Chapter 4
Bioinformatics Tools for Crop Research and Breeding

Jayashree B and Dave Hoisington

Abstract Crop improvement has always been, but will be even more so in the twenty-first century, an information intensive process. For effective and efficient improvement, a range of activities from molecular biology to genetics to indirect selection must now be involved. The rate of progress made by any breeding programme depends as much on the efficient integration of information from these activities as it does on the activities themselves. Plant breeders are now realizing the importance of innovative approaches that include the use of a range of molecular methods and their outputs, and the possibilities of transferring this information from model species to cultivated crops. The use of these high throughput methods in model crops has already generated a large amount of public resources such as databases containing genetic resource, genomic and genetic information; tools for the effective analysis, data mining and visualization of such information; and semantic web resources for data integration. In this chapter, we highlight the role and contributions of bioinformatics to crop research and breeding by focusing on the bioinformatics resources that are available for crop science research and breeding, and indicating gaps that need to be bridged that will allow scientists to access, transfer and integrate data with ease.

4.1 Introduction

The growing world-wide demand for food is placing increasing pressure on crop breeding programs to produce cultivars that can adapt to a range of environments without compromising on quality and yield. As such, crop breeding efforts focus on developing new varieties with improved resistance to fungal, insect or viral diseases, tolerance to abiotic stresses such as drought, cold, salt, dehydration, heavy metal toxicity and numerous quality attributes such as taste, size, shape, color and ease of cooking. In addition there is a growing need to provide for nutritional deficiencies, especially in the developing countries

B. Jayashree and D. Hoisington (✉)
International Crops Research Institute for the Semi Arid Tropics, Patancheru, 502 324,
Andhra Pradesh, India
e-mail: d.hoisington@CGIAR.ORG

S.M. Jain and D.S. Brar (eds.), *Molecular Techniques in Crop Improvement*,
DOI 10.1007/978-90-481-2967-6_4, © Springer Science+Business Media B.V. 2010

through biofortified crops. With an ever-increasing number of desirable traits that must be integrated into a cultivar, crop improvement programs are at an interesting juncture. The combination of existing knowledge and resources with modern structural and functional genomics provides the opportunity to study the genetic, biochemical and physiological basis of complex traits. Efforts currently center on capturing information from model and better-studied crops in order to define genes for important traits. With the advent of high throughput technologies a number of initiatives have emerged, for example large scale mapping studies, genome-wide expression studies and high throughput screening of genotypes and phenotypes along with corresponding bioinformatics resources. It is now becoming clear that crop improvement programs will benefit hugely from a judicious use of these resources coupled with crop genetic resources, which are the basic materials for breeding programs. Genetic resource collections available to breeders are being characterized for diversity so breeders can have access to 'core' collections that contain as much genetic variability as possible. The advances in plant genetics and genomics offer opportunities so far unavailable, for discovering the function of genes and the potential to manipulate them for crop improvement. With so much information being produced that could be of potential use to the breeder, the difficulty is in making sense of all the data so as to facilitate knowledge driven crop selection. Bioinformatics is emerging as the glue that brings these different kinds of data together; as a discipline it spans the realm from scientific software development to meaningful knowledge discovery. A review of current bioinformatics resources, tools and methods available for the purpose of crop improvement gives us an idea of ground covered so far and what is desirable to achieve in the coming years.

4.2 Bioinformatics Resources Available for Crop Research

The burgeoning information from genomics is due to innovative technologies like DNA microarrays, high throughput genotyping and Next Generation Sequencing. Most modern data generation projects have seen a concomitant development of databases to store, access and query data. These data resources are usually made available through the web, store varying kinds of information and are available at different locations. The very distributed nature of this information throws up interesting challenges – that of interoperability of databases that will allow data integration, the use or lack of common vocabularies that will allow comparison of the data and the varying levels of data annotation and curation available that reflects on data quality. Databases can no longer be passive storehouses of information, they need to link to various types of data to be useful.

4.2.1 Data Resources

There are a considerable number of quality databases devoted to crops that allow access to users through GUIs (Graphical User Interfaces). Amongst the online resources listing key databases of value is the *Nucleic Acids Research* online Molecular Biology database collection (http://www.oxfordjournals.org/nar/database/a/).

The number of databases in this collection is 1,170 as of January 2009, with 78 plant specific databases. The collection lists high quality, comprehensive databases with value added in the form of manual curation. The bioinformatics links directory (http://bioinformatics.ca/links directory) is an actively maintained compilation of servers hosting bioinformatics databases with features for improved navigation and accessibility.

Table 4.1 lists popular as well as lesser known crop species and multi-crop species databases covering genotype, phenotype, taxonomy and genomic information. Besides

Table 4.1 Species and clade specific crop databases

Database	Species	Primary site	Database contents
BeanGenes	Phaseolus and Vigna	http://beangenes.cws.ndsu.nodak.edu/	Genetic, germplasm, phenotypic and pathology data
CR-EST (crop EST)	Barley, pea, potato, petunia, tobacco, wheat	http://pgrc.ipk-gatersleben.de/est/index.php	Genomic data
FoggDB	Forage grasses	http://www.igergru.bbsrc.ac.uk/Welcome/IGER/foggdb/foggdb.htm	Genomic data
GDR (genome database for Rosaceae)	Apple, pear, prunus, raspberry, strawberry, prunus	http://www.bioinfo.wsu.edu/gdr/	Genomic data
Graingenes	Wheat, rye, barley, oats, sugarcane and relatives	http://wheat.pw.usda.gov/GG2/index.shtml	Genetic, genomic, expression, phenotypic and taxonomy data
Gramene	Rice, Sorghum, maize, wheat, rye, millets, *Arabidopsis*	http://www.gramene.org/	Genetic, genomic and pathway data
JCVI (TIGR)	25 crops	http://www.tigr.org/	Genomic data
LIS (Legume Information Service)	17 legume species	http://www.comparative-legumes.org/	Genetic and genomic data
MaizeGDB	Maize	http:/www.maizegdb.org	Genetic, genomic and phenotypic data
MIPSPlantsDB	Multispecies	http://mips.gsf.de/projects/plants	Genomic data
Soybase	Soybean	http://soybase.agron.iastate.edu/	Genetic, genomic and phenotypic data
TAIR (The Arabidopsis Information Resource)	Arabidopsis	http://www.arabidopsis.org/	Genetic, genomic and gene expression data
TIGR plant transcript assemblies (TA) database	Multispecies	http://plantta.tigr.org	EST and cDNA data
PlantGDB	Multispecies	http://www.plantgdb.org/	Genomic data
UKCropNet	Central multispecies database querying system	http://ukcrop.net/db.html	Genetic, genomic and pathway data

