Chapter 2
Bioinformatic Tools in Crop Improvement

L. F. De Filippis

Abstract Bioinformatic resources and web databases are essential for the most effective use of genetic, proteomic, metabolomic and phenome information important in increasing agricultural crop productivity. Innovations in web based platforms for omics based research, and application of such information has provided the necessary platform to promote molecular based research in model plants, as well as important crop plants. Combinations of multiple omics web based sites and integration of outcomes is now an important strategy to identify molecular systems promoting comparative genomics, the biological properties in many species, and to accelerate gene discovery and functional analyses. The review details recent advances in plant omics data acquisition sites, together with relevant databases and advance molecular technology under clear biological categories. The information is set out under the molecular biology divisions of; DNA based resources and sequencing, RNA and variation analysis, proteomics, structural proteins, and post-translation modifications, metabolomics, phenome and plant comparative analyses. Tables of relevant web sites are presented under similar headings for convenience, and the application of bioinformation data is reviewed in light of the possible use of these resources for crop improvement. Finally, a long list of future perspectives and research still to be attempted is detailed, which in the fullness of time should enable the full potential of bioinformatics and use in crop improvement programs to be achieved.

Introduction

Sustainable agricultural production and food security are two important issues of concern in response to population increase, environmental degradation and climate change (Brown and Funk 2008; Turner et al. 2009). According to the United Nations, the world population increases by 70–75 million people annually, an aver-

L. F. De Filippis (✉)
Centre for Environmental Sustainability (CENS), Department of Environmental Sciences, University of Technology, Broadway/Sydney NSW 2007,
P O Box 123, Sydney, Australia
e-mail: lou.defilippis@uts.edu.au

K.R. Hakeem et al. (eds.), *Crop Improvement*, DOI 10.1007/978-1-4614-7028-1_2,

age of more than two persons every second; and over 95% of these will live in developing countries (De Filippis 2012). It will be difficult satisfying the needs of this growing population and avoid serious food shortages or even famine from the limited arable land and natural resources available. These factors combined have already resulted in food deficiency and malnutrition, which have become serious health problems. Additionally, recent increased demand for biofuel crops has created a new market for agricultural commodities, causing even more stress on food security (Ozturk et al. 2006; Ozturk 2010; Hakeem et al. 2012). In order to try to resolve these problems and increase crop yields, breeding plants based on a better molecular understanding of gene function, and on the regulatory mechanisms involved in crop production (Pinstrup-Andersen and Cohen 2000; Takeda and Matsuoka 2008) appears to be necessary. Plant molecular biology continues to progress, and important gene sequences and their function have been described; many of which are related to crop yields (production), crop quality (protein and carbohydrate), and tolerance to biotic and abiotic stresses (De Filippis 2012). There are legal, social and political barriers to the full potential use of crop biotechnology and transgenic plants, nevertheless advances in these fields have lead to improvements in agriculture and human life. One vital tool of biotechnology is 'bioinformatics', which is commonly used to genetically type and identify genotypic and phenotypic changes in plants, and this information is important for improvement in performance of crop plants (Ahmad et al. 2011).

The complete genome sequence of the mustard plant *Arabidopsis thaliana* has been available to scientists since 2000 (International Arabidopsis Genome Initiative 2000; Somerville and Dangl 2000). Similarly, the rice (*Oryza sativun* cv. *japonica*) complete genome sequence has been documented since 2005 (International Rice Genome Sequencing Project 2005; Itoh et al. 2007; Hakeem et al. 2012). The rice genome sequencing project in particular with its molecular methods and DNA markers on chromosomes, introduced important developments in mapping populations and chromosome marker resources, which accelerated the isolation of agronomically important quantitative trait loci (QTLs) in crop breeding programs (Ashikari et al. 2005; Konishi et al. 2006; Ma et al. 2006; Kurakawa et al. 2007; Ma et al. 2007; Zhang et al. 2007).

Each biological element that can been measured, can also be represented in a typical plant cell, tissue and organ at various molecular and/or morphological levels, or in other words a conceptual model with layers ranging from the 'genome' to the 'phenome'; a model called 'omic space' (Fig. 2.1) (Toyoda and Wada 2004). Advances in each 'omics' research area have become essential for investigations of gene function and structure, and the type of phenotypic changes present in plants. A schematic presentation of relevant 'omics' resource is shown in Fig. 2.1, together with the current status of available areas of research from Arabidopsis, rice, soybean, corn and *Brassica*; just to cite a few examples. Some of these advances have included improved methods for gene expression, gene modifications, molecular breeding, plant genome and proteome interactions, and metabolite profiling. Large volumes of information in biological resources, mass identification of mutant lines and full-length cDNAs, and the publication of this information in web-based data

