International Journal for Computational Biology (IJCB)
Vol.3, No.1, April 2014, Page. 43-47
ISSN: 2278-8115

Plant Genomic Databases for Oilseeds Crop Improvement

Subhasisa Bal*, Kallamadi Prathap Reddy
Directorate of Oilseeds Research (DOR), Rajendranagar, Hyderabad-500030 (India)

Article Info

Article history:

Received March 29th, 2014
Revised April 11th, 2014
Accepted April 13th, 2014

Keyword:

Plant Genomics
Oilseed Crop
Genomic Database
Crop Improvement
Information Mining

ABSTRACT

Plant genomic databases are collection of huge information on plants, germplasm accessions, descriptors, plant genetics, physical and genomic sequence maps, QTLs, loci, sequence information, molecular markers, references etc. At present more than 100 plants genomic databases are available. These are dedicated to generic genome data focusing on specific crops. Some of the important oilseed plant databases include Castor Bean Genome Database, CGPDB, SoyBase, Legume Information System (LIS), Brassica database, Sinbase etc. Due to availability of number of genomic databases for crop plants the researcher needs to visit most appropriate database and choose suitable parameters for genomic information mining. The wealth of information available to researchers today can be overwhelming hence understanding the plant databases for harnessing genomic information is the need of the hour for crop improvement research programmes.

Corresponding Author:

Subhasisa Bal,
Directorate of Oilseeds Research (DOR), Rajendranagar, Hyderabad
Email: subhasisabal@gmail.com

How to Cite:

Subhaisa Bal *et. al.* Plant Genomic Databases for Oilseeds Crop Improvement. IJCB. 2014; Volume 3 (Issue 1): Page 43-47.

1. INTRODUCTION

Plant genomic databases are collections of huge information on plants, germplasm accessions, descriptors, plant genetics, physical and genomic sequence maps, QTLs, loci and their various alleles, tools, phenotypes, stocks, sequence information, molecular markers, references etc. Most of the databases are available in public domain. Plant genome databases have been designed for a number of agriculturally important crop species. The major oilseeds edible crops are brassica, groundnut, soybean, sunflower, sesamum, linseed, safflower and niger. The major non edible oilseeds crop are castor, jatropha, pongamia, madhuca etc. The paper focuses only on the important oilseeds crop databases; their importance in mining the information is discussed. Some of the important oilseeds plant databases include Castor Bean Genome Database, CGPDB, SoyBase, Legume Information System (LIS), Brassica database, Sinbase etc. Information contained in these databases encompassed all available data from germplasm to genomics and includes molecular mapping data, molecular marker data, germplasm information, trait studies, identified quantitative trait loci, pathogen descriptions, relevant publication citations, images pertaining to all aspects of the crop, and address catalogue of the researchers associated with the crops.

2. SURVEY OF AVAILABLE DATABASE

The goal of these databases is to provide "one-stop shopping" or “single window” information that is relevant to a species or group of species [1]. The genomic information stored in these databases is mainly generated through wet-lab investigation by researchers. The specialized molecular biology techniques, high throughput sequencing techniques and biotechnology experiments are key foundation for the generation of the sequence information. Once the sequence information is generated bioinformatician will mine the useful hidden information in the sequences. After squeezing the information based on the need of the research to be carried

out the sequence information will be submitted to the genomic databases. In this paper web link/addresses of important oilseeds plant genomic databases are listed along with important features, from a researcher's point of view.

2.1 Castor Bean Genome Database

The oilseed plant castor bean (*Ricinus communis*) is a member of the family Euphorbiaceae that includes other important species such as Cassava (*Manihot esculenta*), rubber tree (*Hevea brasiliensis*), ornamental poinsettias (*Euphorbia pulcherrima*), and the weed leafy spurge (*Euphorbia esula*). The castor bean genome database is developed and maintained at J. Craig Venter Institute (JCVI), USA. The sequencing of the castor bean (cultivar HALE) genome with 4.5 X coverage was conducted at JCVI. Sequencing results have shown that the genome is 350 Mb and has an estimated 31,237 genes. The important features of this database includes GBrowse, Blast, TIGR castor bean WGS assembly (assemblies=25,828) which can be browse at http://www.castorbean.jcvi.org/.

2.2 CGPDB (Composite Genome Project Database)

The Michelmore Lab of UC Davis Genome Center maintains the database of composite genome project database which includes oilseeds crops like sunflower and safflower. The objectives of CGPDB database are 1) to develop comprehensive gene catalogs and genomic sequences for economically and taxonomically important species in the Compositae family. 2) to develop detailed genetic maps integrating phenotypic data for agriculturally important traits with sequenced genes. 3) to enhance the introgression of agriculturally useful alleles from wild species. 4) to establish tools and resources for the Compositae that will be the basis of genomic investigations of crops and invasive species in the family. Information on plant genetics, physical and genomic sequence maps, QTLs, loci and their various alleles, phenotypes, stocks, sequence information, molecular markers, references can be accessed from the database site http://www.compgenomics.ucdavis.edu/.

