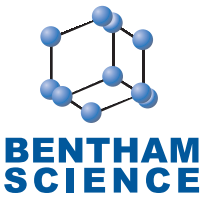
Send Orders for Reprints to reprints@benthamscience.ae

Current Genomics, 2016, *17*, 509-514

REVIEW ARTICLE



Computational Tools for Genomic Studies in Plants

Manuel Martinez[a,*]

[a]*Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Campus Montegancedo, 28223-Pozuelo de Alarcón, Madrid, Spain*

ARTICLE HISTORY

Received: June 17, 2015
Revised: December 09, 2015
Accepted: December 21, 2015

DOI:
10.2174/1389202917666160520103447

Abstract: In recent years, the genomic sequence of numerous plant species including the main crop species has been determined. Computational tools have been developed to deal with the issue of which plant has been sequenced and where is the sequence hosted. In this mini-review, the databases for genome projects, the databases created to host species/clade projects and the databases developed to perform plant comparative genomics are revised. Because of their importance in modern research, an in-depth analysis of the plant comparative genomics databases has been performed. This comparative analysis is focused in the common and specific computational tools developed to achieve the particular objectives of each database. Besides, emerging high-performance bioinformatics tools specific for plant research are commented. What kind of computational approaches should be implemented in next years to efficiently analyze plant genomes is discussed.

M. Martinez

Keywords: Comparative genomics, Computational tools, Genome databases, Genome projects, Plants.

1. OVERVIEW

The establishment of the primary DNA sequence of a species is a key point in modern research. It permits to obtain subsequently further insights on the functionality of the genes present in a genome. The first plant genome sequenced was that of the eudicot model plant *Arabidopsis thaliana*, which was released in 2000. From this date, and mainly after the appearance of Next Generation Sequencing technology in 2005, the number of sequenced plant species has exponentially increased. In recent years, several reviews had targeted the state of art of the plant sequencing projects [1-4]. Among the sequenced species, there are examples from most important clades. Several algae genomes, a moss, a lycophyte, three conifer and examples from the majority of the monocot and eudicot lineages have been sequenced [4, 5]. Interestingly, the vast majority of economical important crops have been sequenced and drafts of their genomes have been produced. Once the genomic sequence has been parsed, assembled and annotated; the corresponding data has been traditionally compiled in a database available for the scientific community. The importance of the creation of an open-access database comes from the rationale of the biological exploitation of the novel genomic resources generated by the researches. Basic considerations on dealing with crop plant databases utilization in the genomic era has been recently reviewed [6]. In this mini-review, the genomic databases and the bioinformatics tools available presently are well described.

2. GENOMIC PROJECTS DATABASES

The first step in the valorization of a genomic project is to know the availability of related projects completed or in progress. These searches can be done by two methods. By searching in the internet motor searches using key words or by exploring the specific databases related to genomic projects. Although the information in these databases is not always up to date, second method is more adequate to obtain the best results. There are four main genomic projects databases that can be used when information on plant genomes is required, GOLD (Genomes Online Database), NCBI genomes, CoGepedia and plaBi (Table **1**).

2.1. Gold [7] is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata. GOLD is now hosted by the JGI DOE institute and the current release is the version 5. The database currently hosts information for more than 20,000 studies, 60,000 biosamples, 60,000 sequencing projects and 50,000 analysis projects. One of the specific features of GOLD is that it includes nuclear and organelle genome projects, but also transcriptomic, methylation, exome and re-sequencing projects. Around one hundred and more than 3,400 finished or ongoing projects involve species from the phyla Chlorophyta and Streptophyta, respectively. GOLD is manually curated, their stored metadata are quality-controlled, and fully supports and follows the Genomic Standards Consortium (GSC) Minimum Information standards.

*Address correspondence to this author at the Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Campus Montegancedo, 28223-Pozuelo de Alarcón, Madrid, Spain; Tel: +3413364564; E-mail: m.martinez@upm.es

1875-5488/16 $58.00+.00 ©2016 Bentham Science Publishers

Current Genomics

Table 1. List of main resources available for plant genomic project information and for plant comparative genomics.