the databases listed in this table, highly specialized databases derived from the research on model crops are available on the web. PathoPlant® is a database on plant–pathogen interactions and signal transduction reactions using microarray gene expression data from *Arabidopsis thaliana* subjected to pathogen infection and elicitor treatment (http://www.pathoplant.de). The cereal small RNA DB (CSRDB) consists of large scale datasets of maize and rice smRNA generated by high throughput pyrosequencing, mapped to the rice and maize genomic sequence (http://sundarlab.ucdavis.edu/smrnas/). Resources for comparative genomics include the POGs/Plant RBP (putative orthologous groups/plant RNA binding proteins, http://plantrbp.uoregon.edu/), ATTED-11 (*A. thaliana* trans factor and *cis* element prediction database) with information on function and regulation of particular genes and gene networks (http://www.atted.bio.titech.ac.jp). The GABI-Kat SimpleSearch is an *Arabidopsis* T-DNA mutant database containing >108,000 mapped FSTs (flanking sequence tags) from ~64,000 lines which cover 64% of all annotated *A. thaliana* protein coding genes (http://www.GABI-Kat.de). The plantTFDB stores information on transcription factors predicted from 22 species: 5 model organisms and 17 plants (http://planttfdb.cbi.pku.edu.cn/). PlantQTL-GE is a database system for identifying candidate genes in rice and *Arabidopsis* by gene expression and QTL information. The database includes genes, gene expression information, ESTs and genetic markers from multiple sources (http://www.scbit.org/qtl2gene/new/). The plant promoter database provides promoter annotations in Arabidopsis and rice (http://www.ppdb.gene.nagoya-u.ac.jp). MetaCrop is a database of crop plant metabolism including pathway diagrams, reactions, transport processes and reaction kinetics besides taxonomy and literature (http://metacrop.ipk-gatersleben.de). All the databases referred to here have been published over the period 2006–2009 and show the differences in resources available on model crops as compared to orphan crops.

4.2.2 Web and Web Services

Most bioinformatics databases and analytical services are available through the Internet. The user may need to interact with many of these in concert to extract different kinds of data, and compare, integrate and format data for submission to an analysis program. Web interfaces are not really suited to handle bulk data export from databases and programmatic access to data is needed to retrieve large quantities of data and format it for submission to analytical tools. Thus data source providers have begun to allow multiple modes of data retrieval and view, from HTML (hypertext markup language), XML (extensible markup language), and SQL (structured query language), to SOAP (simple object access protocol, used in web services) besides allowing hook up to third party analysis tools. Markup languages like HTML and XML provide the means to describe the structure of text-based information. XML defines a way to add markup to information as well as assign meanings to data explicitly, thus facilitating machine readability. Where meaning is implicit only a person with knowledge about the data can understand and interpret it, but where meaning is explicit, the data becomes interpretable by retrieving software. Examples of databases that provide XML access include INSD_v1.4 that provides access to the EMBL/DDBJ/Genbank sequence records in XML, while GrainGenes (http://wheat.pw.usda.gov/cgi-bin/graingenes/sql.cgi) provides SQL access to its database. Web services provide a programmatic interface to databases and web-based tools and are increasingly being used to automate execution of the data retrieval and analysis steps. The users can look up XML-based web service registries that list name, products, locations and services offered by the web service provider on the Internet. Examples of popular bioinformatics web services projects include BioMoby (The BioMoby Consortium 2008) and myGrid (www.mygrid.org.uk). The web service registry here is different from traditional web services in that it uses the meaning of terms in the biological vocabulary (semantics) to mediate web service discovery and invocation. This helps overcome the problem inherent to biological data – that of inconsistent data type. The Virtual Plant Information network hosted at the NCGR is another network of data and service providers based on the semantic web services platform (http://vpin.ncgr.org/). This network differs from paradigmatic web services in that it does not use SOAP for information exchange but instead relies on http and the web ontology language (OWL-DL), a web standard for information processing. VPIN has a web front end that allows users to find disparate data and services based on lexical and semantic criteria. The DAS (Distributed Annotation System) is another data retrieval protocol that can be used for the exchange of biological sequence annotation. It allows a single machine to gather up sequence annotation information from multiple distant web sites, collate the information, and display it to the user in a single view (Prlic et al. 2007). A small number of plant/crop data sources are now beginning to make their data available through web services.

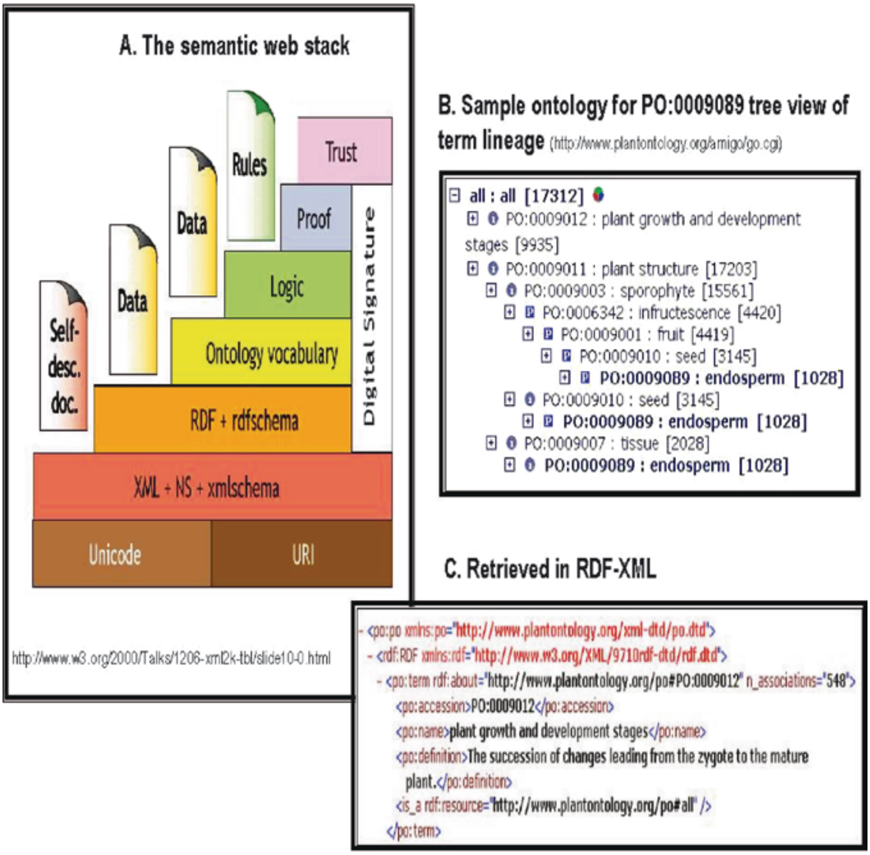
4.2.3 Data Integration and the Semantic Web

The bioinformatics community has been experimenting with two methods of biological database integration. In the data warehouse approach data from different data sources is translated into a local warehouse and all queries are executed on the warehouse. Examples include DataFoundry (Critchlow et al. 2000) and BioWarehouse (Lee et al. 2006). The warehouse needs to be updated frequently to reflect the modifications in the source databases. The second method is the federated database approach, where the query is executed on a single federated schema that is an integration of component database schemas (a schema can be considered to be the layout of a database). A good example of a federated query system designed specifically for use with large datasets is BioMart (http://www.ebi.ac.uk/biomart). Major databases that implement BioMart include Ensembl, a software system that produces and maintains automatic annotation on selected eukaryotic genomes (http://www.ensembl.org/index.html); VEGA (http://vega.sanger.ac.uk/

index.html), the manually annotated Vertebrate Genome Annotation; dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), and the Single Nucleotide Polymorphism database of NCBI.

