International Journal of Plant Genomics

Bioinformatics Tools for Plant Genomics

Guest Editors: Gary R. Skuse and Chunguang Du

Bioinformatics Tools for Plant Genomics

International Journal of Plant Genomics

Bioinformatics Tools for Plant Genomics

Guest Editors: Gary R. Skuse and Chunguang Du

Contents

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 910474, 2 pages
doi:10.1155/2008/910474

Editorial

Bioinformatics Tools for Plant Genomics

Gary R. Skuse[1] and Chunguang Du[2]

[1] *Bioinformatics Program, Department of Biological Sciences, Rochester Institute of Technology, Rochester, NY 14623, USA*
[2] *Science Informatics Program, Department of Biology and Molecular Biology, Montclair State University, Montclair, NJ 07043, USA*

Correspondence should be addressed to Gary R. Skuse, grssbi@rit.edu.

Received 31 December 2008; Accepted 31 December 2008

The articles in this special issue reflect a convergence of developments in the fields of bioinformatics and plant genomics. Bioinformatics has its roots vaguely seated in the early 1980s, a time when personal computers began appearing in research laboratories and researchers began recognizing that those computers could be used as tools to organize, analyze and visualize their data. In the ensuing years bioinformatics tools began appearing at various sites including the European Molecular Biology Laboratory, the Molecular Biology Research Resource at the Dana-Farber Cancer Institute in the mid 1980s, the National Center for Biotechnology Information (NCBI) in 1988, the Genome Database Project at Johns Hopkins University in early 1989, and in countless laboratories throughout the world. These last efforts resulted in the development of many of the tools described in this special issue.

Progress and interest in plant genomics have been accelerating since the time in late 2000 when the genome of *Arabidopsis thaliana* was published. Since then many genome sequencing projects have been undertaken that include poplar *(Populus*), grape (*Vitis*), the moss *Physcomitrella*, the biflagellate algae *Chlamydomonas* and several globally crucial crop plants such as corn (Maize) and rice (*Oryza)*. However, as we have witnessed on numerous occasions, determining the sequence of a genome is only the first step toward understanding genome organization, gene structure, gene expression patterns, disease pathogenesis and a host of other features of both scientific and commercial interests. Computational tools of genomic annotation and comparative genomics must be applied to gain a useful understanding of any genome.

In this special issue we present a collection of papers that together describe a powerful and impactful toolbox of applications and resources for plant genomic analysis. Among those articles you will find a description of research performed by the Mexican headquartered Generation Challenge Programme (GCP) which led to the GCP Platform (Bruskiewich et al.). This research support tool supports a number of data formats and web services and provides access to high performance computing facilities and platform-specific middleware collectively designed to support crop science research.

Probably one of the most promising empirical tools for investigating gene expression developed in the last 15 or so years is that of microarray technology. While the technology has become commonplace, with tools for generating and hybridizing arrays available to all, the analysis of microarray-derived data has been challenging. Many laboratories have struggled not only with this challenge but also with the task of sorting through the plethora of analytical tools available in an effort to find the ones that may be best suited to their own work. In this issue there are two reviews by Page and Coulibaly which examine and describe bioinformatics tools for inferring functional information from plant microarray data. Together these papers step the reader through a collection of tools, and their applications, for analyzing the expression of single and multiple gene expression profiles.

This theme of microarray analysis is continued in the description of the cross chip probe matching tool (CCPMT) by Page et al. Indeed it expands the readers horizons beyond the analysis of individual microarrays with the ability to associate probes across species. And of course, microarray analysis is facilitated by careful experimental design from the start so Robert Tempelman provides a review of statistical methods used to design efficient two-color microarray experiments. Taken together, these microarray

papers provide an overview of the design of microarray experiments and the interpretation of the complex results of those experiments that will be informative for new and experienced laboratorians alike.

Several other novel tools are described herein. One, Blast2GO is a suite of tools for the analysis and functional annotation of plant genomes (Conesa and Goetz). It provides an intuitive interface for identifying functional regions within DNA sequences. Another sequence analysis tool described by da Maia et al. is the SSR locator. That tool enables researchers to identify suitable targets for binding PCR primers in order to ensure that those targets are unique within the genome. It also assists with primer design and has a PCR simulator which facilitates comparisons of hypothetical amplification products among different species.

Another challenge facing scientists today is the need to stay abreast of advances in a field that is progressing rapidly as a consequence of newly available technologies. In order to address this challenge there are two review articles that together provide insights into the discovery of relationships among a varied array of plant species. The first article, by Abdurakhmonov and Abdukarimov, describes the application of association mapping to understanding traits in crop species. Their work is directed toward novices within the crop breeding community in order to expose them to potential problems that they may face and solutions they may employ to overcome those problems. The second article describes the tools available for phylogenetic analyses and the increased use of Bayesian methods in those tools (Aris-Brosou and Xia). Constructing phylogenies has traditionally been a challenge to even the most experienced researcher but modern bioinformatics tools are lowering the bar for those interested in detecting adaptive evolution and estimating divergence among species.

The wealth of information available to researchers today can be overwhelming. In order to address this potential, two papers describe information resources which consolidate and organize related information. PPNEMA is a database resource for those interested in plant-parasitic nematode ribosomal genes (Rubino et al.). That resource allows the user to browse, search and generally explore phytoparasite ribosomal DNA. A second database described in these pages is the MaizeGDB (Lawrence et al.). This resource contains information about *Zea mays* which includes genomic sequences as well as functional information and the tools to explore both.

The body of the papers in this special issue represents the leading edge of plant genomics research. Together they provide the reader with descriptions of the tools and resources necessary to understand and promote advances in this important field.

Gary R. Skuse
Chunguang Du

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 147563, 9 pages
doi:10.1155/2008/147563

Review Article

Bioinformatic Tools for Inferring Functional Information from Plant Microarray Data: Tools for the First Steps

Grier P. Page and Issa Coulibaly

Department of Biostatistics, University of Alabama at Birmingham, 1665 University Blvd Ste 327, Birmingham, AL 35294-0022, USA

Correspondence should be addressed to Grier P. Page, gpage@uab.edu

Received 2 November 2007; Accepted 7 May 2008

Recommended by Gary Skuse

Microarrays are a very powerful tool for quantifying the amount of RNA in samples; however, their ability to query essentially every gene in a genome, which can number in the tens of thousands, presents analytical and interpretative problems. As a result, a variety of software and web-based tools have been developed to help with these issues. This article highlights and reviews some of the tools for the first steps in the analysis of a microarray study. We have tried for a balance between free and commercial systems. We have organized the tools by topics including image processing tools (Section 2), power analysis tools (Section 3), image analysis tools (Section 4), database tools (Section 5), databases of functional information (Section 6), annotation tools (Section 7), statistical and data mining tools (Section 8), and dissemination tools (Section 9).

1. INTRODUCTION

The primary goal of a microarray study is to generate a list of differentially regulated genes and infer pathways that can provide insight into the biological question under investigation. Due to the very high dimensionality of a microarray experiment, running to thousand of genes, bioinformatics, and statistical tools are essential for the analysis of data. This review is written to provide plant investigators with a list of tools and web-based resources designed to help them move from an idea or hypothesis to the conduct of the study, image analysis, generation of expression data, statistical analysis, annotation, and then dissemination of the data.