Database	URL	Reference	Last Update*
Genomic projects			
GOLD	https://gold.jgi.doe.gov/	Reddy *et al.*, 2015 [7]	Nov 2015
NCBI Genomes	http://www.ncbi.nlm.nih.gov/genome/	NCBI Resource Coordinators, 2015 [8]	Nov 2015
NCBI Assembly	http://www.ncbi.nlm.nih.gov/assembly/	Kitts *et al.*, 2015 [9]	Nov 2015
CoGepedia (plant genomes)	https://genomevolution.org/	-	2014
plaBi	http://plabipd.de/	-	Nov 2015
Comparative genomics			
Ensembl Plants	http://plants.ensembl.org/	Bolser *et al.*, 2015 [12]	Oct 2015
Gramene	http://www.gramene.org/	Tello-Ruiz *et al.*, 2015 [13]	Nov 2015
Plant GDB	http://www.plantgdb.org/	Duvick *et al.*, 2008 [14]	Dec 2011
PlantsDB	http://pgsb.helmholtz-muenchen.de/	Spannagl *et al.*, 2015 [15]	2015
Phytozome	http://phytozome.jgi.doe.gov/	Goodstein *et al.*, 2012 [16]	Nov 2015
PLAZA	http://bioinformatics.psb.ugent.be/plaza/	Proost *et al.*, 2015 [17]	2015
GreenPhylDB	http://www.greenphyl.org/	Rouard *et al.*, 2010 [18]	2015
PlantOrDB	http://bioinfolab.miamioh.edu/plantordb/	Li *et al.*, 2015 [19]	2015
SALAD	http://salad.dna.affrc.go.jp/	Mihara *et al.*, 2010 [20]	2012
PlantTribes	http://fgp.bio.psu.edu/tribedb/	Wall *et al.*, 2008 [21]	2007
PlantGenIE.org	http://plantgenie.org/	Sundell *et al.*, 2015 [22]	2015
POGs2	http://pogs.uoregon.edu/	Tomcal *et al.*, 2013 [23]	2014
GenomicusPlants	http://www.genomicus.biologie.ens.fr/	Louis *et al.*, 2015 [24]	2014
PIECE	http://wheat.pw.usda.gov/piece/	Wang *et al.*, 2013 [25]	2012
PlantSEED	http://bioseed.mcs.anl.gov/	Seaver *et al.*, 2014 [26]	2014
PGDBj	http://pgdbj.jp/	Asamizu *et al.*, 2014 [27]	2015

* Data available at November 30, 2015.

2.2. NCBI Genomes [8] is a resource of the NCBI (National Center for Biotechnology Information) which organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations. The NCBI Genome database collects genomic sequencing projects for a given species and provides links to corresponding records in BioProject, Assembly, Nucleotide and Protein databases. Besides, a list of the current status of all genomes annotated at NCBI is provided. Currently, more than 2,200 eukaryotic genomes have been deposited, and from that, more than 200 corresponds to plant species. Recently, NCBI has launched a new database, named Assembly [9]. The database provides a versioned Assembly accession number that tracks changes to assemblies as they are updated by submitting groups over time. The Assembly database reports metadata such as assembly names, simple statistical reports of the assembly as well as the assembly update history. Links in the Assembly resource allow users to easily download sequence and annotations for current versions of genome assemblies from the NCBI genomes FTP site.

2.3. CoGepedia is the official wiki page of CoGe, a platform for performing comparative genomics research which provides a network of interconnected tools to manage, analyze, and visualize next generation sequenced data. Their interest to plant genomics is the existence of a link in this page with much updated information on sequenced plant genomes. This information is structured in a phylogenetic tree of around 100 species sequenced. Individual information on each species with links to the genome publication or the genome project main page is provided.

2.4. plaBi is a plant genomic database with a very updated tool to know which plant species have been sequenced. This information can be visualized in a timeline perspective or a phylogenetic perspective. Links to the research articles where each plant genome has been published are also included.