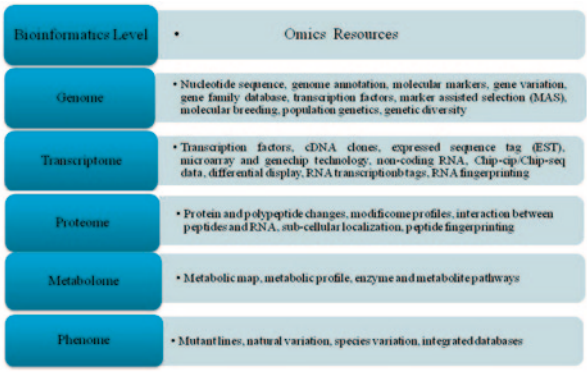

Fig. 2.1 A conceptual model called 'omic space' with layers ranging from the 'genome' to the 'phenome'. (After Toyoda and Wada 2004)

banks have been available for some time (Brady and Provart 2009; Kuromori et al. 2009; Seki and Shinozaki 2009).

Bioinformatic information and web sites have become important for crop scientists in gene data mining, and linking this knowledge to its biological significance (Mochida and Shinozaki 2010). However there needs to be a note of caution. As genomic and proteomic knowledge expands, new forms of electronic data becomes available to help interpret results. Biological data is notoriously variable (even unreliable at times) and 'noisy' in electronic form, due to living systems being complex and measurement and analysis technologies are often imperfect. In my experience two approaches for reducing 'noise' and help reliability of this type of data are required; aggregation and visualisation. Firstly, when combined, multiple forms of evidence become more and more accurate than for example a single source of data, simply because each replicate form of the data reduces overall uncertainty. Secondly, the human mind is an outstanding data analysis tool. It can absorb textual data rather poorly, but it can assimilate visual information in great detail, and the mind can process visual data efficiently to help identify common trends and themes (Cline and Kent 2009).

In this chapter, we provide an overview of the many web-based resources available for use in 'omics' plant research, with particular emphasis on recent progress related to crop species and crop improvement. Therefore we describe DNA and RNA sequence-related resources, molecular markers, whole genome sequencing, protein coding and non-coding transcripts, and provide molecular technology updates. We then review resources important for genetic map-based approaches such as QTL analyses and population genetic (diversity) studies. We also describe the current status of resources and some technologies for transcriptomics, proteomics and metabolomics; however some of these research areas are more comprehensively described in other chapters of this book. We then review molecular developments in each 'omics' field, as well as instances of their combined uses in investigations of particular crop systems. Mutant genotypes for use in 'phenome' research will be discussed, and the integration of 'omics' data between plant species in comparative genomics is dealt with. Throughout this review we provide examples of applica-

tions through available databases in crop plants, and where improvement in crop production has been described.

Bioinformatics and web addresses for plant genomics and proteomics have been reviewed by a number of authors (Rose et al. 2004; Sterck et al. 2007; Takeda and Matsuoka 2008; Zhang 2008; Baginsky 2009; Varshney et al. 2009; Mochida and Shinozaki 2010; Jackson et al. 2011; Memon 2012), and this review will basically cover some new areas in population (breeding) genetics, and topics which require more detail explanation and are updated in crop plants. The excellent review by Mochida and Shinozaki (2010) has provided the framework for this review, and we intend to concentrate on more recent developments, and focus on bioinformation and implications in crop improvement; although the technology, instrumentation and molecular biology achieved in other plants must also be covered.

DNA Based Sequence Resources

Genome Sequencing Projects

Initially, the publication and accumulation of nucleotide sequences for model plants only provided fundamental information, however now these base sequences form the fundamentals of research in functional plant genetics in applied species such as crops and domestic animals. Furthermore, DNA sequence data continues to be central in providing the genomic basis for accelerating molecular level understanding of basic biological mechanisms, and the application of such information to crops. In this section, we describe recently developed plant sequencing advancements. Species-specific nucleotide sequences are now providing information related to phenotypic characters, even when based on genome comparative analyses from the few model plants available (Cogburn et al. 2007; Flicek et al. 2008; Paterson 2008; Tanaka et al. 2008).

The genome sequence of *Arabidopsis thaliana* is now used as a model species in plant molecular biology mainly because of its small size, short generation time and high efficiency of transformation. The genome sequence of rice (*Oryza sativa*), including *japonica* and *indica* (an important staple food and a model monocotyledon) has also been used for comparative studies. These two plants still provide the only model plant systems to date, however several genome sequencing projects involving other plants have been completed, and many others are in progress; these are detailed in Table 2.1. Listed below are six of the most important web-based sites for DNA based genome sequencing and annotation projects, their purpose and their URL are detailed in Table 2.2.

