Genomics and Bioinformatics Resources

6

P. Rajendrakumar and Sujay Rakshit

Contents

P. Rajendrakumar (✉)
Biotechnology, ICAR-Indian Institute of Millets Research, Rajendranagar, Hyderabad, Telangana 500 030, India
e-mail: rajendra@millets.res.in

S. Rakshit
Plant Breeding, ICAR-Indian Institute of Millets Research, Rajendranagar, Hyderabad, Telangana 500 030, India
e-mail: sujay@millets.res.in

Abstract

Rapid generation of genome and transcriptome data from various research projects across the globe resulted in the accumulation of enormous amounts of data of various crop species or groups of crops. These data are organized and stored in different databases, which offers user-friendly search and retrieval of the desired information for further analyses and use. The information in these databases is used to develop DNA-based markers such as SSRs and SNPs, which are the most popular genomics resources applied for QTL mapping and marker-assisted selection. Several bioinformatics resources such as algorithms, stand-alone software, as well as web-based tools are developed by several research groups and made available in the public domain or sold commercially for the rapid and systematic analysis of DNA sequence or gene expression data. Genetic resources such as biparental, multi-parental, natural, as well as mutant populations for various target traits were developed by researchers, which are utilized for the mapping of QTLs as well as identification of candidate genes associated with the traits of interest by the application of genomics tools developed using bioinformatics resources in these mapping populations. This review discusses the most relevant databases useful for sorghum, development of genomics resources such as DNA markers using various bioinformatics tools, and the genetic resources available in sorghum.

R. Madhusudhana et al. (eds.), *Sorghum Molecular Breeding*, DOI 10.1007/978-81-322-2422-8_6

Keywords

Genetic resources • Genomics • Databases • Transcriptomics • Mutant resources

6.1 Introduction

In the current era of genomics and next-generation sequencing, a plethora of information is generated on genome and gene sequences, which is of paramount importance to understand gene function and the regulatory networks involved in plant growth, development, as well as stress tolerance at the molecular level. Plant genomics took a giant leap with the sequencing of the whole genome of *Arabidopsis thaliana* in 2000 (The Arabidopsis Genome Initiative 2000). This was followed by genome sequencing of important food crops like rice (Yu et al. 2002; Goff et al. 2002), sorghum (Paterson et al. 2009), maize (Schnable et al. 2009), barley (The International Barley Genome Sequencing Consortium 2012), pigeon pea (Varshney et al. 2012), chickpea (Varshney et al. 2013), and others. Rapid progress in plant genomics led to the discovery and isolation of important genes that regulate economically important traits and tolerance to biotic and abiotic stresses. With the availability of genome sequences in sorghum, the biggest challenge is to determine the functions of over 30,000 genes and deploy them in a practical genetic improvement program.

Being a C_4 cereal and a drought-tolerant crop, sorghum always remained in the focus of genomics researchers. With the publication of sorghum genome sequence (Paterson et al. 2009), sorghum genomics has taken a paradigm shift and generated enormous information that has been integrated with other related crop species through comparative genomic studies. The completion of sequencing of sorghum genome and the creation of related genomics resources together with advances in the development of mapping populations and molecular marker resources have allowed researchers to accelerate the identification of agronomically important quantitative trait loci (QTLs) (Satish et al. 2009; Aruna et al. 2011; Mace et al. 2012; Nagaraja Reddy et al. 2013; Madhusudhana and Patil 2013). Advances in sequencing technologies such as next-generation sequencing (NGS) have resulted in sequence-based resources and related resource platforms for specific organisms including sorghum. Availability of whole-transcriptome profiling methods, advances in plant proteomics, simultaneous profiling of many metabolites, mutant populations, and biological databases have brought significant change in the approach in dealing with the biological processes. This chapter deals with the available genomics and bioinformatics resources in sorghum, which will help the sorghum researchers to develop useful information for the genetic improvement of sorghum.

6.2 Sequence Databases and Comparative Genomics Resources

Sorghum genome is comparatively small (~730 Mb), making it an attractive model for functional genomics of Saccharinae and other C_4 grasses. It is the first C_4 plant to be sequenced in 2009 by Andrew Paterson and his group involving 20 laboratories across the USA, Germany, China, Switzerland, and India. They observed that sorghum has ~75 % larger heterochromatin DNA as compared to rice. However, sorghum and rice have similar quantities of euchromatin. The net size expansion of the sorghum genome relative to rice predominantly involved long terminal repeat (LTR) retrotransposons. It was found that the sorghum genome contains 55 % retrotransposons, which is intermediate between rice (26 %) and maize (79 %) genomes. Paterson et al. (2009) modeled 34,496 sorghum genes, out of which ~27,640 were bona fide protein-coding genes. The accumulated information/data related to genome, transcriptome, proteome, and metabolome as a result of various high-throughput sequencing projects and proteomic and metabolomic studies are stored in different databases, which are summarized in Table 6.1. The comparative genomic databases like Gramene, PlantGDB, Phytozome, GreenPhylDB, CoGE,

Table 6.1 Database resources for sorghum

Group	Resources	Databases
Genome	Genome sequence, gene annotation	PlantGDB, Phytozome, CoGE, PLAZA
	Molecular markers, DNA variation, quantitative trait locus	Gramene, Phytozome, PIP database, NCBI dbSNP, CSGRqtl, SorGSD
	Genome re-sequencing	$(GIGA)^n$ DB
	Focused gene family database	GRASIUS, Phytozome
Transcriptome	Full-length cDNAs, ESTs	PlantGDB, Phytozome, NCBI dbEST
	Non-coding RNA	NRDR
	microRNA	PMRD, miRBase
Proteome	Proteome/modificome profile	GreenPhylDB, Phytozome
	Sub-cellular localization	Gramene, Phytozome, PlantGDB
Metabolome	Metabolic map	SorghumCyc

Modified from Mochida and Shinozaki (2010)

PLAZA, OrthologID, PlantTribes, SynBrowse, etc., along with associated web portals provide a uniform set of tools and automated analyses across a wider range of plant genomes. Among them, the first six resources deal with sorghum sequences along with other plant species. The salient features of these databases and their utility in comparative genomics are discussed below.

6.2.1 Gramene

Gramene (http://www.gramene.org/) is a curated, open-source, data resource for comparative genome analysis in the grasses developed in recognition of the importance of the grass family and put on public domain in 2002 from Cold Spring Harbor Laboratory (Ware et al. 2002). Initially, the rice sequence was used as base information to facilitate genomics research in other grass families like maize, sorghum, millet, sugarcane, wheat, oats, and barley. Subsequently, Ensembl Genomes at the European Bioinformatics Institute joined the group at the Cold Spring Harbor Laboratory, to further make this database as a resource for plant comparative genomics based on Ensembl technology. Besides rice, the database has information on barley, *Brachypodium*, foxtail millet, maize, oats, pearl millet, rye, wheat, and sorghum. In the recent version, information from other plant species like *Glycine*, *Musa*, *Solanum*, *Brassica*, *Arabidopsis*, *Vitis*, *Populus*, etc., have also been included. The goal of this database is to facilitate the study of cross-species homology relationships using information derived from public projects involved in genomic and EST sequencing, protein structure and function analysis, genetic and physical mapping, interpretation of biochemical pathways, gene and QTL localization, and descriptions of phenotypic characters and mutations. Even though a new version has been launched very recently, the information can be accessed through the old version also, which is organized in a better way. Information in the database organized in different modules (Fig. 6.1) are detailed as follows:

"Genome" Module This contains detailed information about the species, assembly, annotation, structural variation, besides references, and link to other species-related sites on the above-mentioned plant species.

"Genetic Diversity" Module This stores information on genotypes, phenotypes and their environments, germplasm, and association data. It

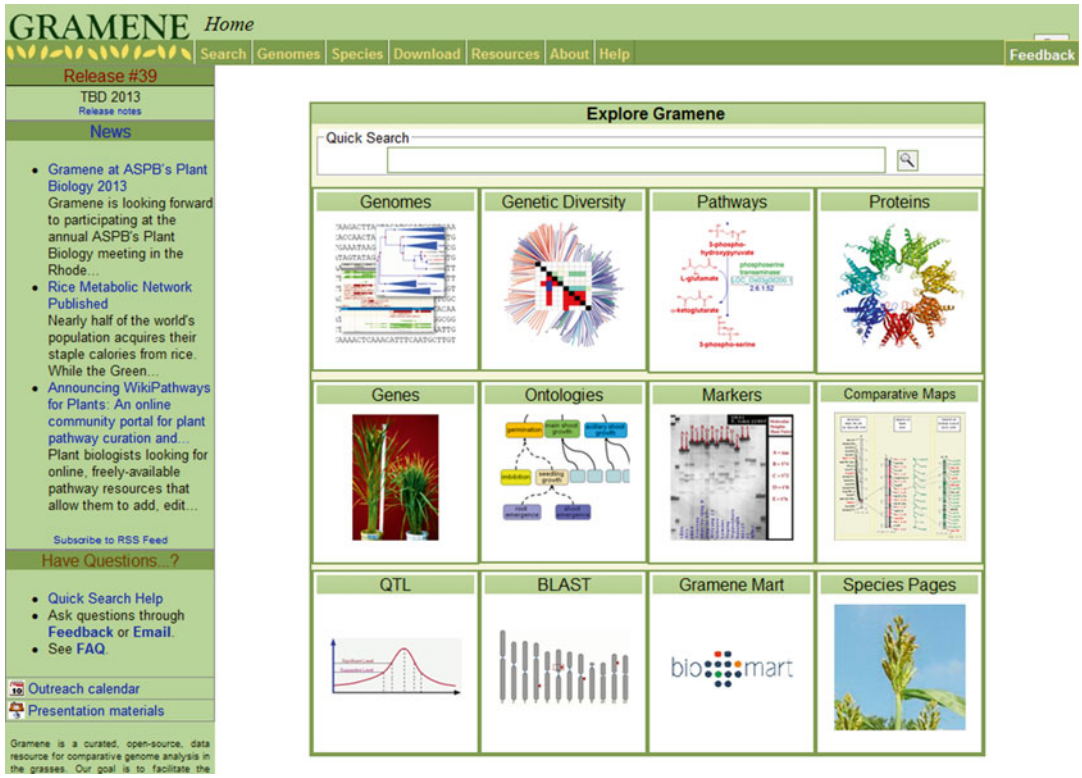

Fig. 6.1 Organization of Gramene database (http://www.gramene.org/)

also contains information from small-scale SSR diversity studies to large-scale SNP/InDel-based genotype-phenotype studies conducted in the mandated crops. With respect to sorghum, the database contains six datasets [Hamblin et al. (2004, 2006, 2007), White et al. (2004) and Casa et al. (2006)] and the results of the Sorghum Diversity Project.

"Pathways" Module This contains four submodules, viz., RiceCyc, MaizeCyc, BrachyCyc, and SorghumCyc, which deals with information on pathway databases of the respective crops. It also provides mirrors of pathway databases from *Arabidopsis*, tomato, potato, pepper, coffee, *Medicago*, *E. coli*, and the MetaCyc and PlantCyc reference databases, thereby enabling comparative genome analysis. In this database, 297 pathways, 1,838 enzymatic reactions, and 9 transport reactions have been described. Known and/or predicted biochemical pathways and genes from sorghum are catalogued in SorghumCyc, which is primarily based on the genome annotations of *Sorghum bicolor* cv. BTx623. Many of the pathways might be incomplete or may contain errors since the functions of many of the sorghum genes are either provided by homology or HMM-based predictions.

"Protein" Module This contains information on Swiss-Prot-TrEMBL protein entries from family Poaceae, which are annotated by the following three concepts of Gene Ontology (GO): (1) molecular function, (2) biological process in which it is involved, and (3) cellular component where it is localized. The associations assigned are based on annotations in the published literatures or generated through in silico approaches. Each association is supported with evidence (reference) and the evidence code (experiment type). On sorghum and related species, information on 35,817 proteins are available.

"Genes" or "Gene and Allele" Module This contains detailed information on publicly available genes in cereal crops. Genes and their alleles associated with morphological, developmental, and agronomically important phenotypes, variants of physiological characters, biochemical functions, and isozymes are described here. Species-wise search for different gene types like "CDS, rRNA, tRNA, miRNA, siRNA, pseudogenes, not classified, sequenced gene loci or all gene types" is possible using wild cards.

"Ontologies" Module This contains a collective information on controlled internationally accepted vocabularies and their associations to various objects such as QTL, phenotype, gene, proteins, and Ensembl rice genes for the following knowledge domains: Plant Ontology (PO), Trait Ontology (TO), Gene Ontology (GO), Environment Ontology (EO), and Gramene's Taxonomy Ontology.

"Markers" Module This contains the basic/primary information on the marker name, synonyms, source species, and a list of map positions of various markers used for mapping. This module has the link to "SSRIT tool," which is useful for the identification of microsatellites.

"Maps" Module This is primarily a visualization tool useful for visualizing the genetic, physical, sequence, and QTL maps for species dealt in the database. Comparative Map Viewer, referred as *CMap*, allows users to construct and compare different maps. All the data including the map sets, maps, features, and correspondences in this module are built from the "Markers" module. In *CMap*, the genomes of rice, sorghum, and *Brachypodium* are compared using syntenic blocks.

"QTL" Module This contains quantitative trait loci (QTL) identified for numerous agronomic traits in the crops dealt in the database. Information on QTL along with associated traits and the mapped locus on the genetic map are available. With respect to sorghum, information on 136 QTL along with details of associated markers, linkage group, trait symbol, etc., are available.

"BLASTView" Module This module provides an integrated platform for homology search against Ensembl plant databases, offering access to both BLAST (Basic Local Alignment Search Tool) and BLAT (BLAST-like Alignment Tool) programs. Species-wise search is possible in both DNA and protein databases using BLASTN (aligns the nucleotide sequences) and BLASTX (aligns translated sequences of any nucleotide sequence in all six reading frames), respectively.

"Gramene Mart" Module This module has four databases, viz., Plant Gene 37, Plant variation 37, Gramene mapping, and Gramene QTL 37. Each database can be searched in 10 datasets of which sorghum is one.

"Species Page" This contains detailed information on all the 11 cereal species dealt in the database with full phylogenetic information.

6.2.2 PlantGDB

PlantGDB was first reported by Dong et al. (2005), in which EST sequences were assembled into contigs that represent tentative unique genes. The functional annotation of these contigs was performed with the information derived from known protein sequences that were highly similar to the putative translation products. Initially, the database started with the data from only two plant species, viz., *Arabidopsis* and rice. Subsequently, PlantGDB (http://www.plantgdb.org) was published as a resource for comparative genomics across 14 plant species by Duvick et al. (2007). The aim of this web resource is to develop robust genome annotation methods, tools, and standard training sets for a number of sequenced or soon to be sequenced plant genomes. PlantGDB has four modules, viz., Sequence module, Genome module, Tools module, and Datasets module. Organization of the PlantGDB is shown in Fig. 6.2.

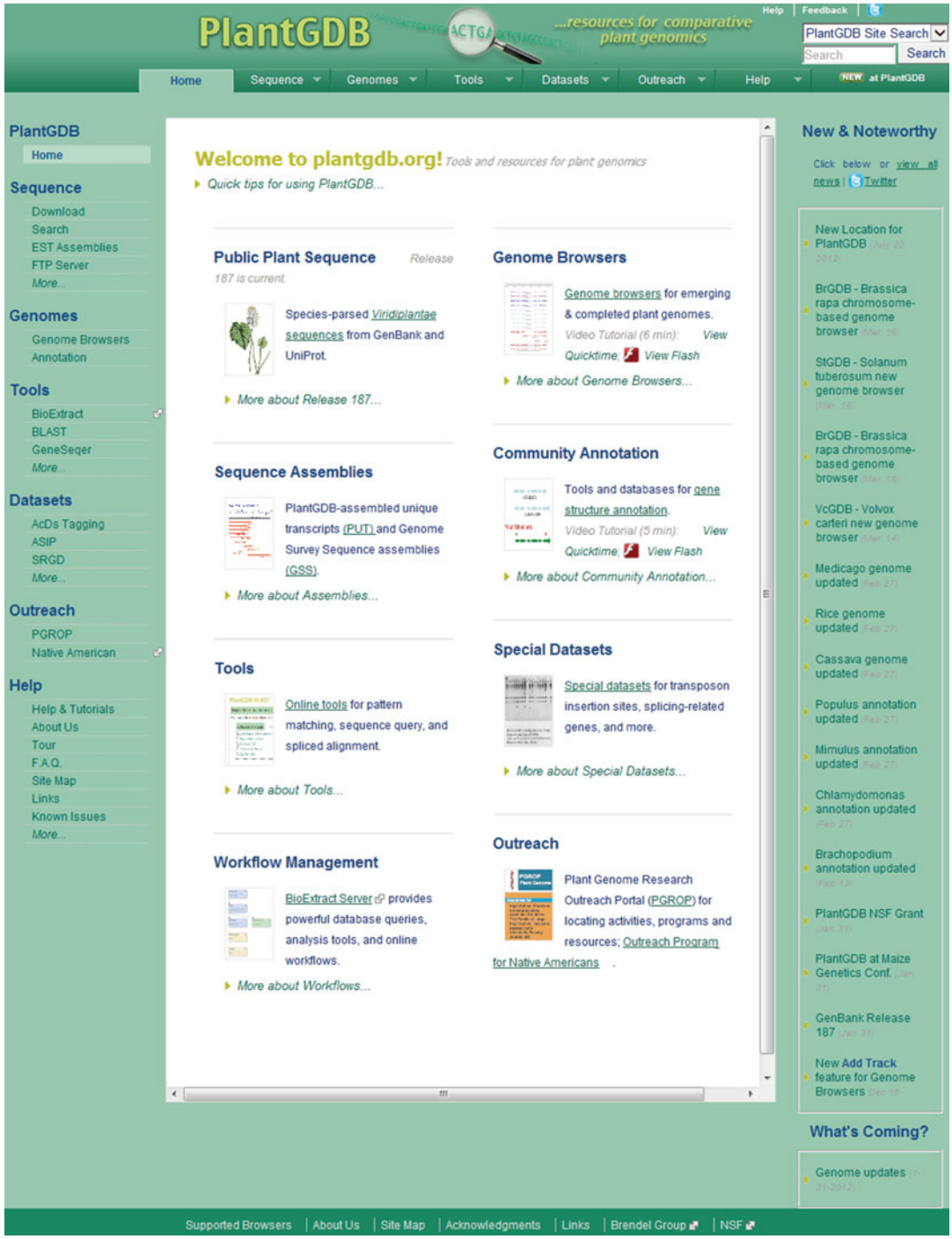

Fig. 6.2 Organization of PlantGDB (http://www.plantgdb.org/)

"Sequence" Module This module can be used to BLAST search or to download nucleotide or protein sequences as well as access custom transcript assemblies. This module contains EST assemblies comprising PlantGDB-derived unique transcripts (PUT) assembled from plant mRNA sequences available at GenBank. Genome survey sequence (GSS) assemblies for maize and sorghum are also available.

