Development of a genome-phenome browser for rice cultivars

ABSTRACT

With the advent of modern sequencing techniques, genome study of Rice becomes more precise and effective. Modern bioinformatics tools have contributed a lot in functional genomics of Rice. Worldwide researches are being carried out to understand biology of rice better. Biological experimental information needs to be inferred more precisely so that further progress can be made in this regard. In this study, a comprehensive analysis is undertaken on functionally validated and characterized rice genes which are found to effect 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 are affecting the traits which will be 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 have 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. 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 out the variation found in rice. Phenotype analysis leads to the better understanding of genes and its interaction with environment. This will improve the introduction of new varieties which will be withstanding more biotic and abiotic stresses. Genome wide Association Studies (GWAS) helps to study phenotype features by identifying causal genes and variants it creates. Many researches have made rice functional genomics to 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 detail information on 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 directly compare traits among various species. A database of rice variation is needed to record and gather various genes- and characteristic 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 form online. SNPs, 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 purpose which enable study rice genes and its 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 biology of Rice and diseases. This kind of browser is required to find out genetic background for a given phenotype. This could help researchers to integrate genotype and phenotypes of rice. There are many rice genome browser Post Rice Genome Annotation project progress has been built to uncover functional diversity of alleles for agriculturally relevant genes which have been obtained from functional genomics researches. We have gathered data of functionally validated rice genes across scientific literatures.Genetic and phenotype information are compiled from these literatures.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 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 browser is developed.Same genes are found to affect multiple traits. 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.

MATERIALS AND METHODS

The categories of information on functionally validated rice genes that have been derived from literatureshave 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 character 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. Entity Relationship diagram is used to design the database. ER diagram 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., their relationship with each other. Genes which translate into protein and their effect on phenotype has been understood by going through the results in the corresponding literatures. The variation found in the number of genes which are functionally validated among the various classes most likely reflect the agronomic importance of each characteristic and the research interests of specific researchers instead of the exact number of genes associated with in each character. In Major category “Morphological” and “Physiological” Mutation method is found to be prevalent. Whereas, in case of “resistance or tolerance “transgenic approaches namely knockdown, overexpression and knockdown-overexpression were applied for functional analysis. The difficulty in identifying mutant and natural population for characters associated with tolerance or resistance in mutant and normal populations may result 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 Gene information table and 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 user will choose to see the gene details in GBrowse. Gene details link has been included in a balloon which will be reflected when user will point the cursor in the desired gene in the GBorwse.A web tool shows the phenotype for a specific gene under particular experiment. Several phenotypic attributes related to genes have been incorporated along with genomic information. Phenotypic information under different stress condition has been shown in browser. The creation of 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 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 helps to understand the disease pathways and biology by providing the inter connection 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 gene 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 for 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 particular experiment. Several phenotypic attributes related to genes have been incorporated along with genomic information. Phenotypic information under different stress condition has been shown in browser. For the purpose of phenome browser development GBrowse genome browser 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 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. 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 data base 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 of 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 desired click. It leads to detailed information table comprising of all the information on the validated rice genes complied in the database extracted from 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 literatures have been summarized will be reflected on the new window in the browser. Database was created in MySQL. MySQL is popular relational database management system. GBrowse is configured to fetch data from 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 literatures 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

