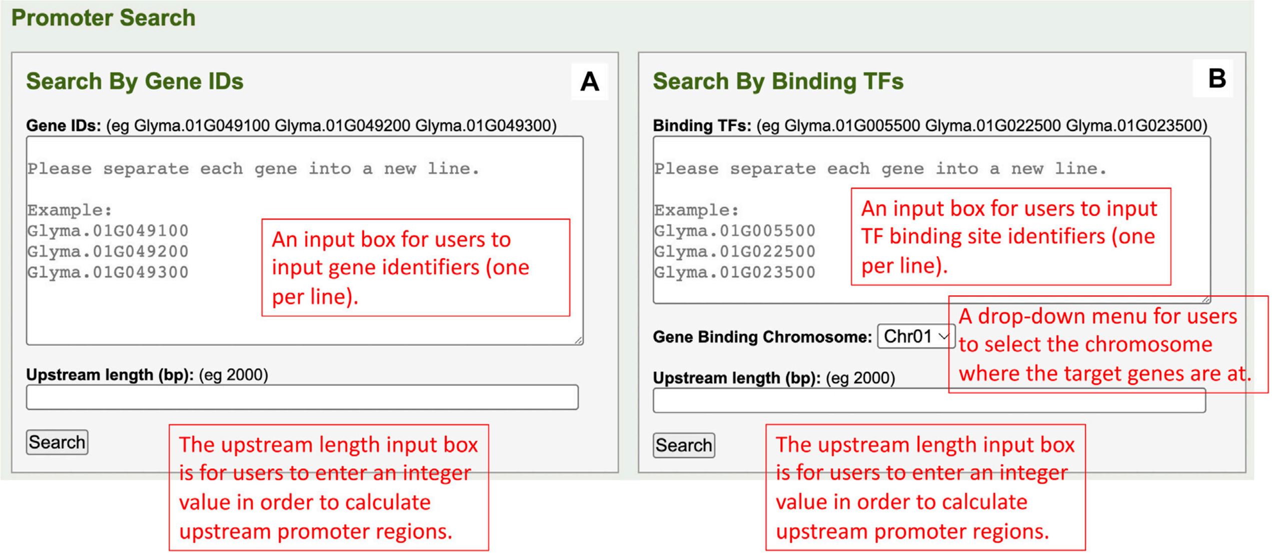
The promoter regions component consists of a data search page, an independent promoter results page, and a phenotype data viewing page. The promoter search windows on the data search page allow users to search by gene IDs or search by binding TFs. Both windows take in user inputs to perform queries and render the results on the promoter results page and the phenotype data viewing page. Here, each search method is discussed separately.

Promoter regions component—Search by Gene IDs

In the Search by Gene IDs of the promoter regions component, there is a window that has one gene identifiers input box, an upstream length input box, and a search button (Figure 1A). The input box allows users to input multiple gene identifiers at one time, and each gene identifier must be separated into a new line. The upstream length input box takes an integer value from users to calculate upstream promoter regions of the inputted genes. When the search button is clicked, the query is done for each gene, that is, searchable in the database. At the same time, TF binding sites that are in the promoter region and all relevant information are also fetched from the database and displayed on the promoter results page.

On the promoter results page, the TF binding sites results are grouped by genes and shown independently for one section per gene (Figure 2). Each section begins with the queried gene identifier along with the chromosome number, coordinates, and strand information. Below the gene identifier, a calculated upstream promoter region is shown. Underneath that, a TF binding sites table with information by row for each TF binding site in the promoter region displays information such as the TF binding site chromosome number, coordinates, strand, TF binding site identifiers, TF family type, and Williams 82 version 2 gene binding sequence. Each TF binding site identifier in the table is a clickable hyperlink. Users can click on the TF binding site identifier that they are interested in to retrieve more details about that TF binding site.

When users click on a TF binding site identifier (Binding_TF), a sequence logo figure and a position-nucleotide table will be loaded onto the respective section of the promoter results page. From the sequence logo figure, users can visualize the possible nucleotides in that TF binding site along with the information entropy of each nucleotide calculated using the Shannon entropy formula (Schneider and Stephens, 1990). The table below the sequence logo figure shows each position and each nucleotide of the Williams 82 version 2 gene binding sequence in a tabular form to assist in comparing the nucleotides in the Williams 82 version 2 gene binding sequence with the possible nucleotides of the TF binding sites shown in the sequence logo figures. The comparison provides insight into the sequence conservation of the nucleotides in



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
```

An input box for users to input gene identifiers (one per line).

Upstream length (bp): (eg 2000)

The upstream length input box is for users to enter an integer value in order to calculate upstream promoter regions.

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
```

An input box for users to input TF binding site identifiers (one per line).

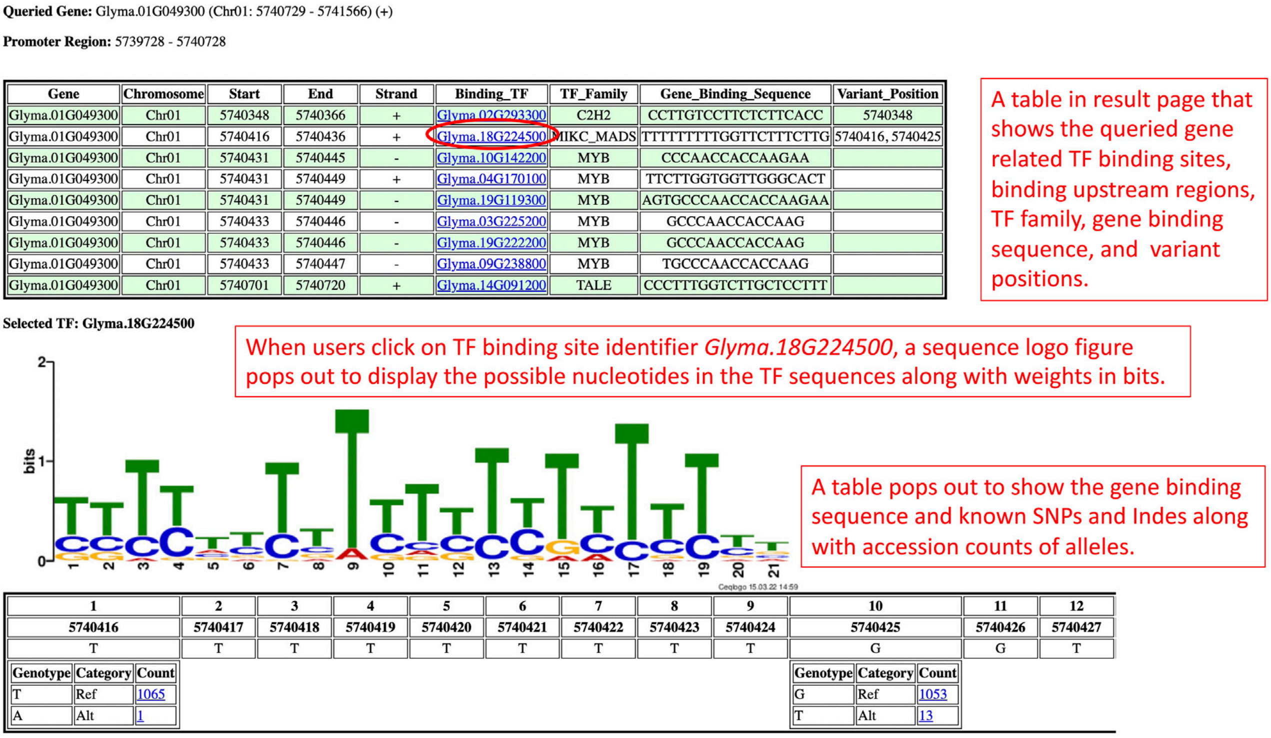
Gene Binding Chromosome: Chr01

A drop-down menu for users to select the chromosome where the target genes are at.

Upstream length (bp): (eg 2000)

The upstream length input box is for users to enter an integer value in order to calculate upstream promoter regions.

FIGURE 1
(A) The Search by Gene IDs window in the promoter regions component. **(B)** The Search by Binding TFs window in the promoter regions component.



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

Promoter Region: 5739728 - 5740728

Gene	Chromosome	Start	End	Strand	Binding_TF	TF_Family	Gene_Binding_Sequence	Variant_Position
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	

A table in result page that shows the queried gene related TF binding sites, binding upstream regions, TF family, gene binding sequence, and variant positions.

Selected TF: Glyma.18G224500

When users click on TF binding site identifier *Glyma.18G224500*, a sequence logo figure pops out to display the possible nucleotides in the TF sequences along with weights in bits.

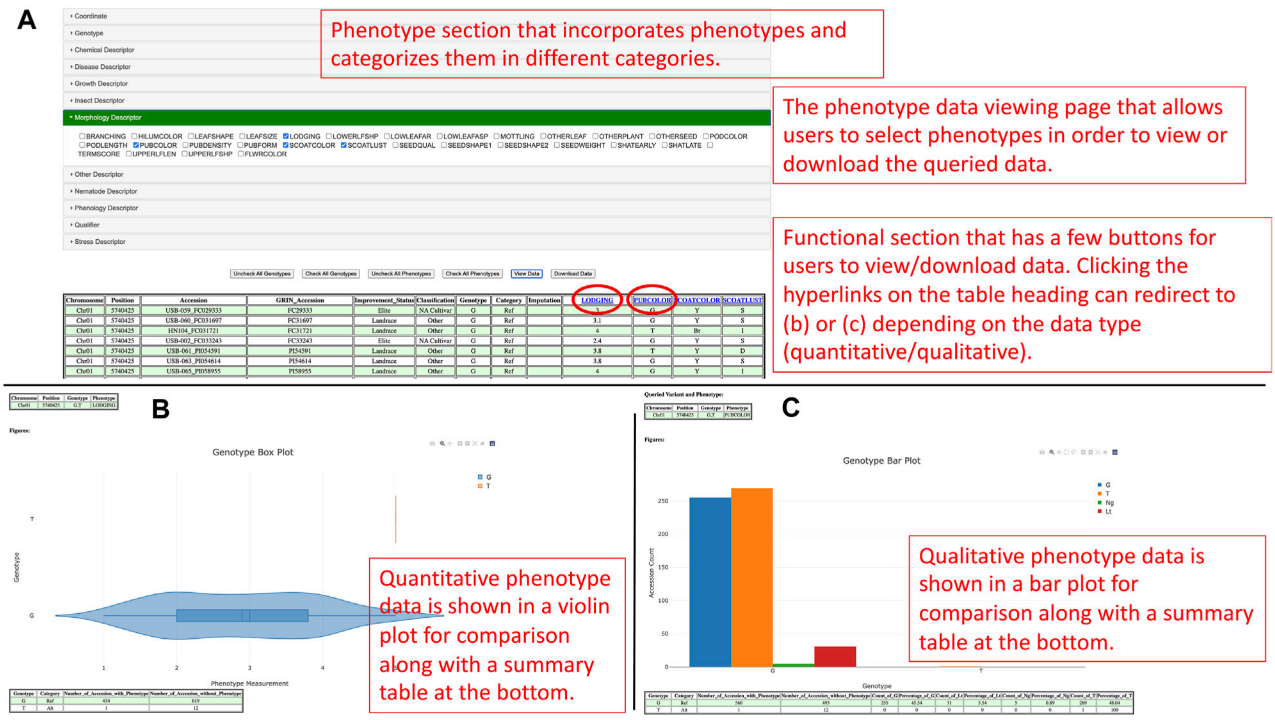
A table pops out to show the gene binding sequence and known SNPs and Indes along with accession counts of alleles.