"Genome" Module This module contains genome sequence information on 16 dicots and seven monocots including sorghum (SbGDB). It has genome browsers to display current gene structure models and transcript evidence from spliced alignments of EST and cDNA sequences. The browsers also link community annotation tools to refine the gene annotations or to identify novel annotations. Each genome assembly is splice-aligned to transcripts as well as proteins from similar species and presented in a simple graphical interface (the *xGDB platform*).

"Tools" Module This module provides a variety of tools for sequence analysis as follows:

- *BioExtract* – a web interface to automate bioinformatics workflows. It is useful to query sequence databases, analyze data with bioinformatics tools, save results, and create and manage workflows.
- *Standard NCBI BLAST* – useful to search against single or multiple BLAST databases simultaneously.
- *Distributed Annotation System* (*DAS*) – useful to access PlantGDB annotations from the remote genome browsers.
- *GeneSeqer* and *GenomeThreader* – useful to develop gene structure models based on spliced alignment to genomic sequences of both native and homologous ESTs, cDNAs, and protein sequences.
- *MuSeqBox* – useful to examine multi-query sequence BLAST output, filter the BLAST hits based on user-defined criteria, and extract the informative parameters in tabular form.
- *PatternSearch* – useful to search the specific patterns in genome sequence, i.e., short matches interspersed with mismatches and InDels.
- *ProbeMatch* – allows the user to query his sequence against PLEXdb Probe Sequences.
- *TableMaker* – an online search tool to access GenBank tables at PlantGDB using MySQL queries.
- *yrGATE* – useful to create gene annotations in an xGDB genome browser itself. It shows all splice junctions revealed by EST/cDNA evidence and helps to create gene models and validate them.

"Datasets" Module This module has datasets on *AcDs* Tagging Project, Alternative Splicing in Plants (ASIP) database, Plant Expression database (PLEXdb), Rescue-Mu tagged maize sequences, Maize-RFLP Full-Length Insert Sequencing Project, Splicing-Related Gene database (SRGD), and Uniform-Mu tagged maize sequences.

6.2.3 Phytozome

Phytozome (http://phytozome.jgi.doe.gov/pz/) is another online resource that was first released in 2008 to facilitate comparative genomic studies among green plants. It enables users with different computational abilities to access annotated plant gene families, navigate their evolutionary history, examine them in genomic context, assign putative function, and provide uniform access to complete genomes, gene and related sequences and alignments, gene functional information, and gene families, either as bulk information or as the result of user-defined queries (Goodstein et al. 2012). A number of commonly used open-source tools like Lucene, GBrowse (Stein et al. 2002), Jalview (Waterhouse et al. 2009), BioMart (Smedley et al. 2009), mView (Brown et al. 1998), and pygr are integrated in this portal which help in the gene family search, inspection, and evaluation. The Phytozome v7.0 contains data and analyses for 25 plant genomes, 18 of which are sequenced, assembled, and partially or completely annotated at the Joint Genome

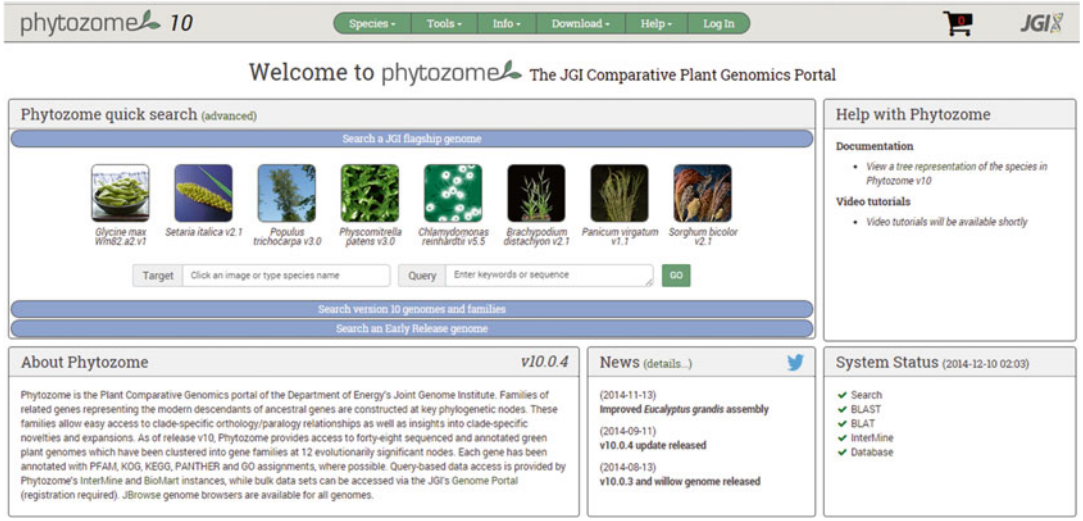

Fig. 6.3 Organization of Phytozome database (http://phytozome.jgi.doe.gov/pz/)

Institute (JGI). However, the recently released Phytozome v10 provides access to 47 sequenced and annotated green plant genomes including early-release genomes. With respect to sorghum, the initial release comprises the Sbi1 assembly and a Sbi1.4 gene set, which are the assembly and annotation reported by Paterson et al. (2009). Now, v2.1 is available as an early release as a result of modern annotation with additional RNA-seq data, comprising v2.0 assembly and v2.1 gene set. The genome is in 10 chromosomes with many short unmapped fragments; some may contain annotated genes (http://www.phytozome.net/sorghum_er.php).

Phytozome contains three important modules, viz., Species, Tools, and Info. "Species" and "Tools" modules have the options for keyword search, BLAST search, BLAT search, JBrowse, and bulk data. In addition to this, the "Tools" module has the options for the InterMine and BioMart, which is useful for data warehousing and construction of customized datasets with information on gene or gene families and annotation. With respect to poplar, *Brachypodium*, eucalyptus, and cassava genomes, both expression and diversity data can be viewed in JBrowse, searched, and downloaded from InterMine, as well as in bulk from the JGI Genome Portal. Screenshot of Phytozome database is depicted in Fig. 6.3.

Keyword and Sequence Similarity Search Information on relevant attributes of gene and gene family like names, symbols, synonyms, external database identifiers, definitions, and functional annotation IDs can be retrieved by keyword search. BLAST and BLAT can be used to identify the genomic regions, gene transcripts, peptides, and gene families most similar to the query sequence. Gene families at a particular evolutionary node and families matching particular phylogenetic profiles can be searched. The database can also be searched for functional annotations to retrieve all matching functional identifiers and gene families.

Information on Gene and Gene Families The *Gene Family view* gives information on each gene family and its constituent members. The default *Genes in this family* tab displays members of a particular gene family along with their source identifier, aliases, synonyms, and gene symbols. The *family page* has a set of lower tabs (Functional Annotation, MSA, and Family History) and upper tabs (Find related families, Align family

members, Get Data, and Display options). The lower tab helps in the exploration of the evolutionary history of the gene family, while the upper tab is useful for analyzing the similarity among the related sub-families of genes. Besides depicting single gene functional annotations and evolutionary history, the *Gene Page* has links to alternatively spliced transcripts, if any; access to genomic, transcript, and peptide sequences associated with the gene; and a graphical view of other Phytozome peptides aligned against the peptide of the gene of interest.

Genome Browser In *GBrowse* module, genome-centric views are provided for all the 41 genomes currently available in the Phytozome. This module can be accessed directly from the Phytozome home page, from individual member gene links available on the Gene Family or Gene Page, and from the BLAST/BLAT results page. Each browser shows a gene prediction track, where homologous peptides from related species, supporting ESTs, and one or more syntenic VISTA tracks identifying regions of this genome are depicted. The gene features and VISTA tracks are hyperlinked to the Gene Page and corresponding regions in the VISTA browser, respectively. On sorghum, 697,578,683 base pairs are arranged, which correspond to 34,496 loci and 36,338 protein-coding transcripts.

Data Retrieval Bulk data files containing the genome assembly sequence, gene structure, transcript, coding, and peptide sequence in FASTA format and general annotation information are available for the genomes hosted in Phytozome database. Repeat-masked genome assemblies as well as supporting annotation data are also available for download. By using BioMart module, customized datasets can be constructed consisting of information on gene or gene family sequences and annotations based on user-defined data filters, attributes, and output formats. This module can be accessed from the "Get Data" tab on the Gene Family page or directly from the main menu of Phytozome.

6.2.4 GreenPhylDB

GreenPhylDB is a database specifically designed for comparative and functional genomics based on completely sequenced genomes. The development of GreenPhylDB v1.0 (5) by Conte et al. (2008) was inspired by the availability of whole-genome sequences of *Arabidopsis thaliana* and *Oryza sativa* genomes that offered opportunity for comparative genomics in plants. Since then, the most popular and reliable approach for the functional annotation of genes is by analyzing genes between species to identify orthologous genes (Kuzniar et al. 2008; Gabaldon et al. 2009). GreenPhylDB v2.0 was published during 2011 by Rouard et al. (2011) by adding 14 new genomes belonging to a major phylum of the plant kingdom including rodophytes, chlorophytes, mosses, lycophytes, and flowering plants with monocotyledons and dicotyledons. Six genomes were added in version 3, while 14 genomes were added in the current version (v4). Currently, the database has the genome information on 37 species. GreenPhylDB is accessible at http://www.greenphyl.org/cgi-bin/index.cgi.

The database represents a catalogue of gene families based on complete genomes. GreenPhylDB comprises complete proteome sequences from the major plant phylum, which are clustered to define a consistent and extensive set of homeomorphic plant families. Lists of plant- or species-specific gene families and several tools are provided to facilitate comparative genomics within plant genomes. The analyses include clustering of gene family followed by a phylogenomic analysis of the generated gene families. Upon validation of a cluster, phylogenetic analyses are performed to predict orthologs and ultraparalogs. Results of clustering are first manually annotated and then analyzed by a phylogenetic-based approach to predict orthologs, which is particularly useful for functional genomics and candidate gene identification of genes affecting agronomic traits of interests. This resource has 2,915 annotated gene families, of which 53 are specific to sorghum. Schematic diagram of the web resource is given in Fig. 6.4.

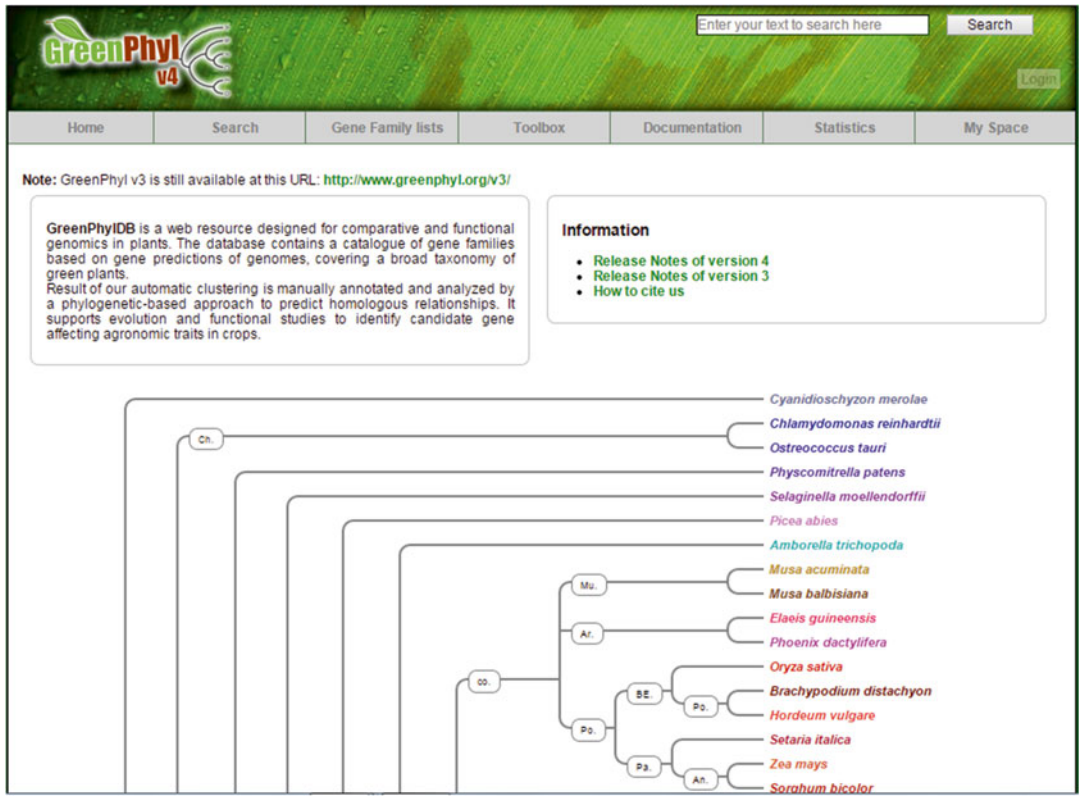

Fig. 6.4 Schematic diagram of GreenPhylDB (http://www.greenphyl.org/cgi-bin/index.cgi)

GreenPhylDB has three important modules that can be used for comparative genomics analysis, viz., "Search" module, "Gene Family Lists" module, and "Tool Box" module.

"Search" Module This module has the options of Quick Search and Advanced Family Search. The former searches based on text/keywords, while the latter on InterPro domains.

"Gene Family Lists" Module This module contains the list of annotated gene families comprising 2,915 clusters and transcription factor families comprising 33 clusters. It has the option to list the gene families specific to a particular species/phylum. A GO Browser was developed as a web interface, which displays a list of terms defined in the Plant GO. By selecting a specific GO entry, the users can access a list of gene families potentially involved in plant growth and development along with the sub-classification of each identified gene family.

"Tool Box" Module This contains the option for BLAST (sequence search with BLASTP or BLASTX), sequences to families (family classification of given sequence), Homolog sequences (get homologs and/or similar sequences with sequence ID inferred from phylogeny), InterPro domain (domain distribution is displayed by sequence and by species), Export sequences (provides a list of sequence ID used in database that can be exported in the selected format), TreePattern (to explore phylogenetic trees), and Create family (to create gene families). It has 32,796 sorghum-specific proteome datasets in it. The biggest advantage of this portal is to construct gene family tree and identify homologs.

6.2.5 CoGe

CoGE stands for Accelerating Comparative Genomics (https://genomevolution.org/CoGe/). It is a unique web resource having many interconnected tools to create open-ended analysis networks. The features of CoGe were published by Lyons and Freeling (2008) and Lyons et al. (2008). CoGe is designed to address four issues: (1) single platform to store multiple versions of multiple genomes from multiple organisms, (2) rapid identification of sequences of interest in genomes of interest (with associated information), (3) comparison of multiple genomic regions using any algorithms, and (4) visualization of the results for easy and quick identification of "interesting" patterns. Organization of CoGe database is given in Fig. 6.5. CoGe database has a set of tools for comparative genome analysis. They are as follows:

- *OrganismView* – searches and gives an overview of an organism and its genomic information
- *CoGeBlast* – BLAST sequences against any number of organisms of user's choice
- *FeatView* – searches for genomic features by name or description
- *SynMap* – generates syntenic dotplots of any two genomes
- *SynFind* – identifies syntenic regions across many genomes
- *GEvo* – compares multiple genomic regions using a variety of sequence comparison algorithms for high-resolution analysis to quickly identify patterns of genome evolution

An Integrative *Orthology Viewer* combines information from different orthology prediction methodologies. Central tools and access points of CoGe allow to find sequences of interest, and "*hub*" points direct from one part of the system to another. For example, if a region with an inversion is identified during the comparison of sorghum with the maize genome using *SynMap*, breakpoints of that region may be compared using *GEvo* in high detail, and the maize sequence can be extracted out using *SeqView*. Subsequently, *FeatView* can be used to identify all the protein-coding regions, and the information generated can be used to find homologs in other plant genomes using *CoGeBlast. GEvo* can be used to validate putative syntenic regions. If, say, a gene with extra copy number is identified in a syntenic region, its sequence may be obtained using *FeatView* once again. Putative intra- and interspecific homologs of it may be obtained using *CoGeBlast*, which will generate a FASTA file using *FastaView*. This can be aligned using *CoGeAlign* and used to build a phylogenetic tree through *TreeView* or exported to more expansive phylogenetic platform such as *CIPRES*. Simultaneously, the codon and protein usage variation of the genes may be checked using *FeatList*. If some interesting variation is observed in some genes, their overall GC content and wobble-position GC content may be checked in *FeatView*. Horizontal transfer of DNA fragments/genes from the mitochondria can be identified using *CoGeBlast* or *GEvo*.

