OPEN ACCESS Freely available online

Global Journal of Life Sciences and Biological Research

Review Article

Update on Plant Genome Databases

Anju Alexander*

Department of Biotechnology, Jawaharlal Institute of Postgraduate Medical Education and Research, Pondicherry, India

ABSTRACT

With the increasing availability and augmented computational capacities to analyze sequencing data, the demand for dedicated databases to store this data is ever growing. This has led to the formation of dedicated plant genome repositories. These repositories are not mere warehouses that store the sequence data, but have huge computational capabilities that can anatomize this data and extract relevant information. This review emphasis the features of plant genome databases.

Keywords: Genome; Plant; Practices; Sequencing; Molecular biology

INTRODUCTION

A Database is a collection of related data organized in a way that data can be easily accessed, managed and updated. Database can be software based or hardware based, with one sole purpose, storing data. There are three main types of biological databases they are large-scale public repositories, community-specific database resources, and project-specific databases and many of the databases are easily accessible for the public. In the ensuing years bioinformatics tools began appearing at various sites including the European Molecular Biology Laboratory, the Molecular Biology Research Resource at the Dana-Farber Cancer Institute in the mid-1980s, the National Center for Biotechnology Information (NCBI) in 1988, the Genome Database Project at Johns Hopkins University in early 1989, and in countless laboratories throughout the world [1]. Sequence of a genome is only the first step toward understanding genome organization, gene structure, gene expression patterns, disease pathogenesis and a host of other features of both scientific and commercial interests. Computational tools of genomic annotation and comparative genomics must be applied to gain a useful understanding of any genome.

PLANT GENOMICS-HISTORY

Plants are the backbones of food chain for all living matters on Earth, which supply the humankind with food, feed products, and clothing and housing materials; balance and earth ecology; provide a medicine and cure for many diseases; and produce energy and biofuels [2]. This has consequently shaped and revolutionized plant sciences, genetics, and crop breeding. A remarkable feature of plant genomics is its ability to bring together more than one species for analysis. According to the Aristotles principles of Taxonomy Theophrastus began the systemic characterization of plants during antiquity. Centuries later, Mendel's studies of the inheritance of traits in pea plants founded the field of genetics, and much of Darwin's work on the evolution of forms by natural selection was supported by experiments on plants. Genomics is the ultimate interdisciplinary approach, as it covers the entire spectrum from DNA sequencing to field based research [3]. Plant genomics is a newly evolved discipline of plant sciences targeting to decode, characterize, and study the genetic compositions, structures, organizations, functions, and interactions of all plant genes in a genome-wide scale. By 2000, the seeds of success were sown in the field of plant genomics with the sequencing of the genome of Arabidopsis thaliana. Being evolved from plant molecular genetics, biology, and biotechnology, Plant genomics represent the key sub-divisions of structural, functional, comparative, evolutionary, physiological, and genetical genomics [4]. Its development and advances are tightly interconnected with plant science sub-disciplines such as proteomics, metabolomics, epigenomics, phenomics, metagenomics, transgenomics, breeding-assisted genomics, bioinformatics, and system biology as well as modern instrumentation and robotics sciences.

PLANT GENOME PROJECT

The availability of recombinant DNA and PCR technologies helped in preparation of molecular maps for plant and animal

Correspondence to: Dr. Anju Alexander, Department of Biotechnology, Jawaharlal Institute of Postgraduate Medical Education and Research, Pondicherry, India, E-mail: anjualexander2298@gmail.com

Received: February 24, 2021; **Accepted**: March 10, 2021; **Published**: March 17, 2021

Citation: Alexander A (2021) Update on Plant Genome Databases. Glob J Lif Sci Biol Res. 7:2.

projects and development of new high-throughput sequencing technologies has increased dramatically the number of successful genomic projects [1,2]. The genomes of the eudicot model plant for plant biology, Arabidopsis thaliana, and the monocot crop model plant rice (Oryza sativa) were the first genomes to be sequenced. Arabidopsis, a weed plant found all over the world, easy to grow, short span of life cycle, one of the smallest genome among dicot plants and Oryza sativa (rice), one of the members from monocot having very simple genome Organisation and similarity with other major cereal plants. The rice genome programme was started in 1991 in Japan At present 10 countries are participating in International Rice Genome Sequencing Programme.

Three American groups namely Meyerowitz, Somerville and Goodman took the first step leading to the project on sequencing of the Arabidopsis genome later known as AGI. According to the Genomes On-Line Database (GOLD), more than 20 plant genomes have been already completed and there are more than 200 ongoing plant genomic projects. The NCBI has genomic sequences of more than 60 plant species. Specific plant comparative genomic databases are become powerful tools for gene family annotation in plant clades. Nowadays, several other plant species from both, eudicot and monocot clades have been completely sequenced and their sequences are publicly available. Main genome centers, such as JGI (Joint Genome Institute), BGI (Beijing Genomics Institute), JCVI (J. Craig Venter Institute) or MSU (Michigan State University), support most of the completed or ongoing plant genomic projects [5]. Formerly, plant genomes have been usually sequenced by the Sanger sequencing technology. And currently NGS technologies made great changes in the field of sequencing technologies.

