# Development of a genome-phenome browser for rice cultivars


## ABSTRACT

*With the advent of modern sequencing techniques, the genome study of Rice has become more precise and effective. Modern bioinformatics tools have contributed a lot to the functional genomics of Rice. Worldwide researches are being carried out to understand the biology of rice better. Biological experimental information needs to be inferred more precisely so that further progress can be made. In this study, a comprehensive analysis is undertaken on functionally validated and characterized rice genes which are found to affect the phenotypes of rice. A database and browser have been developed with the help of information available across literatures which will be used to identify genes responsible for affecting morphological, physiological and resistance or tolerance traits. Such detailed analysis will also provide the annotation of genes that are functionally characterized. It will also help to understand how these genes affect the traits useful for Rice breeding.*



***Keywords:*** Database, Functional Genomics, Genome Browser, Phenotype, Rice (*Oryza sativa*)


**Introduction**

A major chunk of the world's population regularly consumes rice in their diet. Inevitable expansion in global population and climate change has made the demand for higher yield and better quality rice imperative. So, As a result, developing new rice varieties which increase yield and ensure quality is a daunting task. The generation of huge data of rice genome sequence through sequencing technology has made it possible to extract a wealth of data about the variety of genes and alleles that can help to improve beneficial agronomic features. Several researches have been carried out to find the variation in rice. Phenotype analysis leads to a better understanding of genes and their interaction with the environment. This will improve the introduction of new varieties which will be withstanding more biotic and abiotic stresses. Genome-wide Association Studies (GWAS) help to study phenotype features by identifying causal genes and variants it creates. Many researchers have made rice functional genomics achieve a great height during the past two decades. Because rice is a popular food crop and a model plant, functional genomics research findings must be used to improve rice breeding. A comprehensive analysis of the functionally validated rice genes could give detailed information on the natural variation of Rice to collect and analyse phenotypes within the context of a particular species, phenotype ontologies were created. These ontologies have recently been extended with formal class definitions that can be used to combine phenotypic data and make it possible to compare traits among various species directly. A database of rice variation is needed to record and gather various genes- and specific effects. According to Rice Variation Map (Zhao H. *et al.*, 2016), around 1479 rice accessions were sequenced to identify 6551358 single nucleotide polymorphisms (SNPs) and 1214627 insertions/deletions (INDELs). All INDEL/SNP genotypes can be accessed and downloaded from online. SNPs, and INDELs can be searched for using gene names, genomic areas, SNP/INDEL identifiers, and gene annotation keywords. There are many genome browsers of Rice for specific purposes which enable the study of rice genes and their functions precisely. There is no complete phenotypic browser that would allow for extensive annotation of rice phenotypes in the context of mutations, quantitative trait loci, and strains that are utilised as models for the biology of Rice and diseases. This kind of browser is required to find out the genetic background of a given phenotype. This could help researchers to integrate genotypes and phenotypes of rice. There are many rice genome browser Post Rice Genome Annotation project progress has been built to uncover the functional diversity of alleles for agriculturally relevant genes which have been obtained from functional genomics researches. We have gathered data on functionally validated rice genes across scientific literatures. Genetic and phenotype information is compiled from this literature. The data is combined in different categories. In this study, our aim is to provide information on rice gene function to study phenotypes characters due to the effect of various genes in Rice. To accomplish this, a comprehensive search was attempted to gather articles related to functional genomics of rice and prepare a list of genes which are functionally characterized. The information gathered on functionally characterized and validated genes of Rice was compiled to make a new database and finally, a browser is developed. Same genes are found to affect multiple traits. The position of the genes in the genome along with the objective of the study are also compiled. A deep analysis of the functionally characterized and validated genes will help to understand variation which offers a useful additional resource for identifying novel gene functions as well as allelic variants that particularly interact with the genetic makeup and/or environment or alleles showing minor effect on

