Biogen Base - An Interactive Maize Database for Phenomics Platform

Murukarthick. J.
BTech student

Senthil. N
Associate Professor

Raveendran. M
Associate Professor

Prabhakaran.P, Sreedevi.G. S, Sumanth kumar. M, Shobhana. V.G, Dhanya.S, Khushboo.R

Genomics and Proteomics Laboratory, Centre for Plant Molecular Biology,

Arumugasamy, S., Associate Professor, Ravikesavan, R., Associate Professor (Plant Breeding), Maize research station, Vagarai, Palani.

Tamil Nadu Agricultural University, Coimbatore, India.

ABSTRACT

Biogen base serves the maize (*Zea mays L.*) research community by making a wealth of genetics and genomics data available through an intuitive Web-based interface. We have developed an Open access database as a resource to enhance research with the unique data obtained from the genomics and proteomics lab of TNAU. Biogen Base is an interactive database in bringing out the different traits of the inbreds of Tamil Nadu Agricultural University (UMI - University Maize Inbreds's) and the SSR Markers. The database interface is developed in PHP and HTML as the front end and MySQL as the backend tools. The webpage was developed using Dreamweaver and the database in MySQL is connected with the web server. The Current version of this database has four major parts and functions; (1) Germplasm - contains 101 germplasm lines and its 28 corresponding traits with values, (2) Genotype Search – enables the search among 31 SSR markers along with the chromosome number, gel patterns, forward and reverse primer, and allele size, (3) Phenotype Search - which contains text descriptions of all the phenotypic terminologies and their corresponding abbreviations stored in the database and (4) Mutant Phenotype Search – contains 5 mutant phenotypes with its traits and values. In addition, it includes brief description about the terms, and links to other publicly available databases. Images of plants with novel characteristics are also available at the web site. The large and growing body of experimental data on maize germplasm and DNA markers is of enormous value in the Genomics and Proteomics laboratories. The database can be searched using a user friendly web interface. This database is publicly available at http://www.tnaugenomics.com/database/maize. It also facilitates the deposition of new values for processing and inclusion in the database to fulfill the priceless work going on in Genomics and Proteomics Laboratory of TNAU.

KEYWORDS:
University Maize Inbreds, Biogen base, My Sql, Php, TNAU Genomics, Maize germplasm.

I. INTRODUCTION

Maize, Zea mays (2n = 20) belonging to the family Poaceae and subfamily Panicoideae constitutes the staple food for most of the world's population. It is grown under diverse cultural conditions and over wide geographical range. Most of the world's maize is cultivated and consumed in America.

Several maize varieties have been released by Tamil Nadu Agricultural University, Coimbatore. The molecular characterization and fingerprinting of these released varieties using micro-satellite markers will provide sufficient knowledge on diversity among them at the molecular level, which will help the breeders to develop strategies for the further, and the variety specific fingerprints will enable to identify and characterize each variety released. We can find the maize sequence data, phenotype, and genotype data in other databases like MAIZEGDB, GRAMENE QTL Database, TIGR, which also have been designed on the same platform. [1], [2] & [3]

But the genomics and proteomics laboratory in TNAU, assessed about 150 maize germplasm lines, in which 101 lines have 19 morphological traits and 31 SSR marker details have been stored in database initially. It also has the traits being analysed for the mutant phenotype data. This database is more simplified for the users to access the accessions for the Indian Inbreds, which is the Speciality lying under this unique portal.

The front page of the database was developed using Dreamweaver, usually used to design standard front pages. The nineteen morphological traits and thirty one marker details are stored in MySQL (ref) database which is connected with web server. This database can be accessed through the link, www.tnaugenomics.com/database/maize. Hence, the present database is undertaken, to retrieve the genetic diversity among the different germplasm lines through morphological traits, to access or retrieve the marker information along with the gel pictures by querying the

database which is publicly accessible via the World Wide Web. Among all the DNA markers currently available, micro-satellites are considered to be the marker of choice for varietal identification because of their co-dominant segregation and their ability to detect large number of discrete alleles repeatedly, accurately and efficiently SSR markers generate enough allelic diversity to differentiate cultivars within a subspecies or ecotype The information obtained from SSR markers is also useful for making predictions about crossing and selection aimed at increasing the efficiency of parental selection and variety development.