GENE FAMILIES

Gene families can be defined as sets of evolutionary related genes shared by a number of different species and with often similar biological functions, or by a set of homologous genes within one species. Some gene families appear to be more dynamic during evolution and show species-specific gene members. Others are more conserved and consist of genes sharing common ancestry that have diverged by speciation (orthologous genes).Orthologous genes are particularly useful for the characterization of unannotated proteins by identifying annotated counterparts that share high sequence identity [6]. This has lead to the development of traditional signature databases, such as Pfam, since these motifs (or signatures) have been shown to be important for protein functionality and are able to define a family of proteins.

Recently, novel bioinformatics tools have been developed for the analysis of gene families based on comparative genomics. These tools have been integrated in comparative genomic databases that can be used to perform evolutionary and comparative analyses, and to study gene families and genome organization. Based on orthologous genes, comparative genomics provides a powerful approach to translate functional information from model species to crops. The most comprehensive comparative genomic databases that focus on plant gene families are PLAZA, GreenPhylDB and Phytozome (*Martinez.* M). These methods are powerful tools to classify many sequences rapidly, in an automated manner, and with reasonable accuracy, and have allowed discovering novel gene families not covered by signature methods. The plant comparative genomic databases are the best choice for the identification of members of a protein family in different species, which is particularly interesting for phylogenetic analyses and the prediction of gene function.

PLANT GENOME DATABASES

With the increasing availability and augmented computational capacities to analyze sequencing data, the demand for dedicated databases to store this data is ever growing. This has led to the formation of dedicated plant genome repositories. Both general and species-specific plant databases are available that are valuable for plant sequencing data storage and analysis [7]. The genome databases can be generally classified as several types. A good genome database should meet two criteria: (i) integration of various types of genomic data, and (ii) providing genome analysis tools [8]. These databases are updated with the release of new versions or updated roughly each year.

SINGLE SPECIES DATABASE

The most popular Arabidopsis database is the Arabidopsis Information Resource. TAIR provides updated genome sequence (currently V10) and various genomic information, including SNP, transposons, genes, gene families, gene annotations, gene names, proteins, and mutant orderings [9].

The Genome Database for Rosaceae (GDR) is the central repository and data-mining resource for genomics, genetics and breeding data of Rosaceae, an economically and nutritionally important crop family that includes almond, apple, apricot, blackberry, cherry, peach, pear, plum, raspberry, rose and strawberry. It was first established in 2003, GDR initially provided web interfaces and analysis tools for emerging genomic and genetic data such as genus-specific EST unigene sets, linkage maps and genetic markers [10].

EuroPineDB is the largest sequence collection available for a single pine species, Pinus pinaster (maritime pine), since it comprises 951 641 raw sequence reads obtained from non-normalised cDNA libraries and high-throughput sequencing from adult (xylem, phloem, roots, stem, needles, cones, strobili) and embryonic (germinated embryos, buds, callus) maritime pine tissues. Using open-source tools, sequences were optimally pre-processed, assembled, and extensively annotated (GO, EC and KEGG terms, descriptions, SNPs, SSRs, ORFs and InterPro codes). The complete database, which is designed to be scalable, maintainable, and expandable. It can be retrieved by gene libraries, pine species, annotations, UniGenes and microarrays and will be periodically updated [11]. Small assemblies can be viewed using a dedicated visualization tool that connects them with SNPs. Any sequence or annotation set shown on-screen can be downloaded. Pinus pinaster is an economically and ecologically important species that is becoming a woody gymnosperm model. Its enormous genome size makes whole-genome sequencing approaches are hard to apply.

GSAD provides genome size data specifically for Asteraceae (Compositae), which are considered to be one of the largest

plant families (24,000–30,000 species) with a worldwide distribution, except Antarctica. Development of GSAD was initiated by research groups based at the Universitat de Barcelona and Institute Botànic de Barcelona (IBB-CSIC-ICUB) in collaboration with a team from the Université de Paris Sud-CNRS. The aim was to complement the Plant DNA C-values database in the same way that the Index to Chromosome Numbers in Asteraceae (Table 1).

Database	Species	Function
MaizeGDB	Zeamays	SNP, Pedigree
BRAD	Brassica	Annotation ,syntenic /non syntenic orthologus, flanking regions
Spinach Base	Spinach	Gene family classification, SNP
VISTA	Angiosperms	Multiple genome analysis, enhancer prediction
Ensembl Plants v36	Red algae to angiosperms	Homology search, variant effect predictor
PLAZA	Eukaryotic algae seed plants	Gene family, localization, colinearity,

Table 1: Specific function of the databases.