The first step in the conduct of a microarray study is the selection of a microarray platform to use. For many species, there are commercially available arrays from commercial vendors and academic groups. Unfortunately, arrays are not available for all species, while arrays can be used in closely related species, it is usually better to develop arrays based upon the sequence of the species being studied. Section 2 provides a list of tools for generating useful probe sequences from genomic data. Once an array has been developed, it is critical to collect sufficient samples to run an experiment that will generate biologically generalizable results. Section 3 highlights tools for sample size and power analysis for microarray studies. Image analysis tools (Section 4) are used to quantitate the amount of fluorescence for a spot or set of spots. Microarray experiments generate copious amounts of data. The storage and distribution of the data are accomplished by the tools described in Section 5. Databases of gene annotations are provided in Section 6. Sections 7 and 8 describe statistical analysis and annotation tools. The two grouped together for the same tools often provide both functions. In fact, many of the database tools will also provide analytical and annotation functions as well. Finally, in Section 9 we describe web sites for disseminating microarray data and analyses.

2. PROBE DESIGN SOFTWARE

Plant scientists conduct their research on a wide variety of plant taxa. Arrays have been developed for a number of plant species including Arabidopsis, Maize, Populus, Rice, Barley, Grape, Citrus, Cotton, Medicago, Soybean, Sugar Cane, Tomato, and Wheat. While arrays can be used on closely related species, it is often better to design a new array for the species of interest. Several tools have been designed to help design probes for spotting or deposition on arrays, based upon genomic sequence data. The critical stage is to

have high-quality sequence data. The more complete the genome is, the easier it will be to design probes that will not cross hybridize, be subject to SNPs, and query the gene accurately. Table 1 lists a number of tools for probe design; many of them are free, but a number specific to a single array manufacturer.

3. POWER ANALYSIS AND SAMPLE SIZE CALCULATIONS

One of the keys to a successful microarray study is to collect enough data (arrays) in order to derive biologically generalizable results. The key to this is the statistical power of a study. Power is the probability of being able to detect a significant difference between experiment groups when one really exists. There are several factors involved in power, but the main one under the control of an investigator is the sample size. A study with too few samples may not detect real differences, while too many samples will waste resources. Power analysis allows the selection of the optimal sample size. While sample sizes for microarrays can be planned with traditional statistical power calculation tools such as PS (http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize), the unique features of arrays such as the large number of tests and the large number of genes that are different between groups have lead to the development of several methods and tools for calculating power and sample size analysis.

3.1. The Power Atlas

The Power Atlas is a web-based resource to assist investigators in the planning and design of microarray and expression-based experiments. This software currently aims at estimating the power and sample size for a two group comparison based upon pilot data. The methods underlying the web site are reported in Gadbury et al. [1] and the software is described in further detail at Page et al. [2]. The tool may be used in two manners: one may either upload one's own pilot data or select a pilot dataset from over 1 000 public data sets. Output includes graphs of power for a variety of significance and false discovery rates; see http://www.poweratlas.org/ [2].

3.2. Significance analysis of microarrays (SAM)

SAM is a free flexible Excel Addin that includes a number of useful functions for the analysis of microarray data. Tools include statistical analysis for discrete, quantitative, and time series data, adjustments for multiple testing, gene set enrichment analysis, sample size assessment, estimates of False Discovery rate (FDR) and q-value, as well as per gene power analysis; see http://www-stat.stanford.edu/~tibs/SAM/ [3].

4. IMAGE ANALYSIS SOFTWARE

The purpose of image analysis software is to generate a quantified expression score from the scanned microarray images. Some of the tools are specific to particular array types, and thus are not appropriate for all array types. There are a number of tools that are available in this area, many of which are expensive. We present here tools that are still being actively supported and developed. Additional tools are listed in Table 2.

4.1. Affy

This is a package in Bioconductor for processing Affymetrix arrays. A wide variety of image processing, normalization, and quality control procedures are available. As a note, there are a variety of other image processing tools in Bioconductor including PDNN and DCHip that should be considered for use as well; see http://www.bioconductor.org/packages/2.1/bioc/html/affy.html [4].

4.2. Affyprobe miner

Affyprobe miner is used to redefine chip definition files (CDFs) for Affymetrix chips to take into account the more recent genomic sequence information on SNP, alternative splicing, changes in the gene model, exon structure, and other such genomic difference. Precomputed CDFs for several chips are available for download; see http://gauss.dbb.georgetown.edu/liblab/affyprobeminer/ [5].

4.3. Beadarray

This is a function in Bioconductor for reading preprocessed Illumina Bead summary data as well as reconstructing bead-level data using raw TIFF images. Methods for quality control and low-level analysis are also provided; see http://www.bioconductor.org/packages/2.1/bioc/html/beadarray.html [6].

4.4. Genechip operating software (GCOS)

Affymetrix GCOS automates the control of GeneChip Fluidics Stations and Scanners. In addition, GCOS acquires data, manages sample and experimental information, and performs gene expression data analysis. GCOS can quantitate images using MAS 5 and PLIER; see http://www.affymetrix.com/products/software/specific/gcos.affx.

4.5. Gene pix pro 6.0

This software has a number of useful features including imaging, spot finding, quality control, analysis tools, visualizations, and automation capabilities. GenePix can display and process up to four single wavelengths, thus four-channel imaging can be used. This tool can be integrated with a web-accessible database. GenePix is in some ways the default industrial standard microarray image analysis software because of its early development of couple of output file formats, $*$.gpr and $*$.gps that are used by many other applications; see http://www.moleculardevices.com/.

Table 1: Probe design software packages.

Tool and website	Cost and functions of the tool
Array Designer http://www.premierbiosoft.com/dnamicroarray/index.html	Design primers and probes for oligo and cDNA expression microarrays. It can also design probes for SNP detection, single exon, whole gene, tiling, and resequencing arrays. The software is not free.
ArrayScribe http://www.nimblegen.com/products/software/arrayscribe.html	Free, but limited to designing NimbleGen Arrays. The tool can design probes, specify mismatches at specific sequence positions, automatically generate mismatches, generate multiple probes for a gene, and design the placement of spots on an array.
eArray http://earray.chem.agilent.com/earray/login.do	Free, but limited to designing Agilent arrays. Can design probes for expression, CGH, and ChiP for any species with genomic sequence.
Primer3Plus http://www.bioinformatics.nl/cgi-in/primer3plus/primer3plus.cgi	Free software that can design probes for expression detection on arrays, amplification/cloning, and sequencing/resequencing.
Sarani Oligo Design http://www.strandls.com/oligodesign.html	Probe design for expression analysis. The software is not free.
Visual OMP http://www.dnasoftware.com/Products/VisualOMP	Design software for RNA, DNA, single or multiple probe design, microarrays, TaqMan assays, genotyping, single and multiplex PCR, secondary structure simulation, sequencing, genotyping.

Table 2: Other useful image analysis software packages.