1	2	3	4	5	6	7	8	9	10	11	12
5740416	5740417	5740418	5740419	5740420	5740421	5740422	5740423	5740424	5740425	5740426	5740427
T	T	T	T	T	T	T	T	T	G	G	T
Genotype Category Count T Ref 1065 A Alt 1									Genotype Category Count G Ref 1053 T Alt 13		

FIGURE 2
The promoter results page is rendered based on the query from the Search by Gene IDs section. This results page presents the TF binding sites of a gene in a table along with the sequence logo figure and position-nucleotide table.

the Williams 82 version 2 gene binding sequence and the TF binding sites.

Apart from that, the GenVarX toolset also shows SNPs and Indels in allele tables along with the counts of accessions having the corresponding alleles. The counts on the table are clickable and able to redirect users to the phenotype data viewing page (Figure 3A). On the phenotype data viewing page, users not only can see accessions that have a particular allele but also able to connect the accessions with



☐ BRANCHING ☐ HILUMCOLOR ☐ LEAFSHAPE ☐ LEAFSIZE ☑ LODGING ☐ LOWERLFSHP ☐ LOWLEAFAR ☐ LOWLEAFASP ☐ MOTTLING ☐ OTHERLEAF ☐ OTHERPLANT ☐ OTHERSEED ☐ PODCOLOR ☐ PODLENGTH ☑ PUBCOLOR ☐ PUBDENSITY ☐ PUBFORM ☑ SCOATCOLOR ☑ SCOATLUST ☐ SEEDQUAL ☐ SEEDSHAPE1 ☐ SEEDSHAPE2 ☐ SEEDWEIGHT ☐ SHATEARLY ☐ SHATLATE ☐ TERMSCORE ☐ UPPERLFLEN ☐ UPPERLFSHP ☐ FLWRCOLOR

Chromosome	Position	Accession	GRIN_Accession	Improvement_Status	Classification	Genotype	Category	Imputation	LODGING	PUBCOLOR	SCOATCOLOR	SCOATLUST
Chr01	5740425	USB-059_FC029333	FC29333	Elite	NA Cultivar	G	Ref		3	G	Y	S
Chr01	5740425	USB-060_FC031697	FC31697	Landrace	Other	G	Ref		3.1	G	Y	S
Chr01	5740425	HN104_FC031721	FC31721	Landrace	Other	G	Ref		4	T	Br	I
Chr01	5740425	USB-002_FC033243	FC33243	Elite	NA Cultivar	G	Ref		2.4	G	Y	S
Chr01	5740425	USB-061_PI054591	PI54591	Landrace	Other	G	Ref		3.8	T	Y	D
Chr01	5740425	USB-063_PI054614	PI54614	Landrace	Other	G	Ref		3.8	G	Y	S
Chr01	5740425	USB-065_PI058955	PI58955	Landrace	Other	G	Ref		4	G	Y	I

Phenotype section that incorporates phenotypes and categorizes them in different categories.

The phenotype data viewing page that allows users to select phenotypes in order to view or download the queried data.

Functional section that has a few buttons for users to view/download data. Clicking the hyperlinks on the table heading can redirect to (b) or (c) depending on the data type (quantitative/qualitative).

Chromsome	Position	Genotype	Phenotype
Chr01	5740425	G,T	LODGING

Genotype	Category	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype
G	Ref	434	619
T	Alt	1	12

Quantitative phenotype data is shown in a violin plot for comparison along with a summary table at the bottom.

Queried Variant and Phenotype:

Chromsome	Position	Genotype	Phenotype
Chr01	5740425	G,T	PUBCOLOR

Genotype	Category	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype	Count_of_G	Percentage_of_G	Count_of_Lt	Percentage_of_Lt	Count_of_Ng	Percentage_of_Ng	Count_of_T	Percentage_of_T
G	Ref	560	493	255	45.54	31	5.54	5	0.89	269	48.04
T	Alt	1	12	0	0	0	0	0	0	1	100

Qualitative phenotype data is shown in a bar plot for comparison along with a summary table at the bottom.

FIGURE 3
(A) The phenotype data viewing page for users to select genotypes and phenotypes they are interested in and view or download the data. **(B)** The distribution of quantitative phenotype data plotted against genotypes in a violin plot figure. **(C)** The distribution of qualitative phenotype data plotted against genotypes in a bar plot.

phenotype data. In the phenotype accordion drop-down menu, users can select the phenotypes in order to view or download the phenotype data. The phenotype headings on the table are also clickable to plot violin plots or bar plots depending on the data type (quantitative or qualitative) of that phenotype column (Figures 3B, C).

The purpose of generating violin plots and bar plots is to show the distributions of the phenotype data for different alleles. From a violin plot, users can understand the quantitative distribution of phenotype data based on the maximum, minimum, first quantile, third quantile, mean, and median values of that box. Likewise, users can uncover the qualitative distribution of phenotype data in a bar plot according to the counts of different qualitative categories. From the plots, users can also potentially understand the significance and linkage between alleles and phenotypes. Besides the plots, summary tables are also provided at the bottom of each figure to summarize the counts of accessions based on genotypes and either the existence of phenotype data for quantitative measurement or the categories of qualitative measurement. At the bottom of the page, users can also visualize the distribution of improvement status of accessions in different genotypes to understand the linkage between phenotype, improvement status, and genotypes.

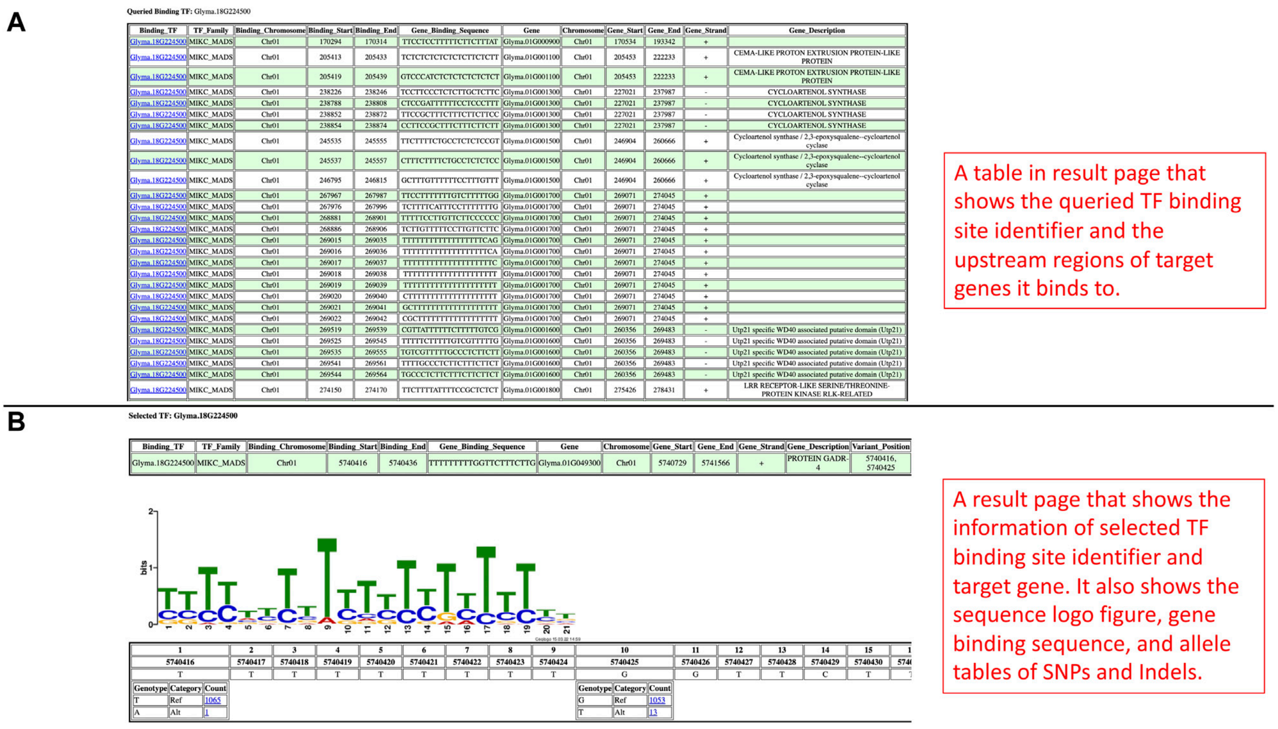
Promoter regions component—Search by Binding TFs

The Search by Binding TFs of the promoter regions component has a window, that is, composed of a TF binding site input box, a gene binding chromosome dropdown list, an upstream length input box, and a search button (Figure 1B). In the TF binding site identifier input box, users can input multiple TF binding site identifiers with each in a new line. The gene-binding chromosome drop-down menu allows users to select one chromosome per search. The upstream length input box is for users to input an integer value for upstream promoter regions of genes' calculations. When users click on the search button, the user input information is utilized in performing queries, and the results are returned to the promoter results page.

In the promoter results page, the information of each TF binding site identifier and its corresponding regulating genes are shown as an independent table (Figure 4A). In each table, the TF binding site identifier is clickable to redirect to a new page to display a sequence logo figure, position-nucleotide table, as well as SNPs and Indels in allele tables (Figure 4B). Similar to the results page in the Search by Gene IDs section, the alleles can also be linked with phenotype data in the phenotype data viewing page, and the functionalities such as data viewing, data downloading, and data plotting are also included as well (Figures 3A–C).