6.2.6 PLAZA

A centralized plant genomics platform is essential for performing evolutionary and comparative analyses of gene families and genome organization, which integrates all the information generated by various sequencing projects along with advanced tools for data mining. PLAZA is a versatile plant comparative genomics resource centralizing genomic data from different genome sequencing initiatives (http://bioinformatics.psb.ugent.be/plaza/) published by Proost et al. (2009). Plant sequence data and comparative genomics methodologies are integrated in an online platform with interactive tools to study gene function and gene and genome evolution within the green plant lineage. It has integrated structural and functional annotation of 25 green plant species, which includes 909,850 genes. Out of these genes, 85.8 % are protein coding, which are clustered in 32,294 multigene families, resulting in 18,547 phylogenic trees. In addition to the basic

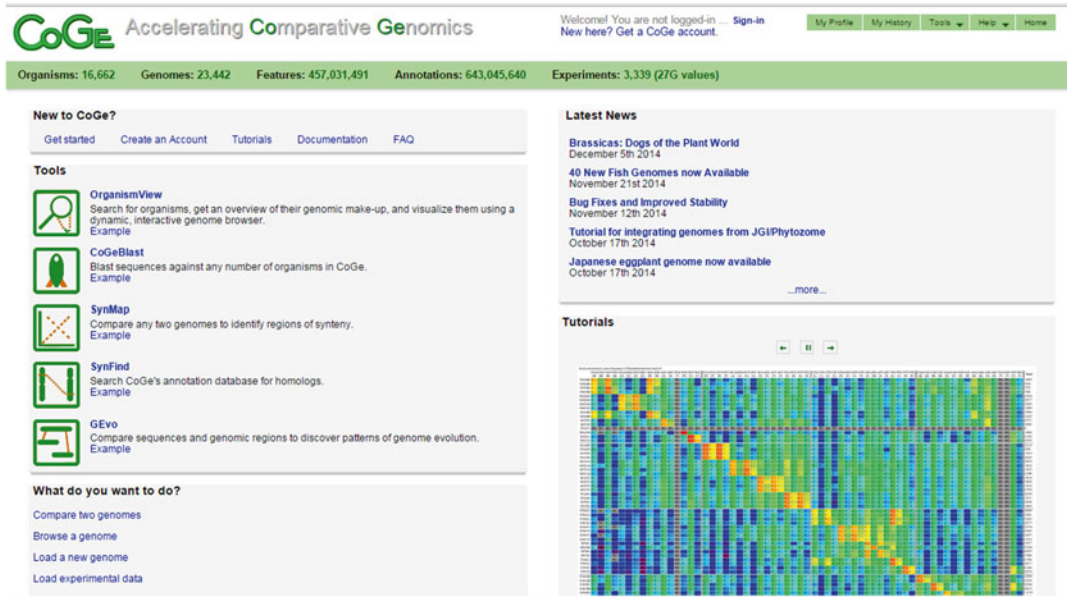

Fig. 6.5 Organization of CoGe database (https://genomevolution.org/CoGe/)

information related to gene structure and function such as genome coordinates, mRNA and protein sequences, and gene description, PLAZA offers various tools to browse genomic data for homology, ranging from local synteny to gene-based colinearity views useful for comparative genomics plant genome evolution. Organization of PLAZA database is given in Fig. 6.6.

- *Synteny plot* – a basic tool that shows all genes from the specified gene family along with their neighboring genes, thereby helping to study genomic homology in comparison to colinearity.
- *WGDotplot* – useful for analyzing genome-wide colinearity leading to the identification of large-scale duplications or to study genomic rearrangements within or between species.
- *Skyline plot* – useful for browsing multiple homologous genomic segments and provides a comprehensive view of the regions that are colinear in the species selected by the user.
- *Workbench* – useful to analyze multiple genes in batch that are uploaded through gene identifiers or based on similarity search and calculate different genome statistics for user-defined gene sets.
- *Whole Genome Mapping tool* – useful to display a selection of genes on the chromosomes and to view the distribution of different classes of genes such as protein coding, pseudogene, or transposable element. It also provides information about the gene duplication.
- *Advanced query system* – useful for the rapid retrieval of relevant information using different data types and research tools.

6.2.7 CSGRqtl

CSGRqtl is a comparative genomic database (http://helos.pgml.uga.edu/qtl/) developed by Zhang et al. (2013a) as a data mining resource specific to crops, weeds, and models of Saccharinae clade. This database complements and supplements the database Gramene, which contains mapping data from a wide spectrum of grasses. CSGRqtl uses sorghum genome sequence as a reference with an aim of anchor-

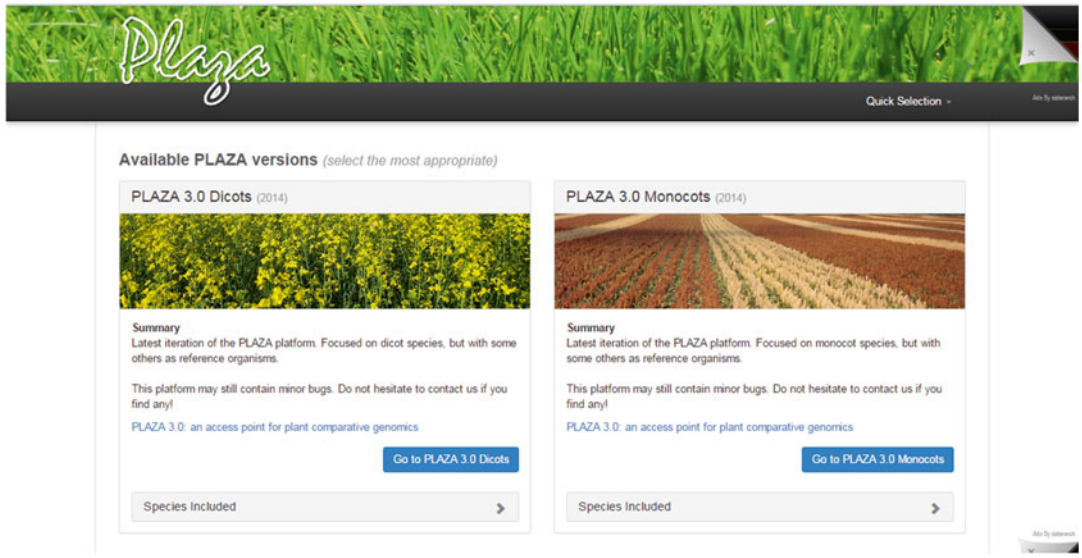

Fig. 6.6 Organization of PLAZA database (http://bioinformatics.psb.ugent.be/plaza/)

ing published QTLs of the Saccharinae clade to the sorghum genome. This database uses the Plant Trait Ontology defined by Gramene and facilitates the data comparisons among the grasses of Saccharinae clade and between Saccharinae and other taxa to categorize quantitative trait loci (QTLs) with their approximate physical positions. CSGRqtl also integrates gene annotations, genetic markers, and paleoduplicated regions, which facilitate QTL mapping and the study of candidate gene underlying the QTLs. Organization of CSGRqtl database is given in Fig. 6.7.

CSGRqtl is equipped with a number of tools for analysis, such as text-based search, trait ontology browser, QTL correspondence, and CMap database, to allow a user to query and visualize the background database.

Text-Based Search QTL search based on the trait of interest gives a set of QTLs underlying the trait. Genome-wide overview of QTL distribution is depicted by a circular plot created by the bioinformatics resource Circos. It also identifies the potential QTL hot spots in the genome. QTL search based on a sorghum gene identifier or annotation gives a list of QTLs containing the queried gene, and the approximate positions of genes and QTLs are depicted by a plot.

Trait Ontology Browser The trait ontology browser displays the hierarchy of trait ontology and lists QTLs associated with each trait accession since each QTL is allotted a trait accession defined by Gramene Plant Trait Ontology.

QTL Correspondence The associations between paleoduplicated regions and QTLs in rice and sorghum are depicted by circular plots to indicate non-overlapping QTLs divided by inter-genomic synteny or narrowed by intra-genomic synteny. The users can download orthologs/paralogs for genes associated with non-overlapping QTLs for the trait of interest.

CMap Database Alignments between genetic maps and the sorghum genome sequence can be viewed by the user. It also gives information on RFLP probe and SSR primer sequences for each anchored marker that is amenable for alignment.

Genome Browser This is implemented using Generic Genome Browser version 2.39 (Stein et al. 2002) and used to associate sorghum QTL

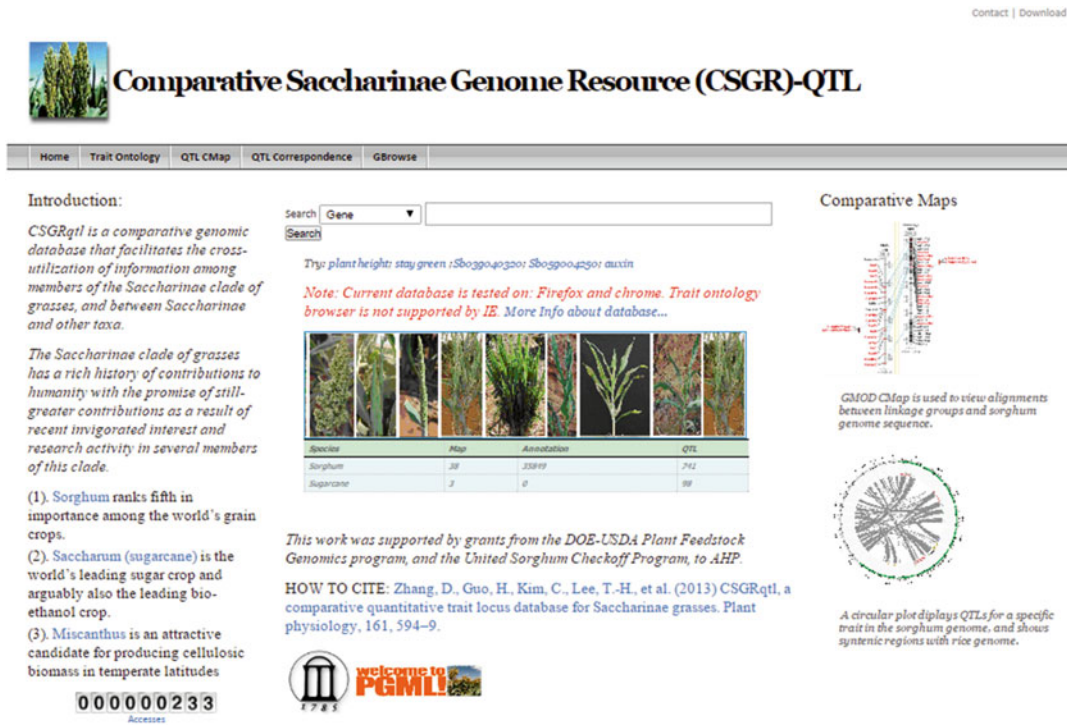

Fig. 6.7 Organization of CSGRqtl database (http://helos.pgml.uga.edu/qtl/)

data with gene annotations. This browser contains gene models from standard sorghum genome annotation version 1.4 and about 209,828 sorghum ESTs from the NCBI. It also contains GC content, six-frame translation, and restriction sites for each genomic region. The user can get information of all annotated genes for a particular QTL region and also can access all QTLs in any genomic region.

6.2.8 SorGSD

DNA sequence variations between diverse sorghum lines are an important pre-requisite for the genetic improvement of sorghum for agronomic traits as well as tolerance to biotic and abiotic stresses through breeding by design and high-efficiency genomic selection. Advances in the next-generation sequencing (NGS) technologies have brought about a surge in the re-sequencing of the diverse sorghum accessions belonging to different categories such as improved inbreds, landraces, wild/weedy sorghums, and wild relatives. Recently, a diverse panel of 48 sorghum accessions which were divided into four groups, including improved inbreds, landraces, wild/weedy sorghums, and a wild relative *Sorghum propinquum*, has been re-sequenced leading to the generation of enormous amount of SNP data (Mace et al. 2013). Proper organization of this SNP data will offer excellent opportunity for researchers to identify variation in their genes of interest, explore evolutionary relationships among cultivated and wild types, develop DNA markers for future genetic studies, and utilize this data for genome-wide association studies (GWAS). With this premise, SorGSD, a web-based large-scale genome variation database, was developed during August 2014 and maintained by the Data Management Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, and the Laboratory for Conservation and Utilization of Bio-resources, Institute of

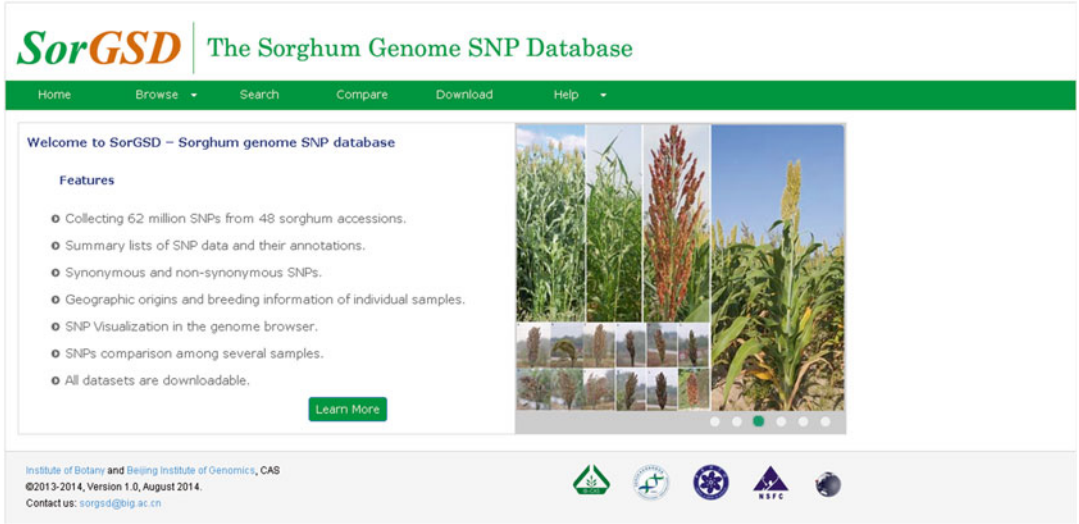

Fig. 6.8 Organization of SorGSD database (http://sorgsd.big.ac.cn/snp/)

Botany, Chinese Academy of Sciences. The database contains 62 million SNPs with annotations assisted by an easy-to-use web interface for users for efficient browsing, searching, and analysis of the SNPs. The pipeline for SNP calling included trimming of adapter and filtering of all low-quality reads, the use of BWA (version 0.6.2-r126) to map clean read to sorghum reference sequence (V1.4), SAMtools package to convert mapping results to BAM format, Picard (version 1.87) program to eliminate duplicated reads generated during the process of library construction, SNP calling by GATK (version 2.5-2-gf57256b) toolkit, SNP identification based on the quality estimation scores generated by GATK (quality value ≥30 and depth of coverage ≥5), and SnpEff program for the annotation of SNPs. This database is a rich repository to molecular breeders for the identification of biomarker, genetic analysis, and marker-assisted breeding of sorghum and other crops. The SorGSD can be accessed from http://sorgsd.big.ac.cn/snp/. The SorGSD has four modules, viz., Browse, Search, Compare, and Download as shown in Fig. 6.8.

Browse Module This module can be used to browse total SNPs as well as gene-wise and chromosome-wise SNPs. SNPs in Gene lists SNPs located in gene and coding regions. The "Coding," "Synonymous," and "Non-synonymous" lists are used to view SNP located in coding region, annotated as synonymous and non-synonymous. Query can be given based on the chromosome number also. The users can browse SNP information and their relevant annotations for each sorghum line. The database uses GBrowse to visualize InDel, SNP, gene, transcript, density information of SNP/300 kb, and allele frequency.

Search Module This module helps in searching SNPs in a single individual by setting parameters such as chromosome location, SNP class, and SNP location in gene region and genotype. Options are provided for the selection of SNP annotation and SNP genotype. The search results can be either visualized graphically in a genome browser or displayed in formatted tables.