At the level of data integration, most methods followed so far are based on syntax; explicit cross references and common contents which heavily rely on manual annotation of data that can be time consuming, error prone and expensive. Several bioinformatics databases are now moving towards a standardized method of describing their data so that data retrieval and integration can be independent of source database schemas. In the semantic web approach to data integration, the web is no longer a network of documents but a network of data and knowledge. The semantic web provides common formats and languages for consistent and standardized data representation and exchange. In the context of databases, it means that data will be encoded with additional meta-information that will provide context to the data which is made available through web services. That encoding makes use of ontologies. The key role of ontologies with respect to database systems is to specify a data modeling representation at a level of abstraction above specific database designs (logical or physical). Due to their independence from lower level data models, ontologies can be used for integrating heterogeneous databases, enabling interoperability among disparate systems, and specifying interfaces that can be queried independently.

Ontologies are part of the Semantic Web architecture (see Fig. 4.1, the W3C or World-Wide Web consortium develops common protocols for the World-Wide Web that promotes its evolution and interoperability). Ontologies define a set of representational classes, attributes and relationships with which to model a domain of knowledge. Take for example the Gene Ontology (GO), a community effort to provide controlled vocabulary to describe gene and gene product attributes in any organism. When one database describes a piece of data as being “a gene as defined by the Gene Ontology”, the data consumer can use or not use the data based on the understanding of “a gene as defined by the Gene Ontology” rather than worry about datasource specific definitions of the ‘gene’. Similarly, the Plant Ontology Consortium (POC) (www.plantontology.org) is a collaborative effort to develop simple yet robust and extensible controlled vocabularies that accurately reflect the biology of plant structures and developmental stages. There is Trait Ontology (TO) for traits and phenotype data (http://www.gramene.org/plant_ontology/trait.ontology). MyGrid and BioMoby ontologies are for the semantic discovery of bioinformatics services. They use ontological reasoning over both data type and service definitions for service discovery. Clients can interact with multiple sources of biological data, regardless of the underlying database format/schema. While ontologies are being implemented only by a small number of data sources, they become relevant to the interoperability of expanding database collections (http://www.gramene.org/resources/plant_databases.pdf). There are published examples to show the application of semantic web technologies to build data warehouses that facilitate integration of genomic/proteomic data (Smith et al. 2007).


A. The semantic web stack
Trust
Rules
Proof
Data
Data
Logic
Self-desc. doc.
Ontology vocabulary
RDF + rdfschema
Digital Signature
XML + NS + xmlschema
Unicode
URI
http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html
B. Sample ontology for PO:0009089 tree view of term lineage (http://www.plantontology.org/amigo/go.cgi)
C. Retrieved in RDF-XML

```
⊟ all : all [17312]
  ⊞ ● PO:0009012 : plant growth and development
  stages [9935]
  ⊞ ● PO:0009011 : plant structure [17203]
    ⊞ ● PO:0009003 : sporophyte [15561]
      ⊞ ▣ PO:0006342 : infructescence [4420]
        ⊞ ▣ PO:0009001 : fruit [4419]
          ⊞ ▣ PO:0009010 : seed [3145]
            ⊞ ▣ PO:0009089 : endosperm [1028]
      ⊞ ● PO:0009010 : seed [3145]
        ⊞ ▣ PO:0009089 : endosperm [1028]
    ⊞ ● PO:0009007 : tissue [2028]
      ⊞ ● PO:0009089 : endosperm [1028]
```

```
- <po:po xmlns:po="http://www.plantontology.org/xml-dtd/po.dtd">
 - <rdf:RDF xmlns:rdf="http://www.w3.org/XML/9710rdf-dtd/rdf.dtd">
  - <po:term rdf:about="http://www.plantontology.org/po#PO:0009012" n_associations="548">
     <po:accession>PO:0009012</po:accession>
     <po:name>plant growth and development stages</po:name>
     <po:definition>The succession of changes leading from the zygote to the mature
      plant.</po:definition>
     <is_a rdf:resource="http://www.plantontology.org/po#all" />
    </po:term>
```

Fig. 4.1 Semantic web for data integration through metadata-based reasoning. (**A**) The semantic web stack (**B**) a sample representation of an ontology for a term derived from the plant ontology consortium website (C) the same ontology retrieved in RDF-XML representation

4.2.4 Bioinformatics Tools for Comparative Genomics

Comparative genomics in silico offers the possibility of linking crops through their sequences and genome maps to provide keys to understanding how genes and genomes are structured, how they function and evolve. Significant synteny amongst the cereal crops has allowed the alignment of major economically important qualitative or quantitative trait loci across specific chromosomal regions. This has facilitated candidate gene and flanking marker identification and their comparisons with annotated sequences from model crops, important for the application of marker-assisted selection. The benefits of transferring genomics information from model to orphan crops could take one of several forms: (a) the identification of potentially useful variants, (b) Marker Assisted Selection (MAS) of desired alleles and allele combinations, and (c) cloning and direct transfer of desirable alleles among taxa (Nelson et al. 2004).

Very large collections of bioinformatics tools have been developed on the open source model, meaning that they are freely available to use and learn from and improve upon. Of the tools available for comparative genomics, sequence alignment tools are the most commonly used. These tools can be used to query databases for sequences similar to an input sequence, find previously characterized sequences, detect relationships amongst sequences, as well as identify possible functions based on similarity to known sequences. There is a considerable amount of literature on sequence alignment tools and their advancements. The advancements reported in the literature relate to algorithms that seek to reduce running time and produce optimal alignments. Pairwise sequence alignment is best accomplished with the Dynamic Programming algorithm, which is slow and time consuming. Several 'shortcuts' to this algorithm have been proposed to improve running time. Best-known variants are the Smith–Waterman for local alignments and the Needleman–Wunsch for global alignments where sequences are related over their full length (Smith and Waterman 1981; Needleman and Wunsch 1970). These algorithms are, however, too compute time intensive to use for database searches. Most sequence databases allow rapid search using BLAST, FASTA (Altschul et al. 1990; Pearson 1990), scanps, MPsrch (http://www.ebi.ac.uk/searches/blitz_input.html); Blast2, PHI-Blast or BLAT (Kent 2002). BLAST is the fastest sequence alignment algorithm, although it compromises some degree of sensitivity in favor of speed. FASTA is slower, but more sensitive.