2. BIOLOGICAL DATABASE

A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. A simple database might be a single file containing many records, each of which includes the same set of information [4&5]. For example, a record associated with a nucleotide sequence database typically contains information such as contact name; the input sequence with a description of the type of molecule; the scientific name of the source organism from which it was isolated; and, often, literature citations associated with the sequence. For researchers to be benefitted from the data stored in a database, two additional requirements must be met:

- Easy access to the information; and
- A method for extracting only that information needed to answer a specific biological question.

Currently, a lot of bioinformatics work is concerned with the technology of databases. These databases include both "public" repositories of gene data like GenBank or the Protein DataBank (the PDB)[4&10], and private databases like those used by research groups involved in gene mapping projects, disease databases or those held by biotech companies. Biological databases have become an important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps the researchers to identify the variations phenotypically and genotypically for better understanding of the allelic variations. An individual column must be created for each type of data you wish to store (i.e., character, values). On the other hand, a row containing the actual values for these specified columns. Each row will have one value for each and every column. For example a table with columns (i.e., character, values) could have a row with the values (plant height -160 cm).

2.1 DATABASE TECHNIQUES FOR BIOLOGICAL DATASETS

Indexing, clustering and mining technology on biological databases are essential to summarize the information of biological data, to efficiently discover knowledge that may be impossible by the traditional methodologies and unexpected patterns which may be meaningful for important biological applications such as protein interaction predictions [6].

- A database index is meant to improve the efficiency of data lookup at rows of a table by a key access retrieval method.
- Clustering is an unsupervised process to group similar objects together based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.
- Classification is a process to model or function to describe and distinguish data classes for the purpose of predicting the class of objects whose class labels are unknown.

Since the biological data becomes tremendous with the growing research interests and the revolution of research approaches, it becomes more and more important and necessary to analyze and understand biological data and the relationships between various data sets using computational approaches [7]. Designing a high-quality database is complicated by the fact that there are several formats for many types of data and a wide variety of ways in which scientists may want to use the data. Many of these databases are best built using relational database architecture, often based on Oracle or Sybase.

A strong background in relational databases is a fundamental requirement for working in database development. Having some background in the molecular biology techniques used to generate the data is also important [8]. Most critical for the bioinformatics specialist is to have a strong working relationship with the researchers who will be using the database and the ability to understand and interpret their needs into functional database capabilities [9&11]. The maize database uses the relational structured database in which the interface is developed in PHP and HTML as the front end and MySQL as the backend tool.

3. SOFTWARES

3.1 HTML

HTML stands for Hyper Text Markup Language which is predominant markup language for web pages. It provides a means to create structure documents by denoting structural semantics for text such as headings, paragraphs, lists etc., as well as links, quotes and other items.

3.2 PHP

Hypertext Preprocessor is a widely used, general-purpose scripting language that was originally designed for web development to produce dynamic web pages. For this purpose, PHP code is embedded into the HTML source document and interpreted by a web server with a PHP processor module, which generates the web page document. As a general-purpose programming language, PHP code is processed by an interpreter application in command-line mode performing desired operating system operations and producing program output on its standard output channel. It may also function as a graphical application. PHP is available as a processor for most modern web servers and as standalone interpreter on most operating systems and computing platforms. PHP is free software released under the PHP License, which is incompatible with the GNU General Public License (GPL) because restrictions exist regarding the use of the term PHP.

3.3 MySQL

MySQL is currently the most popular open source database server in existence. On top of that, it is very commonly used in conjunction with PHP scripts to create powerful and dynamic server-side applications. MySQL has been criticized in the past for not supporting all the features of other popular and more expensive Database Management Systems. However, MySQL continues to improve with each release (currently version 5), and it has become widely popular with individuals and businesses of many different sizes. MySQL is a database and is available in http://www.mysql.com. A database is a data storage feature. It can be used to store, sort, arrange and display information.