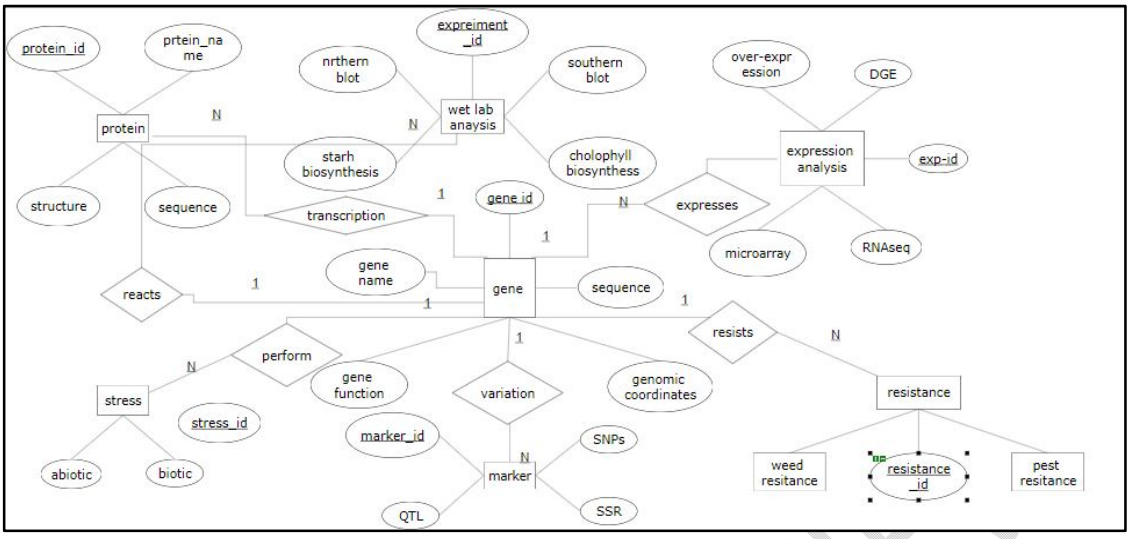
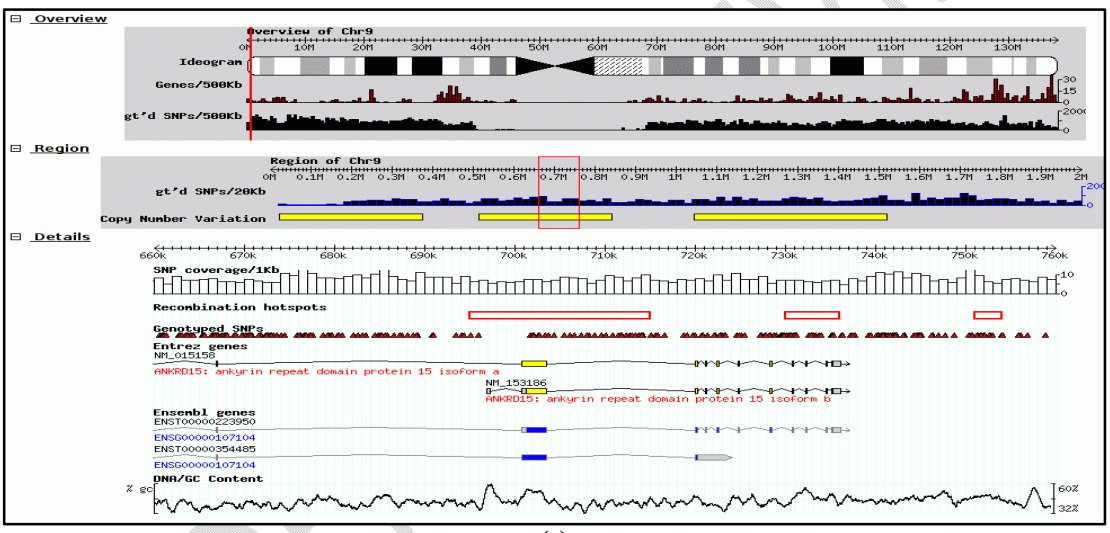
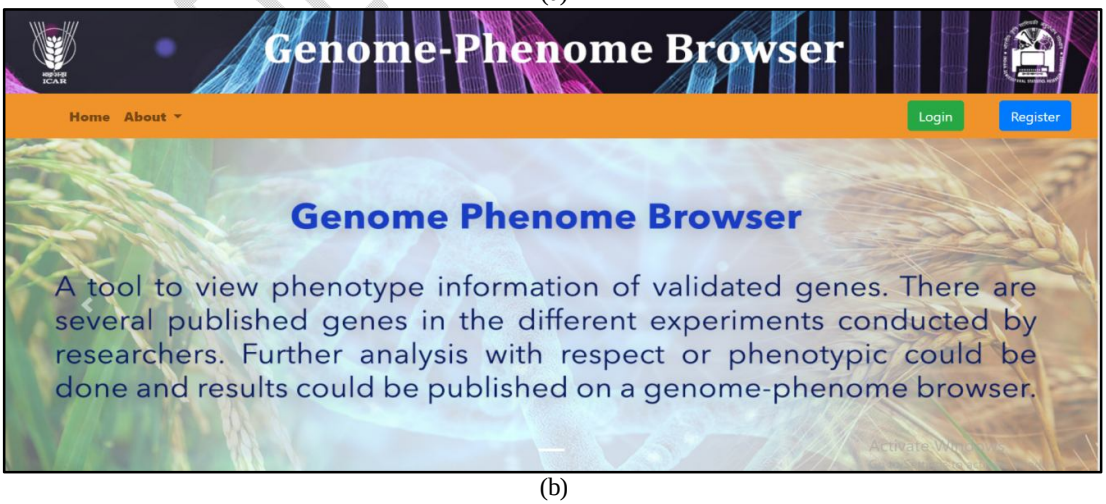