3. PLANT GENOMIC DATABASES

From the first years of genomic sequencing obtained DNA data were compiled in databases. Formerly, species

specific databases were developed. The Arabidopsis Initiative Resource, TAIR (http://www.arabidopsis.org/) was the first database created to harbor the genomic sequence of a plant and to surround it with different tools to explore this sequence. These tools have been increasing with the subsequent releases of the genomic sequence and cover different aspects related not only with the Arabidopsis genome, but with some other features such as germplasm, teaching, laboratory protocols or gene expression experiments. With the increasing of genome projects regarding various related species, databases from specific clades arises. One example is the SGN database for Solanaceae (http://solgenomics.net/). This database comprises the reference genomes of tomato, potato, pepper and *Solanum pennelli*; the draft genomes of *Solanum pimpinellifolium*, *Nicotiana benthamiana* and *Nicotiana tabaccum*; and the genomes of two inbred lines of *Solanum lycopersicum*. As for individual databases, a set of tools have been implemented to analyze the genomic sequences.

The individual/clade databases hosting the approximately 140 plant species sequenced to date can be easily found by writing the word genome and the species/clade in any internet search engine. At the same time, secondary databases have arisen to deal with features regarding comparative plant genomics research. These databases harbor different plant genomes and differ in both, the plant species they host, and the tools and resources they have.

The aim of these projects is to explore how genomics is evolving from largely descriptive to highly predictive driven by quantitative measurements, with algorithms and computation as the domain-adapted language. Several tools are common to all databases. For example, sequence similarity searches are the most reliable strategy to identify homologous proteins or genes by detecting statistically significant similarity [10]. All databases include in their repertoire of bioinformatics tools a search engine to explore the annotated genome sequences, typically using the BLAST algorithm. Likewise, genome browsers, such are the commonly used GBrowse or JBrowse, are key tools offered by all databases to provide a graphical interface for users to browse, search, retrieve and analyze genomic sequence and annotation data [11]. Because of their importance in modern research, several of the comparative genomic databases are described below (Table **1**).

3.1. Ensembl Plants [12] is an integrative resource presenting genome-scale information for 39 sequenced plant species, including 12 eudicots, 21 monocots, one moss, one pseudofern, two green algae, one red algae and the genome of *Amborella trichopoda*, which belongs to a sister clade of the other angiosperm species. Data provided includes genome sequence, gene models, functional annotation, and polymorphic loci. Comparative analyses can be performed on whole genome with available genome alignments. Gene families, based in an all-versus-all BLASTP alignment, are provided. Gene trees showing the evolutionary history of each gene family are available under the "Plant Compara" section. Access to the data is provided through a genome browser with tracks displaying genome sequence and assembly information, additional gene model and variation datasets, and precomputed sequence alignments including ESTs, RNASeq experiments, repeat features, oligo-probe, and marker sets. Ensembl Plants is updated 4–5 times a year and is developed in collaboration with the Gramene database and the transPLANT project (http://www.transplantdb.eu) that aims to facilitate the exchange and integration of plant genome data from distributed resources as well as the development of common standards and protocols.

3.2. Gramene [13] is a curated online resource for comparative functional genomics in crops and model plant species, currently hosting 45 sequenced reference genomes in its build number 48. Since 2009 Gramene has partnered with the Plants division of Ensembl Genomes (http://www.plants.ensemb.orgl/) to jointly produce the genome browser described above, which takes advantage of the Ensembl infrastructure and provides an interface for exploration of genome features, functional ontologies, variation data and comparative phylogenomics. Gramene includes a wide array of potentially useful tools such as genetic and physical maps with genes, ESTs and QTLs locations, genetic diversity data sets, structure-function analysis of proteins, plant pathways databases (BioCyc and Plant Reactome platforms), and descriptions of phenotypic traits and mutations.

3.3. Plant GDB [14] is a web resource devoted to develop robust genome annotation methods, tools, and standard training sets for plant genomes. From 2012 Plant GDB provides access to sequence data from 29 plant species. Plant GDB also provides annotated transcript assemblies for more than 250 plant species, with transcripts mapped to their cognate genomic context where available, integrated with a variety of sequence analysis tools and web services. Plant GDB hosts a plant genomics research outreach portal that facilitates access to a large number of resources for research and training. From July 2015 the Plant GDB's funding has ended and the website is no longer being updated.