Copy number variation (CNV) component

The CNV analysis component consists of three different search sections which are the Search by Gene IDs section, Search By Accession and Copy Numbers section, and Search by Chromosome and Region section. Each section has a search window for users to input data for queries and the results of the



Queried Binding TF: Glyma.18G224500

Binding_TF	TF_Family	Binding_Chromosome	Binding_Start	Binding_End	Gene_Binding_Sequence	Gene	Chromosome	Gene_Start	Gene_End	Gene_Strand	Gene_Description
Glyma.18G224500	MIKC_MADS	Chr01	170294	170314	TTCCTCCTTTTTCTTCTTTAT	Glyma.01G000900	Chr01	170534	193342	+	
Glyma.18G224500	MIKC_MADS	Chr01	205413	205433	TCTCTCTCTCTCTCTTCTCTT	Glyma.01G001100	Chr01	205453	222233	+	CEMA-LIKE PROTON EXTRUSION PROTEIN-LIKE PROTEIN
Glyma.18G224500	MIKC_MADS	Chr01	205419	205439	GTCCCATCTCTCTCTCTCTCT	Glyma.01G001100	Chr01	205453	222233	+	CEMA-LIKE PROTON EXTRUSION PROTEIN-LIKE PROTEIN
Glyma.18G224500	MIKC_MADS	Chr01	238226	238246	TCCTTCCCTCTCTTGCTCTTC	Glyma.01G001300	Chr01	227021	237987	-	CYCLOARTENOL SYNTHASE
Glyma.18G224500	MIKC_MADS	Chr01	238788	238808	CTCCGATTTTTTCCTCCCTTT	Glyma.01G001300	Chr01	227021	237987	-	CYCLOARTENOL SYNTHASE
Glyma.18G224500	MIKC_MADS	Chr01	238852	238872	TTCCGCTTTCTTTCTTCTTCC	Glyma.01G001300	Chr01	227021	237987	-	CYCLOARTENOL SYNTHASE
Glyma.18G224500	MIKC_MADS	Chr01	238854	238874	CCTTCCGCTTTCTTTCTTCTT	Glyma.01G001300	Chr01	227021	237987	-	CYCLOARTENOL SYNTHASE
Glyma.18G224500	MIKC_MADS	Chr01	245535	245555	TTCTTTTCTGCCTCTCTCCGT	Glyma.01G001500	Chr01	246904	260666	+	Cycloartenol synthase / 2,3-epoxysqualene--cycloartenol cyclase
Glyma.18G224500	MIKC_MADS	Chr01	245537	245557	CTTTCTTTTCTGCCTCTCTCC	Glyma.01G001500	Chr01	246904	260666	+	Cycloartenol synthase / 2,3-epoxysqualene--cycloartenol cyclase
Glyma.18G224500	MIKC_MADS	Chr01	246795	246815	GCTTTGTTTTTTCCTTTGTTT	Glyma.01G001500	Chr01	246904	260666	+	Cycloartenol synthase / 2,3-epoxysqualene--cycloartenol cyclase
Glyma.18G224500	MIKC_MADS	Chr01	267967	267987	TTCCTTTTTTTGTCTTTTTGG	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	267976	267996	TCTTTTCATTTCCTTTTTTTG	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	268881	268901	TTTTTCCTTGTTCTTCCCCCC	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	268886	268906	TCTTGTTTTTCCTTGTTCTTC	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269015	269035	TTTTTTTTTTTTTTTTTTCAG	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269016	269036	TTTTTTTTTTTTTTTTTTTCA	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269017	269037	TTTTTTTTTTTTTTTTTTTTC	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269018	269038	TTTTTTTTTTTTTTTTTTTTT	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269019	269039	TTTTTTTTTTTTTTTTTTTTT	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269020	269040	CTTTTTTTTTTTTTTTTTTTT	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269021	269041	GCTTTTTTTTTTTTTTTTTTT	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269022	269042	CGCTTTTTTTTTTTTTTTTTT	Glyma.01G001700	Chr01	269071	274045	+	
Glyma.18G224500	MIKC_MADS	Chr01	269519	269539	CGTTATTTTTTCTTTTTGTCG	Glyma.01G001600	Chr01	260356	269483	-	Utp21 specific WD40 associated putative domain (Utp21)
Glyma.18G224500	MIKC_MADS	Chr01	269525	269545	TTTTTCTTTTTGTCGTTTTTG	Glyma.01G001600	Chr01	260356	269483	-	Utp21 specific WD40 associated putative domain (Utp21)
Glyma.18G224500	MIKC_MADS	Chr01	269535	269555	TGTCGTTTTTGCCCTCTTCTT	Glyma.01G001600	Chr01	260356	269483	-	Utp21 specific WD40 associated putative domain (Utp21)
Glyma.18G224500	MIKC_MADS	Chr01	269541	269561	TTTTGCCCTCTTCTTTCTTCT	Glyma.01G001600	Chr01	260356	269483	-	Utp21 specific WD40 associated putative domain (Utp21)
Glyma.18G224500	MIKC_MADS	Chr01	269544	269564	TGCCCTCTTCTTTCTTCTTCT	Glyma.01G001600	Chr01	260356	269483	-	Utp21 specific WD40 associated putative domain (Utp21)
Glyma.18G224500	MIKC_MADS	Chr01	274150	274170	TTCTTTTATTTTCCGCTCTCT	Glyma.01G001800	Chr01	275426	278431	+	LRR RECEPTOR-LIKE SERINE/THREONINE-PROTEIN KINASE RLK-RELATED

A table in result page that shows the queried TF binding site identifier and the upstream regions of target genes it binds to.

Selected TF: Glyma.18G224500

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.18G224500	MIKC_MADS	Chr01	5740416	5740436	TTTTTTTTTGGTTCTTTCTTG	Glyma.01G049300	Chr01	5740729	5741566	+	PROTEIN GADR-4	5740416, 5740425

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5740416	5740417	5740418	5740419	5740420	5740421	5740422	5740423	5740424	5740425	5740426	5740427	5740428	5740429	5740430
T	T	T	T	T	T	T	T	T	G	G	T	T	C	T

Genotype	Category	Count
T	Ref	1065
A	Alt	1

Genotype	Category	Count
G	Ref	1053
T	Alt	13

A result page that shows the information of selected TF binding site identifier and target gene. It also shows the sequence logo figure, gene binding sequence, and allele tables of SNPs and Indels.

FIGURE 4
(A) The promoter results page is rendered based on the query from the Search by Binding TFs section. It includes tables showing the information on TF binding sites and target genes. **(B)** This page is shown when users select a TF binding site identifier. This page has information related to the selected TF binding site, sequence logo figure, and position-nucleotide table.

queries are rendered on the corresponding results pages. Here, each section is discussed in more detail.

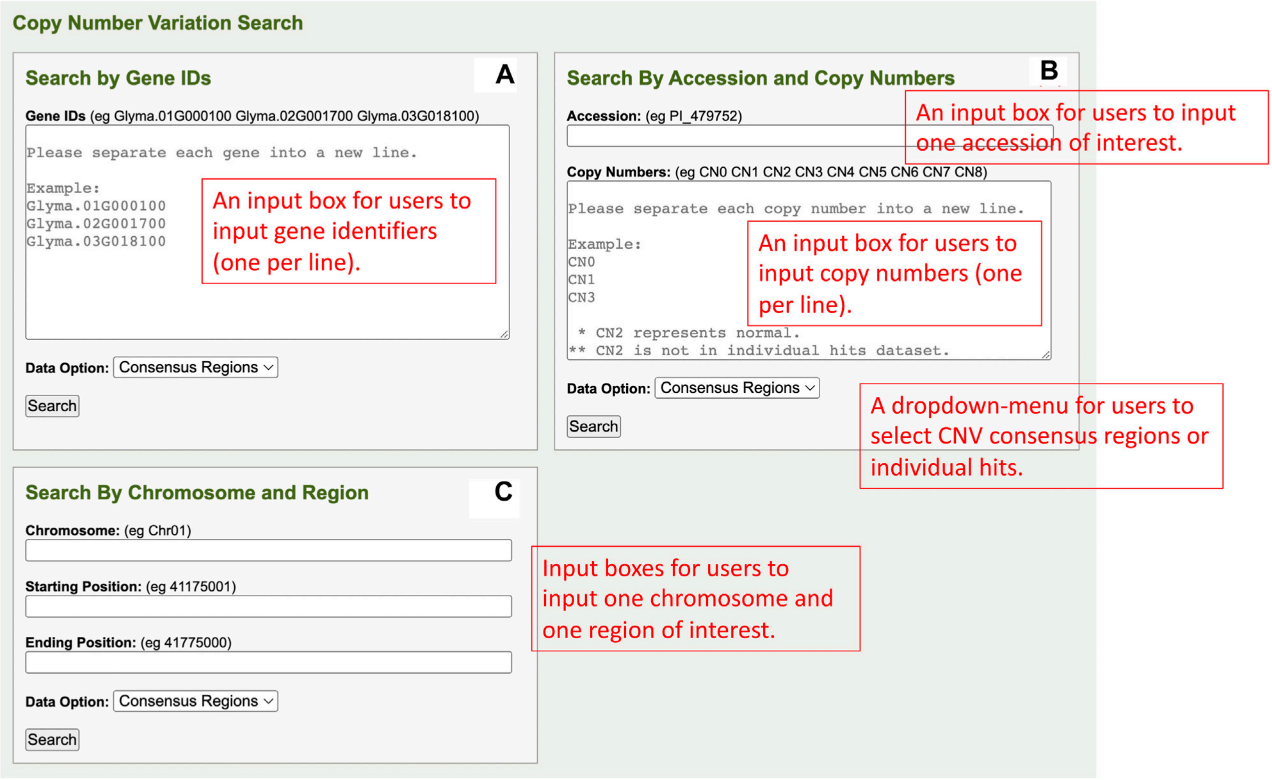
CNV analysis component—Search by Gene IDs

In the Search by Gene IDs of the CNV analysis component, there is one gene ID input box, a data option dropdown menu, and a search button in a window (Figure 5A). Users can input multiple genes of interest into the gene IDs input box, and each gene ID has to be separated into a new line. In the data option dropdown menu, users can select either consensus regions or individual hits. The consensus regions option is for CNV data summarized across accessions, and the individual hits option is for the CNV individual region of each accession. After the user completes the inputs, the user can click on the search button to perform queries and redirect to the results page.