Compare Module This module helps in comparing SNPs in two or more individuals by setting parameters such as chromosome location, SNP class, and SNP location in gene region and genotype. The SNPs in selected individuals can

be compared with that of a single reference genotype or more genotypes. Options are provided for the selection of SNP annotation and SNP genotype.

Download Module This module contains the datasets, viz., SNP files, InDel files, SRA files, and Fastq files, which can be directly downloaded for further analysis.

6.3 Genomics Resources for Analyzing DNA Sequence Variation

In the current era of genomics, large-scale EST as well as genome sequencing projects resulted in the generation of enormous amount of DNA sequence data that are organized and stored in various databases. Such data available in the public domain are the main targets for the development of molecular markers such as simple sequence repeats (SSRs), insertion-deletions (InDels), single nucleotide polymorphisms (SNPs), etc., through various computational approaches and bioinformatics tools available. These molecular markers are the best tools available for the plant geneticists for analyzing the DNA sequence variations through an easy and rapid PCR assay and are useful for assessing the genetic diversity and population structure of germplasm lines, varietal identification, genetic purity testing of hybrids and parental lines, mapping of genetic loci through QTL mapping, and marker-assisted selection.

Assessment of genetic diversity in the germplasm accessions or the parental line gene pool is the primary step in any plant breeding program. Earlier, morphological as well as quantitative traits were used for the assessment of genetic diversity. Due to their inherent limitations in the number as well as environmental influence, molecular markers have become the choice of such genetic diversity assessments since these markers are environmentally neutral. Several studies were undertaken over the years in the assessment of genetic diversity using molecular markers such as RAPD (Ayana et al. 2000; Uptmoor et al. 2003), RFLP (Tao et al. 1993; Ahnert et al. 1996), AFLP (Geleta et al. 2006; Ritter et al. 2007), and ISSR (Aruna et al. 2012). These markers are also used in the mapping of major genes (Knoll et al. 2008; McIntyre et al. 2008) as well as QTL (Srinivas et al. 2009b; Satish et al. 2009).

SSR markers are the widely used PCR-based markers for various genetics and mapping studies in sorghum due to their abundance in the genome, highly polymorphic nature, and easy assay. Prior to the completion of genome sequencing of sorghum in 2009, several research groups have developed a large number of SSR markers which were subsequently used for various genetic studies in sorghum. These studies helped in the development of genomic SSR markers [Xtxp series (Kong et al. 2000; Bhattramakki et al. 2000, http://sorgblast3.tamu.edu/search/marker.htm), XSb series (Taramino et al. 1997), Xgap series (Brown et al. 1996)], SSR markers derived from cDNAs [Xcup series (Schloss et al. 2002)], whole-genome sequence-based SSR markers [SB series (Yonemaru et al. 2009)], expressed sequence tag (EST)-based SSR markers [Xisep series (Ramu et al. 2009), Xiabt series (Arun 2006; Reddy et al. 2008), Stgnhsbm and Dsenhsbm series (Srinivas et al. 2008, 2009a)], SSR markers derived from unigenes [Ungnhsbm series (Srinivas et al. 2009b; Nagaraja Reddy et al. 2012)], $(GATA)_n$ motif-based SSR markers [SbGM series (Jaikishan et al. 2013)], and other SSR markers of an unknown type [gpsb and mSbCIR series (developed at CIRAD, France, and partially published in Mace et al. 2009)]. The details of the SSR markers developed by different sorghum groups are given in Table 6.2.

Even though several studies on the assessment of genetic diversity were reported over the years, very few studies have done a comprehensive analysis and resulted in a set of robust SSR markers that can be used universally across laboratories for this purpose. A set of 38 SSR markers distributed across 10 chromosomes of sorghum selected based on three different linkage maps were used to establish the diversity research set comprising 107 sorghum accessions (Shehzad et al. 2009) from a set of 320 sorghum germplasm

Table 6.2 SSR markers developed by different sorghum research groups

Type of SSR markers	Marker series	No. of markers developed	No. of markers experimentally tested	Reference
Genomic SSR	Xtxp	206	165	Bhattramakki et al. (2000), Kong et al. (2000)
		38	38	
	XSb	15	13	Taramino et al. (1997)
	Xgap	149	149	Brown et al. (1996)
cDNA-derived SSR	Xcup	74	60	Schloss et al. (2002)
Whole-genome SSR	SB	5,599	970	Yonemaru et al. (2009)
EST-derived SSR	Xisep	600	386	Ramu et al. (2009)
	Xiabt	520		Arun (2006)
	Stgnhsbm	50	50	Srinivas et al. (2008, 2009a)
		116	109	
Unigene-derived SSR	Dsenhsbm	50	50	Srinivas et al. (2009b), Nagaraja Reddy et al. (2012)
	Ungnhsbm	1,519	302	
$(GATA)_n$ motif-based SSR	SbGM	110	50	Jaikishan et al. (2013)
Other SSRs	gpsb and mSbCIR	30	24	Mutegi et al. (2011); Billot et al. (2012)

accessions. A diversity analysis kit was developed (Billot et al. 2012), which contains information on 48 robust sorghum SSR markers that can be used to calibrate SSR genotyping data acquired with different technologies and compare them to genetic diversity references. A reference set comprising a wide range of sorghum genetic diversity was screened with 40 EST-SSR markers by Ramu et al. (2013), and the analysis highlighted the greater discriminating power of these markers as compared to the genomic SSR markers.

In the current era of genome sequencing, the discovery of SNPs and insertion and (or) deletions (InDels) through high-throughput methods has led to a revolution in their use as DNA markers (Batley and Edwards 2007; Batley et al. 2007; Edwards et al. 2007). Advancements in sequencing technologies, execution of re-sequencing projects, and availability of the enormous amount of ESTs along with the development of efficient computational platforms have helped in the rapid discovery of SNPs and InDels in sorghum. SNPs may be considered the ultimate genetic marker as they represent the finest resolution of a DNA sequence, generally are abundant in populations, and have a low mutation rate (Syvanen 2001). The mining of readily available sequence data for SNPs through in silico approaches significantly reduces the costs (Taillon-Miller et al. 1998), and several SNP mining tools have been developed (Barker et al. 2003; Batley et al. 2003; Savage et al. 2005; Chagne et al. 2007). Sorghum researchers across the globe have utilized different types of data, such as ESTs, whole-genome re-sequencing data, and genotyping-by-sequencing data for the discovery of SNPs with the help of various computational tools. The detail of the SNPs developed by different sorghum groups is given in Table 6.3.

InDels are next only to SNPs in terms of their abundance in the genome. However, InDels can be converted into PCR-based markers and can be resolved through routine gel electrophoresis systems. InDels exhibit length polymorphisms that have been successfully exploited in sorghum in the mapping of important loci such as waxy (McIntyre et al. 2008) and tannin (Wu et al. 2012). In sorghum, about 99,948 InDels of 1 to 10 bp in length were detected by Zheng et al. (2011) through a genome-wide analysis in sweet and grain sorghum. Potential intron polymorphism (PIP) markers developed by Yang et al.

Table 6.3 SNPs developed by different sorghum research groups

Target data	Computational tool used	No. of SNPs identified	Reference
ESTs	CodonCode Aligner	12,421 SNPs	Girma (2009)
ESTs	HaploSNPer	77,094 potential and 40,589 reliable SNPs	Singhal et al. (2011)
Re-sequencing data	SOAPsnp software	1,057,018 SNPs	Zheng et al. (2011)
Eight genome equivalents to reference genome	SOAP v2 and NovoAlign	283,000 SNPs	Nelson et al. (2011)
GbS data of 971 diverse sorghum accessions	TASSEL 3.0 GBS pipeline	265,487 SNPs	Morris et al. (2013)
Re-sequencing data	realSFS and SOAPsnp	4,946,038 SNPs	Mace et al. (2013)

(2007) are a unique type of markers that targets both the SNPs and InDels. Among the two types of polymorphisms, intron length polymorphism (ILP) can be easily detected by exon-primed intron-crossing PCR (EPIC-PCR) (Palumbi 1995), where primers are designed in exonic regions flanking the target introns. Potential intron polymorphism (PIP) database for plants was developed by Yang et al. (2007) comprising a total of 57,658 PIP markers for 59 plant species, of which 4314 are of sorghum. These markers can be exploited for genetic diversity assessment, cultivar identification, mapping, and marker-assisted selection.

A new high-throughput hybridization-based marker technology that could serve as an efficient alternative to low-throughput gel-based marker systems was reported in sorghum by Mace et al. (2008). This system does not require sequence information and is amenable for high multiplexing. A genotyping array was developed with ~12,000 genomic clones using PstI + BanII complexity with a subset of clones obtained through the suppression subtractive hybridization (SSH) method. About 508 markers were polymorphic and were used for the genetic diversity analysis of 90 diverse sorghum genotypes and for the construction of a genetic linkage map for a cross between R931945-2-2 and IS 8525. These markers are useful for whole-genome profiling and can be used for diversity analyses and construction of medium-density genetic linkage maps.

A robust SNP array platform was developed recently by Bekele et al. (2013) using 2,124 selected Infinium Type II SNPs from a total of over one million high-quality SNPs identified by the alignment of whole-genome sequences (6–12× coverage) of genetically diverse genotypes comprising two grain and three sweet sorghum genotypes of *S. bicolor* and an additional 876 SNPs selected based on their phenotypic association with early-stage chilling tolerance identified by phenotype-based pool sequencing. Testing this array with selected SNPs using 564 genotypes comprising four unrelated RIL and F_2 populations and a genetic diversity collection resulted in the validation of 2,620 robust and polymorphic SNPs. This SNP array platform is very useful for genetic mapping, genome-wide association, and genomic selection.

6.4 Bioinformatics Resources for DNA Sequence Analysis

DNA sequence variations arise either due to point or gross mutations. Point mutations are mostly due to base substitution (transition or transversion). Gross mutation may be insertion or deletion (InDels) of few to large sequences. It also may arise by duplication, inversion, and translocation. Gross mutations involving large sequences can easily be detected cytologically, while it is difficult to detect gross mutations of smaller dimension (say < 500 bases). Sequencing helps us to detect this variation very precisely. For this purpose, the mutant sequence is aligned with the wild-type sequence, and sequence variation is detected.

Sequence alignment is a way of arranging the nucleic acid (DNA, RNA) or protein sequences to identify regions of similarity that may be a

result of functional, structural, or evolutionary relationships between the sequences (Mount 2004). Aligned nucleotide or amino acid sequences are generally represented as a row matrix by inserting gaps between the residues such that identical characters are aligned in successive columns. In the case of two sequences sharing a common ancestor, mismatches are interpreted as point mutations and gaps as InDels. Conservation of base pairs indicates a similar functional or structural role of the target sequence. Very short or very similar sequences can be aligned manually. However, with sequencing projects, large sequences are available, and these need to be aligned in large number, which is not possible manually. Different algorithms have been developed to facilitate high-quality sequence alignments. These include dynamic programming, heuristic algorithms, or probabilistic methods. Computational approaches to sequence alignment are of two categories: global alignments and local alignments.

Global Alignment With this approach, both sequences are aligned along their entire lengths, including every nucleotide or amino acid, and the best alignment is found. This approach is most useful if the query sequences are similar and of roughly equal size. Dynamic programming called Needleman-Wunsch algorithm is very popularly used for this purpose.

Local Alignment In this approach, the best subsequence alignment, which includes only the most similar sequence, is found. This approach is useful, particularly for dissimilar sequences that are expected to contain similar sequence motifs within their larger sequence. The dynamic programming method, namely, Smith-Waterman algorithm, is commonly used for this purpose.

Both the global and local alignments lead to erroneous conclusion when the downstream part of one sequence overlaps with the upstream part of the other sequence. In such situations, hybrid methods such as "semi-global" or "glocal" (short for *global-local*) methods are employed. These methods help in finding the best possible alignment that includes the start and end of one or the other sequence. The sequence alignment may be of two types based on the number of sequences used for alignment, viz., pair-wise sequence alignment (PSA) and multiple sequence alignment (MSA).

PSA This is the comparison of two biological sequences (nucleic acid or protein) at a time to reveal the similarity or homology between them. This helps in finding the best-matching local or global alignments of two query sequences. This method is most useful in situations where extreme precision like searching of database for sequences with high similarity is not required. Dot matrix methods, dynamic programming, and word methods are mostly used for pair-wise alignments. Several PSA tools have been developed by several workers (Table 6.4), many of them are free to use, and some of them are available as commercial software. BLAST is the best example for pair-wise sequence alignment. BLAST searches a query sequence (discovered sequence) against a database of known sequences in order to find similarities. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST searching can be performed using the web-based applications (Web BLAST) or can be run locally in the PC provided it has an existing database to aid searching. There are five types of BLAST:

- BLASTN: Compares a nucleotide query sequence against a nucleotide sequence database
- BLASTP: Compares an amino acid query sequence with a protein database
- BLASTX: Translates a DNA sequence into six protein sequences using all six possible reading frames and then compares each of these proteins to protein database
- TBLASTN: Translates every DNA sequence in a database into six potential proteins and then compares the protein query against each of those translated proteins
- TBLASTX: Translates DNA from both a query and a database into six potential proteins and then performs 36 protein-protein database searches

Table 6.4 Important resources for pair-wise alignment

Name	Sequence type	Alignment type	Availability
AlignMe	P	G, L	http://www.bioinfo.mpg.de/AlignMe
BLAST	P, N	L	http://blast.ncbi.nlm.nih.gov/Blast.cgi
JAligner	P, N	L	http://jaligner.sourceforge.net/
LALIGN	P, N	L (non-overlapping)	http://www.ch.embnet.org/software/LALIGN_form.html
mAlign	N	G, L	ftp://ftp.csse.monash.edu.au/software/m-align/
MCALIGN2	N	G	http://homepages.ed.ac.uk/eang33/mcalign/mcinstructions.html
MUMmer	N	G	http://mummer.sourceforge.net/
needle	P, N	SG	http://www.ebi.ac.uk/Tools/emboss/align/
PatternHunter	N	L	http://www.bioinformaticssolutions.com/products/ph/download_academ.php
PyMOL	P	G (by selection)	www.pymol.org
REPuter	N	L	http://bibiserv.techfak.uni-bielefeld.de/reputer/
SEQALN	P, N	L or G	http://www-hto.usc.edu/software/seqaln/seqaln-query.html
tranalign	N	NA	http://mobyle.pasteur.fr/?form=tranalign
UGENE	P, N	G, L	http://ugene.unipro.ru/

P protein, *N* nucleic acid, *G* global, *L* local, *SG* semi-global

MSA This is the comparison of more number of sequences (three or more) simultaneously. It helps in predicting the structure and function of a given protein sequence due to its ability to detect common features and conserved domains across the sequences analyzed. It also helps to discover novel and related sequences and to construct and search for sequence patterns. Detection of conserved regions among homologous sequences is useful in designing PCR primers. There are a number of MSA programs available in the public domain (Table 6.5).

Clustal is a widely used multiple sequence alignment tool (Chenna et al. 2003). There are three main variations: ClustalW (command line interface, Larkin et al. 2007), ClustalX (this version has a graphical user interface, Thompson et al. 1997), and Clustal Omega [(allows hundreds of thousands of sequences to be aligned in only a few hours). It will also make use of multiple processors, where present. In addition, the quality of alignments is superior to previous versions (Sievers et al. 2011)]. A wide range of input formats, including NBRF/PIR, FASTA, EMBL/Swiss-Prot, Clustal, GCC/MSF, GCG9 RSF, and GDE, are acceptable in this program. The output format can be one or many of the following: Clustal, NBRF/PIR, GCG/MSF, PHYLIP, GDE, or NEXUS. There are three main steps in the alignment process, i.e., pair-wise alignment, followed by the creation of a guide tree (or use a user-defined tree) and finally the use of the guide tree to carry out a multiple alignment. If "Do Complete Alignment" option is selected, all these steps are done automatically, or else the task may be carried out following options, viz., "Do Alignment from guide tree" and "Produce guide tree only." There is an option for default setting or customized settings.

Alignments may be represented graphically and in text format. An asterisk or pipe symbol is commonly used to show identity between two columns. Colors are also used by several programs to display identity and dissimilarity. The sequence alignment results are stored in a variety of text-based file formats. Most common input and output formats are FASTA format and GenBank format.

6.5 Resources for Analyzing Transcriptomes

Transcriptome analysis involves the screening of candidate genes, predicting its function, and discovery of regulatory elements through high-throughput gene expression analysis. Initially, large-scale sequencing of ESTs was used as the main approach for transcriptome analysis. Later, the hybridization-based methods such as microarrays/GeneChips were developed and popularly used for large-scale gene expression analysis. Sequencing-based methods such as serial analysis of gene expression (SAGE) and massively parallel signature sequencing (MPSS) have been successfully employed for the identification of a large number of transcripts along with quantitative comparison of transcriptomes (Velculescu et al. 1995; Brenner et al. 2000). Furthermore, as a next-generation DNA sequencing application, deep sequencing of short fragments of expressed RNAs, including sRNAs, is quickly becoming an efficient tool for use with genome-sequenced species (Harbers and Carninci 2005; de Hoon and Hayashizaki 2008).