phenotype, especially for characteristics related to plant adaptation. **Ohyanag** H. *et. al*., (2015) developed OryzaGenome, a Genome Diversity Database of Wild Oryza Species. This resource can serve as an excellent genotype-phenotype association resource for analyzing rice functional and structural evolution, and the associated diversity of the Oryza genus. Two variant viewers are implemented: SNP Viewer as a conventional genome browser interface and Variant Table as a text-based browser for precise inspection of each variant one by one. Portable VCF (variant call format) file or tab-delimited file download is also available. Rice Stress-Resistant SNP database ( Woldegiorgis S T *et. al*., 2019) (http://bioinformatics.fafu.edu.cn/RSRS) mainly focuses on SNPs specific to biotic and abiotic stress-resistant ability in rice, and presents them in a unified web resource platform. Yao W. *et. al.*, (2017) built comprehensive and accurate dataset of ~2800 functionally characterized rice genes and ~5000 members of different gene families by integrating data from available databases and reviewing every publication on rice functional genomic studies. The dataset accounts for 19.2% of the 39 045 annotated protein-coding rice genes, which provides the most exhaustive archive for investigating the functions of rice genes. They also constructed 214 gene interaction networks based on 1841 connections between 1310 genes. The largest network with 762 genes indicated that pleiotropic genes linked different biological pathways. Increasing degree of conservation of the flowering pathway was observed among more closely related plants, implying substantial value of rice genes for future dissection of flowering regulation in other crops. Yan J. *et. al.,* (2020) developed SnpReady for Rice (SR4R), an Integrative SNP Resource for Genomic Breeding and Population Research in Rice. SR4R presents four reference SNP panels, including 2,097,405 hapmapSNPs after data filtration and genotype imputation, 156,502 tagSNPs selected from linkage disequilibrium-based redundancy removal, 1180 fixedSNPs selected from genes exhibiting selective sweep signatures, and 38 barcodeSNPs selected from DNA fingerprinting simulation. SR4R thus offers a highly efficient rice variation map that combines reduced SNP redundancy with extensive data describing the genetic diversity of rice populations. In addition, SR4R provides rice researchers with a web interface that enables them to browse all four SNP panels, use online toolkits, as well as retrieve the original data and scripts for a variety of population genetics analyses on local computers.

**MATERIALS AND METHODS**

The categories of information on functionally validated rice genes that have been derived from literatures have been listed in Table 1. The descriptions of the phenotypes in each of the articles were used to annotate functionally characterised genes. The phenotypes related to every gene were divided into two categories: "major category" and "minor category". Within each relevant category, genes related to multiple characters were counted. Information gathered from corresponding articles have been divided into specific categories as shown in Table 2. These specific categories are used to make columns in our database. An entity relationship diagram is used to design the database. ER diagram (Fig.1) helps to identify entities, attributes and their relationship exists among entities. In our database, some of the major entities identified are genes, Protein, wet lab analysis, resistance etc., and their relationship with each other. Entity relationship model is represented by ER Diagram. Entities are represented by rectangular box. Attributes are presented by ellipses. In this study, the entities chosen are Protein , Gene, Stress, Marker, Wet lab analysis, expression analysis, resistance . All these entities are related to each other through relation which have been represented by Diamond Box. Every entity has one unique primary key. Like Gene Id is the primary key of gene table. Entities among them different degree of relationship. The degree of relationship may be different. One gene can have one to one relationship with protein and this one to one relationship between entities are indicated with number 1. Similarly, One gene can show multiple types of resistance and there exits one to many relationship with gene and resistance entities. One to Many relationship is indicated by N. Genes which translate into protein and their effect on phenotype have been understood by going through the results in the corresponding literature. The variation found in the number of genes which are functionally validated among the various classes most likely reflects the agronomic importance of each characteristic and the research interests of specific researchers instead of the exact number of genes associated with each character. Transgenic techniques like , knockdown, and knockdown/overexpression were more frequently used for functional study of genes in case of "resistance or tolerance" than for genes in case of "morphological trait" and "physiological trait." The difficulty in identifying mutant and natural populations for characters associated with tolerance or resistance in mutant and normal populations may result in this difference. Overexpression analysis was used to characterise the major numbers

of genes in several categories "cold," "drought," and "salinity" under the major category "resistance or tolerance".