3.4. Plants DB [15] is a database that has been developed by the plant genomics group of the PGSB (Plant Genome and Systems Biology, formerly MIPS). Plants DB aims to provide a data and information resource for individual plant species, especially complex triticeae genomes, and currently hosts 13 monocot and dicot species. Sequence similarity searches are possible against databases from 18 different species. In addition, PlantsDB provides a platform for integrative and comparative plant genome research. The database framework integrates genome data from both model and crop plants and facilitates knowledge transfer between them using state-of-the-art comparative genomics tools such as CrowsNest, created to visualize and investigate syntenic relationships between monocot genomes, and the Genome Zipper concept for an ordered gene annotation in cereals. Plants DB is part of the transPLANT network.

3.5. Phytozome [16] is the plant comparative genomics portal of the Department of Energy's Joint Genome Institute (DOE JGI). In the current release v10.3, Phytozome provides access to sixty-one sequenced and annotated green plant genomes, forty-seven of which have been clustered into gene families at 12 evolutionarily significant nodes. It harbors the individual species genomic sequences compiled in the DOE JGI, and links to the individual pages of the other genomes it hosts. Families of related genes representing the modern descendants of ancestral genes are available. They have been

constructed from an all-versus-all BLASTP alignment used to compute the evolutionary distance between each two proteins, the identification of orthologs via reciprocal best hit or synteny analysis, and the accretion of paralogs using outgroup scores. These families allow easy access to clade-specific orthology/paralogy relationships as well as insights into clade-specific novelties and expansions. Each gene has been annotated with available PFAM, KOG, KEGG, PANTHER and GO assignments, and its evolutionary history at the level of sequence, gene structure, gene family and genome organization is provided. Besides, Phytozome provides access to the plant genomes it hosts using the JBrowse genome browsers available for all genomes.

3.6. PLAZA [17] is a platform designed to make comparative genomics in plants and developed in the University of Ghent. The current version PLAZA 3.0 hosts 37 plant species covering a broad taxonomic range and includes 25 eudicot, 8 monocot, one moss and two algae species, as well as the genome of *Amborella trichopoda*. PLAZA provides detailed structural and functional annotation of the genes, which has been expanded in the new version and now comprises data from Gene Ontology, MapMan, UniProtKB/Swiss-Prot, PlnTFDB and PlantTFDB. From the more than one million genes annotated in the genomes it harbors, gene families and subfamilies have been delineated. First, by computing the protein sequence similarity through an all-against-all BLAST, and then by applying graph-based clustering methods implemented in TribeMCL and OrthoMCL. Other relevant data available in this database consist of phylogenetic trees to identify biologically relevant duplication and speciation events, and detailed information about genome organization to unveil small and large genome duplication events. Furthermore, this database provides tools to transfer functional annotation from well-characterized plant genomes to other plant species.

3.7. GreenPhylDB [18] is a comparative genomics database jointly developed by Biodiversity International and the International Cooperation Center for Agricultural Research for Development (CIRAD). The current version GreenPhylDB 4.0 hosts 37 species of the *plantae* kingdom including one red *alga*, two green *algae*, one moss, one lycophyte, one conifer, the ancestral angiosperm *Amborella*, ten monocots and 20 eudicot species. This database also groups annotated sequences into gene families. The clustering method used is TribeMCL. This software uses different pairwise similarity matrices obtained by running protein-protein BLAST using increasing stringent thresholds. Then, these matrices are used by a Markov cluster algorithm to group proteins in families at different levels of clustering (1 to 4). Results of the automatic clustering are manually annotated using cross reference databases and analyzed by a phylogenetic-based approach to predict homologous relationships. Currently, GreenPhylDB contains 8,347 clusters with more than 5 sequences at level 1, from which 2,939 are annotated and 4,788 have available phylogenetic trees. For each gene cluster, this database offers an easy access to the gene composition by species, and provides protein domains, orthologous gene predictions and relevant external links.

3.8. PlantOrDB [19] is a genome-wide database created to classify genes in families and to find orthologous genes clusters. The recently launched first version of PlantOrDB hosts 41 species including six green algae, one moss, one lycophyte, six monocots and 27 eudicot species. Gene families are available, which have been generated from an all-against-all BLAST search approach. The web interfaces provided by PlantOrDB display information on the evolutionary features of an individual gene and its homolog gene family, which can be deduced from multiple sequence alignments and phylogenetics trees. Tools to provide an accurate classification of a query sequence within a phylogenetic tree and a multiple sequence alignment are also available.