MySQL is a functional feature on its own and we will be using PHP commands to use the functions of a MySQL database. The program phpMyAdmin is a graphical interface that allows you to use the functions of a MySQL database which is available in http://www.phpmyadmin.net/home_page/. MySQL is a data storage area. In this storage area, there are small sections called TABLES. Very similar to a normal HTML table, the MySQL tables consist of rows, columns and cells. MySQL is a relational database management system (RDBMS) that runs as a server providing multi-user access to a number of databases. The MySQL development project has made its source code available under the terms of the GNU General Public License, as well as under a variety of proprietary agreements. MySQL is owned and sponsored by a single for-profit firm, the Swedish Company MySQL AB, now owned by Sun Microsystems, a subsidiary of Oracle Corporation. Free-software projects that require a full-featured database management system often use MySQL. Such projects include Word Press, phpBB, Drupal and other software built on the LAMP software stack. MySQL is also used in many high profiles, large-scale World Wide Web products including Wikipedia, Google and Face book.

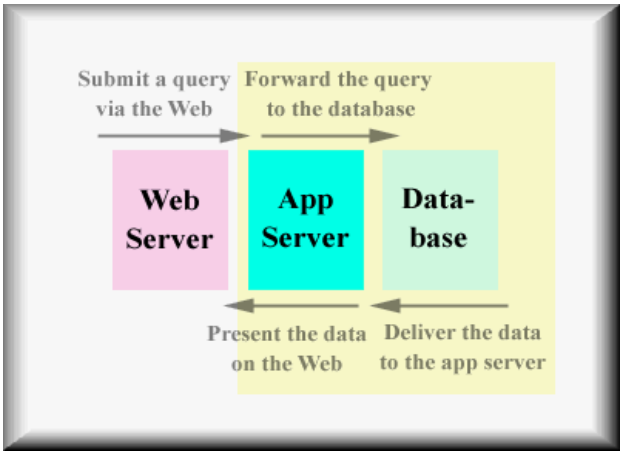


Figure1

4. MATERIALS AND METHODS

DATABASE SCHEMA

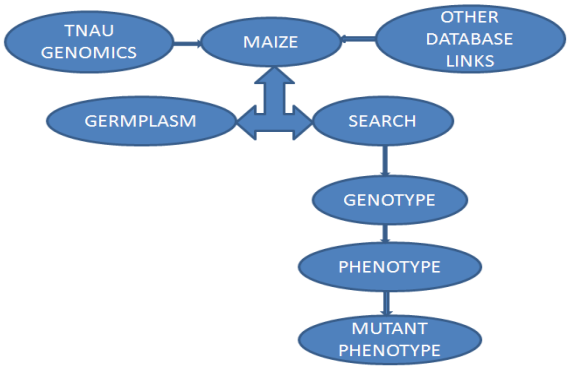


Figure 2

4.1 SYSTEM HARDWARE CONFIGURATION

- **OS:**MS Windows Xp Professional
- **System Manufacturer:** Intel
- **Processor:** Intel Pentium 1.70 GHz
- **RAM:** 1 GB

4.2 SOFTWARES USED

- PHP
- My SQL 5.1 Run Time Environment
- My SQL Query Browser 5.1
- Microsoft Excel
- JDK 1.5 (Java Development Kit)
- HTML Editor
- WAMP Server

4.3 DATA TABLES

Two types of information such as Germplasm and genotype are stored in this database. The germplasm table contains information on the length of the cob (cm), the number of rows per cob, the number of columns per cob, the number of cobs per plant, the length of the leaf (cm), the width of the leaf (cm), the number of leaves, the number of tassel branches, the length of the tassel (cm), the height of the tassel (cm), the height of the plant (cm), the height of the cob (cm), the circumference of the cob (cm), the weight of the cob (g), the weight of hundred seeds (g), the levels of phytate phosphorus (mg/g), total phosphorus (mg/g) and zinc (mg/100g). The assessed data of the UMI lines are tabled and accessed using search options. The genotype table contains information about the gel image, chromosome number, forward sequence, reverse sequence, allelic size with different varieties.