There are three sections on the Search by Gene IDs' results page which are the queried genes section, the CNV regions and accession counts section, and the neighboring genes in different CNV regions section (Figure 6). The queried genes section displays genes' relevant information like chromosomes, coordinates, strands, identifiers, and descriptions. Based on the coordinates of the queried genes, CNV regions that enclosed the queried gene coordinates are displayed in the CNV regions and accession counts section as a table along with the counts of accessions within copy numbers (CN0–CN8). According to the cn. MOPS tool, CN0 and CN1 represent loss, CN2 is normal, and CN3 to CN8 represent gain (Klambauer et al., 2012). Each CN region and the accession counts within copy numbers are organized in a row. At the end of each row, there are view details button and connect phenotypes button that can be clicked to redirect to the detail viewing page and phenotype data viewing page. In the neighboring genes in different CNV regions section, each CNV region and the genes within that CNV region are shown as an independent table.

In the detail viewing page, the distribution of the improvement status in different copy numbers is presented in a bar plot fashion for the selected CNV region (Figure 7). From the bar plot, users can visualize the distributions and uncover significant improvement status that links with a particular copy number if possible. At the bottom of the figure, there is a summary table that summarizes the counts of accessions by improvement status and copy numbers. Within the table, the percentages of accession counts are also calculated out of total accessions. Apart from the figure and summary table, a full table with information such as CNV region, CNV region length, accessions, improvement status, and copy numbers is also provided.

If users would like to gain information about accessions and phenotypes, they can click the connect phenotypes button to redirect to the phenotype data viewing page (Figure 8A). On the phenotype data viewing page, users can select the copy numbers of interest and click the view data button for seeing the data or the download data button to collect the data in a comma-separated values (CSV) file



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
```

An input box for users to input gene identifiers (one per line).

Data Option: Consensus Regions

Search By Accession and Copy Numbers

Accession: (eg PI_479752)

An input box for users to input one accession of interest.

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.
```

An input box for users to input copy numbers (one per line).

Data Option: Consensus Regions

A dropdown-menu for users to select CNV consensus regions or individual hits.

Search By Chromosome and Region

Chromosome: (eg Chr01)

Starting Position: (eg 41175001)

Ending Position: (eg 41775000)

Data Option: Consensus Regions

Input boxes for users to input one chromosome and one region of interest.

FIGURE 5
(A) The Search by Gene IDs window in the CNV analysis component. (B) The Search By Accession and Copy Numbers window in the CNV analysis component. (C) The Search by Chromosome and Region window in the CNV analysis component.

format. Additionally, users can also select copy numbers and phenotypes of interest to overlap and view the data. The data in a tabular fashion allows users to click on a phenotype heading for plotting the distributions of the corresponding phenotype data in a violin plot or a bar chart depending on the data type (quantitative or qualitative) (Figures 8B, C).

In a violin plot, users can visualize the quantitative distributions of the phenotype data by different copy numbers. If users hover the pointer on a box in the figure, they can visualize the maximum, minimum, first quantile, third quantile, mean, and median values of that box. Users can also show and hide copy numbers by toggling the elements in the legend. In a bar plot, users can visualize and compare between counts of qualitative categories in the phenotype data by copy numbers. Similarly, users can also toggle elements in the legend to show or hide categories. At the bottom of the violin plot or bar plot, there is a summary table that summarizes the count of accessions by copy numbers. The summarization for quantitative data is counts of accessions by the existence of phenotype data, whereas the summarization for qualitative data is counts and percentages of accessions of each qualitative category. The last figure on the page is a bar plot that shows the distribution of the improvement status of accessions based on selected copy numbers. The figure aims to provide a linkage between the phenotype, improvement status, and copy numbers so that users can understand the improvement status that leads to the phenotype patterns by copy numbers.

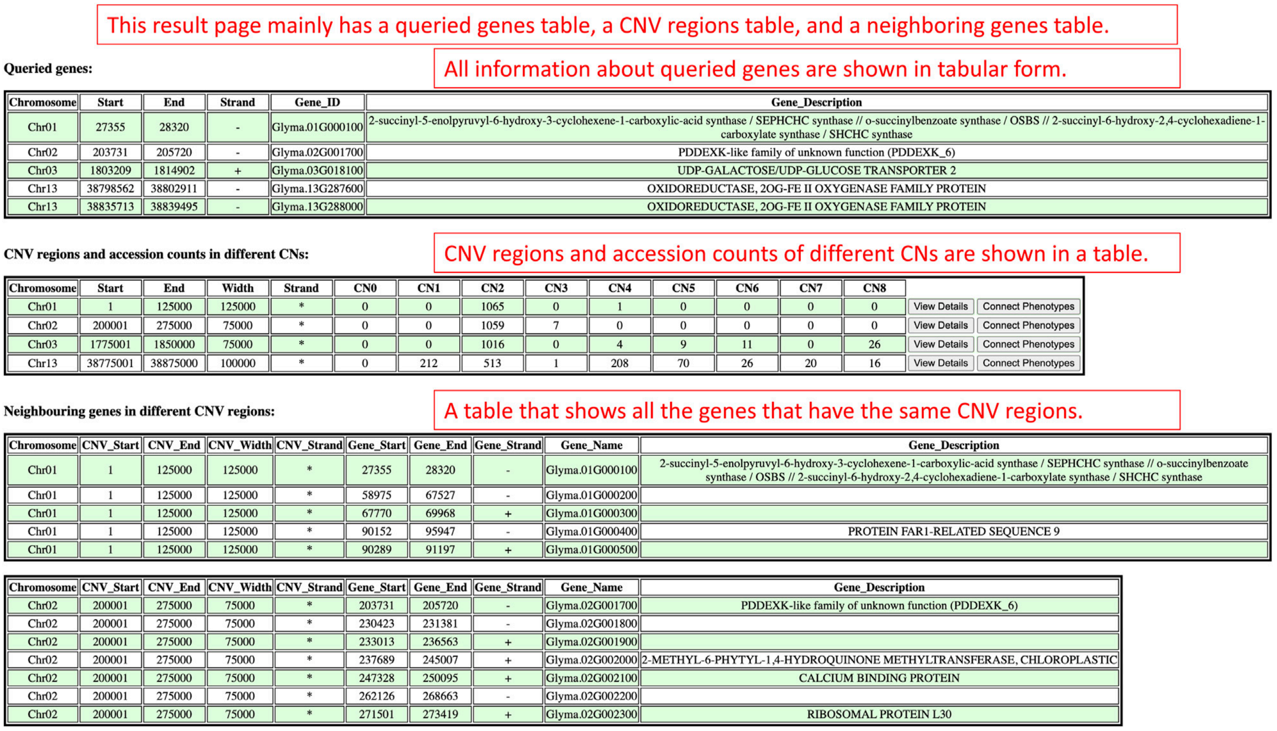
CNV analysis component—Search by Accession and Copy Numbers

The Search By Accession and Copy Numbers section of the CNV analysis component has an accession input box, a copy number input box, a data option dropdown menu, and a search button in a window (Figure 5B). Users can input only one accession into the accession input box, input multiple copy numbers into the copy numbers input box (each in a new line), and select either consensus regions or individual hits in the data option dropdown menu. Upon completing all the fields in the window, users can click on the search button so that queries can be performed and rendered on the results page.

On the results page, there is only one table that has the information related to the inputted accession and copy numbers (Figure 9). The CNV regions on the table have details such as the region chromosome, region start, region end, width, strand, accession, and copy number. The purpose of this table is to show users all the CNV regions of a particular accession and copy numbers.

CNV analysis component—Search by Chromosome and Region

The Search by Chromosome and Region section of the CNV analysis component has one window that consists of a chromosome input box, a starting position input box, an ending position input box, a data option dropdown menu, and a search button (Figure 5C). Users can input the



This result page mainly has a queried genes table, a CNV regions table, and a neighboring genes table.

Queried genes:

All information about queried genes are shown in tabular form.

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
Chr13	38798562	38802911	-	Glyma.13G287600	OXIDOREDUCTASE, 2OG-FE II OXYGENASE FAMILY PROTEIN
Chr13	38835713	38839495	-	Glyma.13G288000	OXIDOREDUCTASE, 2OG-FE II OXYGENASE FAMILY PROTEIN

CNV regions and accession counts in different CNs:

CNV regions and accession counts of different CNs are shown in a table.

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
Chr13	38775001	38875000	100000	*	0	212	513	1	208	70	26	20	16	View Details	Connect Phenotypes

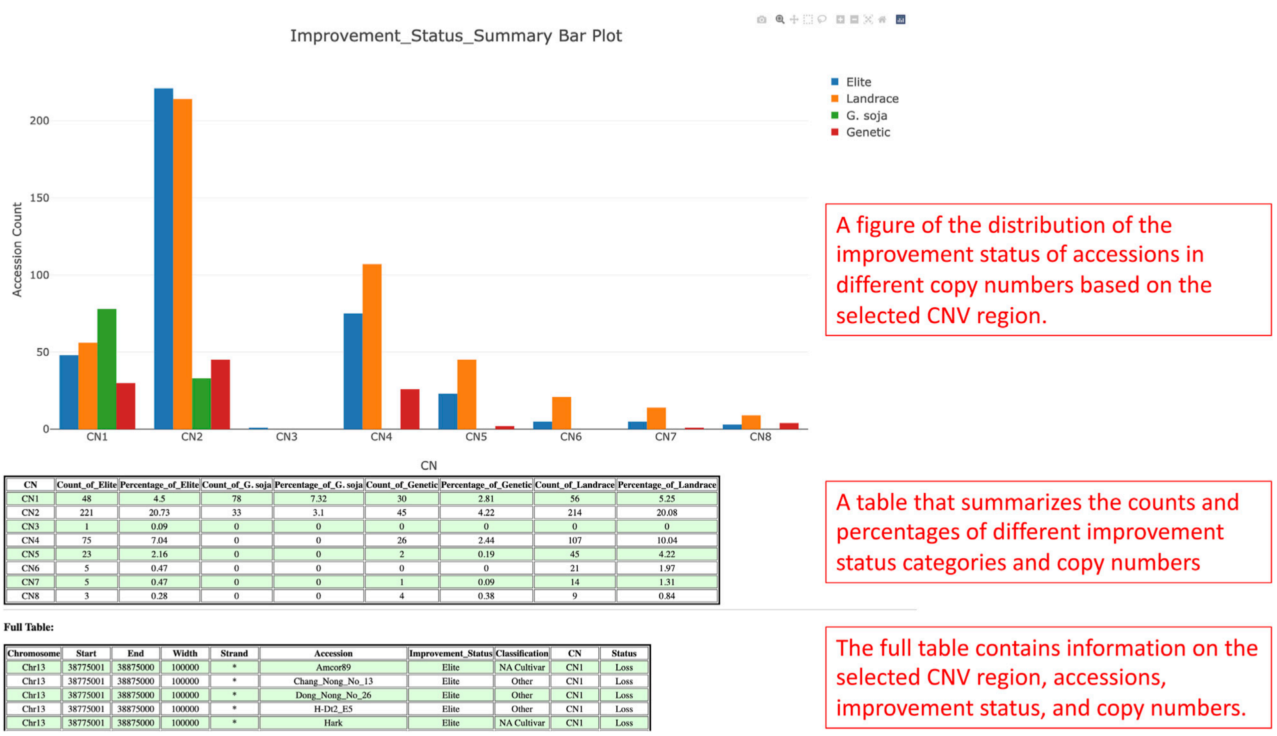
Neighbouring genes in different CNV regions:

A table that shows all the genes that have the same CNV regions.