3.9. SALAD [20] is a comparative genomics database constructed from plant-genome-based proteome data sets. The version 3.0, appeared in 2009, hosts 10 species including a yeast species. The most important singularity of this database is the construction of dendrograms for protein families using the information derived from conserved motifs discovered using the MEME software. Protein clusters are determined from BLASTP searches and, then, MEME motifs are used to select the proteins displayed in the dendrograms. A viewer to see microarray data sets of paralogous genes in the dendrograms is also provided. The website has not been updated since 2012.

3.10. PlantTribes [21] is a gene family resource for comparative genomics in plants that launched its current 2.0 version in 2007. It harbors the proteome of ten plant species, from green algae to angiosperms, from which putative gene families were created by a BLASTP approach and the use of the MCL algorithm for clustering. The database is currently active and permits identify groups of related genes and their expression patterns, using the microarray information existing at the date the current version was released.

3.11. PlantGenIE.org [22] is a collection of web resources for searching, visualizing and analyzing genomics and transcriptomics data recently developed for different plant species. Currently, it includes dedicated web portals for enabling in-depth exploration of poplar, Norway spruce, and Arabidopsis genomes. Standard features, including genome browsers, gene list annotation, BLAST tools and gene information pages are provided. The aim of this database is to continue and develop the resource by inclusion of additional species, maintaining a focus on woody species.

3.12. POGs2 [23] is a relational database designed to facilitate cross-species inferences about gene functions and gene models in plants. In its current 2.0 version, the database integrates data from rice, maize, *Arabidopsis thaliana*, and poplar by placing the complete predicted proteomes into "putative orthologous groups" (POGs). POGs were imported from Gramene's ENSEMBL orthology prediction output and from Plaza 2.5. Each POG entry includes putative orthologs annotated with gene descriptions imported from species-specific databases, graphical representations of conserved protein domains and phylogenetic trees showing closely-related proteins.

3.13. Genomicus Plants [24] is a database that enables users to explore flowering plants genomes. Extant genomes of 16 eudicot and 10 monocot plants can be analyzed, as well as 23 ancestral reconstructed genomes. In the current 16.03 version, all the data on extant species come from the annota-

tions of selected genomes available in Ensembl Plants and Phytozome. Gene families are accessible, based in that existing in Ensembl Plants. A distinctive feature of this database is that the displayed genomic context of a gene is showed in parallel to the genomic context of all its orthologous and paralogous copies evolutionarily ordered in a phylogenetic tree.

3.14. Piece [25] is a comparative genomics database focused in the comparison of exon-intron plant gene structures and their evolutionary and functional relationships. In the first and current version, annotated genes were extracted from 25 species, from green algae to angiosperms, and classified based on Pfam motifs. Phylogenetics trees are available for each gene family, which integrate exon-intron and protein motif information. Both the sequences and gene structure information for each identified gene are also available.

3.15. PlantSEED [26] is a database created to support subsystems-based annotation and metabolic model reconstruction for plant genomes. Annotations of protein families from Ensembl Plants for five eudicot and five monocot species have been used to construct subsystems, which are the integration of metabolic pathways or other biological processes with genome annotations. The manual curation of the subsystems permits a rapid reconstruction and modelling of primary metabolism for all plant genomes in the database.

3.16. PGDBj [27] is the Plant Genome Database Japan, a portal website that aims to integrate plant genome-related information from databases. The main tool developed in this web is the Ortholog DB, which comprises clusters of homologous sequences. Clusters were obtained from reciprocal BLAST searches using the proteome of 20 plant species from the main clades of the Viridiplantae kingdom. PGDBj also provides DNA marker and QTL information of important agronomic traits for many plant species.