4.4 WEB ARCHITECTURE

A GUI interface was developed in HTML with PHP to access the germplasm and genotype data dynamically via web (www.tnaugenomics.com/database/maize). The backend tool for this database is MySQL. This backend can be accessed in the web (www.tnaugenomics.com/database/admin). In the interactive page, the administrator should type the user name and password to add or delete any datum to the database. This MySQL tables and connected PHP front pages are stored in

tnaugenomics web server (ftp.tnaugenomics.com). Finally the front page and MySQL tables are connected using PHP.

4.5 QUERY AND DATA RETRIEVAL

The reason for establishing this database is to facilitate the retrieval of specific germplasm and marker information that the user needs. For example, the user may wish to access specific germplasm information of particular variety, specific marker information, and SSR or RAPD gel patterns. A user-friendly interface is very essential if a database is to be of value to the scientific community. The user interface of this database was designed according to the principles of human computer interactions.

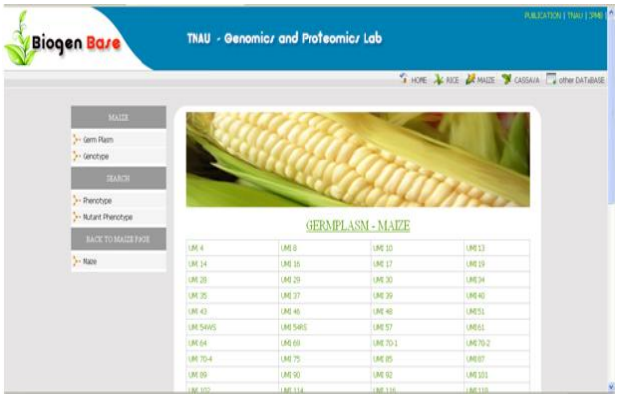
5. RESULTS AND DISCUSSIONS

Homepage

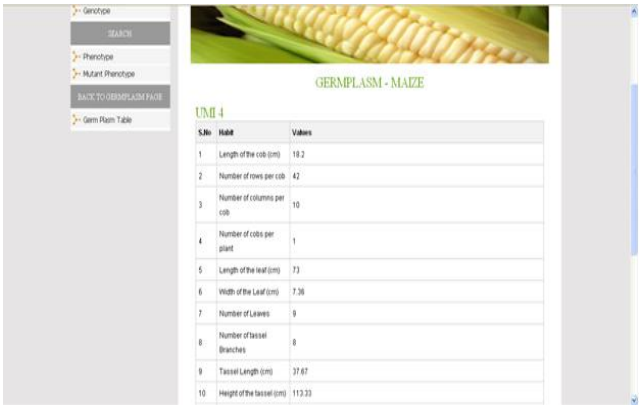
Home page can be accessed through the www.tnaugenomics.com/database/maize. This has four links to germplasm, genotype, phenotype and mutant phenotype.

Germplasm

Entering into the link "Germplasm" one hundred and one (101) varieties are displayed as shown below.



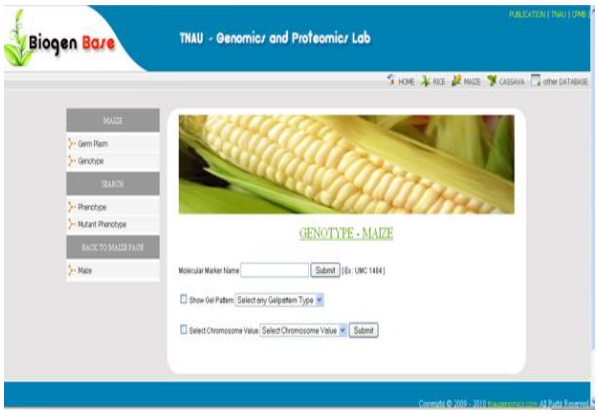
Each entry on its click, gives its own details. For example, by clicking onto the inbred UMI 4, all the phenotypic and marker details of the inbred will be displayed.