2.3 SoyBase

SoyBase, the USDA-ARS soybean genetic database is a comprehensive repository for professionally curated genetics, genomics and related data resources for soybean. SoyBase contains the most current genetic, physical and genomic sequence maps integrated with qualitative and quantitative traits. The quantitative trait loci (QTL) represent more than 18 years of QTL mapping of more than 90 unique traits. SoyBase also contains the well-annotated 'Williams 82' genomic sequence and associated data mining tools which can be searched at http://www.soybase.org/.

2.4 Legume Information System (LIS)

The genomic information related to the groundnut and soybean can be accessed through Legume Information System (LIS). LIS is 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. The web link of the database is http://www.comparative-legumes.org/.

2.5 Brassica database (BRAD)

Brassica database (BRAD) is a web-based resource focusing on genome scale genetic and genomic data for important Brassica crops. BRAD was built based on the first whole genome sequence and on further data analysis of the Brassica A genome species, *Brassica rapa* (Chiifu-401-42). It provides datasets, such as the complete genome sequence of B. rapa, which was de novo assembled from Illumina GA II short reads and from BAC clone sequences, predicted genes and associated annotations, non coding RNAs, transposable elements (TE), B. rapa genes' orthologous to those in A. thaliana, as well as genetic markers and linkage maps. BRAD offers useful searching and data mining tools, including search across annotation datasets, search for syntenic or non-syntenic orthologs, and to search the flanking regions of a certain target, as well as the tools of BLAST and Gbrowse. BRAD allows users to enter almost any kind of information, such as a B. rapa or A. thaliana gene ID, physical position or genetic marker can be browse at http://brassicadb.org/.

2.6 Sinbase

It is a comprehensive *Sesamum indicum* genomics database generated by assembling the whole genome sequence of sesame, and analyzed its evolutionary history and important characteristics. In sesame genome project, whole-genome shotgun (WGS) sequencing strategy, based on paired short reads generated by second-generation Illumina GA sequencing technology, was used to assemble the draft genome of Sesamum indicum. We assembled 274 Mb (77.4% of the estimated genome) of the sesame genome and annotated 27,148 genes. Using a newly constructed genetic map, 150 larger scaffolds were anchored to 16 pseudomolecules,

representing 85.3% of the assembly in size and 91.7% of the predicted genes which is available from the link http://ocri-genomics.org/.

2.7 National genomic resources repository (NBPGR, India)

National genomic resources repository established as an institutional framework for methodical and centralized efforts to collect, generate, conserve and distribute genomic resources for agricultural research in India. The Repository is housed in the premises of National Bureau of Plant Genetics Resources, New Delhi and can be searched from the NBPGR website. The resources which can be deposited includes are:

a) Cloning vectors, expression vectors, binary vectors, RFLP probes
b) Cloned genes, promoters fused to reporter genes
c) Sub-genomic, cDNA , EST, repeat enriched libraries
d) BAC, YAC, PAC clone set from sequencing projects
e) Genomic, mitochondrial or chloroplast DNA
f) Cloned DNA from wild and weedy species produced exclusively for the repository

2.8 NCBI (National Center for Biotechnology Information)

The NCBI database repository is one among the largest and most popular repository of the genomic resources dealing with human, animal plant and microorganisms. The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health, USA. The NCBI is located in Bethesda, Maryland and was founded in 1988 through legislation sponsored by Senator Claude Pepper. The NCBI houses genome sequencing data in GenBank and an index of biomedical research articles in PubMed Central and PubMed, as well as other information relevant to biotechnology. All these databases are available online through the Entrez search engine.

Table 1. List of other important genomic databases related to crop improvement

Database	Weblink
Brachypodium database	http://www.brachypodium.org/
Brassica genome gateway	http://www.brassicagenome.net
DNA Data Bank of Japan (DDBJ)	http://ddbj.sakura.ne.jp
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
EnsEMBL plants	http://plants.ensembl.org
GenBank	http://www.ncbi.nlm.nih.gov/genbank/
Graingenes	http://wheat.pw.usda.gov/
Gramene	http://www.gramene.org/
HapMap	http://www.hapmap.org
International Crop Information System (ICIS)	http://www.icis.cgiar.org
MaizeGDB	http://www.maizegdb.org/
Maize sequence database	http://www.maizesequence.org
Oryzabase	http://www.shigen.nig.ac.jp/rice/oryzabase
Panzea	http://www.panzea.org/
Phytozome	http://www.phytozome.net
PlantsDB	http://mips.helmholtzmuenchen
PlantGDB	http://www.plantgdb.org
The Plant Ontology	http://www.plantontology.org/
SSR Primer	http://flora.acpfg.com.au/ssrprimer2/
SOL Genomics Network (SGN)	http://solgenomics.net/
TAGdb	http://flora.acpfg.com.au/tagdb/
The Triticeae Repeat Sequence Database (TREP)	http://wheat.pw.usda.gov/ITMI/Repeats/
TAIR	http://www.arabidopsis.org/
National Genomic Resources Repository	http://www.nbpgr.ernet.in/repository
The Essential Oil Database	http://www.nipgr.res.in/essoildb.html
Plant Reference Gene Server PlantRGS	http://www.nipgr.res.in/PlantRGS
CastorDB	http://castordb.msubiotech.ac.in/home.htm

The majority of DNA sequence and expressed gene sequence data generated today comes from the next- or second-generation sequencing (NGS/2GS) technologies. NGS technologies produce vast quantities of "short read" data rather than Sanger sequencing at a relatively low cost and short time [1, 2, 3]. Genomics is undergoing a revolution, driven by advances in DNA sequencing technology, and this data flood is having a major impact on approaches and strategies for crop improvement [1]. Numerous databases have been developed for genomic data, on a range of platforms and to suite a variety of different purposes which are enlisted in Table 1. These range from generic DNA sequence or molecular marker databases, QTL, genetic maps, mictoRNA, to those hosting a variety of data for specific species is listed below.