A sorghum cDNA microarray providing data on 12,982 unique gene clusters was used by Buchanan et al. (2005) to examine genome-wide changes in gene expression in sorghum seedlings under high salinity (150 mM NaCl), osmotic stress (20 % polyethylene glycol), or abscisic acid (125 μM ABA). A total of 3,508 cDNAs selected from the two cDNA libraries constructed from a strong greenbug resistance sorghum line (M627) and a susceptible line (Tx7000) with or without infestation were used to develop a cDNA microarray for the identification of sorghum genes responsive to greenbugs (Park et al. 2006).

An Agilent rice gene expression microarray (product number: G2519F, 44 K) was used to study tissue-specific gene expression profiles of *S. propinquum* with special emphasis on rhizome development by Zhang et al. (Zhang et al. 2013b) that contained 45,220 independent probes (60-mer) corresponding to 21,495 *O. sativa* mRNA sequences available in GenBank due to the non-availability of microarray platform in sorghum. Only recently, the first whole-transcriptome

Table 6.5 Important resources for multiple sequence alignment

Name	Sequence type	Alignment type	Availability
AMAP	P, N	G	http://baboon.math.berkeley.edu/mavid/
Base-By-Base	P, N	L or G	http://athena.bioc.uvic.ca/virology-ca-tools/base-by-base/
ClustalW	P, N	L or G	http://www.ebi.ac.uk/Tools/msa/clustalw2/
CodonCode Aligner	N	L or G	http://www.codoncode.com/aligner/
Compass	P	G	http://prodata.swmed.edu/compass/compass_advanced.php
DNA Baser Sequence Assembler	N	L or G	www.DnaBaser.com
FSA	P, N	G	http://orangutan.math.berkeley.edu/fsa/
MSA	P, N	L/G	http://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/msa.html
MULTALIN	P, N	L/G	http://multalin.toulouse.inra.fr/multalin/
MUSCLE	P, N	L/G	http://www.ebi.ac.uk/Tools/msa/muscle/
ePROBALIGN	P	G	http://probalign.njit.edu/index.html
PSAlign	P, N	L/G	http://faculty.cs.tamu.edu/shsze/psalign/
RevTrans	N/P (special)	L/G	http://www.cbs.dtu.dk/services/RevTrans/
SAM-T08	P	L/G	http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html
T-Coffee	P, N	L/G	http://tcoffee.crg.cat/

P protein, *N* nucleic acid, *G* global, *L* local, *SG* semi-global

microarray was developed in sorghum by Shakoor et al. (2013) comprising a gene chip containing 1,026,373 probes covering 149,182 exons (27,577 genes) across the nuclear, chloroplast, and mitochondrial genome along with putative non-coding RNAs to identify tissue-specific genes and novel regulatory sequences. Toward identification and functional characterization of genes in sorghum genome, Shakoor et al. (2014) used the first commercial whole-transcriptome sorghum microarray chip (Sorgh-WTa520972F) to identify tissue- and genotype-specific expression patterns using grain, sweet, and bioenergy sorghums. Microarray dataset was generated using 78 samples involving different tissue types (shoot, root, leaf, and stem) and dissected stem tissues (pith and rind) of six diverse genotypes (R159, Atlas, Fremont, PI152611, AR2400, and PI455230), which revealed tissue- and genotype-specific expression patterns of different metabolic pathways indicating the importance of intraspecies variations in sorghum.

Next-generation sequencing technologies have enabled the researchers to characterize small RNA component of the transcriptomes in many plant species. According to the recent release of the miRBase database (http://www.mirbase.org, release 20: June 2013), about 205 miRNAs are described for sorghum, whereas 592 miRNAs are described for rice. The sorghum genome sequencing consortium identified 149 predicted miRNAs belonging to 27 miRNA families (Paterson et al. 2009). The identification of miRNAs from different target tissues, developmental stages, and stress treatments offer an excellent opportunity to understand the role of miRNAs in the regulation of expression of genes influencing traits of agronomic importance.

Prior to whole-transcriptome microRNA (miRNA) sequencing projects, computational approaches based on homology search were used for the identification of miRNAs in different plant species. Using this approach, Du et al. (2010) identified a total of 17 new miRNAs based on the GSS and the miRNA secondary structure that were distributed unevenly among 11 miRNA families. Analysis of these miRNAs via online software miRU revealed that they might regulate 64 target genes, most of which are involved in RNA processing, metabolism, cell cycle, protein degradation, stress response, and transportation. The small RNA component of the transcriptome of grain and sweet sorghum stems was characterized by Calviño et al. (2011) using F_2 population derived from the cross between BTx623 (grain sorghum) and Rio (sweet sorghum) that segregated for sugar content and flowering time. They reported that the variation in miR172 and miR395 expression correlated with flowering time, while that of miR169 correlated with sugar content in stems.

With the increasing number of whole genomes, large-scale cDNA sequencing, and microarray projects in the recent years, an enormous amount of transcriptome data is generated and stored in public databases. This data serves as a valuable resource for many secondary uses, such as co-expression and comparative transcriptome analyses. NCBI's Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and the European Bioinformatics Institute (EBI)'s ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) are similar such databases, which serve as the primary archives of transcriptome data in the public domain (Parkinson et al. 2007; Barrett et al. 2009). Transcriptome data of *S. propinquum* in relation to rhizome development and the data of grain (BTx623) and sweet (Keller) sorghum are stored in GEO. ArrayExpress database has the transcriptome datasets of tissue-specific transcriptomic profiling of *S. propinquum* using a rice genome array, sorghum gene expression using Agilent custom 4x44K microarray, RNA-Seq of *S. bicolor* 9d seedlings in response to osmotic stress and abscisic acid, and high-throughput sequencing of small RNAs in *S. bicolor*. Another database, namely, EGENES (http://www.genome.jp/kegg-bin/create_kegg_menu?category=plants_egenes), is a multi-species resource integrating the genomic, chemical, and network information comprising genes, molecules, and biological pathways representing the cellular functions. This resource is useful for the comparison and mutual validation of genome-based pathway annotation and EST-based

annotation. The EGENES consists of data on 25 eukaryotic species including sorghum. The sorghum datasets include 190,946 ESTs, 19,597 contigs, 23,171 singletons, 122 pathway maps, and 1,189 mapped contigs.

6.6 Genetic Resources for Mapping Agronomically Important Traits

The majority of the agriculturally important traits such as yield, biotic and abiotic stress tolerance is complex and is governed by quantitative trait loci (QTLs). These QTLs are influenced by the environment and the interaction between QTL and environment. Linkage mapping (biparental mapping) and linkage disequilibrium (LD) mapping (association mapping) are the two most commonly used approaches for the dissection of such complex traits. The first step in QTL mapping is the development of mapping populations by crossing parental lines that are contrasting for the trait of interest (biparental population and multiparental population) or assembling of diverse germplasm lines (natural population), which serve as an important genetic resource.

Biparental Population This approach involves the mating between two parental lines that are contrasting for the trait of interest and advancing them to develop F_2, backcross populations, recombinant inbred lines (RILs), backcross inbred lines (BILs), and double haploid (DH). In addition to this, introgression lines (ILs) and near isogenic lines (NILs) are also developed. These mapping populations are known as a first-generation mapping resource. Even though several preliminary studies used F_2 populations, the use of advanced generations, particularly RILs derived by single-seed descent from F_2 individuals from a cross between two distinct homozygotes, is most commonly used for QTL mapping purposes (Keurentjes et al. 2011) because they are immortal and can be multiplied any number of times (Huang et al. 2011) for phenotyping in different environments/seasons. To tackle the problem of epistasis due to the interaction between multiple loci, the biparental populations such as ILs and NILs are also used (Rakshit et al. 2012). The advantage of employing NIL over RIL is mainly the detection of minor QTL that are missed while using RILs (Keurentjes et al. 2007). Globally, many sorghum research groups have developed and used several RIL populations for various traits (Table 6.6).

Multi-parental Mapping Populations Biparental populations, though popularly used for QTL mapping, have two major limitations such as relying on the recombination events happening in the F_1 and subsequent generation and mapping only the allelic pairs that are present in the two contrasting parents (Rakshit et al. 2012). This affects the map resolution of the QTL since QTL will be placed on a large chromosomal region (Li et al. 2010). In the recent past, the second-generation mapping resources such as association mapping, nested association mapping (NAM), and multi-parent advanced generation inter-cross (MAGIC) were developed to overcome the limitations of biparental populations. NAM populations are developed by crossing a central parent with other diverse parents in a star design (Huang et al. 2011), and such populations have been established in maize (Yu et al. 2008; Buckler et al. 2009; McMullen et al. 2009) and *Arabidopsis* (Bentsink et al. 2010; Brachi et al. 2010). Development of a large NAM population in sorghum was reported recently by Jordan et al. (2012) comprising more than 4,000 lines from 100 sub-populations derived from a large BC_1F_1 population using a single elite line as the recurrent parent resulting in the sampling of the diversity of sorghum including wild relatives. Such populations help in fine mapping of QTL; however, the interaction of QTL with genetic background cannot be analyzed since one parent is common in all sub-populations. Consequently, the concept of MAGIC population was proposed by Cavanagh et al. (2008) to address the major limitations of biparental mapping populations. This concept was used as an additional resource for dissecting the genetics of natural varieties in *Arabidopsis* multi-parent recombinant inbred line (AMPRIL) population (Huang et al. 2011). Recently, another multi-parent mapping population

Table 6.6 Biparental mapping populations in sorghum

Trait	Mapping population	Population size	Reference
Plant phenology			
Plant height			
dw_2	*S. bicolor* × *S. propinquum*	F_2: 320	Lin et al. (1995)
dw_2	Shan Qui Red × SRN39	RIL: 153	Klein et al. (2008)
	296B × IS18551	RIL: 168	Madhusudhana and Patil (2013)
Photoperiod sensitivity	BTx623 × IS3620C	RIL: 127	Childs et al. (1997)
	IS2807 × IS7680	RIL: 85	Chantereau et al. (2001)
Flowering time	Kikuchi Zairai × SC112	F_2: 144	El Mannai et al. (2012)
	BTx642 and Tx7000	RIL: 90	Yang et al. (2014)
Maturity			
Ma_3	100 M × 58 M	RIL: 137	Childs et al. (1997)
Ma_1	*S. bicolor* × *S. propinquum*	F_2: 320	Lin et al. (1995), Klein et al. (2008)
Ma_4	BTx623 × IS3620C	RIL: 137	Hart et al. (2001)
Plant color	IS2807 × 379	RIL: 110	Rami et al. (1998)
	IS2807 × 249	RIL: 91	
	RTx430 × Sureno	RIL: 125	Klein et al. (2001)
	296B × IS18551	RIL: 168	Srinivas et al. (2009b)
Stem	B35 × Tx7000	RIL: 98	Xu et al. (2000)
	BTx623 × IS3620C	RIL: 137	Hart et al. (2001)
	296B × IS18551	RIL: 168	Srinivas et al. (2009b)
Tillering	BTx623 × IS3620C	RIL: 137	Hart et al. (2001)
			Kebrom et al. (2006)
Inflorescence architecture	BTx623 × IS3620C	RIL: 119	Brown et al. (2006)
Nodal root angle	B923296 × SC170-6-8	RIL: 141	Mace et al. (2012)
Agronomic traits			
Agronomically important traits	296B × IS18551	RIL: 168	Srinivas et al. (2009b)
	Sorghum bicolor × *S. sudanense*	$F_{2:3}$: 248	Ping et al. (2011)
	654 × LTR108	RIL: 244	Zou et al. (2012)
	M35-1 × B35	RIL: 245	Nagaraja Reddy et al. (2013)
Agronomic traits and yield components	IS2449 × IS1488	RIL: 100	Phuong et al. (2013)
Grain quality, productivity, morphological and agronomical traits	IS2807 × 379	RIL: 110	Rami et al. (1998)
	IS2807 × 249	RIL: 90	
Hybrid-related traits			
Fertility restoration			
Rf_1	ATx623 × RTx432	F_2: 373	Klein et al. (2001)
Rf_2	R931945-2-2 × IS8525	RIL: 285	Jordan et al. (2010)
	B923296 × SC170-6-8	RIL: 233	
rf_4	(A3Tx398*4/IS1112C//B3Tx398)	BC_3F_1: 378	Wen et al. (2002)
Rf_5	BTx642 × QL12	RIL: 218	Jordan et al. (2011)
Seed-/grain-related traits			
Seed weight	Tx7078 × B35	HILs	Tuinstra et al. (1997)

(continued)

Table 6.6 (continued)

Trait	Mapping population	Population size	Reference
Grain shattering	*S. bicolor* × *S. propinquum*	F_2: 370	Wise et al. (2002)
Grain color	Shan Qui Red × SRN39	RIL: 153	Knoll et al. (2008)
	B35 × Tx7000	RIL: 98	Xu et al. (2000)
Grain testa	IS2807 × 379	RIL: 110	Dufour et al. (1997),
	IS2807 × 249	RIL: 91	Rami et al. (1998)
Grain texture	IS2807 × 379	RIL: 110	Boivin et al. (1999)
	IS2807 × 249	RIL: 91	
	QL39 × QL41	RIL: 160	Tao et al. (2000)
Awns	IS2807 × 379	RIL: 110	Boivin et al. (1999)
	IS2807 × 249	RIL: 91	Tao et al. (2000)
	BTx623 × IS3620C	RIL: 137	Hart et al. (2001)
Glumes	296B × IS18551	RIL: 168	Srinivas et al. (2009b)
Grain pericarp color	QL39 × QL41	RIL: 160	Tao et al. (2000)
Grain protein digestibility	Sureno × P850029	RIL: 277	Winn et al. (2009)
Endosperm texture (waxy)	BTxARG1 x QL39	F_2	McIntyre et al. (2008)
	RTx2907 x QL39		
Biofuel-related traits			
Sugar-related traits	Early Folger × N32B	$F_{2:3}$: 207	Yun-long et al. (2006)
	R9188 × R9403463-2-1	RIL: 184	Ritter et al. (2008)
Brown midrib			
bmr6	Brown County × *bmr6-ref* line	RIL: 218	Saballos et al. (2009)
bmr12	bmr12 × N12	NIL	Bout and Vermerris (2003)
bmr2	AMP11 × Theis	F_2: 200	Saballos et al. (2012)
Biotic stress tolerance			
Head smut resistance	HC325 × RTx7078	F_2: 52	Oh et al. (1994)
Rust resistance	QL39 × QL41	RIL: 160	Tao et al. (1998) McIntyre et al. (2004)
Grain mold tolerance	Sureño × RTx430	RIL: 125	Klein et al. (2001)
Greenbug tolerance	GBIK × Redlan	RIL: 93	Agrama et al. (2002)
	Westland A line × PI550610	F_2: 217	Wu and Huang (2008)
	BTx623 × PI 607900	F_2: 371	Punnuri et al. (2013)
Midge resistance	ICSV745 × 90562	RIL: 120	Tao et al. (2003)
Sorghum aphid resistance	BTx623 × Henong 16	F_3: 64 (homozygous susceptible) and 571 F_4 seedlings	Wang et al. (2013)
Striga resistance	IS9830 × E36-1	RIL: 226	Haussmann et al. (2004)
	N13 × E36-1		
Head bug resistance	Malisor 84–7 × S 34	F_2: 217	Deu et al. (2005)
Anthracnose resistance	HC136 × G73	F_2: 110	Monika Singh et al. (2006)
	BTx623 × SC748-5	$F_{2:3}$	Perumal et al. (2009)
Ergot resistance	R931945-2-2 × IS8525	RIL: 146	Parh et al. (2008)
Stalk rot resistance	IS22380 × E36-1	RIL: 93	Srinivasa Reddy et al. (2008)

(continued)

Table 6.6 (continued)

Trait	Mapping population	Population size	Reference
Shoot fly resistance	296B × IS18551	RIL: 168	Satish et al. (2009)
	27B × IS2122	RIL: 210	Aruna et al. (2011)
Foliar diseases	296B × IS18551	RIL: 168	Murali Mohan et al. (2010)
Abiotic stress tolerance			
Drought tolerance	Tx7078 × B35	$F_{5:8}$ HIFs: 98	Tuinstra et al. (1998)
	B35 × Tx430	RIL: 96	Crasta et al. (1999)
	SC56 × Tx7000	RIL: 125	Kebede et al. (2001)
	B35 × Tx7000	RIL: 98	Sanchez et al. (2002)
	E36-1 × SPV570	RIL: 184	Rajkumar et al. (2013)
Stay green (drought)	B35 × Tx7000	RIL: 98	Xu et al. (2000)
	QL39 × QL41	RIL: 152	Tao et al. (2000)
	IS9830 × E36-1	RIL: 226	Haussmann et al. (2002)
	N13 × E36-1	RIL: 226	
	BTx642 × RTx7000	NIL	Harris et al. (2007)
	296B × IS18551	RIL: 168	Srinivas et al. (2008)
Early-season cold tolerance	Shan Qui Red × SRN39	RIL: 153	Knoll et al. (2008)
Aluminum tolerance	BR007 × SC283	RIL: 354	Magalhaes et al. (2007)
Bloom (drought)	BTx623 × KFS2021	F_2: 220	Burow et al. (2009)
Other traits			
Stem and leaf structural carbohydrates	BTx623 × Rio	RIL: 176	Murray et al. (2008)
Fiber-related traits	SS79 × M71	RIL: 188	Shiringani and Friedt (2011)
Preharvest sprouting resistance	Redland B2 × IS9530	F_2	Lijavetzky et al. (2000)

RIL recombinant inbred line, *NIL* near isogenic line, *HIL* heterogeneous inbred lines of NILs

known as wide diallel population derived from 19 founder lines of sorghum selected from a wide gene pool was used to map the heterotic trait locus and to identify intra-locus interactions underlying hybrid vigor (Ben-Israel et al. 2012).