Tool name	Web site
Able Image Analyser	http://able.mulabs.com/
ArrayVision	http://www.imagingresearch.com/products/ARV.asp
IcononClust	http://www.clondiag.com/frame.php?page=/products/sw/iconoclust/index.php
ImaGene	http://www.biodiscovery.com/index/imagene
Koadarray	http://www.koada.com/koadarray/
Microvigene	http://www.vigenetech.com/MicroVigene.htm
ScanAlyze	http://rana.lbl.gov/EisenSoftware.htm
Spot	http://www.hca-vision.com/productspot.html

4.6. Nimblescan

This is a NimbleGen product designed for the extraction of feature intensity raw values, linkage of the raw intensity values with the corresponding probe parameters, and generation of analysis reports for expression, ChIP-chip and resequencing arrays, and methylation analysis for NimbleGen Arrays; see http://www. nimblegen.com/products/software/nimblescan.html.

4.7. TM4/spotfinder

Spotfinder is part of the larger freely available microarray analysis suite TM4. Spotfinder is designed for the rapid, reproducible, and computer-aided analysis of microarray images, and the quantification of gene expression. Spotfinder can read paired 16-bit or 8-bit TIFF image files generated by most microarray scanners. Automatic, semiautomatic and manual grid construction and adjustments can be made. Two segmentation methods are available. Reusable grid geometry files and automatic grid adjustment allow user to analyze large quantities of images in a consistent and efficient manner. Quality control views allow the user to assess systematic biases in the data; see http://www.tm4.org/spotfinder.html [7, 8].

5. DATABASE TOOLS

Microarray experiments generate a huge amount of data. The handling, storing, sharing, and distribution of the data can be quite complex. As a result a variety of database tools

have been developed for assisting in this aspect of microarray studies. Some of the tools listed below are more than just stand-alone database tools and may include extensive analysis and visualization functionality as well. There are a number of database tools with highly different utility and platform requirements. Table 3 outlines the tools and websites.

6. DATABASES OF FUNCTIONAL INFORMATION

The amount of information about the functions of genes is beyond what any one person can know. Consequently, it is useful to pull in information on what others have discovered about genes in order to fully and correctly interpret an expression study. The following tools are databases of various types on information such as published papers, gene sequences, pathways, and ontologies that might be useful for an investigator who is interpreting an expression study.

6.1. Agbase

AgBase is a curated, open-source, web-accessible resource for functional analysis of agricultural plant and animal gene products. Agbase contains databases of Poplar and Pine gene ontology terms and annotations as well as several animals, microbes, and parasites; see http://www.agbase. msstate.edu) [9, 10].

6.2. Agricola

Agricola is the catalog and index to the collections of the National Agricultural Library. The database covers materials in all formats and periods, dating back to the 15th century. The records include all aspects of agriculture and related disciplines; see http://agricola.nal.usda.gov/.

6.3. Eukaryotic gene orthologues (EGO)

EGO is generated by the pair-wise comparison between the tentative consensus (TC) sequences from individual organisms. The reciprocal pairs of the best match are clustered into individual groups and multiple sequence alignments are displayed for each group. EGO is very useful for connecting homologous genes across species; see http://compbio.dfci.harvard.edu/tgi/ego/ [11].

6.4. Ensembl

Ensembl is a joint project between European Bioinformatics Institute and the Wellcome Trust Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Initially developed for vertebrates, Ensembl has been adapted for use by several plant groups including legume, Gramene, and Arabidopsis; see http://www.ensembl.org/index.html [12].

6.5. Entrez gene

Entrez Gene is an NCBI's database for gene-specific information. Entrez Gene focuses on the genomes that have been completely sequenced, have an active research community to contribute gene-specific information, or that are scheduled for intense sequence analysis. Records are assigned unique, stable and tracked integers as identifiers. The content (nomenclature, map location, gene products and their attributes, markers, phenotypes, and links to citations, sequences, variation details, maps, expression, protein homologs, protein domains and external databases) is updated regularly. There is currently at least some gene information on 113 plant species; see http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene.

6.6. Gene index

The goal of The Gene Index Project is to use the available EST and gene sequences, along with the reference genomes, to provide an inventory of likely genes and variants. Genes are linked to annotation regarding their functions. Currently GI databases have been constructed for 34 plant species; (http://compbio.dfci.harvard.edu/tgi/plant.html) [13, 14].

6.7. Gene ontology

The objective of GO is to provide controlled vocabularies for the description of the molecular function, biological process, and cellular component of gene products. These terms are to be used as attributes of gene products by various collaborating databases such as Gramene and TAIR; see http://www.geneontology.org/ [15].

6.8. Gramene

Gramene is a curated, open-source, data resource for genome analysis in the grasses. The information stored in the database is derived from public sources and includes genomes, EST sequencing, protein structure and function analysis, genetic and physical mapping, interpretation of biochemical pathways, Gene Ontologies, gene and QTL localization and descriptions of phenotypic characters and mutations. Extensive information is provided for *Oryza, Zea, Triticum, Hordeum, Avena, Setaria, Pennisetum, Secale, Sorghum, Zizania,* and *Brachypodium*; see http://www.gramene.org/.

6.9. Kyoto encyclopedia of genes and genomes (KEGG)

KEGG is a database of biological systems, consisting of genes and proteins (KEGG GENES), endogenous and exogenous substances (KEGG LIGAND), pathways (KEGG PATHWAY), and hierarchies and relationships of biological objects (KEGG BRITE). This database is very rich in data with information across hundreds of species including many plants; see http://www.genome.jp/kegg/ [16–18].

6.10. Plant associated microbe gene ontology (PAMGO)

PAMGO is a database of the results of a multiinstitutional collaborative effort, aimed at developing new GO terms and

Table 3: Database tools.

Tool name	Web site
Acuity	http://www.moleculardevices.com/pages/software/gnacuity.html
Array Results Manager ARM	http://www.biodiscovery.com/index/arm
Arraytrack	http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/ [35, 36]
BASE 2	http://base.thep.lu.se/
caArray	http://caarray.nci.nih.gov/
Expressionist	http://www.genedata.com/products/expressionist/index_eng.html
Gene Array Analyzer Software GAAS	http://www.medinfopoli.polimi.it/GAAS/
GeneDirector	http://www.biodiscovery.com/index/genedirector
GeneSpring Workgroup	http://www.chem.agilent.com/scripts/pds.asp?lpage=34668
GeneTraffic	http://www.iobion.com/products/products_GENETRAFFIC.html
Genowiz	http://www.ocimumbio.com/
Longhorn Array Database LAD [37]	http://www.longhornarraydatabase.org/
MaxdLoad2	http://www.bioinf.man.ac.uk/microarray/maxd/index.html
PARTISAN arrayLIMS	http://www.clondiag.com/
Rosetta Resolver System	http://www.rosettabio.com/products/resolver/default.htm
Stanford Microarray Database SMD	http://smd-www.stanford.edu//download/ [38]

relationships for gene products implicated in plant-pathogen interactions. GO terms are currently being developed for the following species: *Erwinia chrysanthemi*, *Pseudomonas syringae* pv tomato and *Agrobacterium tumefaciens*, the fungus *Magnaporthe grisea*, the oomycetes *Phytophthora sojae* and *Phytophthora ramorum*, and the nematode *Meloidogyne hapla*; *see* http://pamgo.vbi.vt.edu/.

6.11. SWISS-PROT

SWISS-PROT is a curated protein sequence database which provides high level of annotations such as the description of the function of a protein, its domains structure, post-translational modifications, variants, and so forth, along with good integration with other databases; see http://www.expasy. ch/sprot/.