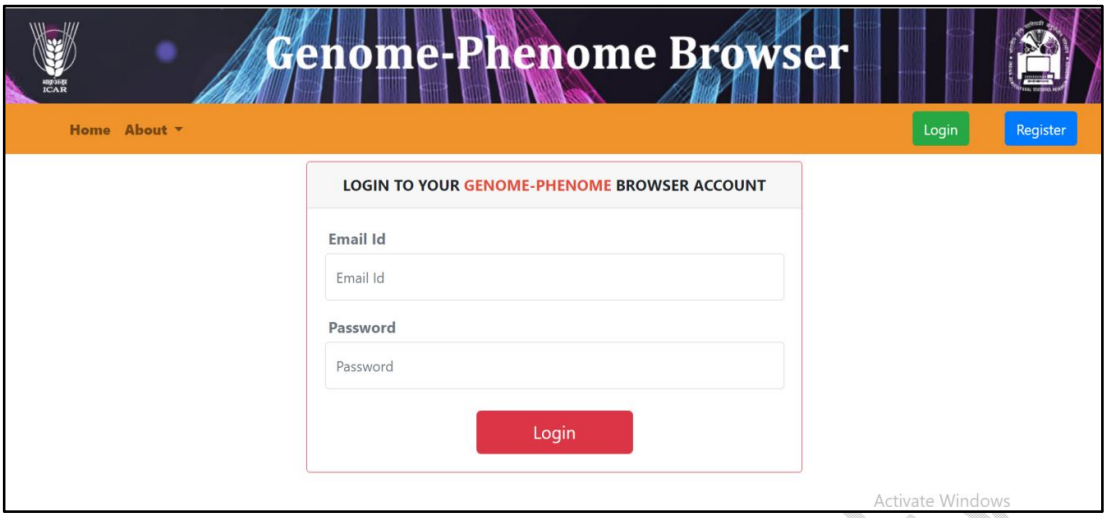
Fig. 1 : ER diagram from different categories of information which will be used to develop database



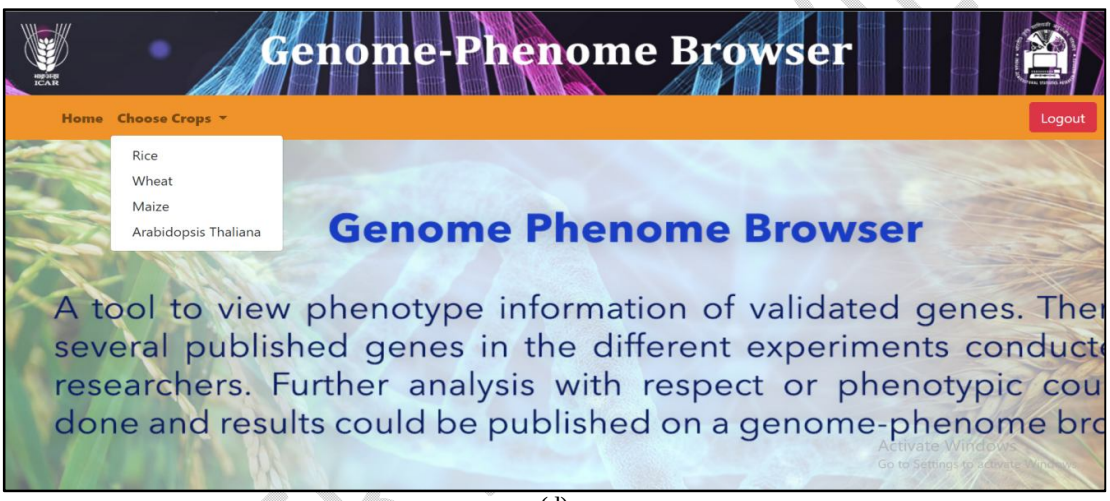
(a)