Natural Populations Linkage analysis involves analyzing a limited number of recombination events that occur during the construction of mapping populations, which results in the localization of QTL in the interval of 10–20 cM. Moreover, cost is involved in the propagation and evaluation of a large number of lines (Doerge 2002; Holland 2007). While several linkage analysis studies have been conducted in sorghum over the past two decades, only a limited number of genes were cloned or tagged at the gene level (Bout and Vermerris 2003; Saballos et al. 2009, 2012). Association mapping, also known as LD mapping, has emerged as an important tool to dissect the genetics of complex traits at the sequence level by exploiting the recombination events accumulated in the natural population (germplasm lines) during the course of evolution (Nordborg and Tavaré 2002; Risch and Merikangas 1996). According to Yu and Buckler (2006), association mapping has three advantages as compared to conventional linkage analysis, viz., (1) better mapping resolution, (2) reduction in research time, and (3) access to greater allele number. In addition, association mapping enables researchers to use next-generation sequencing technologies to exploit the diversity present in the natural populations, the value of

which is known to crop breeders but exploited to a limited extent.

Sorghum is most suitable for association mapping of complex traits since it harbors one fourth sequence diversity that of maize (Hamblin et al. 2006), a 26-fold less population recombination than in maize (Hamblin et al. 2005), and natural homozygosity. LD is extensive enough in sorghum, which allows the simplification of a large number of SNPs into a smaller number of haplotypes resulting in reduced genotyping costs and increased statistical power (Clark 2004). Association mapping studies in sorghum using natural populations have been reported for plant growth and development (Kong et al. 2013), plant height (Murray et al. 2009; Wang et al. 2012), grain quality (de Alencar Figueiredo et al. 2010; Sukumaran et al. 2012), morphological traits (Shehzad et al. 2009), Brix (Murray et al. 2009), kernel weight and tiller number (Upadhyaya et al. 2012a), plant height and maturity (Upadhyaya et al. 2012b), endosperm quality (Hamblin et al. 2007), and agroclimatic traits (Morris et al. 2013).

Assembly of the association mapping panel comprising sorghum germplasm accessions possessing extensive genetic diversity is an important prerequisite for any association mapping study. Few association mapping panels were developed by different sorghum research groups across the world. An association mapping panel comprising 377 accessions representing all major cultivated races and important US breeding lines along with their progenitors was assembled and characterized for genetic and phenotypic diversity (Casa et al. 2008), which serves as an important genetic resource for the sorghum research community. Interested researchers can use this association panel and phenotype for their trait of interest without the need for further genotyping since the genotypic data along with appropriate statistical models are available for ready use.

A mini core comprising 242 accessions was developed from a core collection of 2,247 accessions through hierarchical cluster analysis using the phenotypic distances estimated from 11 qualitative and 10 quantitative traits and selecting about 10 % or a minimum of one accession per cluster covering a total of 21 clusters. Statistical comparisons based on homogeneity of distribution for geographical origin, biological races, qualitative traits, means, variances, phenotypic diversity indices, and phenotypic correlations indicated that the mini core collection represented the core collection (Upadhyaya et al. 2009). In addition to this, a reference set was developed comprising 374 sorghum accessions through the Generation Challenge Program as a means to enhance utilization of genetic resources in crop improvement (http://www.icrisat.org/what-we-do/crops/sorghum/Sorghum_Reference.htm).

A sorghum diversity research set (SDRS) comprising 107 sorghum accessions representing geographically diverse accessions from 27 countries in Asia and Africa was developed by Shehzad et al. (2009) through the analysis of the genetic diversity of 320 sorghum germplasm accessions with a set of 38 selected SSR markers based on three different published SSR linkage maps of sorghum. A sweet sorghum panel (SSP) was assembled by Murray et al. (2009) comprising 125 diverse accessions, which are primarily old and modern sweet sorghum cultivars along with a few grain and forage sorghums.

A core collection composed of 195 sorghum accessions originating from 39 countries and belonging to the five basic and ten intermediate races representative of the genetic diversity of core sample of 210 cultivated sorghum genotypes reported by Deu et al. (2006) was used for the association mapping of grain quality such as amylose content, protein content, lipid content, hardness, endosperm texture, and peak gelatinization temperature along with grain yield (de Alencar Figueiredo et al. 2010). In addition to the landrace collection described by Deu et al. (2006), an additional 45 inbred lines including donors for aluminum tolerance (Caniato et al. 2007) were used for the association mapping for aluminum tolerance to gain insights into the origin and evolution of aluminum tolerance and to detect functional variants (Caniato et al. 2014). Analysis of recombinant haplotypes suggested that causative polymorphisms are in introns and a transposon (MITE) insertion localized to a ~6 kb region,

which is positively correlated with aluminum tolerance. However, the SNP located in the second intron of *SbMATE*, an Al-activated root citrate efflux transporter, exhibited the strongest association signal and recovered 80 % of all the aluminum-tolerant accessions in the association panel.

6.7 Mutant Resources for Analyzing Gene Function

Exploitation of natural or induced genetic variability is considered as a successful strategy in crop improvement in many food crops. Mutagenesis as a tool to create novel variation is particularly important for those crops with limited variability. Over the years, several varieties have been developed in major food crops through mutation breeding programs. Ever since H.J. Muller reported induced mutation in 1937, analysis of mutants remained an effective approach to understand the gene function (Springer 2000; Stanford et al. 2001). The rapid accumulation of genomic sequence information in the past decade has brought the reverse genetic approaches into prominence, thereby directly probing the function of specific genes by testing the in vivo consequence of disruption or overexpression of a gene on the phenotype of an organism (Tierney and Lamour 2005). This is the "reverse" of conventional approach where phenotypes are observed and then the gene responsible for that phenotype is cloned and validated.

Mutant lines are important bioresources in this regard, which can potentially accelerate the understanding of gene function through reverse genetics. Kuromori et al. (2009) while reviewing the available mutant resources for phenome analysis in plant species highlighted the importance of mutant bioresource across crop species. With the availability of various analytical platforms (particularly bioinformatics), it is feasible to discover genes involved in particular phenotypic changes. Logically, these genes need to be functionally tested in collection of mutant resources in a high-throughput manner, called phenome analysis (Alonso and Ecker 2006). Chemical mutagens, like ethyl methanesulfonate (EMS), sodium azide, and methylnitrosourea (MNU), and physical mutagens, like fast neutrons, gamma rays, and ion beam irradiation, have been used extensively to generate mutant populations since the report of the first induced mutation in 1937. However, none of these mutant populations have been systematically annotated and preserved (Sree-Ramulu 1970; Porter et al. 1978; Jenks et al. 1994). Thus, the generated resource could not be combined with genomics tools.

*T*argeting *i*nduced *l*ocal *l*esions *in g*enomes (TILLING) was developed as a general reverse genetic tool to derive an allelic series of induced point mutations in genes of interest (Till et al. 2004, 2006). TILLING allows rapid and low-cost discovery of induced point mutations in populations of chemically mutagenized individuals in a high-throughput manner. This has been utilized to identify mutations in genes of interest in different crops (Till et al. 2004, 2007; Talame et al. 2008; Lababidi et al. 2009) including sorghum (Xin et al. 2008, 2009; Blomstedt et al. 2012). TILLING resources have been created for various crop plants across laboratories (Barkley and Wang 2008). Xin et al. (2008) were first to report the creation of TILLING resource in sorghum through EMS mutagenesis of sorghum cultivar, BTx623. They documented the feasibility of this approach by screening the mutant population for alterations in the genes of agronomic value not associated with cyanogenesis. Recently, Blomstedt et al. (2012) developed an acyanogenic forage line, P414L, with a point mutation in *the CYP79A1 gene* of cyanogenesis biochemical pathway by combining biochemical screen and TILLING approach. In the recent years, several TILLING populations have been developed globally by different sorghum research groups, which can serve as a valuable bioresource useful in understanding gene function and high-throughput SNP discovery.

An Annotated Individually pedigreed Mutated Sorghum (AIMS) library comprising 6144 pedigreed M_4 seed pools developed through single-seed descent from individual mutagenized seeds was established (Xin et al. 2008, 2009), which contains many biologically and agronomically important mutants, such as brown midrib (*bmr*)

mutants for improved biomass digestibility and ethanol yield and erect leaf (*erl*) mutants for improved capture of canopy radiation and hence biomass yield (Xin et al. 2009; Saballos et al. 2012; Sattler et al. 2012). An array of useful mutations harboring a wide range of phenotypic variation, including dwarfness, earliness, high protein digestibility, high lysine, etc., that were reported earlier in sorghum (Singh and Axtell 1973; Quinby 1975; Ejeta and Axtell 1985; Oria et al. 2000) were collected and preserved by the late Dr. Keith Schertz, a former sorghum geneticist with USDA-ARS, and this was released recently as a collection of genetic stocks through Germplasm Resources Information Network (Xin et al. 2013; www.ars-grin.gov), which is a valuable and vital resource for future genomic studies in sorghum.

6.8 Future Prospects

In the current era of crop improvement involving efficient integration of genetic information with the genomic and bioinformatics resources, focus should be on the coordination of various sorghum research groups across the globe on the sharing and utilization of resources available in the public domain. Genomics offers practical advantages for breeding cultivars by providing access to genetic variation through molecular markers and the potential to accurately measure the gene expression. With the initiation of re-sequencing projects in sorghum, whole-genome sequence information of sorghum cultivars possessing different end uses will be available in the near future resulting in the possibility of providing "genotype genomics" services that will contribute to sorghum improvement through "breeding by design." There is a need for the development of a single platform for sorghum, which can integrate the data that are scattered in different databases and also the software tools available for analysis. The main challenges facing the bioinformaticians are the development and management of databases and computational tools for data analysis in such a way that the user can define the target data and select the computational tool in order to get the output in a suitable format. The availability of different platforms for sequencing demands the development of novel algorithms and computational tools for an efficient assembly, annotation, and analysis. The application of molecular markers, comparative genomics, and annotation tools will greatly assist the identification of the genetic variation underlying an increasing number of agronomic traits and assist in the further agronomic improvement of a variety of crops.

References

Agrama HA, Widle GE, Reese JC, Campbell LR, Tuinstra MR (2002) Genetic mapping of QTLs associated with greenbug resistance and tolerance in *Sorghum bicolor*. Theor Appl Genet 104:1373–1378

Ahnert D, Lee M, Austin D, Livini C, Woodman W, Openshaw S, Smith J, Porter K, Dalon G (1996) Genetic diversity among elite sorghum inbred lines assessed with DNA markers and pedigree information. Crop Sci 36:1385–1392

Alonso JM, Ecker JR (2006) Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in *Arabidopsis*. Nat Rev Genet 7:524–536

Arun SS (2006) *In silico* EST data mining for elucidation of repeats biology and functional annotation in sorghum [*Sorghum bicolor* (L.) Moench.]. M.Sc. thesis. University of Agricultural Sciences, Dharwad

Aruna C, Bhagwat VR, Madhusudhana R, Sharma V, Hussain T, Ghorade RB, Khandalkar HG, Audilakshmi S, Seetharama N (2011) Identification and validation of genomic regions that affect shoot fly resistance in sorghum [*Sorghum bicolor* (L.) Moench]. Theor Appl Genet 122:1617–1630

Aruna C, Priya AR, Neeraja CN, Patil JV, Visarada KBRS (2012) Diversity analysis using ISSR markers for resistance to shoot pests in sorghum. Crop Prot 35:110–117

Ayana A, Bryngelsson T, Bekele E (2000) Geographic and altitudinal allozyme variation in sorghum [*Sorghum bicolor* (L.) Moench] landraces from Ethiopia and Eritrea. Hereditas 135:1–12

Barker G, Batley J, O'Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. Bioinformatics 19:421–422

Barkley NA, Wang ML (2008) Application of TILLING and EcoTILLING as reverse genetic approaches to elucidate the function of genes in plants and animals. Curr Genomics 9:212–226

Barrett T, Troup DB, Wilhite SE et al (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res 37:D885–D890

Batley J, Edwards D (2007) SNP applications in plants. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De

Silva HN (eds) Association mapping in plants. Springer, New York, pp 95–102
Batley J, Barker G, O'Sullivan H, Edwards KJ, Edwards D (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. Plant Physiol 132:84–91
Batley J, Jewell E, Edwards D (2007) Automated discovery of single nucleotide polymorphism (SNP) and simple sequence repeat (SSR) molecular genetic markers. In: Edwards D (ed) Plant bioinformatics. Humana Press, Totowa, pp 473–494
Bekele WA, Wieckhorst S, Friedt W, Snowden RJ (2013) High-throughput genomics in sorghum: from whole-genome resequencing to a SNP screening array. Plant Biotechnol J 11(9):1112–1125
Ben-Israel I, Kilian B, Nida H, Fridman E (2012) Heterotic trait locus (HTL) mapping identifies intra-locus interactions that underlie reproductive hybrid vigor in *Sorghum bicolor*. PLoS One 7:e38993
Bentsink L, Hanson J, Hanhart CJ et al (2010) Natural variation for seed dormancy in *Arabidopsis* is regulated by additive genetic and molecular pathways. Proc Natl Acad Sci U S A 107:4264–4269
Bhattramakki D, Dong J, Chhabra AK, Hart G (2000) An integrated SSR and RFLP linkage map of *Sorghum bicolor* (L.) Moench. Genome 43:988–1002
Billot C, Rivallan R, Sall MN et al (2012) A reference microsatellite kit to assess for genetic diversity of *Sorghum bicolor* (Poaceae). Am J Bot 99:e245–e250
Blomstedt CK, Gleadow RM, O'Donnell N et al (2012) A combined biochemical screen and TILLING approach identifies mutations in *Sorghum bicolor* L. Moench resulting in acyanogenic forage production. Plant Biotechnol J 10:54–66
Boivin K, Deu M, Rami JF, Trouche G, Hamon P (1999) Towards a saturated sorghum map using RFLP and AFLP markers. Theor Appl Genet 98:320–328
Bout S, Vermerris W (2003) A candidate-gene approach to clone the sorghum Brown midrib gene encoding caffeic acid O-methyltransferase. Mol Genet Genomics 269:205–214
Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, Nordborg M, Bergelson J, Cuguen J, Roux F (2010) Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS Genet 6:1–17
Brenner S, Johnson M, Bridgham J et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630–634
Brown SM, Hopkins MS, Mitchell SE, Senior ML, Wang TY, Duncan RR, Gonzalez-Candelas F, Kresovich S (1996) Multiple methods for the identification of polymorphic simple sequence repeats (SSRs) in sorghum [*Sorghum bicolor* (L.) Moench]. Theor Appl Genet 93:190–198
Brown NP, Leroy C, Sander C (1998) MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 14:380–381
Brown PJ, Klein PE, Bortiri E, Acharya CB, Rooney WL, Kresovich S (2006) Inheritance of inflorescence architecture in sorghum. Theor Appl Genet 113:931–942
Buchanan CD, Lim S, Salzman RA et al (2005) Sorghum bicolor's transcriptome response to dehydration, high salinity and ABA. Plant Mol Biol 58:699–720
Buckler ES, Holland JB, Bradbury PJ et al (2009) The genetic architecture of maize flowering time. Science 325:714–718
Burow GB, Franks CD, Acosta-Martinez V, Xin ZG (2009) Molecular mapping and characterization of BLMC, a locus for profuse wax (bloom) and enhanced cuticular features of Sorghum [*Sorghum bicolor* (L.) Moench.]. Theor Appl Genet 118:423–431
Calviño M, Bruggmann R, Messing J (2011) Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems. BMC Genom 12:356
Caniato FF, Guimaraes CT, Schaffert RE, Alves VMC, Kochian LV, Borem A, Klein PE, Magalhaes JV (2007) Genetic diversity for aluminum tolerance in sorghum. Theor Appl Genet 114:863–876
Caniato FF, Hamblin MT, Guimaraes CT, Zhang Z, Schaffert RE, Kochian LV, Magalhaes JV (2014) Association mapping provides insights into the origin and the fine structure of the sorghum aluminum tolerance locus, Alt_{SB}. PLoS One 9(1):e87438
Casa AM, Mitchell SE, Jensen JD, Hamblin MT, Paterson AH, Aquardo CF, Kresovich S (2006) Evidence for a selective sweep on chromosome 1 of cultivated sorghum. Crop Sci 46:S27–S40
Casa AM, Pressoir G, Brown PJ, Mitchell SE, RooneyWL TMR, Franks CD, Kresovich S (2008) Community resources and strategies for association mapping in sorghum. Crop Sci 48:30–40
Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11:215–221
Chagne D, Forster JW, Cogan NOI, Batley J, Edwards D (2007) Single nucleotide polymorphism discovery. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds) Association mapping in plants. Springer, New York, pp 53–76
Chantereau J, Trouche G, Rami JF, Deu M, Barro C, Grivet L (2001) RFLP mapping of QTLs for photoperiod response in tropical sorghum. Euphytica 120:183–194
Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31:3497–3500
Childs KL, Miller FR, Cordonnier-Pratt MM, Pratt LH, Morgan PW, Mullet JE (1997) The sorghum photoperiod sensitive gene, Ma_3, encodes a phytochrome B. Plant Physiol 113:611–619
Clark AG (2004) The role of haplotypes in candidate gene studies. Genet Epidemiol 27:321–333
Conte MG, Gaillard S, Lanau N, Rouard M, Perin C (2008) GreenPhylDB: a database for plant comparative genomics. Nucleic Acids Res 36:D991–D998
Crasta OR, Xu WW, Rosenow DT, Mullet J, Nguyen HT (1999) Mapping of post-flowering drought resistance traits in grain sorghum: association between QTLs

influencing premature senescence and maturity. Mol Gen Genet 262:579–588

de Alencar Figueiredo L, Sine B, Chantereau J, Mestres C, Fliedel G, Rami JF, Glaszmann JC, Deu M, Courtois M (2010) Variability of grain quality in sorghum: association with polymorphism in *Sh2*, *Bt2*, *SssI*, *Ae1*, *Wx* and *O2*. Theor Appl Genet 121:1171–1185

de Hoon M, Hayashizaki Y (2008) Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. Biotechniques 44:627–628