***Database and web tool***

All data of the Rice genes which are functionally validated were compiled to create a database. The database is consisting of a Gene information table and a genome-phenome viewer. GBrowse (Donlin*M J* , 2007) is used for the purpose of the Genome viewer. Interactive web pages have been developed to see Gene information table on the desired click. The gene information will be displayed in a separate window when the user will choose to see the gene details in GBrowse. The gene details link has been included in a balloon which will be reflected when user will point the cursor to the desired gene in the GBorwse.A web tool shows the phenotype for a specific gene under a particular experiment. Several phenotypic attributes related to genes have been incorporated along with genomic information. Phenotypic information under different stress conditions has been shown in the browser. The creation of an integrated PheGenI database (Ramos *et al.,* 2009) has made it possible to further study GWAS results in the context of the genome and link them to characteristics of SNP, gene, and eQTL data. These data are easily accessible in a user-friendly format made available to the scientific community who are interested to look into genetic association results in more depth. This is achieved by including functionalities to download data tables, personalise the view, and dynamically browse features of the genomic sequence. Web tools help to understand disease pathways and biology by providing the interconnection between the genetic variants with many phenotypes. RPAN (Sun C. *et al* .2016) database is used to search and analyze the rice pan-genome obtained from 3KRGP. RPAN is a database that contains general information about 3010 accessions of rice, such as sequences, expression data, gene annotations and genes from the rice pan-genome. RPAN also includes a variety of search and visualization features. PhenoScanner (Staley JR et al. 2016) is a curated library of results from extensive genetic association research that are readily accessible to all. This programme performs "phenome scans," or the comparison of genetic variations with a broad range of phenotypes, to enhance knowledge of biological processes and disease pathways. Similarly, a genome-phenome browser shows the phenotype for a specific gene under a particular experiment. Several phenotypic attributes related to genes have been incorporated along with genomic information. Phenotypic information under different stress conditions has been shown in browser. For the purpose of phenome browser development GBrowse genome browser has have been used. Through the customization facility, GBrowse is configured to visualize the phenotype data. Rice annotated data is used which is added in GBrowse. Annotation tracks in the Genome Browser are comprised of files in line-oriented format. Every line in the file defines a track display characteristic or a data item within the track. There are three types of lines in annotation files: browser lines, track lines, and data lines.

**RESULTS AND DISCUSSION**

In this study, we have a list of rice genes which are functionally characterized and validated. The information contains the Gene name, Gene symbol, Major characteristics including Morphological traits, physiological traits, resistance or tolerance, methods of isolation, doi etc. All these information helps to better understand the phenotypes and how it is affected by genes. A genome browser is a graphical user interface for visualizing genomic data obtained from a biological database. The genome browser shows the biological data in a user-friendly format. It generally takes very huge datasets, such as whole genome FASTA files, and exhibits them in a way that users can understand the information contained within. A comprehensive database of functionally characterized rice Genes is developed. This database consists of two types of information one is genome information and other phenotype exhibited by the gene. Genome Browser GBrowse has been configured for Rice Annotation data. All the genomic information gathered on functionally characterized genes of Rice have been displayed in GBrowse. GBrowse displays biological information in different coordinates of the genes. It contains several information like genomic position of genes, locus ids, positions of the genes in the chromosomes, SNPs, Flanking regions, copy number variations etc. Further information on specific genes that how it is functionally affecting different traits can be viewed on the desired click. It leads to a detailed information table comprising of all the information on the validated rice genes compiled in the database extracted from the literatures. In the table, data regarding the phenotypes of the validated genes will be displayed. All the data gathered on functionally validated rice genes from corresponding literature have been summarized and will be reflected on the new window in the browser. The database was created in MySQL. MySQL is a popular relational database management system. GBrowse is configured to fetch data from a

database which enables GBrowse to integrate phenotypes and provide a channel to search, browse, retrieve and analyze genomic as well as phenotype data of queried genes.