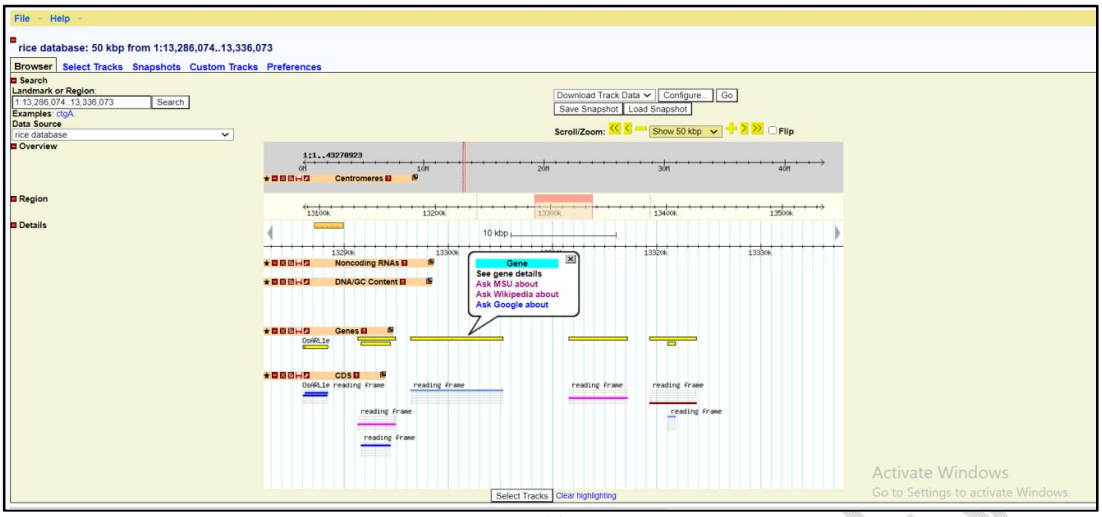
(b)


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

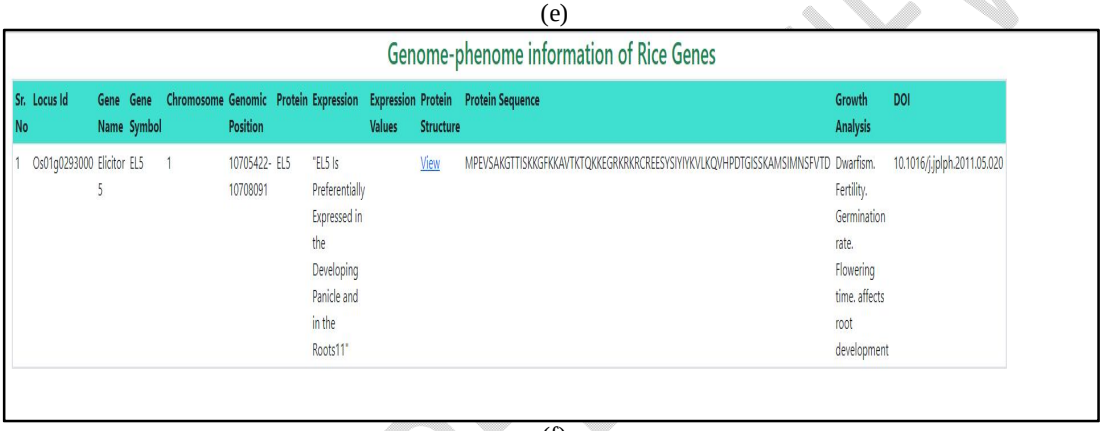
(c)


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
Activate Windows

(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 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. Over view 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 repserent 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

ETHICAL APPROVAL:This article does not contain any studies with human participants or animals performed by any of the authors.

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).