A multiple sequence alignment (MSA) is an alignment of three or more protein, DNA or RNA sequences and the purpose of creating such an alignment is to highlight their similarity or differences, which might reflect the biological relationship between them. Generation of MSA is a very useful exercise and needs special care when being used in phylogenetic tree construction, for identification of profiles and structure prediction, or in degenerate primer design. Computing exact MSAs is computationally almost impossible, and in practice approximate algorithms (heuristics) are used to align multiple sequences, by maximizing their similarity. Many MSA algorithms are in use, including the popular matrix-based methods ClustalW (Thompson et al. 1994) and Muscle (Edgar 2004), and the consistency-based methods T-Coffee (Notredame et al. 2000) and PCMA (Pei et al. 2003). Consistency-based methods are evaluated superior to matrix-based methods of alignment though they require cpu time several times higher than the matrix methods (Notredame and Abergel 2003). With the availability of so many quality methods and the growing importance of MSA generation, the development of meta-methods that can seamlessly combine the output of several methods, and also incorporate structure information, was the next milestone (Pei and Grishin 2006). Emerging advances in this area include template-based alignment, an extension of consistency-based methods. Under this new model, the purpose of an MSA is not to squeeze a dataset and extract all the information it may contain, but rather to use the dataset as a starting point for exploring and retrieving all the related information contained in public databases. This information is used to drive the MSA computation. Such a usage of sequence and related information is seen as a major step toward global biological data integration (Notredame 2007). Jalview(http://www.jalview.org/download.html), BioEdit(http://www.mbio.ncsu.edu/BioEdit/BioEdit.html) and Genedoc(http://www.genedoc.us/gdsrc.htm) are popular freeware to edit multiple sequence alignments.

Several web-based tools are now available to browse and analyze genome alignments. These include the comparative genome viewers SynBrowse (Pan et al. 2005), SYBIL (http://sybil.sourceforge.net) and VISTA (Frazer et al. 2004). The VISTA family of tools includes a browser and rVISTA that combines a transcription factor binding site database search (using Blast) with comparative sequence analysis along with PHYLO-VISTA for phylogeny. Sybil is a web-based software package for comparative genomics, developed by the Bioinformatics group at J. Craig Venter Institute (formerly TIGR). This package includes several tools and browsers for genome comparisons and ortholog detection. FISH (Fast Identification of Segmental Homologies) is another useful algorithm available to explore the extent and distribution of conserved synteny between two species (Calabrese et al 2003). The Lagan Toolkit is a set of alignment programs for comparative genomics (Brudno et al. 2003) while the Staden Package is a suite of tools for sequence assembly, analysis, and mutation detection (Staden et al. 1998). The Gbrowse is a very popular viewer for manipulating and displaying annotations on genomes and was developed as part of the GMOD or Generic model organism database project. The tool is easy to use, fast, allows cross species comparisons, customizable and is freely available (http://www.gmod.org). The Ensembl Genome Browser is a software system using which a large selection of annotated eukaryotic genomes can be browsed and compared (http://www.ensembl.org/). Other comparative genomics tools include VisGenome (Jakubowska et al. 2007) and the SGN comparative map viewer (Mueller et al. 2008). cMAP (http://www.gramene.org/cmap/) allows comparisons of genetic and physical, sequence and QTL (Quantitative Trait Loci) maps, while CMTV (http://www.ncgr.org/cmtv/) allows comparative viewing of genetic and QTL maps and their integration to generate consensus maps.

4.2.5 Bioinformatics Tools for Functional Genomics

Functional genomics came of age when a shift of emphasis occurred from genome mapping and sequencing to determining how genes work together to produce traits. Current structural genomic approaches (i.e., mapping) generally focus on traits controlled by one or only a few genes, and often they provide information regarding the location of one or more genes only. Where functional information is available the scientist is equipped to a large extent to create varieties with exact combinations of traits. Most of available functional genomics resources are in the model crops, but since the genes that code for scores of plant traits and processes are quite similar across many species, this knowledge can be applied to genetic research on other crops. Functional genomics as it is being applied in the plant sciences includes functional annotation, gene expression, and elucidation of protein structure that can help link genome and proteome with phenotype, protein–protein interaction, intracellular localization and posttranslational regulation. Rapid improvements in innovations such as microarray and RNA interference technology, allow simple, low-cost, high-throughput screening of phenotypes, as opposed to looking at just a few specific "candidate genes." The predominant methods for sequence-based expression analysis are SAGE (Serial Analysis of Gene Expression) and MPSS (Massively Parallel Signature Sequencing) of which SAGE is more widely used, while for model crops MPSS resources are available (http://mpss.dbi.udel.edu/).

Functional annotation is the process of collecting information about and describing a gene's biological identity – its various aliases, molecular function, biological role(s), subcellular location and its expression domains within the plant. The association between sequence and functional phenotype can be predicted using homology search tools based on sequence alignment. Larger data sources like TAIR (The Arabidopsis Information Resource) use a combination of published literature, solicited contributions from the research community as well as computational analyses of the sequence as part of the functional annotation process (Swarbreck et al. 2007). Pattern recognition programs, tools to transfer annotation to GO terms, as well as available controlled vocabulary add value to the annotation. Software such as *GeneTools*, allows users to rapidly extract gene annotation data, to add "user defined" GO annotation to gene products and to perform hypothesis testing using *e*GOn (Beisvag et al. 2006). B2GO is a single tool for the functional annotation of sequence data that uses BLAST to find homologous sequences to fasta formatted input sequences. The program extracts GO terms to each obtained hit and assigns GO terms to the query sequence using an annotation rule. Annotation and functional analysis can be visualized in graph form (http://www.blast2go.de/). Whichever the tool of choice, the user should be aware that the annotation is only an approximation that must be further validated computationally and/or through wet lab experimentation.

Existing open source software generated by the bioinformatics community for fragment assembly and mapping are well known and widely used (Phrap (http://www.phrap.com/), cap3 (http://genome.cs.mtu.edu/cap/cap3.html), PCAP (http://seq.cs.iastate.edu/) and TGICL(http://compbio.dfci.harvard.edu/tgi/software/)), while feature prediction tools like Genscan for gene structure prediction have versions suitable for crops such as maize and *Arabidopsis*. The NetPlantGene web server (http://www.cbs.dtu.dk/services/NetPGene/) provides tools for the prediction of splice sites in *Arabidopsis* besides modelling and structure prediction tools. The Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) provides a comprehensive set of functional annotation tools. AutoFACT is another fully automated and customizable annotation tool that assigns biologically informative functions to a sequence (Koski et al. 2005). Other functional genomics platforms are also becoming available such as the Purdue Ionomics Information Management System (PiiMS) that provides integrated workflow control, data storage and analysis to facilitate high-throughput phenotypic data acquisition, along with integrated tools for data search, retrieval and visualization for hypothesis development. PiiMS is deployed as a web-enabled system, allowing for integration of distributed workflow processes and open access to raw data for analysis by numerous laboratories (Baxter et al. 2007). This platform is currently being used to integrate high throughput phenotypic data with functional genomics data in *Arabidopsis*. TraitMill is an automated plant evaluation platform allowing high throughput testing of the effect of plant-based transgenes on agronomically valuable traits. The platform offers high throughput function prediction, allows selection of candidate trait improvement genes among annotated genes and is currently being used for rice (Reuzeau et al. 2006). The Generation Challenge Program (GCP) with the CGIAR centers, Advanced Research Institutes and a number of National Agricultural Research and Education Systems is also developing a platform for functional genomics customizing the MAXD database for rice gene expression data along with data mining and analysis pipelines (Takeya et al. 2006).