**Conclusion**

It is concluded that all the data gathered on functionally validated rice genes from corresponding literature have been summarized and will be reflected on the new window in the browser. The databasewas created in MySQL. MySQL is a popular relational database management system. GBrowse is configured to fetch data from a database which enables GBrowse to integrate phenotypes and provide a channel to search, browse, retrieve and analyze genomic as well as phenotype data of queried genes.



**Table 1: Different categories of information for rice genes**

| Particulars of Gene Information | Remarks |
|---|---|
| Gene | Name of the gene |
| Gene Symbol | Abbreviated Gene Name |
| Major | Major character |
| Minor | Minor character |
| Chromosome | No. of chromosome where gene lies |
| Start | Start of gene on the Chromosome |
| End | End of the gene on the chromosome |
| Locus ID | Locus id of the gene |
| Isolation | Method of isolation |
| Objective | Phenotypes studied in each of the research articles |
| Reference | Digital Object Identifier (doi) |

**Table2**: **Minor characters studied in literature classified under each of the Major Characters**

| Character Major | Character Minor |
|---|---|
| **Morphological** | Dwarf |
| | Culm Leaf |
| | Seed |
| | Shoot seedling |
| | Panicle flower |
| | root |
| **Physiological traits** | Eating quality |
| | Flowering |
| | Germination dormancy |
| | Sterility |
| | Lethality |
| | Source activity |
| **Resistance** | Bacterial Blight Resistance |
| | Blast Resistance |
| | Lodging resistance |
| **Tolerance** | Cold tolerance |
| | Drought tolerance |
| | Salinity tolerance |
| | Sub mergence tolerance |

**Table 3: Broad categorization of data found from literatures**

| | | | | | |
|---|---|---|---|---|---|
| Gene name | Gene Id | Gene sequence | Genomic position | Protein ID | Protein structure |
| Protein sequence | Expression pattern | Gene function | Growth analysis | Northern blot analysis | Southern blot analysis |
| Hormonal sensitivity | Stress response | Transcriptional analysis | PCR | Biochemical analysis | Metal concentration |
| Enzyme activity | Pest resistance | Chemical treatment | Over expression | QTL | cDNA analysis |

| | | | | | |
|---|---|---|---|---|---|
| Physiochemical properties | Starch biosynthesis | Chlorophyll biosynthesis | Herbicide resistance | Weed resistance | Transgenic lines |
| Agricultural traits | Mycorrhiza formation | Nutrient deficiency | SSR | SNPs | DGE |

*PCR = Polymerase Cahin Reaction, QTL = Quantative Trait Loci, SSR= Single Sequece Repeat, DGE= Differential gene Expression Analysis , cDNA= complementary DNA