4. EMERGING HIGH-PERFORMANCE COMPUTING WEB RESOURCES

The new high-throughput genomic methods implies the generation of massive data. Computation capability is a major challenge for the handling of these data. The accessibility to computational tools easy to manage for a scientist without programming or informatics expertise is a solution to address this problem. In recent years, several web-based platforms have been created to store next-generation sequencing data and to host different applications to work with these data to extract biological relevant information. One of the first platforms developed is Galaxy [28] (https://usegalaxy.org/). Galaxy is an open web-based platform for genomic research with many tools to work with computation-reliant results. Recently, an open-source project specific for plants, the iPlant, has been launched.

4.1. The iPlant Collaborative [29] (https://de.iplantcollaborative.org/) is a United States National Science Foundation (NSF) funded project created as an innovative, comprehensive, and foundational cyberinfrastructure in support of plant biology research. iPlant includes the use of high-performance computing, use of large shared data storage, and the establishment of collaborations and virtual organizations around shared analysis tools and analyzed data. iPlant hosts bioinformatics tools for most of the modern research challenges, such as data importers, sequence alignments and phylogenetic tree building, phylogenetic and evolutionary analyses, QTL mapping and genome-wide association studies, ultrahigh-throughput sequence processing, functional analyses, clustering and network analyses, variant detection and annotation, RNAseq analyses and ChIP-seq studies. In particular, to manage and serve published plant genome data, iPlant has developed a workspace named DNA Subway. DNA subway has been conceived to be a space for gene annotation and genome analysis that currently hosts tools to predict and annotate genes, prospect entire plant genomes for related genes and sequences, determine sequence relationships and analyze RNA-Seq reads to measure differential expression. To prospect genomes, it harbors the TARGeT program that permits a search for homologous sequences in 18 plant genomes, covering different clades from algae to angiosperms, and builds multiple sequence alignments and phylogenetic trees with the obtained sequences.

5. CONCLUDING REMARKS

Bioinformatic programs and algorithms have arisen as key tools for modern research. In plants, the number and relevance of these tools have been increased in recent years, mainly those related to genomics aspects. New science implies the existence of databases where the genomes are hosted and exhaustive analysis can be performed. Although these databases should be continuously updated, the assembly and annotation of many of the draft genomic sequences have not been improved from its first release. This is an actual challenge, since it is easy now to produce a new draft sequence but more complicated to improve its quality. For example, the pre-release of the version number 11 for the Arabidopsis genome has appeared in Araport (https://www.araport.org/) in October 2015 with genomic sequences assembled in its five chromosomes. On the contrary, the first release of the genome of the lycophyte *Selaginella moellendorffii* appeared in 2007 assembled into 768 scaffolds and has not been subsequently updated. As a next step in plant genomics, new databases have been developed to deal with the issue of comparing the increasing number of sequenced genomes. Most comparative genomic databases are periodically updated and have become powerful tools to extract shared information coming from evolutionarily related plant species. However, the extremely high new information obtained in modern laboratories using computational approaches makes necessary the existence of new high-performance computational developments that uses the computational power of the internet and the storage of resources in the cloud. As a new challenge, actual projects in Arabidopsis and rice are directed to the comparison of thousands of genomes obtaining from different genotypes of the same species. Innovative solutions for storage and visualization, such as the Rice SNP-Seek Database (http://oryzasnp.org/iric-portal/), have to be developed to use efficiently the massive new information obtained.

NOTE ADDED IN PROOF

As an example of the continuous evolution of genomic databases, the iPlant Collaborative resource has evolved into

the CyVerse cyberinfrastructure, which provides tools not only for plant research (http://www.cyverse.org/).

CONFLICT OF INTEREST

The author confirms that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by the Ministerio de Economía y Competitividad of Spain (project BIO2014-53508-R).

REFERENCES

[1] Feuillet, C.; Leach, J.E.; Rogers, J.; Schnable, P.S.; Eversole, K. Crop genome sequencing: lessons and rationales. *Trends Plant Sci*., **2011**, *16* (2), 77-88.

[2] Hamilton, J.P.; Buell, C.R. Advances in plant genome sequencing. *Plant J.*, **2012**, *70* (1), 177-90.

[3] Hirsch, C.N.; Buell, C.R. Tapping the promise of genomics in species with complex, nonmodel genomes. *Annu. Rev. Plant Biol.*, **2013**, *64*, 89-110.