4.2.6 Availability of High Performance Clusters and Grid

The problems of biological datasets have only grown in scale and complexity with high throughput technology. Single experiments may generate gigabytes of data and a single gene product may have several thousand interactions that create more functions than one can imagine. So there is a continual demand for increased computation speed from a computer system. High performance compute (HPC) systems have been available since the mid-1970s to users with large budgets. For the others with limited budgets and large computing needs, hardware parallelism can be achieved by connecting several independent computers. The idea being that *n* computers can provide up to *n* times the computational speed of a single computer. The popular beowulf clusters are created through networking a group of computers running linux. Continual improvements in execution speeds of single processors and their availability has made such clusters faster and cheaper to build. There are a number of approaches available to creating effective parallel computers with different levels of effectiveness for different kinds of problems. For programmes to show an increase in speed a substantial fraction of the computation needs to be executed in parallel. Software parallelism is the ability to find well-defined areas in a problem that can be broken down into self contained parts. The distributed processing of these parts speeds the programme up. Such parallel programmes are increasingly being used in the agricultural domain for data mining, comparative genomics, phylogenetics and population genetics analysis applications as well as in breeding simulation programmes. Parallel systems are also being used for fault tolerant applications such as hosting distributed databases (high availability clusters). While with high performance clusters one can deploy a solution with a fixed number of nodes (processors) on dedicated hardware, Grid computing brings several clusters together with the flexibility of using standard non-heterogeneous hardware where nodes can be added on demand and is not limited to the local LAN (Local Area Network), meaning that they could be geographically distributed. Through the Generation Challenge Progam, an HPC grid is becoming available that connects HPCs from four geographically distributed member institutions (http://hpc.cip.cgiar.org/webeval/), hosting several analysis software. Projects like myGrid allow biologists to design and execute in silico experiments on their desktop/laptop accessing datasources and tools available through the grid using the Taverna workflow bench. MyGrid uses the Feta web services discovery engine that is very similar in function to Moby Central of BioMoby (mygrid.org.uk).

4.2.7 Bioinformatics and Molecular Marker Technology

4.2.7.1 In silico Marker Mining Tools

Growing sequence information in databases has seen a corresponding increase in bioinformatics tools available to mine this information usefully. In the crop sciences, sequence data are useful sources of molecular markers like SSRs (simple sequence repeats), SNPs (single nucleotide polymorphisms), annotated ESTs, anchor markers, TRAPs (target region amplification polymorphisms), CISPs (conserved intron spanning primers) and conserved ortholog sets. Table 4.2 gives a compilation of the more popular tools available to researchers for the purpose of mining sequence data for putative molecular markers. Bioinformatics methods also allow the identification of functional markers that are more relevant and superior to random markers because they are linked to functional motifs and trait locus alleles. They rely on comparative genomics and phylogeny and elucidate the nature of genes conserved. Tools are available for the design of degenerate oligonucleotides for PCR for gene isolation and subsequent development of gene markers (Rose et al. 2003). The markers mined can then be applied to genetic trait mapping (Morgante and Salamini 2003). One can use the annotated genome of any one species to transfer knowledge to another genome. The identification of genes and related markers through computational methods is currently employed as a component of the marker development process.

4.2.7.2 Data Acquisition Software

Rapid data generation through high throughput methods has also led to the development of several systems for the capture, storage and retrieval of this data. Some freely available information management systems have been developed for genotyping, such as software to manage TaqMan SNP genotyping data (Monnier et al. 2005), the GenoDB (Li et al. 2001), AGL-LIMS (Jayashree et al. 2006b), PacLIMS (Donofrio et al. 2005) and SNPP (Zhao et al. 2005) each with different levels of

Table 4.2 Bioinformatics tools and pipelines available for in silico marker mining from sequence data

Tool	Marker	URL	Programming language
AutoSNP	SNP	http://www.cerealsdb.uk.net/discover.htm	Perl
CISPrimerTool	CISP	http://www.icrisat.org/gt-bt/softwares_downloads.htm	Java
GeMprospector	Cross species marker candidates	http://cgi-www.daimi.au.dk/cgi-chili/GeMprospector/main	Python, CGI
MISA	SSR	http://pgrc.ipk-gatersleben.de/misa	Perl
Polybayes	SNP	http://genome.wustl.edu/tools/software/polybayes.cgi	Perl
SNPdetector	SNP	http://lpg.nci.nih.gov	C and Perl
SNPpipeline	SNP	http://www.icrisat.org/gt-bt/softwares_downloads.htm	Parallel programme with an MPI wrapper (C++ and Python)
SSRIT	SSR	http://www.gramene.org/db/searches/ssrtool	Perl
Tandem Repeat Finder	SSR	http://tandem.bu.edu/trf/trf.download.html	Perl
TROLL (Tandem repeats occurrence locator)	SSR	http://sourceforge.net/projects/finder	C++

dependencies and functionalities. While GenoDB is a data management system for microsatellite markers and linkage analysis with functionalities tuned to human genotyping projects running on Windows platform, AGL-LIMS is a genotyping workflow management system for high throughput crop genotyping, platform independent and web enabled. Such systems, while serving as electronic notebooks for lab personnel, also help provide a measure of the quality of data being generated in the laboratory, better traceability and centralization of data. The ability to track data and communicate quality information gives the marker laboratory the tools to improve methods and work practices.

4.2.7.3 Molecular Marker Data Repositories and Visualization Tools

PlantMarkers is a genetic marker database that contains a comprehensive pool of predicted molecular markers (Rudd et al. 2005). The database contains putative single nucleotide polymorphism (SNP); simple sequence repeat (SSR) and conserved orthologue set (COS) markers. The database is derived from a systematic approach to identify a broad range of putative markers by screening the available openSputnik unigene consensus sequences from over 50 plant species. Cereal marker repositories include Gramene (Liang et al. 2008) and MaizeGDB (Lawrence 2008) while legume marker repositories exist at LIS(Legume Information System) (Gonzales et al. 2005). Besides there are other multi-species marker databases published online as a result of individual institutional efforts such as the CUGI plant SSR database (http://www.genome.clemson.edu/projects/ssr/), SSRDB (Jayashree et al. 2006a) and TOGsDB (http://intranet.icrisat.org/gt1/tog/homepage.htm). The high-throughput marker discovery protocol – Diversity Arrays Technology (DArT) is sequence-independent (Jaccoud et al. 2001; Wenzl et al. 2004). As it becomes more accessible, there will soon be highly populated DArT marker databases. Major marker repositories also provide tools for the visualization of maps and comparisons with linkage maps from related species. The cereal markers repository Gramene provides cMAP, and the LIS allows the use of both cMAP and CMTV (Fig. 4.2). CMAP is