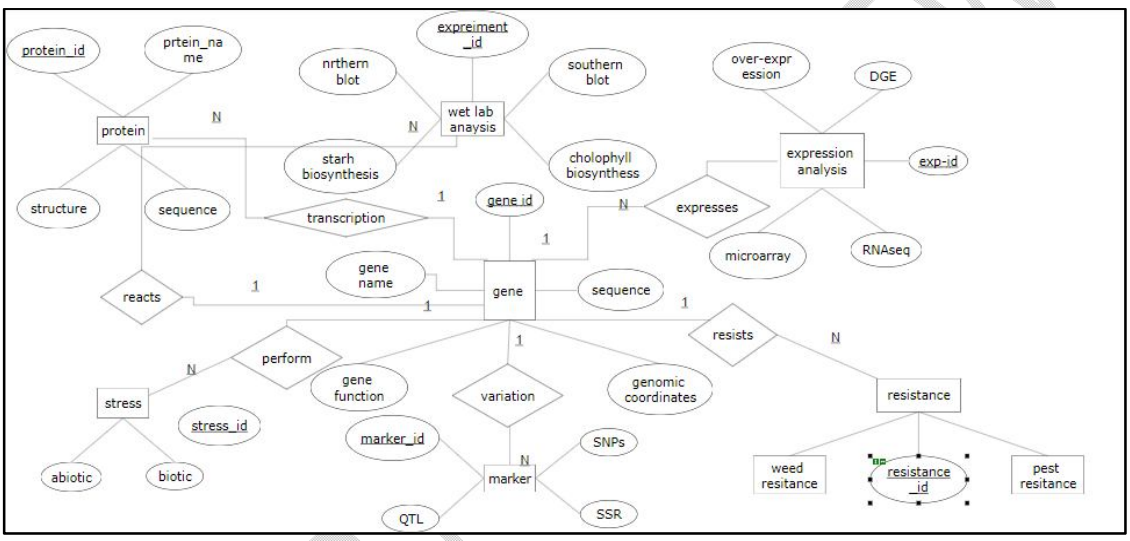
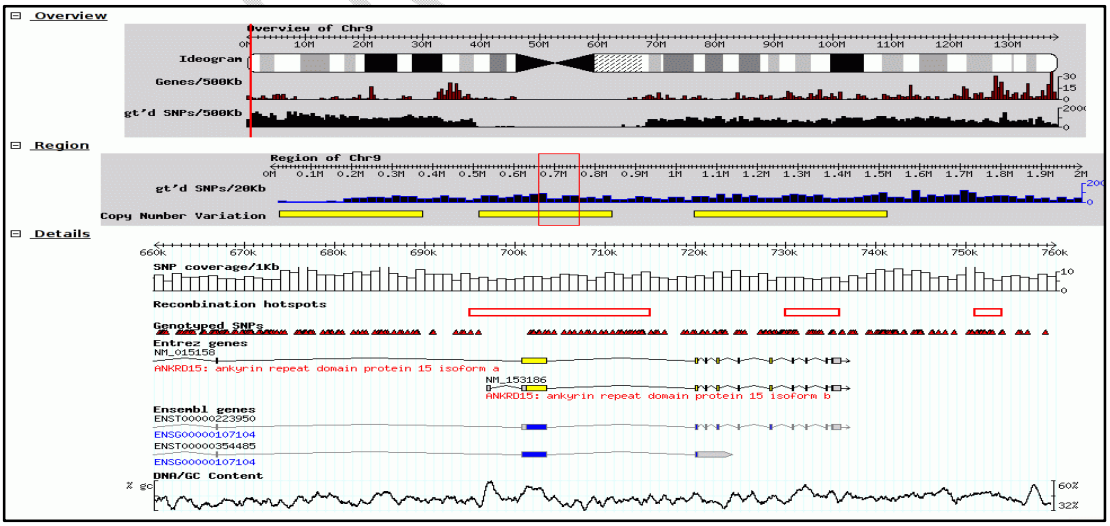
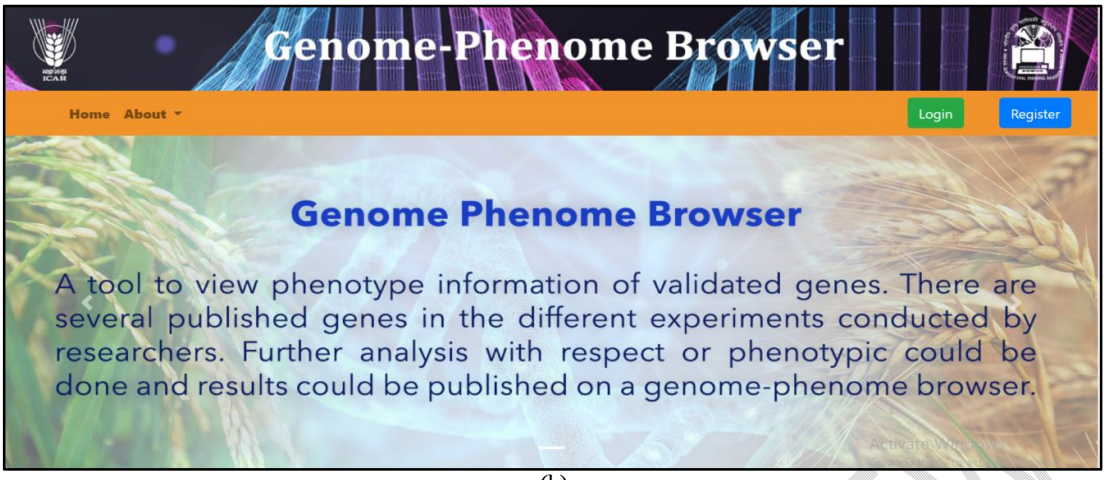