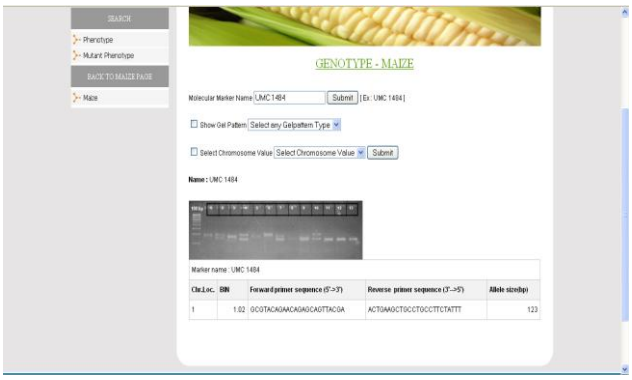
The germplasm accessions in this database is unique to the other databases existing, as it provides information unique from TNAU , which can be accessed more user friendly, in a GUI interface.

Genotype

The "Genotype" page helps the user to query chromosome number, Gel image or the marker name. The Gel pattern can be SSR or RAPD.



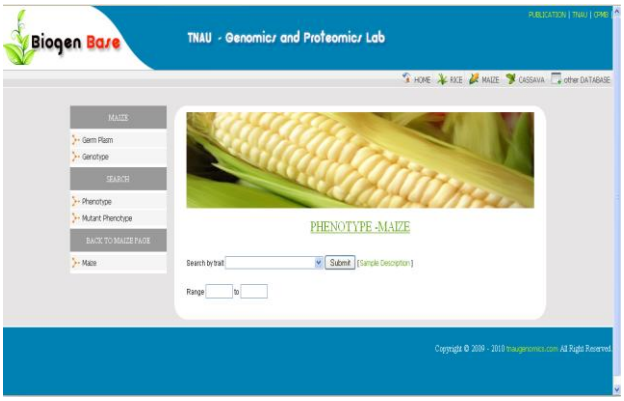
The name of the marker is queried and the output is displayed as follows with the gel pattern, the location on the chromosome, the BIN, the primer sequences for both the forward and the reverse primers and the allele size.



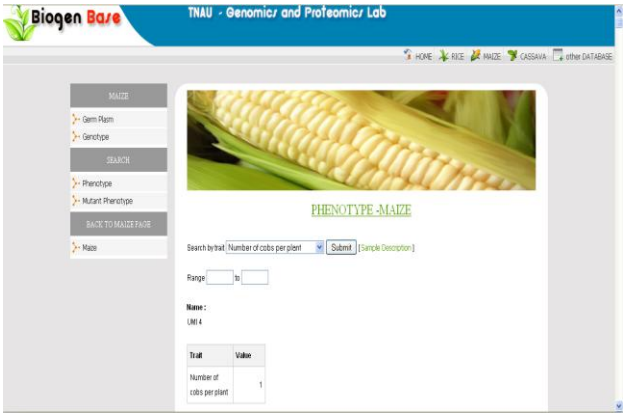
The Genotype search in this database is more advantageous than MAIZEGDB, and other databases by providing the specific information about the markers for the available accssions.

Phenotype

The phenotype is queried by selecting the trait from the drop down menu.

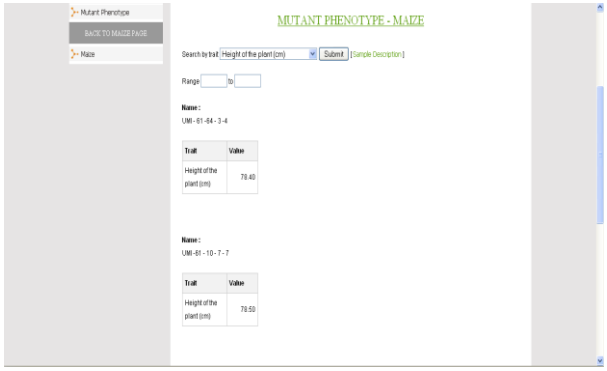


The number of cobs per plant is queried and all the germplasm showing only the number of cobs per plant are given as the output.