6.12. TAIR

The Arabidopsis Information Resource (TAIR) maintains a database of genetic and molecular biology data for *Arabidopsis thaliana.* Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, and publications; see http://www.arabidopsis.org/.

7. ANNOTATION TOOLS

The databases described in Section 6 can provide data in a variety of forms, which makes merging the annotations with the expression difficult. To deal with this heterogeneity a number of tools have been developed to increase the ease of annotating genes in expression studies.

7.1. CiteXplore

CiteXplore combines literature search with text mining tools for biology. Search results are cross referenced to European Bioinformatics Institute applications based on publication identifiers. Links to full text versions are provided where available; see http://www.ebi.ac.uk/citexplore/.

7.2. Database for annotation, visualization, and integrated discovery (DAVID)

DAVID provides a huge set of functional annotation tools for investigators to understand biological meaning behind a large list of genes. The key is the DAVID Knowledgebase which provides a comprehensive, high-quality collection of gene annotation resource, the flexibility to cross-reference gene identifiers, and heterogeneous annotations from almost all databases. The DAVID tools are able to identify enriched biological themes, particularly GO terms, cluster redundant annotation terms, visualize genes on Baccarat and KEGG pathway maps, display related many-genes-to-many-terms on 2D view, search for other functionally related genes not in the list, list interacting proteins, highlight protein functional domains and motifs, redirect to related literatures, and convert gene identifiers from one type to another; see http://david.abcc.ncifcrf.gov/ [19].

7.3. MatchMiner

MatchMiner translates between several gene identifier types for the same list of hundreds or thousands of genes. The gene identifier types supported by the tool includes GenBank accession numbers, IMAGE clone IDs, common gene names, HUGO names, gene symbols, UniGene clusters, FISH-mapped BAC clones, Affymetrix identifiers, and chromosome locations. MatchMiner can also find the intersection

of two lists of genes specified by different identifiers; see http://discover.nci.nih.gov/matchminer/index.jsp [20].

7.4. Medminer

MedMiner searches and organizes the biomedical literature on genes, gene-gene relationships, and gene-drug relationships. It uses GeneCards, PubMed, and syntactic analysis, truncated-keyword filtering of relational and user-controlled sculpting of Boolean queries to generate key sentences from pertinent abstracts. Abstracts selected can be automatically entered into EndNote; see http://discover.nci.nih.gov/textmining/main.jsp [21].

8. DATA ANALYSIS SOFTWARE

There is an incredible breadth of tools in this area with many tools providing very slick interfaces and very useful functions; however, you really do not need any of these tools. Most statistical packages such as SAS, SPSS, JMP, and R can be used to analyze microarray data and will do most of the functions the following tools will do, for there are few statistical methods that are 100% unique to expression studies. Nonetheless many of the following tools are much easier to use and often have better visualization functions than the pure statistical programs. Typically the tools have been designed for ease of use, often too easy. Regardless of the tool you use, strive to understand the function and analyses provided and the assumption that are made when you choose to use them for analysis. For example, in cluster analysis you need to make a choice of link and weight functions and the clusters that result will be quite different based on methods which are chosen. There are similar issues to learn and understand for all statistical methods and most visualization methods. Additional tools are listed in Table 4.

8.1. Bioconductor

Bioconductor is a multicenter effort to develop tools in the R programming environment for analyzing genomic data, especially microarray data. There are a large number of different packages available to conduct many types of analyses; currently there are over 115 microarray applications. Tools are still in very active development, and are all freely available. Some of the most relevant tools are affy, maanova, genefilter, limma, mulltest, annotate, geneplotter, marray to name a few. A couple of the packages are described elsewhere in this document, but for more details of specific tools see the Bioconductor web site; see http://www.bioconductor.org/ [22].

8.2. Biometric research branch (BRB) arrays tools

BRB ArrayTools is a free integrated package for the visualization and statistical analysis of DNA microarray gene expression data. It functions as an Excel Addin. It was developed by professional statisticians experienced in the analysis of microarray data. It is probably the best tool available for discriminate analysis and has a variety of other statistical and cluster methods included; see http://linus.nci.nih.gov/BRB-ArrayTools.html.

8.3. Expression profiler

Expression Profiler is a set of tools for cluster analysis, pattern discovery, pattern visualization, study and search for gene ontology categories. The tool also generates sequence logos, extracts regulatory sequences, studies protein interactions, and links analysis results to external tools and databases; see http://ep.ebi.ac.uk/.

8.4. Genepattern

GenePattern puts sophisticated computational methods into the hands of the biomedical research community. A simple application interface gives a broad audience access to a growing repository of analytic tools for genomic data, while an API supports computational biologists. GenePattern is a powerful analysis workflow tool developed to support multidisciplinary genomic research programs and designed to encourage rapid integration of new techniques; see http://www.broad.mit.edu/cancer/software/genepattern/index.html [23].

8.5. GeneXpress

GeneXPress is a visualization and analysis tool for gene expression data, integrating clustering, gene annotation, and sequence information. GeneXPress allows the user to load clustering results and automatically analyze them for significance of functional groups through correlation with functional annotations (e.g., Gene Ontology) and for enrichment of motif binding sites (e.g., TRANSFAC motifs); see http://genexpress.stanford.edu/.

8.6. GEPAS (gene expression pattern analysis suite)

GEPAS is an integrated web-based tool for the analysis of gene expression data. GEPAS includes tools for normalization, many clustering methods, supervised analysis, differential analysis, differential gene expression, predictors, array CGH and functional annotation; see http://gepas.bioinfo.cipf.es/ [24, 25].

8.7. High-dimensional biology statistics (HDBStat!)

HDBStat is a free java application that allows for the normalization, transformation, and statistical analysis of expression data. HDBStat also has a number of unique quality control procedures included. HDBStat has implemented reproducible research design to allow for analysis to be readily repeated; (http://www.ssg.uab.edu/hdbstat/) [26].

8.8. JMP genomics

JMP genomics leverages many statistical tools in JMP, a statistical analysis package, as a result it has over 100 different analytical procedures that can be run. It also includes

Table 4: Other useful statistical analysis and data-mining tools.

Tool name	Web site
Amiada (analyzing microarray data)	http://dambe.bio.uottawa.ca/amiada.asp [39]
ArrayAssist Enterprise	http://www.stratagene.com/
caGEDA	http://bioinformatics.upmc.edu/GE2/GEDA.html
Cluster	http://rana.lbl.gov/EisenSoftware.htm
dChip	http://www.dchip.org/
GeneMaths XT	http://www.applied-maths.com/genemaths/genemaths.htm
INCLUSive	http://homes.esat.kuleuven.be/~dna/Biol/Software.html
J-Express Pro	http://www.molmine.com/software.htm
MAExplorer	http://maexplorer.sourceforge.net/
NIA Array analysis	http://lgsun.grc.nia.nih.gov/ANOVA/
Onto-Tools	http://vortex.cs.wayne.edu/projects.htm
Probe Profiler	http://www.corimbia.com/Pages/ProductOverview.htm
TableView	http://ccgb.umn.edu/software/java/apps/TableView/
Venn Mapper	http://www.gatcplatform.nl//vennmapper/index.php

extensive visualization tools. Scripts can be written for the development of standard analytical procedures; see http://www.jmp.com/software/genomics/.