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	

Chromosome	CNV_Start	CNV_End	CNV_Width	CNV_Strand	Gene_Start	Gene_End	Gene_Strand	Gene_Name	Gene_Description
Chr02	200001	275000	75000	*	203731	205720	-	Glyma.02G001700	PDDEXK-like family of unknown function (PDDEXK_6)
Chr02	200001	275000	75000	*	230423	231381	-	Glyma.02G001800	
Chr02	200001	275000	75000	*	233013	236563	+	Glyma.02G001900	
Chr02	200001	275000	75000	*	237689	245007	+	Glyma.02G002000	2-METHYL-6-PHYTYL-1,4-HYDROQUINONE METHYLTRANSFERASE, CHLOROPLASTIC
Chr02	200001	275000	75000	*	247328	250095	+	Glyma.02G002100	CALCIUM BINDING PROTEIN
Chr02	200001	275000	75000	*	262126	268663	-	Glyma.02G002200	
Chr02	200001	275000	75000	*	271501	273419	+	Glyma.02G002300	RIBOSOMAL PROTEIN L30

FIGURE 6
The results render on the results page when users use the Search by Gene IDs window in the CNV analysis component.



A figure of the distribution of the improvement status of accessions in different copy numbers based on the selected CNV region.

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
CN1	48	4.5	78	7.32	30	2.81	56	5.25
CN2	221	20.73	33	3.1	45	4.22	214	20.08
CN3	1	0.09	0	0	0	0	0	0
CN4	75	7.04	0	0	26	2.44	107	10.04
CN5	23	2.16	0	0	2	0.19	45	4.22
CN6	5	0.47	0	0	0	0	21	1.97
CN7	5	0.47	0	0	1	0.09	14	1.31
CN8	3	0.28	0	0	4	0.38	9	0.84

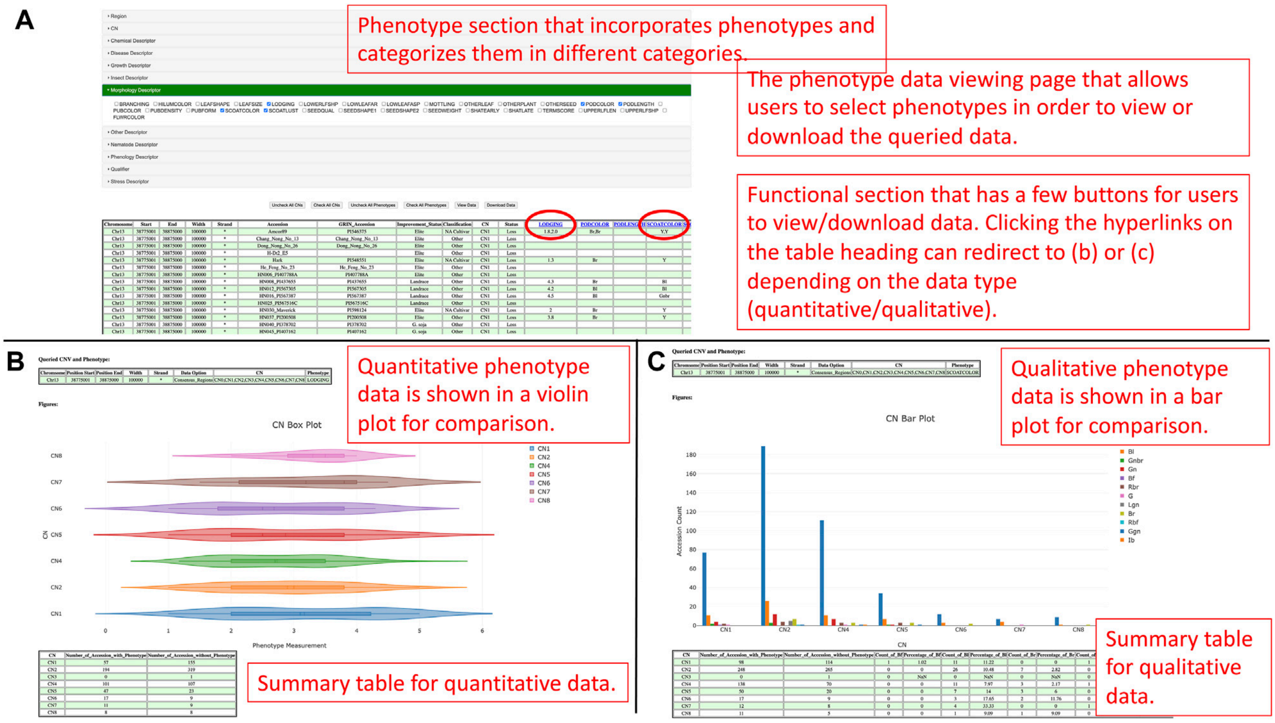
A table that summarizes the counts and percentages of different improvement status categories and copy numbers

Full Table:

Chromosome	Start	End	Width	Strand	Accession	Improvement_Status	Classification	CN	Status
Chr13	38775001	38875000	100000	*	Amcor89	Elite	NA Cultivar	CN1	Loss
Chr13	38775001	38875000	100000	*	Chang_Nong_No_13	Elite	Other	CN1	Loss
Chr13	38775001	38875000	100000	*	Dong_Nong_No_26	Elite	Other	CN1	Loss
Chr13	38775001	38875000	100000	*	H-Dt2_E5	Elite	Other	CN1	Loss
Chr13	38775001	38875000	100000	*	Hark	Elite	NA Cultivar	CN1	Loss

The full table contains information on the selected CNV region, accessions, improvement status, and copy numbers.

FIGURE 7
The detail viewing page shows the distribution of the improvement status of the selected CNV region along with a summary table of the figure and a full table of the whole CNV region, accessions, improvement status, and copy numbers data.



Phenotype section that incorporates phenotypes and categorizes them in different categories.

The phenotype data viewing page that allows users to select phenotypes in order to view or download the queried data.

Functional section that has a few buttons for users to view/download data. Clicking the hyperlinks on the table heading can redirect to (b) or (c) depending on the data type (quantitative/qualitative).

Quantitative phenotype data is shown in a violin plot for comparison.

Qualitative phenotype data is shown in a bar plot for comparison.

CN	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype
CN1	57	155
CN2	194	319
CN3	0	1
CN4	101	107
CN5	47	23
CN6	17	9
CN7	11	9
CN8	8	8

Summary table for quantitative data.

CN	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype	Count_of_Bf	Percentage_of_Bf	Count_of_Bl	Percentage_of_Bl	Count_of_Br	Percentage_of_Br	Count_of
CN1	98	114	1	1.02	11	11.22	0	0	1
CN2	248	265	0	0	26	10.48	7	2.82	0
CN3	0	1	0	NaN	0	NaN	0	NaN	0
CN4	138	70	0	0	11	7.97	3	2.17	1
CN5	50	20	0	0	7	14	3	6	0
CN6	17	9	0	0	3	17.65	2	11.76	0
CN7	12	8	0	0	4	33.33	0	0	1
CN8	11	5	0	0	1	9.09	1	9.09	0

Summary table for qualitative data.

FIGURE 8
(A) The phenotype data viewing page for users to select copy numbers and phenotypes they are interested in and view or download the data. **(B)** The distribution of a quantitative trait is plotted against copy numbers in a violin plot figure. **(C)** The distribution of a qualitative trait is plotted against copy numbers in a bar plot.

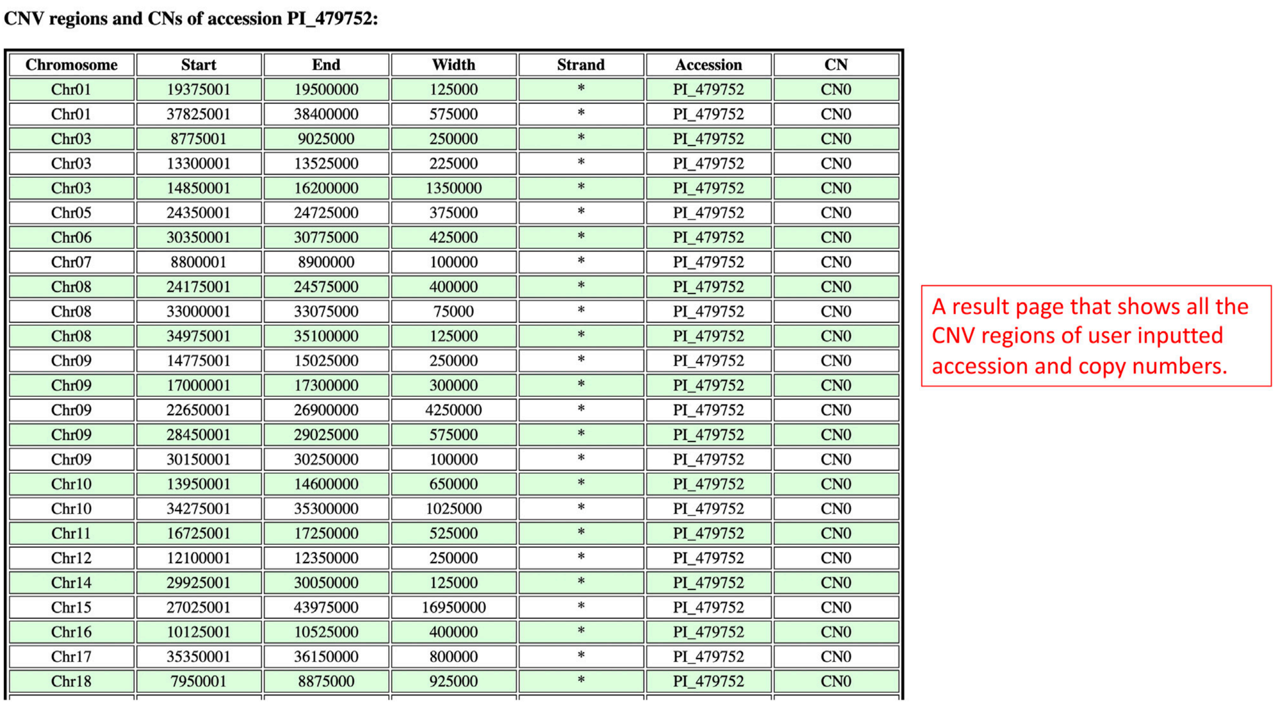
region of interest with chromosome, starting position, and ending position into the respective input boxes and select either consensus region or individual hits in the data option dropdown menu. When users are ready to perform queries, they can click the search button, and, simultaneously, the users are redirected to the results page.