Deu M, Ratnadass A, Hamada MA, Noyer JL, Diabate M, Chantereau J (2005) Quantitative trait loci for headbug resistance in sorghum. Afr J Biotechnol 4:247–250

Deu M, Rattunde F, Chantereau J (2006) A global view of genetic diversity in cultivated sorghums using a core collection. Genome 49:168–180

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3:43–52

Dong Q, Lawrence CJ, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C, Brendel V (2005) Comparative plant genomics resources at PlantGDB. Plant Physiol 139:610–618

Du JF, Wu YJ, Fang XF, Cao JX, Zhao L, Tao SH (2010) Prediction of sorghum miRNAs and their targets with computational methods. Chinese Sci Bull 55:1263–1270

Dufour P, Deu M, Grivet L, D'Hont A, Paulet F, Bouet A, Lanaud C, Glaszmann JC, Hamon P (1997) Construction of a composite sorghum genome map and comparison with sugarcane, a related complex polyploïd. Theor Appl Genet 94:409–418

Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, Lushbough C, Brendel V (2007) PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res 36:D959–D965

Edwards D, Batley J, Cogan NOI, Forster JW, Chagne D (2007) Single nucleotide polymorphism discovery. In: Oraguzie NC, Rikkerink EHA, Gardiner SE, De Silva HN (eds) Association mapping in plants. Springer, New York, pp 53–76

Ejeta G, Axtell J (1985) Mutant gene in sorghum causing leaf "reddening" and increased protein concentration in the grain. J Hered 76:301–302

El Mannai Y, Shehzad T, Okuno K (2012) Mapping of QTLs underlying flowering time in sorghum [*Sorghum bicolor* (L.) Moench]. Breed Sci 62:151–159

Gabaldon T, Dessimoz C, Huxley-Jones J, Vilella A, Sonnhammer E, Lewis S (2009) Joining forces in the quest for orthologs. Genome Biol 10:403

Geleta N, Labuschangne MT, Chris D, Viljoen CD (2006) Genetic diversity analysis in sorghum germplasm as estimated by AFLP, SSR and morpho-agronomical markers. Biodivers Conserv 15:3251–3265

Girma Y (2009) Mining genomic resources for SNP and SNP-CAPS markers and divergence for drought tolerance in sorghum [*Sorghum bicolor* (L.) Moench]. M.Sc. (Ag) Thesis submitted to the University of Agricultural Sciences, Dharwad, India

Goff SA, Ricke D, Lan TH et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *japonica*). Science 296:92–100

Goodstein DM, Shu S, Howson R et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186

Hamblin MT, Mitchell SE, White GM, Gallego J, Kukatla R, Wing RA, Paterson AH, Kresovich S (2004) Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics 167:471–483

Hamblin MT, Salas Fernandez MG, Casa AM, Mitchell SE, Paterson AH, Kresovich S (2005) Equilibrium processes cannot explain high levels of short- and medium-range linkage disequilibrium in the domesticated grass *Sorghum bicolor*. Genetics 171:1247–1256

Hamblin MT, Casa AM, Sun H, Murray SC, Paterson AH, Aquardo CF, Kresovich S (2006) Challenges of detecting directional selection after a bottleneck: lessons from *Sorghum bicolor*. Genetics 173:953–964

Hamblin MT, Fernandez MGS, Tuinstra MR, Rooney WL, Kresovich S (2007) Sequence variation at candidate loci in the starch metabolism pathway in sorghum: prospects for linkage disequilibrium mapping. Crop Sci 47:S125–S134

Harbers M, Carninci P (2005) Tag-based approaches for transcriptome research and genome annotation. Nat Methods 2:495–502

Harris K, Subudhi PK, Borrell A, Jordan D, Rosenow D, Nguyen H, Klein P, Klein R, Mullet J (2007) Sorghum stay-green QTL individually reduce post-flowering drought-induced leaf senescence. J Exp Bot 58:327–338

Hart GE, Schertz KF, Peng Y, Syed NH (2001) Genetic mapping of *Sorghum bicolor* (L.) Moench QTLs that control variation in tillering and other morphological characters. Theor Appl Genet 103:1232–1242

Haussmann BIG, Mahalakshmi V, Reddy BVS, Seetharama N, Hash CT, Geiger HH (2002) QTL mapping of stay-green in two sorghum recombinant inbred populations. Theor Appl Genet 106:133–142

Haussmann BIG, Hess DE, Omanya GO, Folkertsma RT, Reddy BVS, Kayentao M, Welz HG, Geiger HH (2004) Genomic regions influencing resistance to the parasitic weed *Striga hermonthica* in two recombinant inbred populations of sorghum. Theor Appl Genet 109:1005–1016

Holland JB (2007) Genetic architecture of complex traits in plants. Curr Opin Plant Biol 10:156–161

Huang X, Paulo MJ, Boer M, Effgen S, Keizer P, Koornneef M, Eeuwijk FV (2011) Analysis of natural

allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. Proc Natl Acad Sci U S A 108:4488–4493

International Barley Genome Sequencing Consortium, Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, Close TJ, Wise RP, Stein N (2012) A physical, genetic and functional sequence assembly of the barley genome. Nature 491:711–716

Jaikishan I, Paik GR, Madhusudhana R, Elangovan M, Rajendrakumar P (2013) Development of microsatellite markers targeting $(GATA)_n$ motifs in sorghum [*Sorghum bicolor* (L.) Moench]. Mol Breed 31:223–231

Jenks MA, Joly RJ, Peters PJ, Rich PJ, Axtell JD, Ashworth EN (1994) Chemically induced cuticle mutation affecting epidermal conductance to water vapor and disease susceptibility in *Sorghum bicolor* (L.) Moench. Plant Physiol 105:1239–1245

Jordan DR, Mace ES, Henzell RG, Klein PE, Klein RR (2010) Molecular mapping and candidate gene identification of the Rf_2 gene for pollen fertility restoration in sorghum [*Sorghum bicolor* (L.) Moench]. Theor Appl Genet 120:1279–1287

Jordan DR, Klein RR, Sakrewski KG, Henzell RG, Klein PE, Mace ES (2011) Mapping and characterization of Rf_5: a new gene conditioning pollen fertility restoration in A_1 and A_2 cytoplasm in sorghum [*Sorghum bicolor* (L.) Moench]. Theor Appl Genet 123:383–396

Jordan D, Mace E, Cruikshank AW, Hunt CH, Hammer GL, Henzell RG (2012) Development and use of a sorghum nested association mapping population. Poster presented at International Plant and Animal Genome Conference, Jan., 14–18, 2012, San Diego, CA

Kebede H, Subudhi PK, Rosenow DT, Nguyen HT (2001) Quantitative trait loci influencing drought tolerance in grain sorghum (*Sorghum bicolor* L. Moench). Theor Appl Genet 103:266–276

Kebrom TH, Burson BL, Finlayson SA (2006) Phytochrome B represses Teosinte Branched1 expression and induces sorghum axillary bud outgrowth in response to light signals. Plant Physiol 140:1109–1117

Keurentjes JJ, Bentsink L, Alonso-Blanco C, Hanhart CJ, Blankestijn-De Vries H, Effgen S, Vreugdenhil D, Koornneef M (2007) Development of a near-isogenic line population of *Arabidopsis thaliana* and comparison of mapping power with a recombinant inbred line population. Genetics 175:891–905

Keurentjes JJB, Willems G, van Eeuwijk F, Nordborg M, Koornneef M (2011) A comparison of population type used for QTL mapping in *Arabidopsis thaliana*. Plant Genet Res 9:185–188

Klein RR, Rodriguez-Herrera R, Schlueter JA, Klein PE, Yu ZH, Rooney WL (2001) Identification of genomic regions that affect grain-mould incidence and other traits of agronomic importance in sorghum. Theor Appl Genet 102:307–319

Klein RR, Mullet JE, Jordan DR, Miller FR, Rooney WL, Menz MA, Franks CD, Klein PE (2008) The effect of tropical sorghum conversion and inbred development on genome diversity as revealed by high-resolution genotyping. Crop Sci 48(S1):S12–S26

Knoll J, Gunaratna N, Ejeta G (2008) QTL analysis of early-season cold tolerance in sorghum. Theor Appl Genet 116:577–587

Kong L, Dong J, Hart GE (2000) Characteristics linkage-map positions and allelic differentiation of *Sorghum bicolor* (L.) Moench DNA simple sequence repeats (SSRs). Theor Appl Genet 101:438–448

Kong W, Jin H, Franks CD, et al. (2013) Genetic analysis of recombinant inbred lines for *Sorghum bicolor* × *Sorghum propinquum*. G3 (Bethesda) 3:101–108

Kuromori T, Takahashi S, Kondou Y, Shinozaki K, Matsui M (2009) Phenome analysis in plant species using loss-of-function and gain-of-function mutants. Plant Cell Physiol 50:1215–1321

Kuzniar A, van Ham RC, Pongor S, Leunissen JA (2008) The quest for orthologs: finding the corresponding gene across genomes. Trends Genet 24:539–551

Lababidi S, Mejlhede N, Rasmussen SK, Backes G, Al-Said W, Baum M, Jahoor A (2009) Identification of barley mutants in the cultivar Lux at the Dhn loci through TILLING. Plant Breed 128:332–336

Larkin MA, Blackshields G, Brown NP et al (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948

Li H, Hearne S, Banziger K, Li Z, Wang J (2010) Statistical properties of QTL linkage mapping in biparental genetic populations. Heredity 105:257–267

Lijavetzky D, Martinez MC, Carrari F, Hopp HE (2000) QTL analysis and mapping of pre-harvest sprouting resistance in sorghum. Euphytica 112:125–135

Lin Y, Schertz K, Paterson A (1995) Comparative analysis of QTLs affecting plant height and maturity across the Poaceae, in reference to an interspecific sorghum population. Genetics 141:391–411

Lyons E, Freeling M (2008) How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J 53:661–673

Lyons E, Pedersen B, Kane J, Freeling M (2008) The value of non-model genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the Rosids. Trop Plant Biol 1:181–190

Mace ES, Xia L, Jordan DR, Halloran K, Parh DK, Huttner E, Wenzl P, Kilian A (2008) DArT markers: diversity analyses and mapping in *Sorghum bicolor*. BMC Genom 9:26

Mace ES, Rami JF, Bouchet S, Klein PE, Klein RR, Kilian A, Wenzl P, Xia L, Halloran K, Jordan DR (2009) A consensus genetic map of sorghum that integrates multiple component maps and high-throughput diversity array technology (DArT) markers. BMC Plant Biol 9:13

Mace ES, Singh V, Van Oosterom EJ, Hammer GL, Hunt CH, Jordan DR (2012) QTL for nodal root angle in

sorghum (*Sorghum bicolor* L. Moench) co-locate with QTL for traits associated with drought adaptation. Theor Appl Genet 124:97–109

Mace ES, Tai S, Gilding EK et al (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. Nat Commun 4:2320

Madhusudhana R, Patil JV (2013) A major QTL for plant height is linked with bloom locus in sorghum [*Sorghum bicolor* (L.) Moench]. Euphytica 191:259–268

Magalhaes JV, Liu J, Guimaraes CT et al (2007) A gene in the multidrug and toxic compound extrusion (MATE) family confers aluminum tolerance in sorghum. Nat Genet 39:1156–1161

McIntyre CL, Hermann SM, Casu RE et al (2004) Homologues of the maize rust resistance gene Rp1-D are genetically associated with a major rust resistance QTL in sorghum. Theor Appl Genet 109:875–883

McIntyre CL, Drenth J, Gonzalez N, Henzell RG, Jordan DR (2008) Molecular characterization of the waxy locus in sorghum. Genome 51:524–533

McMullen MD, Kresovich S, Villeda HS et al (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740

Mochida K, Shinozaki K (2010) Genomics and bioinformatics resources for crop improvement. Plant Cell Physiol 51:497–523

Morris GP, Ramu P, Deshpande SP et al (2013) Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A 110:453–458

Mount DM (2004) Bioinformatics: sequence and genome analysis, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor

Muller HJ (1937) The biological effects of radiation, with especial reference to mutation. Act Sci Ind 11:477–494

Murali Mohan S, Madhusudhana R, Mathur K, Chakravarthi DVN, Rathore S, Nagaraja Reddy R, Satish K, Srinivas G, Sarada Mani N, Seetharama N (2010) Identification of quantitative trait loci associated with resistance to foliar diseases in sorghum [*Sorghum bicolor* (L.) Moench]. Euphytica 176:199–211

Murray SC, Rooney WL, Mitchell SE, Sharma A, Klein PE, Mullet JE, Kresovich S (2008) Genetic improvement of sorghum as a biofuel feedstock: II. QTL for stem and leaf structural carbohydrates. Crop Sci 48:2180–2193

Murray SC, Rooney WL, Hamblin MT, Mitchell SE, Kresovich S (2009) Sweet sorghum genetic diversity and association mapping for brix and height. Plant Gen 2:48–62

Mutegi E, Sagnard F, Semagn K, Deu M, Muraya M, Kanyenji B, de Villiers S, Kiambi D, Herselman L, Labuschagne M (2011) Genetic structure and relationships within and between cultivated and wild sorghum [*Sorghum bicolor* (L.) Moench] in Kenya as revealed by microsatellite markers. Theor Appl Genet 122:989–1004

Nagaraja Reddy R, Madhusudhana R, Murali Mohan S, Chakravarthi DVN, Seetharama N (2012) Characterization, development and mapping of unigene-derived microsatellite markers in sorghum [*Sorghum bicolor* (L.) Moench]. Mol Breed 29:543–564

Nagaraja Reddy R, Madhusudhana R, Murali Mohan S, Chakravarthi DV, Mehtre SP, Seetharama N, Patil JV (2013) Mapping QTL for grain yield and other agronomic traits in post-rainy sorghum [*Sorghum bicolor* (L.) Moench]. Theor Appl Genet 126:1921–1939

Nelson JC, Wang S, Wu Y, Li X, Antony G, White F, Yu J (2011) Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. BMC Genom 12:352

Nordborg M, Tavaré S (2002) Linkage disequilibrium: what history has to tell us? Trends Genet 18:83–90

Oh BJ, Frederiksen RA, MagiIl CW (1994) Identification of molecular markers linked to head smut resistance gene (*Shs*) in sorghum by RFLP and RAPD analyses. Phytopathology 84:830–833

Oria MP, Hamaker BR, Axtell JD, Huang CP (2000) A highly digestible sorghum mutant cultivar exhibits a unique folded structure of endosperm protein bodies. Proc Natl Acad Sci U S A 97:5065–5070

Palumbi SR (1995) Nucleic acids II: the polymerase chain reaction. In: Hillis D, Moritz C (eds) Molecular systematics, 2nd edn. Sinauer Associates Inc., Sunderland, pp 205–247

Parh DK, Jordan DR, Aitken EAB, Mace ES, Junai P, McIntyre CL, Godwin ID (2008) QTL analysis of ergot resistance in sorghum. Theor Appl Genet 117:369–382

Park SJ, Huang Y, Ayoubi P (2006) Identification of expression profiles of sorghum genes in response to greenbug phloem-feeding using cDNA subtraction and microarray analysis. Planta 223:932–947

Parkinson H, Kapushesky M, Shojatalab M et al (2007) ArrayExpress – a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 35:D747–D750

Paterson AH, Bowers JE, Bruggmann R et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Perumal R, Menz MA, Mehta PJ et al (2009) Molecular mapping of *Cg1*, a gene for resistance to anthracnose (*Colletotrichum sublineolum*) in sorghum. Euphytica 165:597–606

Phuong N, Stutzel H, Uptimoor R (2013) Quantitative trait loci associated to agronomic traits and yield components in a *Sorghum bicolor* L. Moench RIL population cultivated under pre-flowering drought and well-watered conditions. Agric Sci 4(12):781–791