8.9. Onto-tools

Onto-Tools are a series of freely available tools for the analysis of microarray data. Tools are available for array design (onto-design), gene class testing (onto-express), comparing the content of arrays (onto-compare), mapping gene information across databases (onto-translate), annotation (onto-miner), and pathway analysis (pathway-express); see http://www.vortex.cs.wayne.edu [27].

8.10. Partek genomic suite

Partek Genomics Suite can be used for gene expression analysis, exon expression analysis, chromosomal copy number analysis, and promoter tiling array analysis, and analysis of SNP arrays. Partek includes a large number of statistical, visualization, and annotation tools that can be tied together using workflow tools for rapid repetition of analysis and for reproducible research; see http://www.partek.com/software/.

8.11. R/maanova

Maanova stands for MicroArray ANalysis Of VAriance. It provides a complete work flow for microarray data analysis including data-quality checks and visualization, data transformation, ANOVA model fitting for both fixed and mixed effects models, statistical tests including permutation tests, confidence interval with bootstrapping, and cluster analysis. R/maanova is available in Bioconductor/R; refer to http://www.jax.org/staff/churchill/labsite/software/Rmaanova/index.html [28].

8.12. SAM (significant analysis of microarrays)

SAM can be used on any type of array data: oligo or cDNA arrays, SNP arrays, protein arrays, and so forth. Both parametric and nonparametric tests are available for correlating expression data to clinical parameters including treatment, diagnosis categories, survival time, paired data, quantitative (e.g., tumor volume), and one-class. SAM can also implement imputation methods for missing data via nearest neighbor algorithm; see http://www-stat.stanford.edu/~tibs/SAM/.

8.13. TM4

The TM4 suite of tools consists of four major applications, Microarray Data Manager (MADAM), TIGR_Spotfinder, Microarray Data Analysis System (MIDAS), and Multiexperiment Viewer (MeV), as well as a MySQL database, all of which are freely available. Although these software tools were developed for spotted two-color arrays, many of the components can be easily adapted to work with single-color formats such as filter arrays and GeneChips; see http://www.tm4.org/index.html.

9. DISSEMINATION

Early in the use of microarray in research, it became common practice for many journals to require investigators to submit expression data for publication in a public database. This sharing of data has allowed the mining of these rich resources that many investigators have used to help their research. A number of the public databases exist that contain and accept plant data.

9.1. ArrayExpress

ArrayExpress is a public repository for microarray data, which is aimed at storing MIAME-compliant data in

accordance with MGED recommendations. This database is a bit less biomedical in focus than GEO with a good representation of plant expression data; see http://www.ebi.ac.uk/arrayexpress [29, 30].

9.2. GEO

Gene Expression Omnibus is a gene expression/molecular abundance repository supporting MIAME compliant data submissions, and a curated, online resource for gene expression data browsing, query and retrieval. This is supported by the US National Library of Medicine, but contains a good amount of plant expression data; see http://www.ncbi.nlm.nih.gov/projects/geo/ [31, 32].

9.3. NASC (nottingham arabidopsis stock center) arrays

NASC runs a database of its own arrays as well as other data that has been deposited in the database. The database primarily contains Arabidopsis array data; see http://affymetrix.arabidopsis.info/ [33].

9.4. Plant expression database (PlexDB)

PLEXdb is a unified public resource for gene expression for plants and plant pathogens. PLEXdb serves as a portal to integrate gene expression profile data sets with structural genomics and phenotypic data. Data from seven species is contained in the database; see http://www.plexdb.org/index.php [34].

10. CONCLUSIONS

We hope this listing of tools, which only dip the surface of the possible tools, will assist you in conducting, analyzing, and interpreting expression studies. We suggest exploring several tools in an area and understanding the principles of the methods implemented before settling on one or a few to use regularly. By exploring several tools you will understand the potential of the various tools, how easy (or difficult) they are to use, and determine what you really want and need for your microarray analysis.

ACKNOWLEDGMENT

The work on this grant was supported by NSF grant 0501890 and NIH grant U54 AT100949.

REFERENCES

[1] G. Gadbury, G. P. Page, J. Edwards, et al., "Power analysis and sample size estimation in the age of high dimensional biology," *Statistical Methods in Medical Research*, vol. 13, pp. 325–338, 2004.

[2] G. P. Page, J. W. Edwards, G. L. Gadbury, et al., "The *PowerAtlas*: a power and sample size atlas for microarray experimental design and research," *BMC Bioinformatics*, vol. 7, article 84, 2006.

[3] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[4] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "Affy—analysis of *Affymetrix GeneChip* data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

[5] H. Liu, B. R. Zeeberg, G. Qu, et al., "AffyProbeMiner: a web resource for computing or retrieving accurately redefined Affymetrix probe sets," *Bioinformatics*, vol. 23, no. 18, pp. 2385–2390, 2007.

[6] M. J. Dunning, M. L. Smith, M. E. Ritchie, and S. Tavaré, "Beadarray: R classes and methods for Illumina bead-based data," *Bioinformatics*, vol. 23, no. 16, pp. 2183–2184, 2007.

[7] A. I. Saeed, N. K. Bhagabati, J. C. Braisted, et al., "TM4 microarray software suite," *Methods in Enzymology*, vol. 411, pp. 134–193, 2006.

[8] A. I. Saeed, V. Sharov, J. White, et al., "TM4: a free, open-source system for microarray data management and analysis," *BioTechniques*, vol. 34, no. 2, pp. 374–378, 2003.

[9] F. M. McCarthy, S. M. Bridges, N. Wang, et al., "AgBase: a unified resource for functional analysis in agriculture," *Nucleic Acids Research*, vol. 35, database issue, pp. D599–D603, 2007.

[10] F. M. McCarthy, N. Wang, G. B. Magee, et al., "AgBase: a functional genomics resource for agriculture," *BMC Genomics*, vol. 7, article 229, 2006.

[11] Y. Lee, J. Tsai, S. Sunkara, et al., "The TIGR Gene Indices: clustering and assembling EST and know genes and integration with eukaryotic genomes," *Nucleic Acids Research*, vol. 33, database issue, pp. D71–D74, 2005.

[12] T. J. P. Hubbard, B. L. Aken, K. Beal, et al., "Ensembl 2007," *Nucleic Acids Research*, vol. 35, database issue, pp. D610–D617, 2007.

[13] J. Quackenbush, J. Cho, D. Lee, et al., "The TIGR Gene Indices: analysis of gene transcipt sequences in highly sampled eukaryotic species," *Nucleic Acids Research*, vol. 29, no. 1, pp. 159–164, 2001.

[14] J. Quackenbush, F. Liang, I. Holt, G. Pertea, and J. Upton, "The TIGR Gene Indices: reconstruction and representation of expressed gene sequences," *Nucleic Acids Research*, vol. 28, no. 1, pp. 141–145, 2000.

[15] M. Ashburner, C. A. Ball, J. A. Blake, et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[16] M. Kanehisa, "The KEGG database," *Novartis Foundation Symposium*, vol. 247, pp. 91–101, 2002.

[17] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, "The KEGG databases at GenomeNet," *Nucleic Acids Research*, vol. 30, no. 1, pp. 42–46, 2002.