On the results page, there are two sections which are the queried CNV regions and accession counts section, and the accessions and copy numbers within the queried CNV region section (Figure 10). The queried CNV regions and accession counts section shows a table that contains CNV regions that are bounded between the region of interest input by users. In this section, the table also displays the accession counts in each copy number within each CNV region. At the end of the table, users can also access the view details page and phenotype data viewing page with buttons to understand the improvement status distribution in that CNV region or connect copy numbers and accessions with phenotype data. In order to show more details of each CNV region, the accessions and copy numbers within the queried CNV region section present each CNV region along with all the accessions and copy numbers of that CNV region in a table. If users are interested in knowing the copy number variations within a region of interest, the results presented on this results page will suit their needs.

Case studies

Analysis of CNV distribution by improvement status offers insight into soybean domestication-related gene gain that impacts plant height.

Gibberellin acid oxidase 2 (*GA2ox*) is an enzyme that, besides other enzymes, converts bioactive phytohormone gibberellins (GA) into inactive forms (Thomas et al., 1999). Modulation of GAs metabolism genes played an important role in the green revolution of crop improvement. In soybean, reducing trailing growth and shoot length were observed in soybean *GA2ox8* overexpressing mutants. Furthermore, plant height of wild *G. soja* (*Glycine soja*) ancestor W05 was associated with less copy number of *GA2ox8A* (*Glyma.13G287600*) and *GA2ox8B* (*Glyma.13G288000*) in comparison to generally shorter cultivated *G.* max *C08* (Wang et al., 2021). Thus, this suggests an important role in the *GA2ox8* copy number increase during domestication. Here, we analyzed CNV in the *GA2ox8*-related cn. MOPS predicted CNV associated region, that is, shown in our data on chromosome 13 (Chr13:38,798,562-38,802,911). There are 513 accessions with normal CN, 212 accessions with CN loss, and 341 accessions with CN gain (Figure 11A). Figure 11B illustrates the distribution of CNVs by improvement status in the soybean 1066 accessions and demonstrates that all G. soja accessions possess either CN1 or CN2 that are considered as loss or normal CNV whereas accessions with the other improvement status are all *G.* max (*Glycine max*) and can potentially bear more *GA2ox8* copies. This result is in accordance with Wang et al. (2021) and thus, supports the hypothesis of the *GA2ox8* gain during soybean domestication (Wang et al., 2021). We further associated the observed CNV with soybean plant height phenotype (Figure 12A). Since there is no phenotype data available for any G. soja accessions for plant height in the

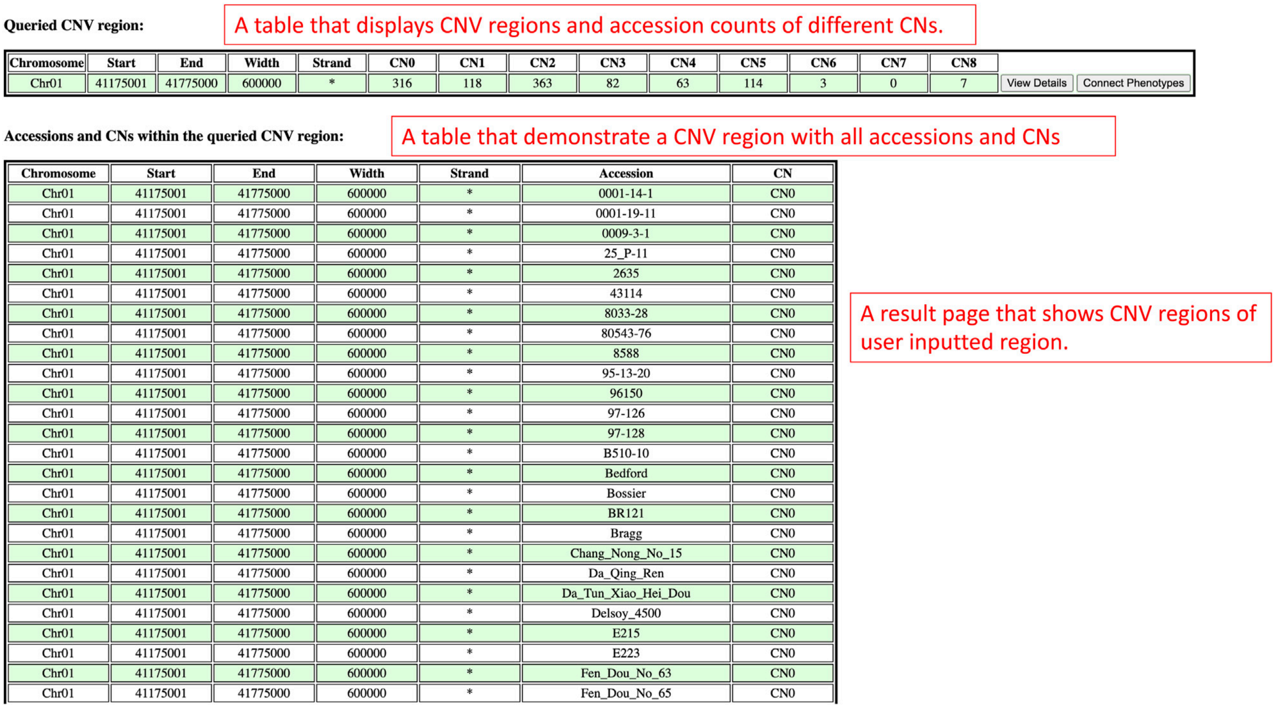


CNV regions and CNs of accession PI_479752:

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
Chr11	16725001	17250000	525000	*	PI_479752	CN0
Chr12	12100001	12350000	250000	*	PI_479752	CN0
Chr14	29925001	30050000	125000	*	PI_479752	CN0
Chr15	27025001	43975000	16950000	*	PI_479752	CN0
Chr16	10125001	10525000	400000	*	PI_479752	CN0
Chr17	35350001	36150000	800000	*	PI_479752	CN0
Chr18	7950001	8875000	925000	*	PI_479752	CN0

A result page that shows all the CNV regions of user inputted accession and copy numbers.

FIGURE 9
The results page redirected from the Search By Accession and Copy Numbers window to show CNV regions related to the inputted accession and copy numbers.



Queried CNV region:

A table that displays CNV regions and accession counts of different CNs.

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

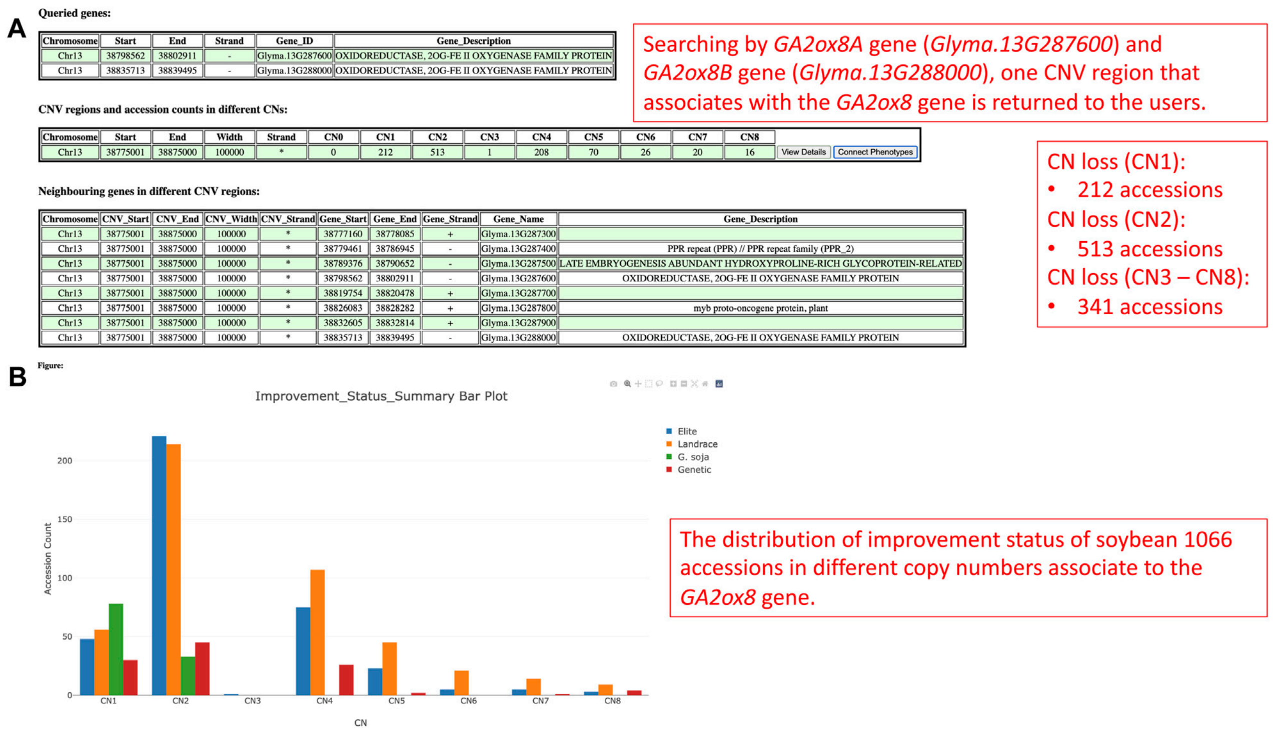
Accessions and CNs within the queried CNV region:

A table that demonstrate a CNV region with all accessions and CNs

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
Chr01	41175001	41775000	600000	*	Bragg	CN0
Chr01	41175001	41775000	600000	*	Chang_Nong_No_15	CN0
Chr01	41175001	41775000	600000	*	Da_Qing_Ren	CN0
Chr01	41175001	41775000	600000	*	Da_Tun_Xiao_Hei_Dou	CN0
Chr01	41175001	41775000	600000	*	Delsoy_4500	CN0
Chr01	41175001	41775000	600000	*	E215	CN0
Chr01	41175001	41775000	600000	*	E223	CN0
Chr01	41175001	41775000	600000	*	Fen_Dou_No_63	CN0
Chr01	41175001	41775000	600000	*	Fen_Dou_No_65	CN0

A result page that shows CNV regions of user inputted region.