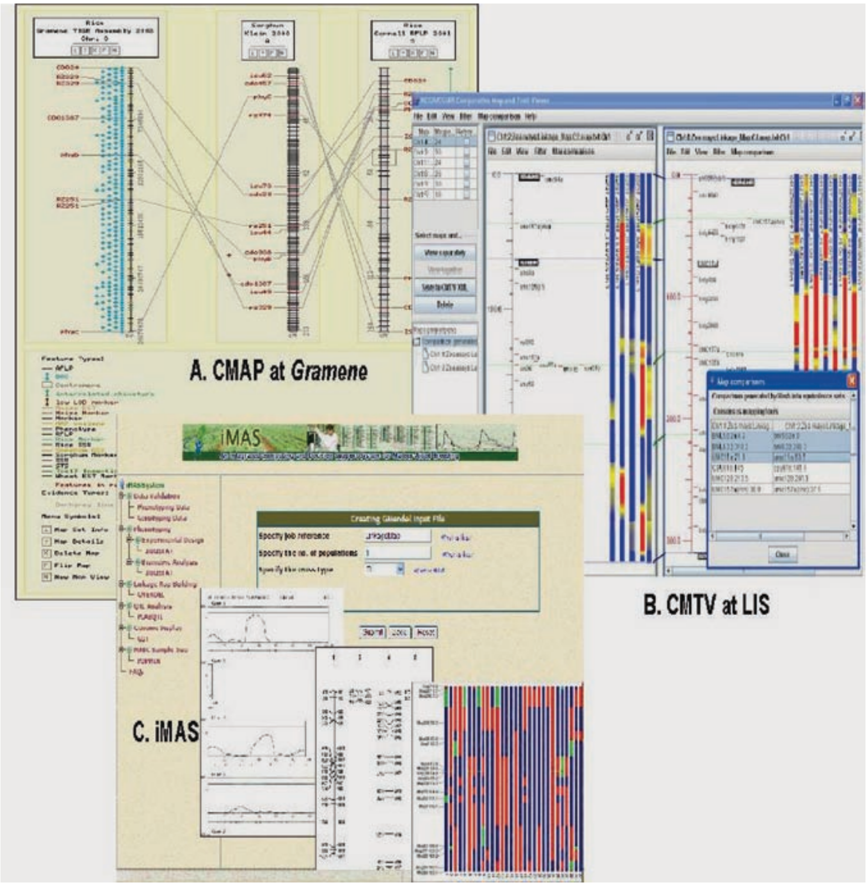


Fig. 4.2 Tools for map generation and comparisons. (**A**) The cMAP tool available at the gramene website (**B**) CMTV available from the Legume Information Service website (**C**) The desktop application iMAS

available under an open source license. CMTV allows viewing of multiple maps, the identification of correspondences as well as the combining of maps from different experiments to produce aggregate maps. This desktop application is also freely available (http://www.ncgr.org/cmtv/).

4.2.7.4 Software for Mapping and Association Analysis

The analysis of phenotypic and genotypic datasets leading to QTL maps, marker-aided selection and breeding involves the use of a number of different computing software. The last few years have seen a deluge of tools for map generation, association analysis and visualization. Many of these tools are available as freeware and some of them are open source (Table 4.3). There are several publications citing simulation software available to the plant breeder. Such tools have been used to investigate the introgression of one or several superior QTL alleles into a recipient line, to compare selection strategies based on proportion of recurrent parent genome recovered, and to investigate the effect of varying population size, marker density, marker positions, and required number of marker data points. Simulation approaches predict cross performance, compare different selection methods, and identify best performing crosses and breeding strategies. Software like PBMASS (pedigree-based marker assisted selection system) for MAS and recurrent parent recovery in wheat and barley has been published although the software is not publicly available (Eisemann et al. 2004).

4.3 Closing the Gap to Meet Molecular Breeding Requirements

Molecular breeding calls for integration of various kinds of information: genetic resource information with phenotype information linked to the allelic profiles of specific germplasm accessions coupled with results arising out of comparative and functional genomics experimentation. The goal is to rapidly assay the genetic makeup of individual plants or varieties in breeding populations and make accurate phenotypic predictions. This knowledge can be used to design a genoytpe that is targeted to perform well under a given set of environmental conditions. Marker assisted breeding programs typically involve information gathering over a prolonged period of time, need a management system to keep track of this information, and require a suite of analysis tools to help the scientist/breeder make decisions regarding which individuals to use from a segregating progeny. There is a need for systems that allow information to be carried forward and backward between the steps of the breeding program, allowing the user to choose breeding schema, to identify markers for foreground and background selection, to track inheritance and to serve as an information repository for data pertaining to the parental source materials, linkage maps, loci and genotyping data, polymorphism information for background and foreground markers in the parents and recombinants. Such systems will also serve as a link between the field books, the MAS and marker

Table 4.3 Software tools for mapping, association analysis and breeding simulation. The list is not extensive and includes only software available in the public domain

Tool	URL	Application
Adegenet	http://pbil.univ-lyon1.fr/software/adegenet	Related to ADE4, a R package for population genetics data analysis
Arlequin	http://cmpg.unibe.ch/software/arlequin3/	Implements a variety of population genetics methods that can be conveniently selected through the graphical interface
Blossoc	http://www.birc.dk/~mailund/Blossoc/	Linkage disequilibrium association mapping tool
CPSIM, BCSIM	http://www.plantbreeding.wur.nl/UK/software_cpsim.html	Simulation software for cross-pollinated population data or backcross simulation
GeneRecon	http://www.daimi.au.dk/~mailund/GeneRecon	LD mapping, based on a Bayesian MCMC method for fine scale linkage-disequilibrium gene mapping using high-density marker maps and association mapping
GGT	http://www.dpw.wau.nl/pv/PUB/ggt/	Graphical genotyping software
ICIM	http://www.isbreeding.net/software.html	Inclusive CIM, that provides an improvement over existing methods
IMAS	http://www.icrisat.org/gt-bt/download(bm)_iMAS.htm	Package of several integrated software for tasks from experimental design to map generation, qtl analysis and visualization along with a decision support platform
MADMAPPER	http://www.atgc.org/Xlinkage/MadMapper	Quality control of genetic markers, inference of linear order of markers on linkage groups
MAPL	http://lbm.ab.a.u-tokyo.ac.jp/software.html	QTL analysis by interval mapping and ANOVA, graphical genotyping
Mapmaker and Mapmaker/QTL	http://linkage.rockefeller.edu/soft/mapmaker/	QTL analysis, biologist friendly user interface
MapQTL	http://www.mapqtl.nl	Interval mapping, mapping QTLs for several types of mapping populations, Composite interval mapping, non-parametric mapping through a MS-Windows interface
PlabQTL	https://www.uni-hohenheim.de/plantbreeding/software/	Implement composite interval mapping besides others
PLABSIM	http://www.uni-hohenheim.de/~frisch/software.html	Plant breeding simulation software
PYPOP	http://www.pypop.org/	Software for the analysis of large-scale multi locus genotype data
QTLcartographer	http://statgen.ncsu.edu/	Implement composite interval mapping besides others
Qu-gene	http://www.uq.edu.au/lcafs/index.html?page=59974	Simulation platform for quantitative analysis of genetic models
Qu-Line	http://www.uq.edu.au/lcafs/index.html?page=59974	A component of Qu-gene, it is a simulation programme for the development of final advanced lines
STRAT	http://pritch.bsd.uchicago.edu/software/STRAT.html	Companion programme to Structure written for use in association mapping
Tassel	http://sourceforge.net/projects/tassel	Association mapping software