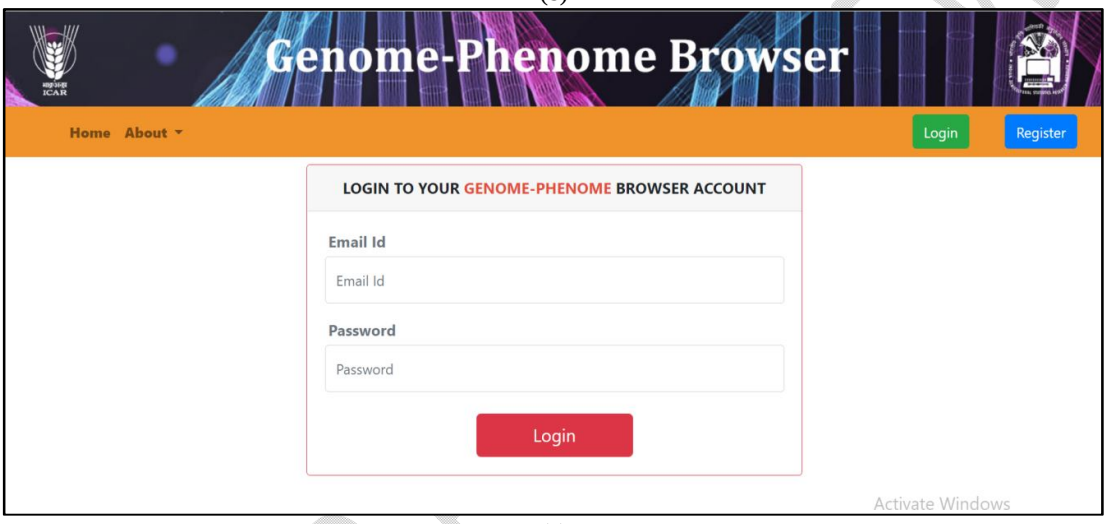

**Fig. 1 :** ER diagram from different categories of information which will be used to develop database



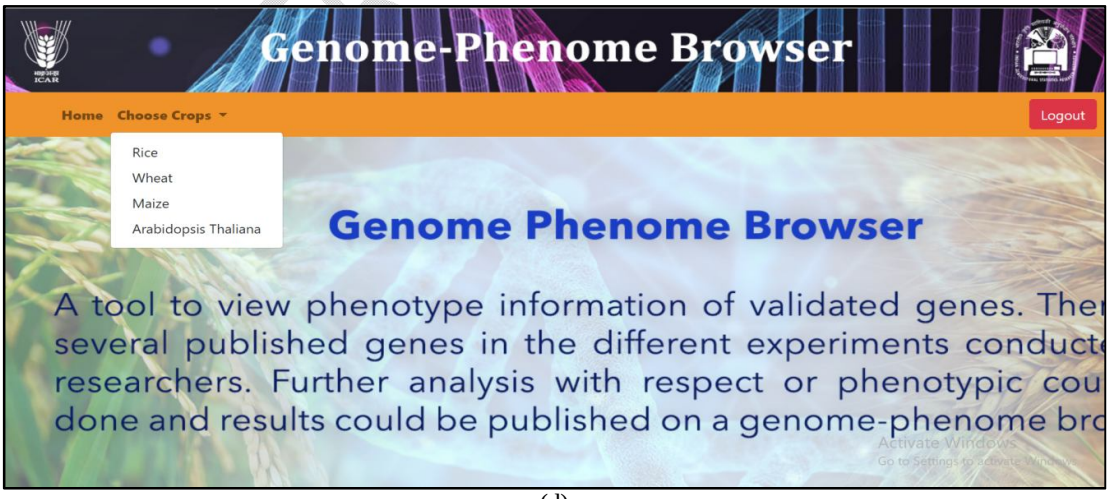

(a)


Genome-Phenome Browser
Home
About
Login
Register
Genome Phenome Browser
A tool to view phenotype information of validated genes. There are several published genes in the different experiments conducted by researchers. Further analysis with respect or phenotypic could be done and results could be published on a genome-phenome browser.