COMPREHENSIVE DATABASE

Phytozome is a large plant genomic portal sponsored by the USA Department of Energy (DOE). In addition to BLAST and Gbrowse tools, Phytozome also provide Biomart which allow users to annotate plant gene families, to study the evolution of plant gene families, to display genes in the genomic context which is valuable for a wide range of scientists who are interested in gene family evolution. However, considering that the genomes of 236 angiosperm species have been sequenced, <1 third of all sequenced angiosperms are included by Phytozome, suggesting the presence of a major gap in the availability of most angiosperm genomes at Phytozome [12].

The Plant Genome Duplication Database (PGDD) is a database currently hosting 43 angiosperm genomes, with tools to identify the intra genome and cross-genome synteny relationships. Synonymous substitutions of homologs inferred from syntenic alignments could be calculated from this database. By the syntenic comparison, PGDD facilitates their identification of evolutionary analysis of gene and genome duplication [13].

PlantGDB is a database which provides molecular sequence data for all plant species .The goal of PlantGDB is to determine the set of genes common to all plants or specific to particular species by integrating a number of bioinformatics tools that facilitate gene prediction and cross-species comparisons. PlantGDB provides genome browsing capabilities for species with large scale genome sequencing efforts. PlantGDB also seeks to make database technology available to individual molecular biology and genomics research groups. This aim is based on the premise that software tools should be as widely distributed as are significant laboratory techniques [12,13].

CLADE-ORIENTED DATABASE

Gramene is a clade-oriented database that provides comparative genomic data for plants. As the first completely sequenced crop genome, rice continues to be the best-annotated genome for monocots and offers a wealth of information on the structure and function of genes, polymorphisms and other functional elements anchored to the genome Gramene is a curated resource for genetic, genomic and comparative genomics data for the major crop species, including rice, maize, wheat and many other plant (mainly grass) species. Gramene is an open-source project. All data and software are freely downloadable through the ftp site and available for use without restriction [14]. Gramene's core data types include genome assembly and annotations, other DNA/mRNA sequences, genetic and physical maps/markers, genes, Quantitative Trait Loci (QTLs), proteins, ontologies, literature and comparative mappings.

The Sol Genomics Network is a clade-oriented database (COD) containing biological data for species in the Solanaceae and their close relatives, with data types ranging from chromosomes and genes to phenotypes and accessions. SGN hosts several genome maps and sequences, including a pre-release of the tomato (Solanum lycopersicum cv Heinz 1706) reference genome. SGN is also an open source software project, continuously developing and improving a complex system for storing, integrating and analyzing data. All code and development work is publicly visible on GitHub. The database architecture combines SGN-specific schemas and the community-developed Chado schema for compatibility with other genome databases [15].Since it is a community driven, using simple web tools researchers can add and edit information in the SGN curation model.

The Legume Information System (LIS), developed by the National Center for Genome Resources in cooperation with the USDA Agricultural Research Service (ARS), is a comparative legume resource that integrates genetic and molecular data from multiple legume species enabling cross-species genomic and transcript comparisons and also gene expression and biochemical pathways. Transcript libraries are represented as images of plant organs in different developmental stages, which are selected to query the analyzed and annotated data. Complex queries can be sort it out by adding keywords and sequence names. SoyBase, a resource familiar to the soybean community, has been partially integrated into the LIS, capturing all of the soybean map and biochemical pathway data.

DATABASE FOR PLANT PEPTIDE

PlantPepDB is a manually curated database that consists of 3848 plant-derived peptides among which 2821 are experimentally validated at the protein level and 458 have experimental evidence at the transcript level. Incorporation of

physicochemical properties and tertiary structure into PlantPepDB will help the users to study the therapeutic potential of a peptide, thus, it's a powerful resource for therapeutic research [16]. Overall, PlantPepDB is the first database comprising detailed analysis and comprehensive information of phyto-peptides from a broad functional range which will be useful for peptide-based applied research.

DATABASE FOR STRESS-RESPONSIVE GENES

This database comprises the genes that are differentially regulated under stress conditions. The Stress-Responsive Transcription Factor DataBase (STIFDB) is as well a collection of abiotic and biotic stress responsive genes from Arabidopsis and *Oryza sativa* (rice) including the possibility to identify putative binding sites in the promoters of these genes. In the same line, there is a database of rice transcription factors under stress conditions, RiceSRTFDB, giving information about expression, *cis*-elements and information about mutant phenotypes. Moreover, there is a general 'Plant Stress Gene Database' including 259 stress-related genes from 11 species and for some genes also information about orthologs is included. However, all these databases are neither specific for drought stress nor dedicated to genes with experimentally proven function under stress conditions. PPNEMA is a database that represents plant-parasitic nematode ribosomal genes. That resource allows the user to browse, search and generally explore phytoparasite ribosomal DNA [17-19].