3. RESOURCES FOR CROP IMPROVEMENT

Genomic information available in the databases either for plant, animal or microorganisms is increasing with alarming rate. Thanks to internet and online mode by which a researcher sitting at the remote place can access the information as per wish and need, is great achievement of the information technology in 21st century. The computational biologist can utilize this genomic information for in silico studies particularly for the mining of information. The freely available sequence information in databases will speed up the comparative genomic studies in orphan crops where genomic resources are lacking [4]. Availability of all the reference genome sequence information in respective crop will greatly enhance the crop improvement. The sequence information may increase our ability to decipher the underlying molecular mechanisms of a complex traits, understand the gene regulatory mechanisms, determine gene expression differences and variations in expressed gene sequences, and other structural variations such as copy number variations (CNV) and presence-absence variations (PAV). From the reference genome, genome wide candidate gene marker for economically important genes, simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers can be developed which can be used for genetic mapping applications [5].

Another important application of reference genome is in association genetics or population biology, where either genomes or pools of PCR products of thousands of candidate genes can be sequenced in hundreds of individuals using barcodes [6]. The sequence data obtained could then be used to identify SNPs or haplotypes across genes or genomes for use in association genetics or population biology. Another area i.e. microRNA identification from reference genome databases is gaining importance.

Michael, and Jackson, [7] in their paper revealed that fifty-five plant genomes (sequence information) have been published representing 49 different species. One of the key take homes from the first 49 sequenced plant species is that we still have a lot to learn about the organization of genomes, function of genes, and how to characterize the non-coding space. Once the genome gets published each new genome uncovers novel genes specific to a species, and a vast amount of non-coding space that requires methods for ab initio and functional annotation. They also pointed out the specific challenge that how we will leverage a growing number of high throughput technologies, otherwise referred to as "omics" approaches, to functionally annotate features of the plant genome. Hence, for the computational biologist or crop bioinformatics researchers it is important to understand the intricacies of the genome sequence before downloading and mining of the information.

4. CONCLUSION AND FUTURE DIRECTIONS

At present more than 100 plants genomic databases are available [8]. These are dedicated to generic genomic data focusing on specific crops or family. Both the type and volumes of data have increased greatly over the last few years due to advancement in the NGS techniques and this trend looks to continue. As plant genome technology continues to advance and an increasing number of crop genomes become available, an expanding number of the plant genomic databases will be developed in near future. This situation is equally true for animals and microbes. One of the main challenges facing crop bioinformatics researchers and computational biologist is to make the ever increasing volume and types of sequence data available in a suitable format for genomic analysis [1]. Hence it is important for computational biologist or bioinformatician to have sound background of the basic biology along with computational tools. Due to availability of number of genomic databases for crop plants the researcher needs to visit most appropriate database and choose suitable parameters for genomic information mining. The wealth of information available to researchers today can be overwhelming therefore understanding the plant databases for harnessing genomic information is the need of the hour for crop improvement research programmes.

REFERENCES

[1] Lai K, Lorenc MT, Edwards D, Genomic databases for crop improvement. Agronomy 2, 62-73 (2013).
[2] Dudhe MY, Sarada C, Plant genomic databases for harnessing genomic information for oilseed crop improvement , DOR News Letter, 18(4):1-4 (2012a).
[3] Dudhe MY, Sarada C, In-silico development of microsatellite markers by using microsatellite identification tools available in public domain. Invited speaker lecture delivered in National Symposium on Currents Trends in Biotechnology, 7th and 8th Dec., 2013. Organized by Department of Botany Osmania University, Hyderabad, pp35 (2013).
[4] Dudhe MY, Meena HP, Ranganatha ARG, Mukta N, Lavanya C, In silico -identification of conserved domains from EST database in safflower. J. Oilseeds Res. 29 (Special issue): 178-181(2012).
[5] Mochida K, Shinozaki K, Genomics and bioinformatics resources for crop improvement. Plant Cell Physiol. 51(4): 497–523 (2010).
[6] Varshney RK, Nayak SN, May GD, Jackson SA, Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends in Biotechnology 27: 522–530 (2009).
[7] Michael TP, Jackson SA, The first 50 plant genomes. Plant Gen. vol. 6, doi:10.3835/plantgenome2013.03.0001in (2013).
[8] Dudhe MY, Sarada C, Comparative assessment of microsatellite identification tools available in public domain, DOR News Letter, 18(2 & 3):8-9 (2012b).