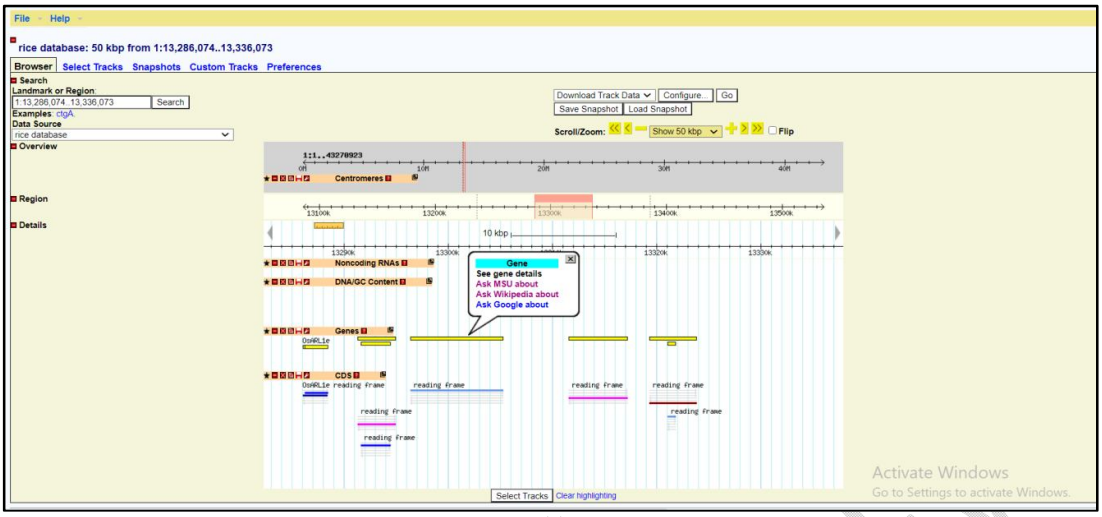
(b)


Genome-Phenome Browser
Home
About
Login
Register
LOGIN TO YOUR GENOME-PHENOME BROWSER ACCOUNT
Email Id
Email Id
Password
Password
Login
Activate Windows

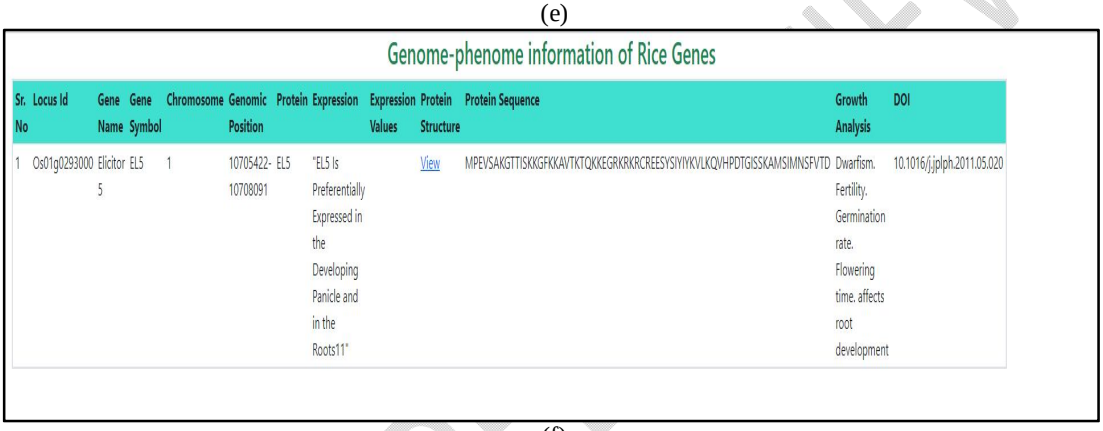
(c)