DATABASE FOR FOREST TREES

The Dendrome Project and associated TreeGenes database serve the forest genetics research community through a curated and integrated web-based relational database. TreeGenes was developed to provide a centralized web resource with analysis and visualization tools to support data storage and exchange. The TreeGenes database is a resource for all forest trees. Because of the very large size and complexity of the conifer genomes, greater emphasis has been placed on the expressed portion of the genome. TreeGenes functions through a semi-automated PostgreSQL version 8.1.4-based database that consists of modules to hold a broad range of data and information for trees.

DATABASE BASED ON PLANT GENOME SIZE

The Plant DNA C-values database was first launched in 2001 to provide a user-friendly searchable database where both published and unpublished values of plant genome size could be readily found. Overall, analysis of the data available in the Plant DNA C-values database illustrates the considerable diversity in genome sizes between the different land plant and algal groups, both in terms of the range of genome sizes encountered and the distribution of genome sizes. It helps to understand plant lineages and argue strongly for the need to continue to collate and analyze genome sizes across the plant tree of life to form a more holistic understanding of plant genomic diversity [12-15].

CONCLUSION

Creation of databases, all have generated a wealth of meaningful creation information. The knowledge gained from these efforts could be utilized to further understand plant genome design and intricacy, ultimately leading to a) the determination of functions of genes involved in environmental stress resistance, b) the generation of plants with better fruit quality, c) the superior use of genetic diversity that could help produce better plants and crops for the future. The main function of genome databases has evolved from data storage to online analysis. The scope, tools, and data of each type of databases and their features are concisely discussed. While a timely-updated comprehensive database is more powerful for address of major scientific mysteries at the genome scale. In general, the plant genome database will become a new biological branch. The super computing equipment, bioinformatics algorithms, and tool development need to be introduced and upgraded. In addition, a user-developer interactive than user-friendly interface is required.

REFERENCES

1. Alter S, Bader KC, Spannagl M, Wang Y, Bauer E, Schön CC, et al. Drought DB: An expert-Curated compilation of plant drought stress genes and their homologs in nine species. Database. 2015.
2. Bombarely A, Menda N, Tecle IY, Buels RM, Strickler S, York TF, et al. The Sol genomics network (solgenomics.net): Growing tomatoes using Perl. Nucleic Acids Res. 2011;39(S1):D1149-D1155.
3. Cheng F, Liu S, Wu J, Fang L, Sun S, Liu B, et al. BRAD, the genetics and genomic database for Brassica plants. BMC Plant Biology. 2011.
4. Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, et al. The sequenced angiosperm genomes and genome databases. Front Plant Sci. 2018.
5. Das D, Jaiswal M, Khan FN, Ahamad S. PlantPepDB: A manually curated Plant Peptide Database, Scientific Reports. 2020;10.
6. Gonzales MD, Farmer A, Archuleta E, Gajendran K. The Legume Information System (LIS): An integrated information resource for comparative biology. Nucl Acids Res. 2005;33:D660-5.
7. Garcia S. Recent updates and development to plant genome size databases. Nucl Acids Res. 2014;42(D1):D1159-D1166.
8. Gary R. Bioinformatics tools for plant genomics. Int J Plant Genomics. 2008.
9. Khurana JP. An update on chloroplast genomes. Plant Systematics and Evolution. 2008;271(1):101-122.
10. Jung S. 15 Years of GDR: New data and functions in the genome database for Rosaceae. Nucl Acids Res. 2019; 47(D1):D1137-D1145.
11. Lai K. Wheat Genome info: An integrated Database and portal for wheat genome organization. Plant Cell Physio. 2012;53(2):e2.
12. Maslen G. Ensembl Genomes 2020-Enabling non-vertebrate Genomic Research. Nucl Acids Res. 2020;48(D1):D689–D695.
13. Pachter L. VISTA: Computational tools for comparative genomics. Nucl Acids Res. 2004;32(S2):W273-W279.
14. Proost S. PLAZA: A Comparative Genomic Resources to study gene and genome evolution in plants. Special Series in Large Scale Biology. 2009;21(12):3718-3731.
15. Rhee Y. Biological databases for plant research. Plant Physiol. 2005;138(1):1-3.
16. Singh A. An update on bioinformatics resources for plant genomics. Curr Plant Biol. 2017.
17. Stein G. Phytozome: A comparative platform for green plant genomics. Nucl Acids Res. 2012;40(D1):D1178-D 1186.

18. Ware D. Gramene: A growing plant comparative genomic resource. Nucl Acids Res. 2008;36(D1):D947-D953.

19. Wegrzyn L. Tree Genes: A Forest Tree Genome Data Base. Int J of Plant Genomics. 2008.