[4] Michael, T.P.; VanBuren, R. Progress, challenges and the future of crop genomes. *Curr. Opin. Plant Biol.*, **2015**, *24*, 71-81.

[5] Martinez, M. From plant genomes to protein families: computational tools. *Comp. Struct. Biotechnol. J.*, **2013**, *8*, e201307001.

[6] Dhanapal, A.P.; Govindaraj, M. Unlimited thirst for genome sequencing, data interpretation, and database usage in genomic era: the road towards fast-track crop plant improvement. *Genet. Res. Int.*, **2015**, *2015*, 684321.

[7] Reddy, T.B.; Thomas, A.D.; Stamatis, D.; Bertsch, J.; Isbandi, M.; Jansson, J.; Mallajosyula, J.; Pagani, I.; Lobos, E.A.; Kyrpides, N.C. The Genomes On Line Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Res.*, **2015**, *43* (Database issue), D1099-106.

[8] NCBI Resource Coordinators, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **2015**, *43* (Database issue), D6-17.

[9] Kitts, P.A.; Church, D.M.; Thibaud-Nissen, F.; Choi, J.; Hem, V.; Sapojnikov, V.; Smith, R.G.; Tatusova, T.; Xiang, C.; Zherikov, A.; DiCuccio, M.; Murphy, T.D.; Pruitt, K.D.; Kimchi, A. Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.*, **2016**, *44* (Database issue), D73-80.

[10] Pearson, W.R. An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinformatics*, **2013**, Chapter 3, Unit 3.1.

[11] Wang, J.; Kong, L.; Gao, G.; Luo, J. A brief introduction to web-based genome browsers. *Brief Bioinformatics*, **2013**, *14* (2), 131-43.

[12] Bolser, D.; Staines, D.M.; Pritchard, E.; Kersey, P. Ensembl Plants: integrating tools for visualizing, mining, and analyzing plant genomics data. *Methods Mol. Biol*., **2016**, *1374*, 115-40.

[13] Monaco, M.K.; Stein, J.; Naithani, S.; Wei, S.; Dharmawardhana, P.; Kumari, S.; Amarasinghe, V.; Youens-Clark, K.; Thomason, J.; Preece, J.; Pasternak, S.; Olson, A.; Jiao, Y.; Lu, Z.; Bolser, D.; Kerhornou, A.; Staines, D.; Walts, B.; Wu, G.; D'Eustachio, P.; Haw, R.; Croft, D.; Kersey, P.J.; Stein, L.; Jaiswal, P.; Ware, D. Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res.*, **2014**, *42* (Database issue), D1193-9.

[14] Duvick, J.; Fu, A.; Muppirala, U.; Sabharwal, M.; Wilkerson, M. D.; Lawrence, C.J.; Lushbough, C.; Brendel, V. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.*, **2008**, *36* (Database issue), D959-65.

[15] Spannagl, M.; Nussbaumer, T.; Bader, K.C.; Martis, M.M.; Seidel, M.; Kugler, K.G.; Gundlach, H.; Mayer, K.F. PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic Acids Res.*, **2016**, *44* (Database issue), D1141-7.

[16] Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; Rokhsar, D.S. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **2012**, *40* (Database issue), D1178-86.

[17] Proost, S.; Van Bel, M.; Vaneechoutte, D.; Van de Peer, Y.; Inzé, D.; Mueller-Roeber, B.; Vandepoele, K. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res.*, **2015**, *43* (Database issue), D974-81.

[18] Rouard, M.; Guignon, V.; Aluome, C.; Laporte, M.A.; Droc, G.; Walde, C.; Zmasek, C.M.; Périn, C.; Conte, M.G. GreenPhylDB v2.0: comparative and functional genomics in plants. *Nucleic Acids Res.*, **2011**, *39* (Database issue), D1095-102.

[19] Li, L.; Ji, G.; Ye, C.; Shu, C.; Zhang, J.; Liang, C. PlantOrDB: a genome-wide ortholog database for land plants and green algae. *BMC Plant Biol.*, **2015**, *15*, 161.

[20] Mihara, M.; Itoh, T.; Izawa, T. SALAD database: a motif-based database of protein annotations for plant comparative genomics. *Nucleic Acids Res.*, **2010**, *38* (Database issue), D835-42.