NCBI—BioProject

The NCBI site provides genome sequences and information for many plant species (Viridiplantae) designed to facilitate comparative genomic studies amongst the

Table 2.1 List of plant species in which partial or whole genomes have been sequenced. (Data extracted from the following internet sites: http://www.ncbi.nlm.nih.gov/genomes/PLANTS/PlantList.html; http://www.arabidopsis.org/portals/genAnnotation/other_genomes/index.jsp; http://www.ildis.org/)

Division	Class	Species
Non Vascular	Algae	Chlamydomonas reinhardtii Chlorella variabilis Coccomyxa sp. Cyanidioschyzon merolae Ectocarpus siliculosus Micromonas pusilla Micromonas sp. Ostreococcus lucimarinus Ostreococcus tauri Volvox carteri Zostera marina
	Bryophytes	Physcomitrella patens Selaginella moellendorffii
Vascular	Dicotyledons	Amborella trichopoda Aquilegia sp. Arabidopsis lyrata Arabidopsis thaliana Arachis hypogaea Asclepias syriaca Beta vulgaris Boechera holboellii Brassica napus Brassica napa Brassica rapa Caffea caneophora Cajanus cajan Cannabis sativa Capsella rubella Carica papaya Castanea mollissima Citrullus lanatus Citrus clementine Corchorus olitorius Cucumis sativus Eucalyptus grandis Fragaria vesca Glycine max Gossypium hirsutum Gossypium raimonddi Hordeum vulgare Jatropha curcas Lactuca sativa Linum usitatissinun Lotus japonicas Malus domestica Manihot esculenta Medicago truncatula Mimulus guttatus Phaseolus vulgaris Pinus taeda Populus tremula Ricinus communis Theobroma cacao Populus nigra Populus trichocarpa Prunus avium Prunus persica Pyrus bretschneideri Rubus idaeus Salix purpure Solanum lycopersicum Solanum pimpinellifolium Solanum tuberosum Spirodella polyrhiza Thellungiella parvula Vitis vinifera
	Monocotyledons	Brachypodium distachyon Elaeis guineensis Miscanthus giganteus Musa acuminate malaccensis Oryza sativa Oryza glaberrima Panicum hallii Panicum virgatum Phoenix dactylifera Seratia italic Sorghum bicolor Triticum aestivum Zea mays

many other records of plants there. The current version consists of documentation in at least 115 different plants with partial sequences, and about 40,000 Expressed Sequence Tags (ESTs). It also contains separate sites and resources for other web based tools, data banks and other web servers, including agronomically important crops for food or fruit, medicinal plants, a number of green algae, pathogenic bacteria and fungi, viruses and animals. It will be important to become familiar and navigate through this very important site.

Phytozome

The site includes genome sequences and data sets for various crop species designed to facilitate comparative genomic studies amongst other green plants. The current version consists of 31 plant species wholly or partially sequenced, and is set-up into 10 evolutionary significant nodes.

Table 2.2 Integrative databases for DNA, Gene Sequences and Population Genetics analysis in plants

Database Name	Plant Species/Purpose	URL
Home—BioProject—NCBI	Multi-Purpose site; Over 1000 genomes; plants, animals, bacteria, fungi, virus	http://www.ncbi.nlm.nih.gov/sites/entrez?db=bioproject
Phytozome v8.0: Details	Over 31 species of plants; some software	http://www.phytozome.net/Phytozome_info.php
Gramene	Over 29 species of mainly monocots	http://www.gramene.org/
BLAST: Basic Local Alignment Search Tool	Multi-Purpose site for genome comparison; plants, animals, bacteria, fungi, virus	http://blast.ncbi.nlm.nih.gov/Blast.cgi
GrainGenes Class Browser: Marker	Triticeae and Avena site; nearly 200 species	http://wheat.pw.usda.gov/cgi-bin/graingenes/browse.cgi?class=marker
PlantGDB—Resource Plant Comparative Genomics	Multi Plant site; 15 dicot, 7 monocot, 3 other plant species	http://www.plantgdb.org/
TreeView	Phylogenetic tree software	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
Rod Page	Population genetics, gene diversity software	http://taxonomy.zoology.gla.ac.uk/rod/rod.html
Software	Softwere site; BLAST, sequence alignment	http://evolve.zoo.ox.ac.uk/Evolve/Software.html
LALIGN Server	Sequence alignment software	http://www.ch.embnet.org/software/LALIGN_form.html
PopGene	Population genetics software	http://www2.unil.ch/popgen/softwares/fstat.htm
Arlequin 3.11	Population genetics program	http://cmpg.unibe.ch/software/arlequin3/
ANU Bot Zool	Population genetics statistics, software	http://www.anu.edu.au/BoZo/GenAlEx/genalex_download_6_1.php
SOPH U AB	Population genetics programs	http://www.soph.uab.edu/ssg/linkage/population
Francis Yeh	PopGen software	http://www.ualberta.ca/~fyeh/
IBD	Gene and Distance relations statistics	http://www.bio.sdsu.edu/pub/andy/IBD.html
PRIMER-E	Population software and gene diversity statistics and indices	http://www.primer-e.com/