Genome-Phenome Browser
Home
Choose Crops
Logout
Rice
Wheat
Maize
Arabidopsis Thaliana
Genome Phenome Browser
A tool to view phenotype information of validated genes. Ther
several published genes in the different experiments conducte
researchers. Further analysis with respect or phenotypic cou
done and results could be published on a genome-phenome bro

(d)




(e)



Genome-phenome information of Rice Genes

| Sr. No | Locus Id | Gene Name | Gene Symbol | Chromosome | Genomic Position | Protein | Expression | Expression Values | Protein Structure | Protein Sequence | Growth Analysis | DOI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Os01g0293000 | Elicitor 5 | EL5 | 1 | 10705422- 10708091 | EL5 | "EL5 Is Preferentially Expressed in the Developing Panicle and in the Roots11" | | View | MPEVSAKGTTISKKGFKKAVTKTQKKEGRKRKRCREESYSIYIYKVLKQVHPDTGISSKAMSIMNSFVTD | Dwarfism. Fertility. Germination rate. Flowering time. affects root development | 10.1016/j.jplph.2011.05.020 |

(f)

**Fig. 2. Overview genome-phenome browser. A) GUI of GBrowse B)Home page of the Genome-phenome browser C)Login page where user has to login for accessing the browser D )After login use can choose desired crop, in this study Rice is used E) for chosen crop GBrowse will appear along with all the genomic details including the validated genes that have been used in this study, yellow bar represent the gene F) gene details consists of the genome –phenome phenotype information's of the particular gene**

**STUDY AREA / SAMPLE COLLECTION:** Division of Bioinformatics, ICAR- Indian Agricultural Statistics Research Institute, New Delhi 110012

## REFERENCES

Zhao Hu, Wen Y., Ouyang Y., Yang W., Wang G., Lian X., Xing X., Chen L., and Xie W.2014. RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Research.43.

Erin M Ramos, Douglas Hoffman, Heather A Junkins, Donna Maglott, Lon Phan, Stephen T Sherry, Mike Feolo& Lucia A Hindorff.2013. Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources.European Journal of Human Genetics volume 22, pages144–147

Maureen J Donlin .2007 . Current protocol in Bioinfroamtics.Using the Generic Genome Browser (GBrowse)

Chen Sun, Zhiqiang Hu, Tianqing Zheng, Kuangchen Lu, Yue Zhao,Wensheng Wang, Jianxin Shi, Chunchao Wang, Jinyuan Lu, Dabing Zhang, Zhikang Li and Chaochun Wei.2016RPAN: rice pan-genome browser for ~3000 ricegenomes

Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, Paul DS, Freitag D, Burgess S, Danesh J, Young R, Butterworth AS. PhenoScanner: a database of human genotype-phenotype associations.2016. Bioinformatics. 32(20).

Jun Yan , Dong Zou , Chen Li , Zhang Zhang , Shuhui Song , Xiangfeng Wang .2020. SR4R: An Integrative SNP Resource for Genomic Breeding and Population Research in Rice. *Genomics Proteomics Bioinformatics*.18 (2)

Wen Yao, Guangwei Li, Yiming Yu and Yidan Ouyang. 2018. funRiceGenes dataset for comprehensive understanding and application of rice functional genes. *GigaScience*.

Samuel Tareke Woldegiorgis , Shaobo Wang , Yiruo He , Zhenhua Xu , Lijuan Chen , Huan Tao , Yu Zhang , Yang Zou , Andrew Harrison , Lina Zhang , Yufang Ai , Wei Liu and Huaqin He. Rice Stress-Resistant SNP Database.Rice

Hajime Ohyanagi, Toshinobu Ebata , Xuehui Huang , Hao Gong , Masahiro Fujita , Takako Mochizuki , Atsushi Toyoda , Asao Fujiyama, Eli Kaminuma, Yasukazu Nakamura, Qi Feng , Zi-Xuan Wang, Bin Han and Nori Kurata.2015.*Plant & Cell Physiology*.