FIGURE 10
The results page redirected from the Search by Chromosome and Region window to display CNV regions between users' region of interest.



Queried genes:

Chromosome	Start	End	Strand	Gene_ID	Gene_Description
Chr13	38798562	38802911	-	Glyma.13G287600	OXIDOREDUCTASE, 2OG-FE II OXYGENASE FAMILY PROTEIN
Chr13	38835713	38839495	-	Glyma.13G288000	OXIDOREDUCTASE, 2OG-FE II OXYGENASE FAMILY PROTEIN

Searching by *GA2ox8A* gene (*Glyma.13G287600*) and *GA2ox8B* gene (*Glyma.13G288000*), one CNV region that associates with the *GA2ox8* gene is returned to the users.

CNV regions and accession counts in different CNs:

Chromosome	Start	End	Width	Strand	CN0	CN1	CN2	CN3	CN4	CN5	CN6	CN7	CN8		
Chr13	38775001	38875000	100000	*	0	212	513	1	208	70	26	20	16	View Details	Connect Phenotypes

CN loss (CN1):
- 212 accessions

CN loss (CN2):
- 513 accessions

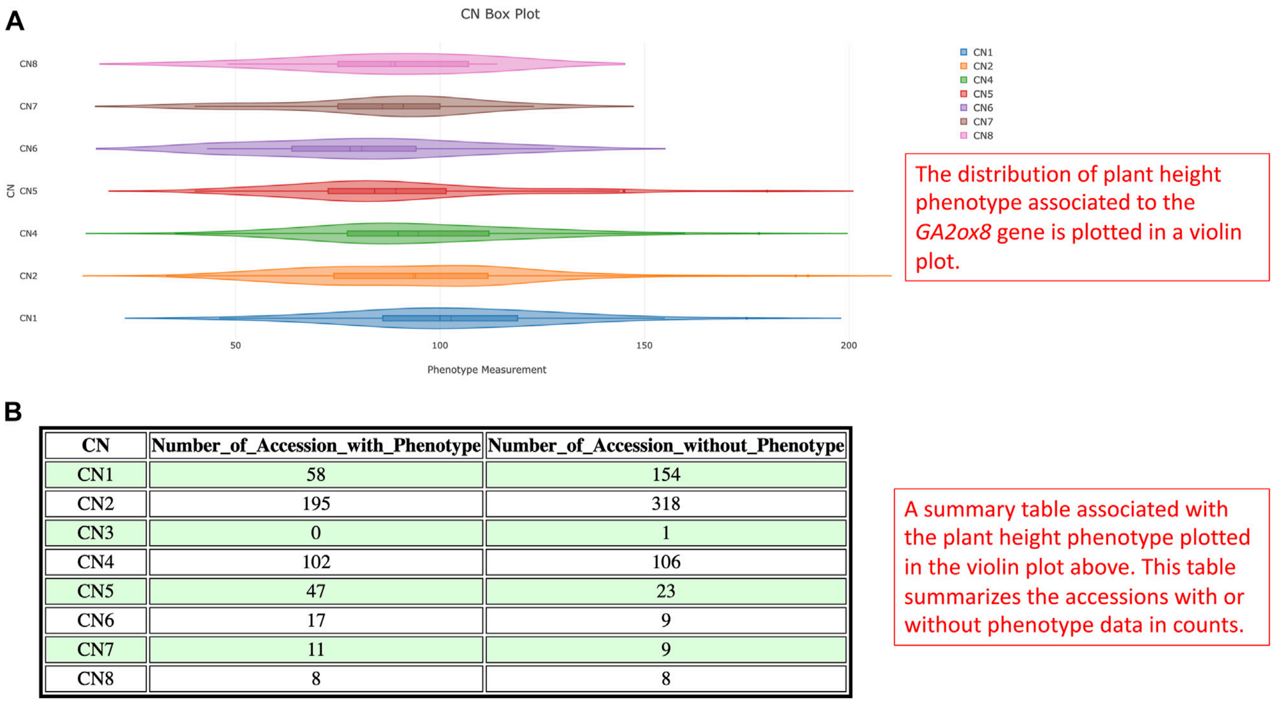
CN loss (CN3 – CN8):
- 341 accessions

Neighbouring genes in different CNV regions:

Chromosome	CNV_Start	CNV_End	CNV_Width	CNV_Strand	Gene_Start	Gene_End	Gene_Strand	Gene_Name	Gene_Description
Chr13	38775001	38875000	100000	*	38777160	38778085	+	Glyma.13G287300	
Chr13	38775001	38875000	100000	*	38779461	38786945	-	Glyma.13G287400	PPR repeat (PPR) // PPR repeat family (PPR_2)
Chr13	38775001	38875000	100000	*	38789376	38790652	-	Glyma.13G287500	LATE EMBRYOGENESIS ABUNDANT HYDROXYPROLINE-RICH GLYCOPROTEIN-RELATED
Chr13	38775001	38875000	100000	*	38798562	38802911	-	Glyma.13G287600	OXIDOREDUCTASE, 2OG-FE II OXYGENASE FAMILY PROTEIN
Chr13	38775001	38875000	100000	*	38819754	38820478	+	Glyma.13G287700	
Chr13	38775001	38875000	100000	*	38826083	38828282	+	Glyma.13G287800	myb proto-oncogene protein, plant
Chr13	38775001	38875000	100000	*	38832605	38832814	+	Glyma.13G287900	
Chr13	38775001	38875000	100000	*	38835713	38839495	-	Glyma.13G288000	OXIDOREDUCTASE, 2OG-FE II OXYGENASE FAMILY PROTEIN

The distribution of improvement status of soybean 1066 accessions in different copy numbers associate to the *GA2ox8* gene.

FIGURE 11
(A) The CNV region on chromosome 13 that is, associated with the GA2ox8 gene. **(B)** The soybean 1066 accessions' improvement status distribution in different copy numbers.



The distribution of plant height phenotype associated to the *GA2ox8* gene is plotted in a violin plot.

CN	Number_of_Accession_with_Phenotype	Number_of_Accession_without_Phenotype
CN1	58	154
CN2	195	318
CN3	0	1
CN4	102	106
CN5	47	23
CN6	17	9
CN7	11	9
CN8	8	8

A summary table associated with the plant height phenotype plotted in the violin plot above. This table summarizes the accessions with or without phenotype data in counts.

FIGURE 12
(A) The violin plot illustrates the statistical distribution of plant height of soybean 1066 accessions for CNV in the GA2ox8 gene-associated region on chromosome 13. **(B)** A summary table associated with the violin plot documents the corresponding phenotype counts and missing information.

GRIN database, the conclusions made based on this analysis might be influenced by this fact. However, when comparing the three CNVs with the highest accession counts/known phenotype (CN2 - normal, CN1—loss, and CN4—gain, Figure 12B) here we can see median shifts to increase plant height of accessions with CNV loss (CNV1) in comparison to slightly reduced plant height of normal CNV (CNV2) and CNV gain (CNV4). Most importantly, among 208 accessions with CN4, 141 are Elite accessions with 83 out of the 102-known phenotype accessions in this group. Thus, this result indicates that the *GA2ox8* CN gain might be responsible for plant height in cultivated soybean varieties as demonstrated by Wang et al. (2021); Wang et al. (2021).

Discussion

The GenVarX toolset offers tools that enable online analyses of the associations of pre-defined phenotypes with their variations in promoter regions and CNV on soybean, rice, and *Arabidopsis*. The toolset provides query, association testing, visualization, and download capabilities for users to obtain new insight from the promoter and CNV data. Using the toolset, users can interact with the web interfaces to do their research without the need of spending long-time processing and running big data to gain promoter and CNV results. It also provides opportunities for users who do not have large-scale servers and large storage space to have access to the promoter and CNV data.

In the development of the GenVarX toolset, we faced some challenges in SNP data processing and phenotype data retrieval. SNP data in a VCF file usually consists of many positions and accessions in a single file. Processing the SNP data in a VCF file and uploading the processed SNP data into a database are usually time-consuming processes. A divide-and-conquer strategy is usually required in code implementation to achieve a certain speed-up in the processing and uploading. Furthermore, multi-processing and multi-threading can also be helpful when processing and uploading the data programmatically. Besides that, retrieving the SNP data from the database could also be a slow process. Therefore, a database indexing method is required to increase the data retrieval speed. Apart from the SNP data processing problem, we faced another challenge that was caused by the limited availability of phenotype data for rice and *Arabidopsis*. To solve this problem, online data explorations are required in order to find suitable data for our GenVarX toolset.

The GenVarX toolset is developed with the incorporation of extensive capabilities that are related to genotype and phenotype. The extensions of showing gene binding sequences, sequence logo figures, mutative variant positions, as well as the linkage of the genotype data to phenotype data are the strengths of this toolset. Although PlantTFDB and PlantRegMap are the main sources for the datasets of the GenVarX toolset, these new capabilities are not developed in their web portals. Another promoter database is the Eukaryotic Promoter Database (EPD) (Périer et al., 2000) which also allows users to search for promoters. However, important crop species like soybean and rice are not available in that database. In terms of CNV, there is a lack of plant and crop related CNV databases for users to perform queries, visualize CNV, and link CNV to phenotypes. Thus, our research group developed the GenVarX toolset to assist the research community to advance their research.

In future development, our research group will focus on expanding the GenVarX toolset to support more organisms. Besides that, we will also incorporate more phenotypic data into the database to allow users to visualize different phenotypes more easily. Furthermore, integration of CNV results from other tools can also be done to make the CNV component of the GenVarX toolset capable of a more enriched comparative analysis. As mentioned in Gabrielaite et al., which shows the comparisons of different CNV tools in terms of outcomes, metrics, and performance, there are several tools such as GATK gCNV, Lumpy, and DELLY, that were developed with different methodologies that outperform many other CNV tools (Gabrielaite et al., 2021). These CNV tools can be used to further analyze our data and integrate into the GenVarX toolset so that users can select CNVs from different methods to link with genotype and phenotype.