[18] M. Kanehisa, S. Goto, M. Hattori, et al., "From genomics to chemical genomics: new developments in KEGG," *Nucleic Acids Research*, vol. 34, database issue, pp. D354–D357, 2006.

[19] G. Dennis Jr., B. T. Sherman, D. A. Hosack, et al., "DAVID: database for annotation, visualization, and integrated discovery," *Genome Biology*, vol. 4, no. 5, article P3, 2003.

[20] K. J. Bussey, D. Kane, M. Sunshine, et al., "MatchMiner: a tool for batch navigation among gene and gene product identifiers," *Genome Biology*, vol. 4, no. 4, article R27, 2003.

[21] L. Tanabe, U. Scherf, L. H. Smith, J. K. Lee, L. Hunter, and J. N. Weinstein, "MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling," *BioTechniques*, vol. 27, no. 6, pp. 1210–1217, 1999.

[22] R. C. Gentleman, V. J. Carey, D. M. Bates, et al., "Bioconductor: open software development for computational biology

and bioinformatics," *Genome Biology*, vol. 5, no. 10, article R80, 2004.

[23] M. Reich, T. Liefeld, J. Gould, J. Lerner, P. Tamayo, and J. P. Mesirov, "GenePattern 2.0," *Nature Genetics*, vol. 38, no. 5, pp. 500–501, 2006.

[24] J. Herrero, F. Al-Shahrour, R. Díaz-Uriarte, et al., "GEPAS: a web-based resource for microarray gene expression data analysis," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3461–3467, 2003.

[25] J. M. Vaquerizas, L. Conde, P. Yankilevich, et al., "GEPAS, an experiment-oriented pipeline for the analysis of microarray gene expression data," *Nucleic Acids Research*, vol. 33, web server issue, pp. W616–W620, 2005.

[26] P. Trivedi, J. W. Edwards, J. Wang, et al., "HDBStat!: a platform-independent software suite for statistical analysis of high dimensional biology data," *BMC Bioinformatics*, vol. 6, article 86, 2005.

[27] P. Khatri, P. Bhavsar, G. Bawa, and S. Draghici, "Onto-Tools: an ensemble of web-accessible ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments," *Nucleic Acids Research*, vol. 32, web server issue, pp. W449–W456, 2004.

[28] M. K. Kerr, M. Martin, and G. A. Churchill, "Analysis of variance for gene expression microarray data," *Journal of Computational Biology*, vol. 7, no. 6, pp. 819–837, 2000.

[29] A. Brazma, H. Parkinson, U. Sarkans, et al., "ArrayExpress—a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research*, vol. 31, no. 1, pp. 68–71, 2003.

[30] H. Parkinson, M. Kapushesky, M. Shojatalab, et al., "ArrayExpress—a public database of microarray experiments and gene expression profiles," *Nucleic Acids Research*, vol. 35, database issue, pp. D747–D750, 2007.

[31] T. Barrett, D. B. Troup, S. E. Wilhite, et al., "NCBI GEO: mining tens of millions of expression profiles—database and tools update," *Nucleic Acids Research*, vol. 35, database issue, pp. D760–D765, 2007.

[32] T. Barrett, T. O. Suzek, D. B. Troup, et al., "NCBI GEO: mining millions of expression profiles—database and tools," *Nucleic Acids Research*, vol. 33, database issue, pp. D562–D566, 2005.

[33] D. J. Craigon, N. James, J. Okyere, J. Higgins, J. Jotham, and S. May, "NASCArrays: a repository for microarray data generated by NASC's transcriptomics service," *Nucleic Acids Research*, vol. 32, database issue, pp. D575–D577, 2004.

[34] L. Shen, J. Gong, R. A. Caldo, et al., "BarleyBase—an expression profiling database for plant genomics," *Nucleic Acids Research*, vol. 33, database issue, pp. D614–D618, 2005.

[35] W. Tong, S. Harris, X. Cao, et al., "Development of public toxicogenomics software for microarray data management and analysis," *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, vol. 549, no. 1-2, pp. 241–253, 2004.

[36] W. Tong, X. Cao, S. Harris, et al., "Array track—supporting toxicogenomic research at the U.S. Food and Drug Administration National Center for Toxicological Research," *Environmental Health Perspectives*, vol. 111, no. 15, pp. 1819–1826, 2003.

[37] P. J. Killion, G. Sherlock, and V. R. Iyer, "The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD)," *BMC Bioinformatics*, vol. 4, article 32, 2003.

[38] J. Demeter, C. Beauheim, J. Gollub, et al., "The Stanford Microarray Database: implementation of new analysis tools and open source release of software," *Nucleic Acids Research*, vol. 35, database issue, pp. D766–D770, 2007.

[39] X. Xia and Z. Xie, "AMADA: analysis of microarray data," *Bioinformatics*, vol. 17, no. 6, pp. 569–570, 2001.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 893941, 13 pages
doi:10.1155/2008/893941

Review Article

Bioinformatic Tools for Inferring Functional Information from Plant Microarray Data II: Analysis Beyond Single Gene

Issa Coulibaly and Grier P. Page

Department of Biostatistics, University of Alabama at Birmingham, 1665 University Blvd Ste 327, Birmingham, AL 35294-0022, USA

Correspondence should be addressed to Grier P. Page, gpage@uab.edu

Received 2 November 2007; Accepted 5 May 2008

Recommended by Gary Skuse

While it is possible to interpret microarray experiments a single gene at a time, most studies generate long lists of differentially expressed genes whose interpretation requires the integration of prior biological knowledge. This prior knowledge is stored in various public and private databases and covers several aspects of gene function and biological information. In this review, we will describe the tools and places where to find prior accurate biological information and how to process and incorporate them to interpret microarray data analyses. Here, we highlight selected tools and resources for gene class level ontology analysis (Section 2), gene coexpression analysis (Section 3), gene network analysis (Section 4), biological pathway analysis (Section 5), analysis of transcriptional regulation (Section 6), and omics data integration (Section 7). The overall goal of this review is to provide researchers with tools and information to facilitate the interpretation of microarray data.

1. INTRODUCTION

Microarray analysis is exploratory and very high dimensional, and the primary purpose is to generate a list of differentially regulated genes that can provide insight into the biological phenomena under investigation. However, analysis should not stop with a list, it should be the starting point for secondary analyses that aim at deciphering the molecular mechanisms underlying the biological phenotypes analyzed. Combining microarray data with prior biological knowledge is a fundamental key to the interpretation of the list of genes. This prior knowledge is stored in various public and private databases and covers several aspects of genes functions and biological information such as regulatory sequence analysis, gene ontology, and pathway information. In this review, we will describe the tools and places where to find prior accurate biological information and how to incorporate them into the analysis of microarray data. The plant genome outreach portal (http://www.plantgdb.org/PGROP/pgrop.php?app=pgrop) list many of these resources and other tools and resources such as EST resources and BLAST that are not covered in this review. We also address some theoretical aspects and methodological issues of the algorithms implemented in the tools that have been recently developed for bioinformatic and what needs to be considered when selecting a tool for use.