Mutant Phenotype

The mutant phenotype is also queried by selecting the traits from the drop down menu. The height of the plant (cm) is queried and all the mutants containing plant height as a mutant trait are obtained.



6. CONCLUSION

The maize database, containing the data obtained from the University Maize Inbreds (UMI) is a wonderful piece of equipment for storing large quantities of data efficiently. Through the process of database normalization, we have brought our schema's tables into conformance with progressive normal forms. As a result each of our tables represents a single entity and so there is decreased redundancy, fewer anomalies and ultimately the efficiency are improved. The methods, thus, appear to be effective and may be of interest for use in other scientific databases too.

Ideally, the scientists of TNAU Genomics and Proteomics lab will deposit their newer data that become accessible on the web. In the future, we have plans to address the issue of data quality further by establishing a mechanism for data checking and curation. It may also be useful to allow users to search germplasm in search box, as the number of germplasm will be increasing in the days to come. Thus the user may face some difficulties to find out their need. It will also be important to integrate phenotype search for germplasm phenotype characters with specific ranges. Finally, as the database grows, it may be necessary to optimize performance to minimize response times. This can be done by indexing the commonly searched attributes, such as names and keywords.

7. ACKNOWLEDGEMENT

This work was supported by Department of Biotechnology, New Delhi under Programme support for research and development in agricultural biotechnology to Genomics and Proteomics Laboratory, Centre for Plant Molecular Biology, Tamil Nadu Agricultural University, Coimbatore, India.

8. REFERENCES

[1] T.Z. Sen, C.M. Andorf, M.L. Schaeffer, L.C. Harper, M.E. Sparks, J. Duvick, V.P. Brendel, E. Cannon, D.A. Campbell and C.J. Lawrence. MaizeGDB becomes 'Sequence-cenric'. 2010.

[2] M.C. Costanzo, M.S. Skrzypek, R. Nash, E. Wong, G. Binkley, S.R. Engel, B. Hitz, E.L. Hong and J.M. Cherry. The Saccharomyces Genome Database Project. New mutant phenotype data curation system in the Saccharomyces Genome Database. 2009.

[3] N. Junjian, P. Anuradha, K. Youens-Clark, Immanuel, Y., Pankaj, J., Isaak, T., Chih, W.T., Liya, R., William, S., Xuehong, W., Shuly, A., Doreen, W., Lincoln, S. and Susan, M. Gramene QTL Database: Development, content and applications. 2009.

[4] S. Hunter, R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, D. Das, L. Daugherty, and Duquenne, L. InterPro: The integrative protein signature database. Nucleic Acids Res., 2009. D211 – D215.

[5] H. Mcwilliam, F. Valentin, M. Goujon, W. Li, M. Narayanasamy, J. Martin, T. Miyar and R. Lopez. European Bioinformatics Institute, EMBL Outstation, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Web services at the European Bioinformatics Institute. 2009.

[6] L. Wissler, E. Dattolo, A.D. Moore, T.B.H. Reusch, J.L. Olsen, M. Migliaccio, E. Bornberg-Bauer and G. Procaccini, Dr. Zompo. An online data repository for Zostera marina and Posidonia oceanic ESTs. 2009.

[7] K. McLeod and A. Burger, Heriot-Watt University and 2MRC Human Genetics Unit, Edinburgh, UK. Towards the use of argumentation in bioinformatics: A gene expression case study. 2007.

[8] Co te', R.G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R. and Hermjakob, H. The Protein Identifier Cross-referencing (PICR) service: Reconciling protein identifiers across multiple source databases. BMC Bioinform, 2007. 8. 401.

[9] P. Rice, I. Longden, and Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet., 2000. 16.

[10] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and Bourne, P.E. The Protein Data Bank. Nucleic Acids Res., 2000. 28. 235 – 242.

[11] R.T. Fielding. Architectural styles and design of network-based software architectures. Ph.D. Thesis. UC Irvine. 2000.