Gramene

An information resource established as a portal for grass species and grass genomics; including genome sequence information. The current version provides data on

24 plants; including 12 wild and domesticated rice genomes. An organelle data bank is also available from this site (Sect. 4.5).

NCBI—Entrez

Tracks over 800 whole genome projects from biological organisms, and the 115 species of Viridiplantae; including agronomically important crops for food and fruit, medicine and a number of green algae. The Entrez database can be accessed through the home page of NCBI.

NCBI—BLAST

One of the most important sites and tools available, in determining base similarities between nucleotide sequences in databanks. It also contains protein searches and queries. It includes searches for translated nucleotide sequences, conserved domains, multiple alignment tools, evolutionary relationships, and can be applied to all organisms, or limited to specific plants.

GrainGenes and PlantGDB

GrainGenes is a specific database for Triticeae and *Avena* genes, markers, maps and germplasm. PlantGDB contains sequences and a search engine linked to NCIB BLAST for 15 Dicotyledon, 7 Monocotyledon and 3 other plant species. It contains more limited information than the NCBI site over most plants, but is especially useful for agricultural grain species.

DNA Sequencing and UltraHigh-Throughput

Genome sequence information aids researchers in identify genes and gene families, including the identification of coding or non-coding regions, regulatory genes, and repetitive sequences within the genome (e.g. simple sequence repeats—SSRs); all of these are important in molecular biology. This type of information has become primary material for the design of genome advancements, such as microarrays, tilling arrays or molecular and chromosome markers, and these methods are important in whole plant genomic sequencing (Sect. 2.3 below). Pyrosequencing, massive parallel DNA sequencing and single molecule sequencing are adaptations of existing methods, which have become available in recent years (Margulies et al. 2005; Ansorge 2009). These new technologies have provided researchers with new methods to addressed web information in an entirely different way, and ongoing innovations in next-generation sequencing technology (Sect. 2.3), and the release of new

genome sequenced plants (listed in Table 2.1) is expected to accelerate the use of the DNA based web-information considerably in crop plants.

Whole Genome Sequencing

Information obtained from whole-genome sequencing in plants allows attempts at chromosome-scale genetic comparisons, thereby identifying conserved genetic areas, which can facilitate identification and documentation of similar genomic sequences in related plant species (Haas et al. 2004; De Bodt et al. 2005). Whole-genome comparisons identifying chromosomal duplication of alleles among related species for example can provide comparative evolutionary histories and diversification of species in ecology, taxonomy and plant breeding (Paterson et al. 2009; Schnable et al. 2009). Next-generation sequencing will allow identification of even more fundamental diversity and variation in genes amongst and between individuals, strains and/or populations. Single nucleotide polymorphisms (SNPs) have been central to these advancements, and SSR fragments have been shown to map consistently in many non-sequenced plant species; a capability that is of immensely important in genetic research. A genome re-sequencing project to identify whole-genome sequence variations in 1001 strains (accessions) of Arabidopsis is in progress. On completion this data will become an important resource for future genetics and population studies to identify alleles associated with phenotypes and diversity across entire plant species (http://1001genomes.org/) (Weigel and Mott 2009). In the same way a high-throughput method for genotyping recombinations in populations of rice, using whole-genome resequencing data generated by the Illumina Genome Analyzer has already been initiated (Huang et al. 2009).