laboratory while providing easy to use interfaces and graphical visualization tools to view recombinant data. Thus, efficient use of DNA markers for crop improvement depends as much on computational tools as on laboratory technology. While information systems are becoming available for the acquisition, storage and retrieval of data derived from high throughput experimentation procedures, systems for integrating them with other data sources for the benefit of the plant breeder are as yet lacking. Software remains to be implemented that caters to the data integration needs of a plant breeder. Software specific to the management of information in marker assisted breeding programs is unavailable in the public domain. There is information available about the existence of LIMS for sample handling and databases specific to plant breeding operations, but these are private software packages developed for industry operated MAS programs that are neither licensed nor sold. Efforts are now being made to develop such information management systems (an ongoing project at ICRISAT).

For genomics to be applied to plant breeding, there is need for high throughput techniques, cost effective protocols, precise determination of quantitative trait expression, besides bioinformatics platforms that provide for the ability to combine outputs from these along with curated data on allelic variation annotated with alterations in phenotype. Thus, a high degree of curation for annotation polymorphisms with phenotypic variations in different genetic backgrounds is required along with high quality sequence annotation in selected germplasm resources. The Information Systems must also link to model crop data sources like genomic, genetic maps and functional genomics data sources. Figure 4.3 indicates the desired flow of information and integration of data sources. Crop improvement programmes can incorporate the results of genomics projects if they were available to those involved, namely the breeders. This calls for the coming together of a common platform for various disciplines at various locations. An example of one such succesfull disparate data/location integration initiative is PlaNet, a collaborative network of bioinformatics groups and plant molecular biologists from several plant genome data centres in Europe (JIC, NASC, CNB/CSIC, VIB, PRI and MIPS). The PlaNET approach to data integration reduces the strain on individual resources, distributes the burden of data curation and maximizes the value of individual data collections (Schoof et al. 2004). The established platform interconnects several databases, gathering external data into PlaNET through integration tools that allow flexible migration of data from various representations. This project uses BioMoby for interoperability. For crop improvement programmes to benefit from the various genomic resources and data collections, efforts such as these are needed that bring into the picture individual data sources held by groups that are involved in generating quality genotype, phenotype and genomics information for germplasm collections. Since curation is a long term effort, a consortium of dedicated data providers who are willing to share quality data across a common informatics platform accessible to breeders is a required investment. Careful annotation of DNA polymorphisms is required, whether the variation is indeed linked to an alteration in phenotype or whether it is a neutral sequence variation. The existing disparities in resources available to model crops research relative to orphan crops that are important to a large section of people in the developing world also needs to be closed. Increased investment in such crops will undoubtedly see a concomitant increase in bioinformatics

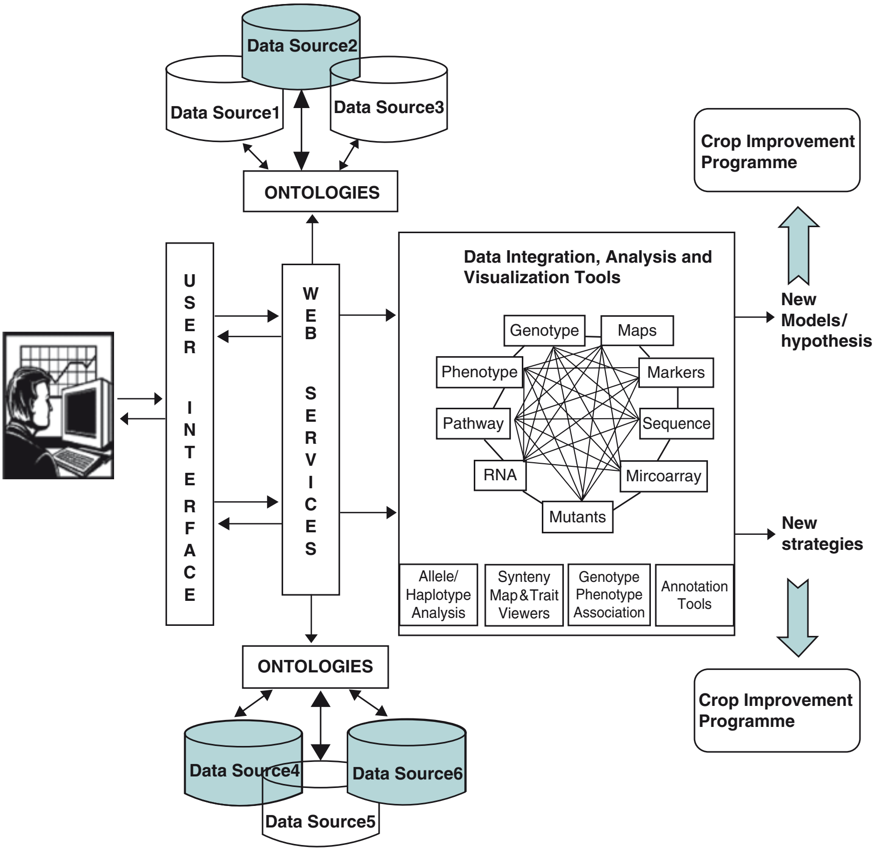


Fig. 4.3 Information and desirable data integration requirements for crop improvement programmes

data sources and adaptation/customization of tools developed for model crops. The availability of all this data through an integrated network of information to breeders who have been empowered to use it will provide the means to apply the outputs of modern technologies in crop improvement programmes.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman D.J (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Baxter I, Ouzzani M, Orcun S, Kennedy B, Jandhyala SS, Salt DE (2007) Purdue ionomics information management system. an integrated functional genomics platform. Plant Physiol 143:600–611

Beisvag V, Jünge FKR, Hallgeir B, Jølsum L, Lydersen S, Günther C-C, Ramampiaro H, Langaas M, Sandvik AK, Lægreid A (2006) GeneTools – application for functional annotation and statistical hypothesis testing. BMC Bioinform 7:470

Brudno M, Chuong Do, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S (2003) LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. Genome Res 13:721–731

Calabrese PP, Chakravarty S, Vision TJ (2003) Fast identification and statistical evaluation of segmental homologies in comparative maps. Bioinformatics 19:i74–i80

Critchlow T, Fidelis K, Ganesh M, Musick R, Slezak T (2000) DataFoundry: information management for scientific data. IEEE Trans Inform Technol Biomed 4:52–57