2. CLASS LEVEL FUNCTIONAL ANNOTATION TOOLS

The goal of these class level functional annotation tools is to relate the expression data to other attributes such as cellular localization, biological process, and molecular function for groups of related genes. The most common way to functionally analyze a gene list is to gather information from the literature or from databases covering the whole genome. The recent developments in technologies and instrumentation enabled a rapid accumulation of large amount of in silico data in the area of genomics, transcriptomics, and proteomics as well. The gene ontology (GO) consortium was created to develop consistent descriptions of gene products in different databases [1]. The GO provides researchers with a powerful way to query and analyze this information in a way that is independent of species [2]. GO allows for the

annotation of genes at different levels of abstraction due to the directed acyclic graph (DAG) structure of the GO. In this particular hierarchical structure, each term can have one or more child terms as well as one or more parent terms. For instance, the same gene list is annotated with a more general GO term such as "cell communication" at a higher level of abstraction, whereas the lowest level provides a more specific ontology term such as "intracellular signaling cascade."

In recent years, various tools have been developed to assess the statistical significance of association of a list of genes with GO annotations terms, and new ones are being regularly released [3]. There has been extensive discussion of the most appropriate methods for the class level analysis of microarray data [4–6]. The methods and tools are based on different methodological assumptions. There are two key points to consider: (1) whether the method uses *gene sampling* or *subject sampling* and (2) whether the method uses *competitive* or *self-contained* procedures. The subject sampling methods are preferred and the competitive versus self-contained debate continues. Gene sampling methods base their calculation of the p-value for the geneset on a distribution in which the gene is the unit of sampling, while the subject sampling methods take the subject as the sampling unit. The latter is more valid for the unit of randomization is the subjects not the genes [7–9].

Competitive tests, which encompass most of the existing tools, test whether a gene class, defined by a specific GO term or pathway or similar, is overrepresented in the list of genes differentially expressed compared to a reference set of genes. A *self-contained* test compares the gene set toa fixed standard that does not depend on the measurements of genes outside the gene set. Goeman et al. [10, 11], Mansmann and Meister [7], and Tomfohr et al. [9] applied the self-contained methods.

Another important aspect of ontological analysis regardless of the tool or statistical method is the choice of the reference gene list against which the list of differentially regulated genes is compared. Inappropriate choice of reference genes may lead to false functional characterization of the differentiated gene list. Khatri and Drăghici [3] pointed out that only the genes represented on the array, although quite incomplete, should be used as reference list instead of the whole genome as it is a common practice. In addition correct, up to date, and complete annotation of genes with GO terms is critical. The competitive and gene sample-based procedures tend to have better and more complete databases. GO allows for the annotation of genes at different levels of abstraction due to the directed acyclic graph (DAG) structure of the GO. In this particular hierarchical structure, each term can have one or more child terms as well as one or more parent terms. For instance, the same gene list is annotated with a more general GO term such as "cell communication" at a higher level of abstraction, whereas the lowest level provides a more specific ontology term such as "intracellular signaling cascade." It is important to integrate the hierarchical structure of the GO in the analysis since various levels of abstraction usually give different p-values. The large number (hundreds or thousands) of tests performed during ontological analysis may lead to spurious associations just by chance, thus correction for multiple testing is a necessary step to take. We present here a nonexhaustive list of tools available that can be used to perform functional annotation of gene list and attempt to compare their functionalities (Table 1). All tools accept input data from *Arabidopsis thaliana*, the most used model organism in plant studies, as well as many animal organism models.

Onto-Express (OE): http://vortex.cs.wayne.edu/projects.htm#Onto-Express

Onto-Express is a software application used to translate a list of differentially regulated genes into a functional profile [12, 13]. Onto-Express constructs a profile for each of the GO categories: cellular component, biological process, molecular function, and chromosome location as well. Onto-Express implements hypergeometric, binomial, $\mathcal{X}^2$ and Fisher's exact tests. The results are displayed in a graphical form that allows the user to collapse or expand GO node and visualize the p-values associated with each level of GO abstraction. Onto-Express performs Bonferroni, Holm, Sidak, and FDR corrections to adjust for multiple testing. Users have an option of either providing their own reference gene list or selecting a microarray platform as reference gene list. An extensive list of up to date annotations is provided for many arrays.

FuncAssociate: http://llama.med.harvard.edu/cgi/func/funcassociate

FuncAssociate is a web-based tool that characterizes large sets of genes with GO terms using the Fisher's exact test [14]. Among all annotation tools FuncAssociate stands out in that it implements a Monte Carlo simulation to correct for multiple testing. In addition the tools can conduct analysis on ranked list of query genes. Although FuncAssociate supports 10 organisms, it does not provide visualization or level information for the GO annotation.

SAFE (Significance Analysis of Function and Expression)

SAFE is a Bioconductor/R algorithm that first computes gene-specific statistics in order to test for association between gene expression and the phenotype of interest [15]. Gene-specific statistics are used to estimate global statistics that detects shifts in the local statistics within a gene category. The significance of the global statistics is assessed by repeatedly permuting the response values. SAFE implements a rank-based global statistics that enables a better use of marginally significant genes than those based on a p-value cutoff.

Global test

Global test is a Bioconductor/R package that tests the association of expression pattern of a group of genes with selected phenotypes of interest using self-contained methods [10]. The method is based on a penalized regression model

Table 1: Recapitulative list of GO annotations tools.

Tool name	Statistical model	GO abstraction level	GO visualization	Multiple testing	Type of array	Other annotation	OS
Onto-Express	hypergeometric, Fisher's exact test, binomial, $\mathcal{X}^2$	Available	DAG	Bonferroni, Holm, Sidak, FDR	172 commercial arrays	Chromosomal position	Any
FatiGO+	Fisher's exact test	Available	One level at a time	FDR	User-provided	KEGG pathways, SwissPROT keywords	Any
FuncAssociate	Fisher's exact test	Not available	Not available	Monte Carlo simulation	User-provided	Not available	Web-based
GoToolBox	hypergeometric test, Fisher's exact test or binomial	Available	One level at a time	Bonferroni	User-provided	Not available	Any
CLENCH2	Hypergeometric, binomial, $\mathcal{X}^2$	Static global	DAG	None	User-provided	Not available	Windows
BiNGO	Hypergeometric, binomial	Available, GOSlim	DAG	FDR, Bonferroni	commercial arrays	Not available	
GoSurfer	$\mathcal{X}^2$	Lowest level	DAG	FDR	Affymetrix only	Not available	Windows

that shrinks regression coefficient between gene expression and phenotype toward a common mean. The algorithm allows the users to testbiological hypothesis or to search GO databases for potential pathways. The results of gene lists of various sizes can be compared.

FatiGO+ (Fast Assignment and Transference of Information): http://babelomics2.bioinfo.cipf.es/fatigoplus/cgi-bin/fatigoplus.cgi

FatiGO+ tests for significant difference in distribution of GO terms between any two groups of genes (ideally a group of interest and a reference set of genes) using a Fisher's exact test for 2 by 2 contingency table [16]. FatiGO+ implements an inclusive analysis in which at a given level in the GO DAG hierarchy, genes annotated with child GO terms take the annotation from the parent. This increases the power of the test. The software returns adjusted p-values using the FDR method [17].