Conclusion

In the GenVarX toolset development, we have collected and processed publicly available data from various sources and platforms. Having the data, we have built the GenVarX toolset that has promoter regions and CNV analysis components for soybean, rice, and *Arabidopsis*. The soybean GenVarX toolset is deployed on the SoyKB website (https://soykb.org/SoybeanGenVarX/) whereas the universal GenVarX toolset for other organisms is deployed on the KBCommons website (https://kbcommons.org/system/tools/GenVarX/Osativa and https://kbcommons.org/system/tools/GenVarX/Athaliana). The broad plant research community can utilize the GenVarX toolset as a comprehensive source of information to gain insights into soybean, rice, and *Arabidopsis* promoter regions or CNV analysis outcomes. More specifically, a better understanding of variations in the TF binding sites and CNV can be predicted with the toolset. Hence, it serves as a valuable pre-experimental step for further gene transcription studies.

Available and requirements

Project Name: GenVarX
Project Homepage:

- 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

Programming Languages:

- Data Analytics: Python and R

- Web Development: PHP, HTML, CSS, and JavaScript

Other Requirements:

- Data Analytics:
  o Python 3.7.0 or higher
  o R 3.6.0 or higher
  o SQLAlchemy 1.4.41 or higher
  o cn.MOPS 1.40.0 or higher
  o Burrows-Wheeler Aligner (BWA) 0.7.17
  o Genome Analysis Toolkit (GATK) 4.2.6.1
- Web Development:
  o PHP 8
- Web Browsing:
  o Google Chrome (Recommended), Firefox, or Microsoft Edge

Source Code:

- GenVarX Data Processing Scripts: https://github.com/yenon118/GenVarX_Data_Processing
- Soybean GenVarX Toolset Source Code: https://github.com/yenon118/SoybeanGenVarX
- Rice and *Arabidopsis* Toolsets Source Code: https://github.com/yenon118/GenVarX

License:

- Soybean GenVarX Toolset: MIT License
- Rice GenVarX Toolset: MIT License
- *Arabidopsis* GenVarX Toolset: MIT License

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

YC collected the transcription factor related data, sequencing data, annotation data, genotypic, and phenotypic data. YC created scripts to generate the copy number variation data and developed the GenVarX toolset. YC wrote the first draft of the manuscript. TJ, MŠ, and KB proposed the tool functionalities. YC, AM, ND, MŠ, KB, and TJ participated in the GenVarX toolset design process. MŠ and JB performed numerous tests in the tool development cycles, conducted analyses, and assembled presentable results. YC, KB, MŠ, and TJ revised and edited the manuscript. TJ provided expert guidance on the conceptualization and development process of the toolset in SoyKB and KBCommons. All authors contributed to the article and approved the submitted version.

Funding

The research was supported using Missouri soybean farmers' checkoff dollars provided by the United Soybean Board (USB). The principal investigators, project title, and grant ID information are as below: 1. Dr. Trupti Joshi and Dr. Kristin Bilyeu; Applied Genomics to Improve Soybean Seed Protein; #1920-152-0131-C. 2. Dr. Trupti Joshi and Dr. Kristin Bilyeu; Enhancing Soybean Applied Genomics Tools for Improving Soybean; #2220-152-0202. 3. Dr. Trupti Joshi and Dr. Kristin Bilyeu; Leveraging Genomics to Enhance The US Soybean Quality Reputation; #2332-201-0101.

Acknowledgments

Our research group would like to acknowledge Sai Preethi Induri for assisting us in setting up the RiceKB on the KBCommons website. We also would like to thank Ajay Kumar and Adama Tukuli for providing assistance in searching for rice phenotypic datasets.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alonso-Blanco, C., Andrade, J., Becker, C., Bemm, F., Bergelson, J., Borgwardt, K. M., et al. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. *Cell* 166, 481–491. doi:10.1016/j.cell.2016.05.063

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res.* 43, W39–W49. doi:10.1093/nar/gkv416

Bayer, M. (2012). *SQLAlchemy. Mountain view: aosabook.org*.

Bolger, M., Schwacke, R., Gundlach, H., Schmutzer, T., Chen, J., Arend, D., et al. (2017). From plant genomes to phenotypes. *J. Biotechnol.* 261, 46–52. doi:10.1016/j.jbiotec.2017.06.003

Castro-Mondragon, J. A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu lemma, R., Turchi, L., Blanc-Mathieu, R., et al. (2021). JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 50, D165–D173. doi:10.1093/nar/gkab1113

Gabrielaite, M., Torp, M. H., Rasmussen, M. S., Andreu-Sánchez, S., Vieira, F. G., Pedersen, C. B., et al. (2021). A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data. *Cancers* 13, 6283. doi:10.3390/cancers13246283

Goff, S., Vaughn, M., Mckay, S., Lyons, E., Stapleton, A., Gessler, D., et al. (2011). The iPlant Collaborative: cyberinfrastructure for plant biology. *Front. Plant Sci.* 2, 34. doi:10.3389/fpls.2011.00034

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., et al. (2011). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40, D1178–D1186. doi:10.1093/nar/gkr944

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., et al. (2016). PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45, D1040–D1045. doi:10.1093/nar/gkw982

Joshi, T., Fitzpatrick, M. R., Chen, S., Liu, Y., Zhang, H., Endacott, R. Z., et al. (2013). Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. *Nucleic Acids Res.* 42, D1245–D1252. doi:10.1093/nar/gkt905

Joshi, T., Patil, K., Fitzpatrick, M. R., Franklin, L. D., Yao, Q., Cook, J. R., et al. (2012). Soybean Knowledge Base (SoyKB): a web resource for soybean translational genomics. *BMC Genomics* 13, S15. doi:10.1186/1471-2164-13-S1-S15

Joshi, T., Wang, J., Zhang, H., Chen, S., Zeng, S., Xu, B., et al. (2017). "The Evolution of Soybean Knowledge Base (SoyKB)," in *Plant genomics databases: methods and protocols*. Editor A. D. J. Van Dijk (New York, NY: Springer New York), 149–159.

Kim, M. Y., Lee, S., Van, K., Kim, T.-H., Jeong, S.-C., Choi, I.-Y., et al. (2010). Whole-genome sequencing and intensive analysis of the undomesticated soybean (Glycine soja Sieb. and Zucc.) genome. *Proc. Natl. Acad. Sci.* 107, 22032–22037. doi:10.1073/pnas.1009526107

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U., et al. (2012). cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* 40, e69. doi:10.1093/nar/gks003

Li, A., Liu, A., Wu, S., Qu, K., Hu, H., Yang, J., et al. (2022). Comparison of structural variants in the whole genome sequences of two Medicago truncatula ecotypes: jemalong a17 and r108. *BMC Plant Biol.* 22, 77. doi:10.1186/s12870-022-03469-0

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760. doi:10.1093/bioinformatics/btp324

Liu, Y., Du, H., Li, P., Shen, Y., Peng, H., Liu, S., et al. (2020). Pan-Genome of Wild and Cultivated Soybeans. *Cell* 182, 162–176. doi:10.1016/j.cell.2020.05.023

Liu, Y., Khan, S. M., Wang, J., Rynge, M., Zhang, Y., Zeng, S., et al. (2016). PGen: large-scale genomic variations analysis workflow and browser in SoyKB. *BMC Bioinforma.* 17, 337. doi:10.1186/s12859-016-1227-y

Mckenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., et al. (2016). The iPlant Collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biol.* 14, e1002342. doi:10.1371/journal.pbio.1002342

Périer, R. C., Praz, V., Junier, T., Bonnard, C., and Bucher, P. (2000). The eukaryotic promoter database (EPD). *Nucleic Acids Res.* 28, 302–303. doi:10.1093/nar/28.1.302

Sakai, H., Lee, S. S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013). Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54, e6. doi:10.1093/pcp/pcs183

Samarakoon, P. S., Sorte, H. S., Stray-Pedersen, A., Rødningen, O. K., Rognes, T., and Lyle, R. (2016). cnvScan: a CNV screening and annotation tool to improve the clinical utility of computational CNV prediction from exome sequencing data. *BMC Genomics* 17, 51. doi:10.1186/s12864-016-2374-2

Schneider, T. D., and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18, 6097–6100. doi:10.1093/nar/18.20.6097

The 3,000 rice genomes project (2014). The 3,000 rice genomes project. *GigaScience* 3, 7. doi:10.1186/2047-217X-3-7

Thomas, S. G., Phillips, A. L., and Hedden, P. (1999). Molecular cloning and functional expression of gibberellin 2- oxidases, multifunctional enzymes involved in gibberellin deactivation. *Proc. Natl. Acad. Sci.* 96, 4698–4703. doi:10.1073/pnas.96.8.4698

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., and Gao, G. (2019). PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* 48, D1104-D1113–D1113. doi:10.1093/nar/gkz1020

Valliyodan, B., Brown, A. V., Wang, J., Patil, G., Liu, Y., Otyama, P. I., et al. (2021). Genetic variation among 481 diverse soybean accessions, inferred from genomic re-sequencing. *Sci. Data* 8, 50. doi:10.1038/s41597-021-00834-w

Valliyodan, B., and Nguyen, H. T. (2006). Understanding regulatory networks and engineering for enhanced drought tolerance in plants. *Curr. Opin. Plant Biol.* 9, 189–195. doi:10.1016/j.pbi.2006.01.019

Wang, X., Li, M.-W., Wong, F.-L., Luk, C.-Y., Chung, C.Y.-L., Yung, W.-S., et al. (2021). Increased copy number of gibberellin 2-oxidase 8 genes reduced trailing growth and shoot length during soybean domestication. *Plant J.* 107, 1739–1755. doi:10.1111/tpj.15414

Xie, C., and Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinforma.* 10, 80. doi:10.1186/1471-2105-10-80

Yevshin, I., Sharipov, R., Valeev, T., Kel, A., and Kolpakov, F. (2017). GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res.* 45, D61-D67–d67. doi:10.1093/nar/gkw951

Zeng, S., Lyu, Z., Narisetti, S. R. K., Xu, D., and Joshi, T. (2018). "Knowledge Base Commons (KBCommons) v1.0: A multi OMICS' web-based data integration framework for biological discoveries," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, December 6 2018, 589–594.

Zeng, S., Lyu, Z., Narisetti, S. R. K., Xu, D., and Joshi, T. (2019). Knowledge Base Commons (KBCommons) v1.1: a universal framework for multi-omics data integration and biological discoveries. *BMC Genomics* 20, 947. doi:10.1186/s12864-019-6287-8

Zhou, Z., Jiang, Y., Wang, Z., Gou, Z., Lyu, J., Li, W., et al. (2015). Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* 33, 408–414. doi:10.1038/nbt.3096

Żmieńko, A., Samelak, A., Kozłowski, P., and Figlerowicz, M. (2014). Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* 127, 1–18. doi:10.1007/s00122-013-2177-7