Molecular (DNA) Markers

Identification and location of available molecular DNA markers have contributed significantly to marker-assisted studies and selection (MAS) in plant breeding, and in a wider range of research, including species identification and evolution. Genetic markers constructed to cover the complete genome may allow identification of individual genes associated with complex traits by QTL analysis, and the identification of genetic diversity and induced variations (Feltus et al. 2004; Varshney et al. 2005; Caicedo et al. 2007). Genome sequencing and large-scale EST databanks (Sect. 3) have become important for the construction of molecular markers, and a number of genome-wide rice DNA polymorphic markers have been constructed based on co-alignment between *japonica* and *indica* rice ESTs (Han and Xue 2003; Shen et al. 2004). Computer assisted EST-base single-nucleotide polymorphisms (SNPs) and/or EST-SNP markers for the purpose of identifying sequence-tagged sites (STS) has progressed for numerous species; including the crop plants of barley, wheat, maize, melon, *Brassica*, common bean, sunflower, potato, citrus and

grapevine (Mullins et al. 2006; Torada et al. 2006; Jaillon et al. 2007; Heesacker et al. 2008; Kota et al. 2008; Talon and Gmitter 2008; Blair et al. 2009; Deleu et al. 2009; Kaur et al. 2009; Li et al. 2009).

Some molecular markers identified this way allow the indirect selection of interesting genotypes (i.e. breeding lines in crops), and these cultivars constitute an essential tool for the development of marker-assisted selection (MAS) in plant breeding. The use of DNA markers (and indirectly EST markers from RNA) for direct selection offers greater potential gains in breeding for QTL and traits with low heritability, and these can be the most difficult to work with in crop breeding. However these low heritability traits are also amongst the most interesting and the most difficult to develop.

When a locus has many variants, or alleles, it is referred to as being polymorphic. Mutation(s) at a number of loci generate multiple alleles, most of which are eliminated from the population by genetic drift or breeding selection. Only a small number of alleles are incorporated into the population by chance or selection. Most polymorphisms can be genetically straightforward, with two alleles directly determining two versions of the same protein (gene), however, some can be highly complex, with multiple, related genes in a complex system of metabolic differences. Crop breeders have known the complexity of multiple alleles for decades. However with the advent of molecular markers, genetic diversity and other forms of genetic structure in breeding populations is possible. Listed in Table 2.2 are the most important web-based sites for DNA markers and some of the population statistics programs and web resources commonly in use. Molecular markers fall into a number of types listed below, each having positive and negative features, and careful consideration is required before they are adopted in any type of research (Hoang et al. 2009; De Filippis 2012).

Restriction Fragment Length Polymorphism (RFLP)

RFLP requires hydrolysis of probe DNA from samples. RFLP can provide high quality data but has severe restrictions on throughput because large amounts of DNA are required, and because it is not based on amplification of the target DNA via the polymerase chain reaction (PCR).

Random Amplified Polymorphic DNA (RAPD)

RAPD is a method based on PCR but uses arbitrary short primers (10 bases long) to identity plant DNA regions. No knowledge of the genome is needed, but by the same token markers can target many places on the genome. Results can be inconsistent and only dominant genes can be identified.

Simple Sequence Repeats (SSR)

SSR are high quality and consistent DNA markers, but they are the most expensive to develop. SSR markers require extensive band sequencing data for each marker developed, and often the markers are species and even cultivar specific. However they are molecular markers of choice in crop plants.

Amplified Fragment Length Polymorphism (AFLP)

AFLP requires enzymatic degradation of DNA and careful fragment separation, where only a sub-fraction of the population genetic data is sampled by PCR. It can provide too much information at any time. It is more technically demanding and information can be difficult to interpret. It produces very good high quality data, which is suitable for high output sources and automation.

Single Nucleotide Polymorphism (SNP)

SNP relies on the fact that the vast majority of differences in eukaryotic organisms are surprising point mutations in their DNA. So there are a vast number of polymorphisms that are SNPs. The biggest advantage is automation and techniques that do not require electrophoresis to separate fragments. However it does require DNA sequencing which can be costly. SNPs are becoming more and more important as molecular markers for genome information and advancement in crop plants.

Expressed Sequence Tags (EST)

ESTs require cDNA synthesis from RNA, and therefore are the only markers listed which are based on RNA. Preferences for this method should be for crop species where there is already extensive sequencing, and part or full EST data present (Sect. 3).

NCBI—Plant Markers

A genetic marker web database that contains molecular markers such as SNP, SSR and conserved ortholog set cosmid (COS) markers and primers from various plant resources (Heesacker et al. 2008).

GrainGenes

Theweb site for Triticeae genomics, and provides considerable detail of DNA markers and chromosome linkage map data on wheat, barley, rye and oat (Carollo et al. 2005).

http://www.springer.com/978-1-4614-7027-4

Crop Improvement
New Approaches and Modern Techniques
Hakeem, K.R.; Ahmad, P.; Öztürk, M. (Eds.)
2013, XVII, 493 p. 27 illus., 15 illus. in color., Hardcover
ISBN: 978-1-4614-7027-4