[21] Wall, P.K.; Leebens-Mack, J.; Müller, K.F.; Field, D.; Altman, N.S.; dePamphilis, C.W. PlantTribes: a gene and gene family resource for comparative genomics in plants. *Nucleic Acids Res.*, **2008**, *36* (Database issue), D970-6.

[22] Sundell, D.; Mannapperuma, C.; Netotea, S.; Delhomme, N.; Lin, Y.C.; Sjödin, A.; Van dePeer, Y.; Jansson, S.; Hvidsten, T.R.; Street, N.R. The Plant Genome Integrative Explorer Resource: PlantGenIE.org. *New Phytol.*, **2015**, *208*(4), 1149-56.

[23] Tomcal, M.; Stiffler, N.; Barkan, A. POGs2: a web portal to facilitate cross-species inferences about protein architecture and function in plants. *PLoS One*, **2013**, *8*(12), e82569.

[24] Louis, A.; Murat, F.; Salse, J.; Crollius, H.R. Genomicus Plants: a web resource to study genome evolution in flowering plants. *Plant Cell Physiol.*, **2015**, *56*(1), e4.

[25] Wang, Y.; You, F.M.; Lazo, G.R.; Luo, M.C.; Thilmony, R.; Gordon, S.; Kianian, S.F.; Gu, Y.Q. PIECE: a database for plant gene structure comparison and evolution. *Nucleic Acids Res.*, **2013**, *41*(Database issue), D1159-66.

[26] Seaver, S.M.; Gerdes, S.; Frelin, O.; Lerma-Ortiz, C.; Bradbury, L.M.; Zallot, R.; Hasnain, G.; Niehaus, T.D.; El Yacoubi, B.; Pasternak, S.; Olson, R.; Pusch, G.; Overbeek, R.; Stevens, R.; de Crécy-Lagard, V.; Ware, D.; Hanson, A.D.; Henry, C.S. High-throughput comparison, functional annotation, and metabolic modeling of plant genomes using the PlantSEED resource. *Proc. Natl. Acad. Sci. U.S.A.*, **2014**, *111*(26), 9645-50.

[27] Asamizu, E.; Ichihara, H.; Nakaya, A.; Nakamura, Y.; Hirakawa, H.; Ishii, T.; Tamura, T.; Fukami-Kobayashi, K.; Nakajima, Y.; Tabata, S. Plant Genome DataBase Japan (PGDBj): a portal website for the integration of plant genome-related databases. *Plant Cell Physiol.*, **2014**, *55*(1), e8.

[28] Goecks, J.; Nekrutenko, A.; Taylor, J.; Team, G. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **2010**, *11* (8), R86.

[29] Goff, S.A.; Vaughn, M.; McKay, S.; Lyons, E.; Stapleton, A.E.; Gessler, D.; Matasci, N.; Wang, L.; Hanlon, M.; Lenards, A.; Muir, A.; Merchant, N.; Lowry, S.; Mock, S.; Helmke, M.; Kubach, A.; Narro, M.; Hopkins, N.; Micklos, D.; Hilgert, U.; Gonzales, M.; Jordan, C.; Skidmore, E.; Dooley, R.; Cazes, J.; McLay, R.; Lu, Z.; Pasternak, S.; Koesterke, L.; Piel, W.H.; Grene, R.; Noutsos, C.; Gendler, K.; Feng, X.; Tang, C.; Lent, M.; Kim, S.J.; Kvilekval, K.; Manjunath, B.S.; Tannen, V.; Stamatakis, A.; Sanderson, M.; Welch, S.M.; Cranston, K.A.; Soltis, P.; Soltis, D.; O'Meara, B.; Ane, C.; Brutnell, T.; Kleibenstein, D.J.; White, J.W.; Leebens-Mack, J.; Donoghue, M.J.; Spalding, E.P.; Vision, T.J.; Myers, C.R.; Lowenthal, D.; Enquist, B.J.; Boyle, B.; Akoglu, A.; Andrews, G.; Ram, S.; Ware, D.; Stein, L.; Stanzione, D. The iPlant Collaborative: cyber infrastructure for plant biology. *Front. Plant Sci.*, **2011**, *2*, 34.