Ping LX, Feng YJ, Ping GC, Acharya S (2011) Quantitative trait loci analysis of economically important traits in *Sorghum bicolor*×*S. sudanense* hybrid. Can J Plant Sci 91:81–90

Porter KS, Anxtell JD, Lechtenberg VL, Colenbrander VF (1978) Phenotype, fiber composition, and *in vitro* dry matter disappearance of chemically induced brown midrib (*bmr*) mutants of sorghum. Crop Sci 18:205–208

Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K (2009) PLAZA: a compara-

tive genomics resource to study gene and genome evolution in plants. Plant Cell 21:718–3731

Punnuri S, Huang Y, Steets J, Yanqi W (2013) Developing new markers and QTL mapping for greenbug resistance in sorghum [*Sorghum bicolor* (L.) Moench]. Euphytica 191:191–203

Quinby JR (1975) The genetics of sorghum improvement. J Hered 66:56–62

Rajkumar FB, Kavil SP et al (2013) Molecular mapping of genomic regions harbouring QTLs for root and yield traits in sorghum [*Sorghum bicolor* (L.) Moench]. Physiol Mol Biol Plants 19(3):409–419

Rakshit S, Rakshit A, Patil JV (2012) Multiparent intercross populations in analysis of quantitative traits. J Genet 91:111–117

Rami JF, Dufour P, Trouche G, Fliedel G, Mestres C, Davrieux F, Blancard P, Hamon P (1998) Quantitative trait loci for grain quality, productivity, morphological, and agronomical traits in sorghum (*Sorghum bicolor* L. Moench). Theor Appl Genet 97:605–616

Ramu P, Kassahun B, Senthilvel S, Kumar CA, Jayashree B, Folkertsma RT, Reddy LA, Kuruvinashetti MS, Haussmann BIG, Hash CT (2009) Exploiting rice-sorghum synteny for targeted development of EST-SSRs to enrich the sorghum genetic linkage map. Theor Appl Genet 119:1193–1204

Ramu P, Billot C, Rami JF, Senthilvel S, Upadhyaya HD, Ananda Reddy L, Hash CT (2013) Assessment of genetic diversity in the sorghum reference set using EST-SSR markers. Theor Appl Genet 126:2051–2064

Reddy PS, Fakrudin B, Rajkumar, Punnuri SM, Arun SS, Kuruvinashetti MS, Das IK, Seetharama N (2008) Molecular mapping of genomic regions harboring QTLs for stalk rot resistance in sorghum. Euphytica 159:191–198

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Ritter KB, Lynne McIntyre C, Godwin ID, Jordan DR, Chapman SC (2007) An assessment of the genetic relationship between sweet and grain sorghums, within *Sorghum bicolor* ssp. *bicolor* (L.) Moench, using AFLP markers. Euphytica 157:161–176

Ritter KB, Jordan DR, Chapman SC, Godwin ID, Mace ES, McIntyre CL (2008) Identification of QTL for sugar-related traits in sweet x grain sorghum (*Sorghum bicolor* L. Moench) recombinant inbred population. Mol Breed 22:367–384

Rouard M, Guignon V, Aluome C, Laporte MA, Droc G, Walde C, Zmasek CM, Perin C, Conte MG (2011) GreenPhylDB v2.0: comparative and functional genomics in plants. Nucleic Acids Res 39:D1095–D1102

Saballos A, Ejeta G, Sanchez E, Kang C, Vermerris W (2009) A genome-wide analysis of the cinnamyl alcohol dehydrogenase family in sorghum [*Sorghum bicolor* (L.) Moench] identifies *SbCAD2* as the brown midrib6 gene. Genetics 181:783–795

Saballos A, Sattler SE, Sanchez E, Foster TP, Xin Z, Kang C, Pedersen JF, Vermerris W (2012) Brown midrib2 (*bmr2*) encodes the major 4-coumarate: coenzyme A ligase involved in lignin biosynthesis in sorghum (*Sorghum bicolor* (L.) Moench). Plant J 70:818–830

Sanchez AC, Subudhi PK, Rosenow DT, Nguyen HT (2002) Mapping QTLs associated with drought resistance in sorghum (*Sorghum bicolor* L. Moench). Plant Mol Biol 48:713–726

Satish K, Srinivas G, Madhusudhana R, Padmaja PG, Nagaraja Reddy R, Murali Mohan S, Seetharama N (2009) Identification of quantitative trait loci for resistance to shoot a fly in sorghum [*Sorghum bicolor* (L.) Moench]. Theor Appl Genet 119:1425–1439

Sattler SE, Palmer NA, Saballos A, Greene AM, Xin Z, Sarath G, Vermerris W, Pedersen JF (2012) Identification and characterization of four missense mutations in brown midrib 12 (*bmr12*), the Caffeic O-Methyltransferase (*COMT*) of sorghum. Bioenergy Res 5:855–865

Savage D, Batley J, Erwin T, Logan E, Love CG, Lim GAC, Mongin E, Barker G, Spangenberg GC, Edwards D (2005) SNPServer: a real-time SNP discovery tool. Nucleic Acids Res 33:W493–W495

Schloss SJ, Mitchell SE, White GM, Kukatla R, Bowers JE, Paterson AH, Kresovich S (2002) Characterization of RFLP probe sequences for gene discovery and SSR development in *Sorghum bicolor* (L.) Moench. Theor Appl Genet 105:912–920

Schnable PS, Ware D, Fulton RS et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Shakoor N, Crasta OR, Nair R, Huang SW, Fan Z, Morris G, Kresovich S (2013) Whole genome gene expression profiling of multiple tissue types and developmental stages in *Sorghum bicolor*. Abstract: Plant and Animal Genome XXI Symposium, January 12–16, 2013, San Diego, CA, USA

Shakoor N, Nair R, Crasta OR, Morris G, Feltus A, Kresovich S (2014) A *Sorghum bicolor* expression atlas reveals dynamic genotype-specific expression profiles for vegetative tissues of grain, sweet and bioenergy sorghums. BMC Plant Biol 14:35

Shehzad T, Okuizumi H, Kawase M, Okuno K (2009) Development of SSR-based sorghum (*Sorghum bicolor* L. Moench) diversity research set of germplasm and its evaluation by morphological traits. Genet Resour Crop Evol 56:809–827

Shiringani AL, Friedt W (2011) QTL for fibre-related traits in grain×sweet sorghum as a tool for the enhancement of sorghum as a biomass crop. Theor Appl Genet 123:999–1011

Sievers F, Wilm A, Dineen D et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539

Singh R, Axtell JD (1973) High lysine mutant gene that improves protein quality and biological value of grain sorghum. Crop Sci 13:535–539

Singh M, Chaudhary K, Singal KR, Magill CW, Boora KS (2006) Identification and characterization of RAPD and SCAR markers linked to anthracnose resistance

gene in sorghum [*Sorghum bicolor* (L.) Moench]. Euphytica 149:179–187
Singhal D, Gupta P, Sharma P, Kashyap N, Anand S, Sharma H (2011) *In-silico* single nucleotide polymorphisms (SNP) mining of *Sorghum bicolor* genome. Afr J Biotechnol 10:580–583
Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A (2009) BioMart – biological queries made easy. BMC Genom 10:22
Springer PS (2000) Gene traps: tools for plant development and genomics. Plant Cell 12:1007–1020
Sree-Ramulu K (1970) Sensitivity and induction of mutations in sorghum. Mutat Res 10:197–206
Srinivas G, Satish K, Murali Mohan S, Nagaraja Reddy R, Madhusudhana R, Balakrishna D, Bhat BV, Howarth CJ, Seetharama N (2008) Development of genic-microsatellite markers for sorghum staygreen QTL using a comparative genomic approach with rice. Theor Appl Genet 117:283–296
Srinivas G, Satish K, Madhusudhana R, Seetharama N (2009a) Exploration and mapping of microsatellite markers from subtracted drought stress ESTs in *Sorghum bicolor* (L.) Moench. Theor Appl Genet 118:703–717
Srinivas G, Satish K, Madhusudhana R, Nagaraja Reddy R, Murali Mohan S, Seetharama N (2009b) Identification of quantitative trait loci for agronomically important traits and their association with genic-microsatellite markers in sorghum. Theor Appl Genet 118:1439–1454
Srinivasa Reddy P, Fakrudin B, Rajkumar, Punnuri SM, Arun SS, Kuruvinashetti MS, Das IK, Seetharama N (2008) Molecular mapping of genomic regions harboring QTLs for stalk rot resistance in sorghum. Euphytica 159:191–198
Stanford WL, Cohn JB, Cordes SP (2001) Gene-trap mutagenesis: past, present and beyond. Nat Rev Genet 2:756–768
Stein LD, Mungall C, Shu S et al (2002) The Generic Genome Browser, a building block for a model organism system database. Genome Res 12:1599–1610
Sukumaran S, Xiang W, Bean SR, Pedersen JF, Kresovich S, Tuinstra MR, Tesso TT, Hamblin MT, Yu J (2012) Association mapping for grain quality in a diverse sorghum collection. Plant Gen 5:126–135
Syvanen AC (2001) Accessing genetic variation genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942
Taillon-Miller P, Gu ZJ, Li Q, Hillier L, Kwok PY (1998) Overlapping genomic sequences: a treasure trove of single nucleotide polymorphisms. Genome Res 8:748–754
Talame V, Bovina R, Sanguineti MC, Tuberosa R, Lundqvist U, Salvi S (2008) TILLMore, a resource for the discovery of chemically induced mutants in barley. Plant Biotechnol J 6:477–485
Tao Y, Manners J, Ludlow M, Henzell R (1993) DNA polymorphisms in grain sorghum (*Sorghum bicolor* (L.) Moench). Theor Appl Genet 86:679–688
Tao YZ, Jordan DR, Henzell RG, McIntyre CL (1998) Identification of genomic regions for rust resistance in sorghum. Euphytica 103:287–292
Tao YZ, Henzell RG, Jordan DR, Butler DG, Kelly AM, McIntyre CL (2000) Identification of genomic regions associated with stay green in sorghum by testing RILs in multiple environments. Theor Appl Genet 100:1225–1232
Tao YZ, Hardy A, Drenth J, Henzell RG, Franzmann BA, Jordan DR, Butler DG, McIntyre CL (2003) Identifications of two different mechanisms for sorghum midge resistance through QTL mapping. Theor Appl Genet 107:116–122
Taramino G, Tarchini R, Ferrario S, Lee M, Pe ME (1997) Characterization and mapping of simple sequence repeats (SSRs) in *Sorghum bicolor*. Theor Appl Genet 95:66–72
The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815
Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882
Tierney MB, Lamour KH (2005) An introduction to reverse genetic tools for investigating gene function. Plant Health Instructor. doi:10.1094/PHI-A-2005-1025-01
Till BJ, Reynolds SH, Weil C et al (2004) Discovery of induced point mutations in maize genes by TILLING. BMC Plant Biol 4:12
Till BJ, Zerr T, Comai L, Henikoff S (2006) A protocol for TILLING and Ecotilling in plants and animals. Nat Protoc 1:2465–2477
Till BJ, Cooper J, Tai TH, Colowit P, Greene EA, Henikoff S, Comai L (2007) Discovery of chemically induced mutations in rice by TILLING. BMC Plant Biol 7:19
Tuinstra MR, Grote EM, Goldsbrough PB, Ejeta G (1997) Genetic analysis of post-flowering drought tolerance and components of grain development in sorghum. Mol Breed 3:439–448
Tuinstra MR, Ejeta G, Goldsbrough P (1998) Evaluation of near-isogenic sorghum lines contrasting for QTL markers associated with drought tolerance. Crop Sci 38:835–842
Upadhyaya HD, Pundir RPS, Dwivedi SL, Gowda CLL, Reddy VG, Singh S (2009) Developing a mini core collection of sorghum for diversified utilization of germplasm. Crop Sci 49:1769–1780
Upadhyaya H, Wang YH, Sharma S, Singh S, Hasenstein K (2012a) SSR markers linked to kernel weight and tiller number in sorghum identified by association mapping. Euphytica 187:401–410
Upadhyaya HD, Wang YH, Sharma S, Singh S (2012b) Association mapping of height and maturity across five environments using the sorghum mini core collection. Genome 55:471–479
Uptmoor R, Wenzel W, Friedt W, Donaldson G, Ayisi H, Ordon F (2003) Comparative analysis on the genetic relatedness of *Sorghum bicolor* accessions from

Southern Africa by RAPDs, AFLPs and SSRs. Theor Appl Genet 106:1316–1325

Varshney RK, Chen W, Li Y et al (2012) Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol 30:83–89

Varshney RK, Song C, Saxena RK et al (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol 31:240–246

Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487

Wang YH, Bible P, Loganantharaj R, Upadhyaya H (2012) Identification of SSR markers associated with height using pool-based genome-wide association mapping in sorghum. Mol Breed 30:281–292

Wang F, Zhao S, Han Y, Shao Y, Dong Z, Gao Y, Zhang K, Liu X, Li D, Chang J, Wang D (2013) Efficient and fine mapping of RMES1 conferring resistance to sorghum aphid *Melanaphis sacchari*. Mol Breed 31:777–784

Ware D, Jaiswal P, Ni J et al (2002) Gramene: a resource for comparative grass genomics. Nucleic Acids Res 30:103–105

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2 – a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189–1191

Wen L, Tang HV, Chen W, Chang R, Pring DR, Klein PE, Childs KL, Klein RR (2002) Development and mapping of AFLP markers linked to the sorghum fertility restorer gene rf_4. Theor Appl Genet 104:577–585

White GM, Hamblin MT, Kresovich S (2004) Molecular evolution of the phytochrome gene family in sorghum: changing rates of synonymous and replacement evolution. Mol Biol Evol 21:716–723

Winn JA, Mason RE, Robbins AL, Rooney WL, Hays DB (2009) QTL mapping of a high protein digestibility trait in *Sorghum bicolor*. Int J Plant Genomics 2009:471853

Wise MG, Schulze SR, Lin Y-R, Bowers JE, Okuizumi H, Schertz KF, Paterson AH (2002) Progress towards the positional cloning of the sorghum grain shattering gene. Plant and Animal Genome X, January 12–16, 2002, San Diego, USA

Wu Y, Huang Y (2008) Molecular mapping of QTLs for resistance to the greenbug *Schizaphis graminum* (Rondani) in *Sorghum bicolor* (Moench). Theor Appl Genet 117:117–124

Wu Y, Li X, Xiang W et al (2012) Presence of tannins in sorghum grains is conditioned by different natural alleles of Tannin1. Proc Natl Acad Sci U S A 109:10281–10286

Xin Z, Wang ML, Barkley NA, Burow G, Franks C, Pederson G, Burke J (2008) Applying genotyping (TILLING) and phenotyping analyses to elucidate gene function in a chemically induced sorghum mutant population. BMC Plant Biol 8:103

Xin Z, Wang ML, Burow G, Burke J (2009) An induced sorghum mutant population suitable for bioenergy research. Bioenergy Res 2:10–16

Xin Z, Burow G, Woodfin C, Franks CD, Klein RR, Schertz KF, Pederson GA, Burke JJ (2013) Registration of a diverse collection of sorghum genetic stocks. J Plant Reg 7:119–124

Xu W, Subudhi PK, Crasta OR, Rosenow DT, Mullet JE, Nguyen NT (2000) Molecular mapping of QTLs conferring stay-green in grain sorghum [*Sorghum bicolor* (L.) Moench]. Genome 43:461–469

Yang L, Jin G, Zhao X, Zheng Y, Xu Z, Wu W (2007) PIP: a database of potential intron polymorphism markers. Bioinformatics 23:2174–2177

Yang S, Weers BD, Morishige DT, Mullet JE (2014) *CONSTANS* is a photoperiod-regulated activator of flowering in sorghum. BMC Plant Biol 14:148

Yonemaru JI, Ando T, Mizzubayashi T, Kasuga S, Matsumoto T, Yano M (2009) Development of genome–wide simple sequence repeat markers using whole-genome shot-gun sequences of sorghum [*Sorghum bicolor* (L.) Moench]. DNA Res 16:187–193

Yu J, Buckler E (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160

Yu J, Hu S, Wang J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp *indica*). Science 296:79–92

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Yun-long B, Seiji Y, Maiko I, Wei CH (2006) QTLs for sugar content of stalk in sweet sorghum (*Sorghum bicolor* L. Moench). Agric Sci China 5:736–744

Zhang D, Guo H, Kim C, Lee TH, Li J, Robertson J, Wang X, Wang Z, Paterson AH (2013a) *CSGRqtl*, a comparative quantitative trait locus database for Saccharinae grasses. Plant Physiol 161:594–599

Zhang T, Zhao X, Huang L, Liu X, Zong Y, Zhu L, Yang D, Fu B (2013b) Tissue-specific transcriptomic profiling of sorghum propinquum using a rice genome array. PLoS One 8(3):e60202

Zheng LY, Guo XS, He B et al (2011) Genome-wide patterns of genetic variation in sweet and grain sorghum (*Sorghum bicolor*). Genome Biol 12:R114

Zou G, Zhai G, Feng Q et al (2012) Identification of QTLs for eight agronomically important traits using an ultra-high-density map based on SNPs generated from high-throughput sequencing in sorghum under contrasting photoperiods. J Exp Bot 63:5451–5462