Donofrio NM, Rajagopalan R, Brown DE, Diener SE, Windham DE, Nolin S, Floyd A, Mitchell TK, Galadima N, Tucker S, Orbach MJ, Patel G, Farman ML, Pampanwar V, Soderlund C, Lee Y-H, Deen RA (2005) PACLIMS: a component LIM system for high throughput functional genomic analysis. BMC Bioinformatics 6:94

Edgar RC (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Eisemann B, Banks P, Butler D, Christopher M, Delacy I, Jordan D, Mace E, McGowan P, McIntyre L, Poulsen D, Rodgers D, Sheppard J (2004) Pedigree based genome mapping for marker assisted selection and recurrent parent recovery in wheat and barley. In: new directions for a diverse planet. Proceedings 4th International Crop Science Congress, Brisbane, 26 September–1 October, 2004)

Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32:W273–W279

Gonzales MD, Archuleta E, Farmer A, Gajendran K, Grant D, Shoemaker R, Beavis WD, Waugh ME (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. Nucleic Acids Res 33:D660–D665

Jaccoud D, Peng K, Feinstein D, Kilian A (2001) Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:e25

Jakubowska J, Hunt E, Chalmers M, McBride M, Dominiczak AF (2007) VisGenome: visualization of single and comparative genome representations. Bioinformatics 23:2641–2642

Jayashree B, Crouch JH, Prasad PVNS, Hoisington D (2006a) A database of annotated tentative orthologs from crop abiotic stress transcripts. Bioinformation 1:225–227 (http://www.bioinformation.net/1/57–1–2006.htm)

Jayashree B, Reddy PT, Leeladevi Y, Crouch JH, Mahalakshmi V, Buhariwalla HK, Eshwar KE, Mace E, Folkertsma R, Senthilvel S, Varshney RK, Seetha K, Rajalakshmi R, Prasanth VP,Chandra S, Swarupa L, SriKalyani P, Hoisington DA (2006b) Laboratory Information Management Software for genotyping workflows: applications in high throughput crop genotyping. BMC Bioinformatics 7:383

Kent WJ (2002) BLAT – the Blast like alignment tool. Genome Res 12:656–664

Koski LB, Gray MW, Lang BF, Gertraud B (2005) AutoFACT: an automatic functional annotation and classification tool. BMC Bioinformatics 6:151

Lawrence CJ (2008) MaizeGDB, The maize genetics and genomics database. In: Methods in molecular biology: plant bioinformatics methods and protocols. Humana Press, Springer, pp 331–345

Lee TJ, Pouliot Y,Wagner V, Gupta P, Stringer-Calvert DWJ, Tenenbaum JD, Karp PD (2006) BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics 7:170

Li J-L, Deng H, Dong-Bing L, Fuhua X, Chen J, Gao G, Recker R, Deng H-W (2001) Toward high-throughput genotyping: dynamic and automatic software for manipulating large-scale genotype data using fluorescently labeled dinucleotide markers. Genome Res 11:1304–1314

Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, Ravenscroft D, Ren L, Spooner W, Tecle I, Thomason J, Tung CW, Wei X, Yap I, Youens-Clark K, Ware D, Stein L (2008) Gramene: a growing plant comparative genomics resource. Nucleic Acids Res (Database Issue) D947–D953

Monnier S, Cox DG, Albion T, Canzian F (2005) T.I.M.S: Taqman Information Management System, tools to organize data flow in a genotyping laboratory. BMC Bioinformatics 6:246

Morgante M, Salamini F (2003) From plant genomics to breeding practice. Curr Opin Biotechnol 14:214–219

Mueller L, Mills A, Skwarecki B, Buels R, Menda N, Tanksley S (2008) The SGN comparative map viewer. Bioinformatics Advance Access, published 17 January 2008

Needleman S, Wunsch C (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Nelson RJ, Naylor RL, Jahn MM (2004) The role of genomics research in improvement of "orphan" crops. Crop Sci 44:1901–1904

Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comp Biol 3:e123. doi:10.1371/journal.pcbi.0030123

Notredame C, Abergel C (2003) Using multiple alignment methods to assess the quality of genomic data analysis. In: Andrade M (ed) Bioinformatics and genomes: current perspectives. Horizon Scientific Press, Wymondham, UK, pp 30–55

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302:205–217

Pan X, Stein L, Brendel V (2005) SynBrowse: a synteny browser for comparative sequence analysis. Bioinformatics 21:3461–3468. doi:10.1093/bioinformatics/bti555)

Pearson WR (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63–98

Pei J, Grishin NV (2006) MUMMALS: Multiple sequence alignment improved by using hidden Markov models with local structural information. Nucleic Acids Res 34:4364–4374

Pei J, Sadreyev R, Grishin NV (2003) PCMA: fast and accurate multiple sequence alignment based on profile consistency. Bioinformatics 19:427–428

Prlic A, Down TA, Kulesha E, Finn RD, Kahari A, Hubbard TJP (2007) Integrating sequence and structural biology with DAS. BMC Bioinformatics 8:333

Reuzeau C, Frankard V, Hatzfeld Y, Sanz A,Van Camp W, Lejeunne P, de Vilde C, Herve P, Peerbolte R, Broekaert W (2006) TraitMill: a functional genomics platform for the phenotypic analysis of cereals. Plant Genetic Resources: characterization and utilization 4:20–24

Rose T, Henikoff J, Henikoff S (2003) CODEHOP (COnsensus Degenerate Hybrid Oligonucleotide Primer) PCR primer design. Nucleic Acids Res 31:3763–3766

Rudd S, Schoof H, Mayer K (2005) PlantMarkers – a database of predicted molecular markers from plants. Nucleic Acids Res 33:D628–D632

Schoof H, Ernst R, Mayer KF (2004) The PlaNet consortium: A network of European plant databases connecting plant genome data in an integrated biological knowledge resource. Comp Funct Genomics 5:184–189

Smith AK, Kei-Hoi Cheung, Yip KY, Schultz M, Gerstein MB (2007) LinkHub: a Semantic Web system that facilitates cross-database queries and information retrieval in proteomics. Published in Semantic Web Approach to Database Integration in the Life Sciences. Baker CJO, Kei-Hoi Cheung (eds) Springer US, pp 11–30

Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Staden R, Beal KF, Bonfield JK (1998) The Staden package. Meth Mol Biol 132:115–130

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2007)The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucleic Acids Res. doi:10.1093/nar/gkm965

Takeya M, Suzuku K, Doi K, Kikuchi S, Bruskiewich R (2006) Development of a platform for functional genomics under the Generation Challenge Program. www.jsbi.org/modules/journal1/index.php/GIW06/GIW06P095.pdf

The BioMoby Consortium (2008) Interoperability with Moby 1.0 – It's better than sharing your toothbrush! Brief Bioinform 9:220–231

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc Natl Acad Sci 101:9915–9920

Zhao L-J, Li M-X, Guo Y-F, Xu F-H, Li J-L, Deng H-W (2005) SNPP: automating large-scale SNP genotype data management. Bioinformatics 21:266–268