GOToolBox: http://burgundy.cmmt.ubc.ca/GOToolBox/

GOToolBox identifies over-or under-represented GO terms in a gene set using either hypergeometric distribution-based tests or binomial test [18]. The user has the option of choosing between the total set of genes in the genome as reference or provides his own list of reference genes. The software implements Bonferroni correction to adjust for multiple testing. Its also allows the user to select a specific level of GO abstraction prior to the analysis.

CLENCH2 (CLuster ENriCHment): http://www.stanford.edu/~nigam/cgi-bin/dokuwiki/doku.php?id=clench

Clench is used to calculate cluster enrichment for GO terms [19]. The program accepts two lists of genes: a reference set of genes and the list of changed genes. CLENCH performs hypergeometric, binomial and $\mathcal{X}^2$ tests to estimate GO terms enrichment. The program allows the user to choose an FDR cutoff in order to account for multiple testing.

BiNGO (Biological Network Gene Ontology tool): http://www.psb.ugent.be/cbd/papers/BiNGO/

BiNGO is a Java-based tool to determine which gene ontology (GO) categories are statistically overrepresented in a set of genes or a subgraph of a biological network [20]. BiNGO is implemented as a plugin for Cytoscape, which is an open source bioinformatics software platform for visualizing and integrating molecular interaction networks. The program implements hypergeometric test and binomial test and performs FDR to control multiple testing. BiNGO maps predominant functional themes of the tested genes on the GO hierarchy. It allows a customizable visual representation of the results. One limitation is that the user can only choose between the whole genome or the network under study as reference set of gene for the enrichment test.

GoSurfer: http://bioinformatics.bioen.uiuc.edu/gosurfer/

GoSurfer is used to visualize and compare gene sets by mapping them onto gene ontology (GO) information in the form of a hierarchical tree [21]. Users can manipulate the tree output by various means, like setting heuristic thresholds or using statistical tests. Significantly important GO terms resulting from a $\mathcal{X}^2$ test can be highlighted. The software controls for false discovery rate.

3. GENE COEXPRESSION ANALYSIS TOOLS

In most microarray studies, gene expressions are measured on a small number of arrays or samples; however, large collections of arrays are available in microarray database

that contain transcript levels data from thousands of genes across a wide variety of experiments and samples. These tools provide scientists with the opportunity to analyze the transcriptome by pooling gene expression information from multiple data sets. This meta-analytic approach allows biologists to test the consistency of gene expression patterns across different studies. Most importantly, the analysis of concerted changes in transcript levels between genes can lead to biological function discovery. It has been demonstrated that genes which protein products cooperate in the same pathway or are in a multimeric protein complex display similar expression patterns across a variety of experimental conditions [22, 23]. Using the guilt-by-association principle, investigators can functionally characterize a previously uncharacterized gene when it displays expression pattern similar to that of known genes. The coexpression relationship between two genes is usually assessed by computing the Pearson's correlation coefficient or other distance measures. Prior to the coexpression analysis, a set of "bait-genes" is selected based on previous biological or literature information. Then the genes which expression is significantly correlated with bait-genes expression are analyzed to identify new potential actors in a given pathway or biological process. However, coexpression between two genes does not necessarily translate into similar function between both genes. Some statistically significant correlations may occur by chance. Some authors suggest that to be sustainable the gene coexpressions observed in one species should be confirmed in other evolutionary close species [24]. Tools have been developed that make use of the large sample size available in these databases to identify more reliable concerted changes in transcripts levels as well as to examine the coordinated change of gene expression levels.

Cress-express:http://www.cressexpress.org/

Cress-express estimates the coexpression between a user-provided list of genes and all genes from Affymetrix Ath1 platform using up to 1779 arrays. Cress-express also performs pathway-level coexpression (PLC) [25]. PLC identifies and ranks genes based on their coexpression with a group of genes. Cress-express also delivers results in "bulk" formats suitable for downstream data mining via web services. The tool generates files for easy import into Cytoscape for visualization. The tool has the data processed with a variety of image processing methods: RMA, MAS5, and GCRMA. Investigators can select which of over 100 experiments to include in coexpression analysis.

ATTED-II (Arabidopsis thaliana transfactor and cis-element prediction database): http://www.atted.bio.titech.ac.jp/

ATTED-II provides coregulated gene relationships in *Arabidopsis thaliana* to estimate gene functions. In addition, it can predict overrepresented cis-elements based upon all possible heptamers. There is also several visualization tools and databases of annotations attached to the coexpression.

Genevestigator: http://www.genevestigator.ethz.ch/

Genevestigator is a web-based discovery tool to study the expression and regulation of genes, pathways, and networks [26, 27]. Among other applications, the software allows the user to look at individual gene expression or group of genes coexpression in many different tissues, at multiple developmental stages, or in response to large sets of stimuli, diseases, drug treatments, or mutations. In addition, electronic northern blots and other analyses may be conducted.

BAR (the botany array resource) expression ANGLER: http://www.bar.utoronto.ca/

The expression anger allows the user to identify genes with similar expression profile with the user provided gene across multiple samples [28]. The user can specify the Pearson correlation coefficient threshold and the array database to use for the coexpression analysis.

AthCor@CSB.DB (A. thaliana coresponse database): http://csbdb.mpimp-golm.mpg.de/csbdb/dbcor/ath.html

AthCor is a coexpression tool that allows the use of functional ontology filter to identify genes coexpressed with a gene of interest filtering the search by functional ontologies [29]. The user can select between parametric and nonparametric correlation tests.

PLEXdb (Plant Expression Database): http://www.plexdb.org/

PLEXdb serves as a comprehensive public repository for gene expression for plants and plant pathogens [30]. PLEXdb integrates new gene expression datasets with traditional genomics and phenotypic data. The integrated tools of PLEXdb allow plant investigators to perform comparative and functional genomics analyses using large-scale expression data sets.

ACT (Arabidopsis Coexpression Data mining Tool): http://www.arabidopsis.leeds.ac.uk/act/index.php

ACT estimates the coexpression of 21 891 *Arabidopsis* genes based on Affymetrix ATH1 platform using a simple correlation test [31]. The web server includes a database that stores precalculated correlation results from over 300 arrays of the NASC/GARNet dataset. A "clique finder" tool allows the user to identify groups of consistently coexpressed genes within a user-defined list of genes. The identification of genes with a known function within a cluster allows inference to be made about the other genes. Users can also visualize the coexpression scatter plots of all genes against a group of genes.

4. GENE NETWORK ANALYSIS

Genes and their protein products are related to each other through a complex network of interactions. In higher metazoa, on average each gene is estimated to interact with five

other genes [32], and to be involved in ten different biological functions during development [33]. On a molecular level, the function of a gene depends on its cellular context, and the activity of a cell is determined by which genes are being expressed and which are not and how they interact with each other. In such high interconnectedness, analyzing a network as a whole is essential to understanding the complex molecular processes underlying biological systems. The traditional reductionist approach that investigates biological phenomena by analyzing one gene at a time cannot address this complexity. By using systems biology approach and network theories, investigators can analyze the behavior and relationships of all of the elements in a particular biological system to arrive at a more complete description of how the system functions [34]. High-throughput gene expression profiling offers the opportunity to analyze gene interrelationships at the genome scale. Clustering analysis on microarray expression data only extracts lists of coregulated genes out of a large-scale expression data. It does not tell us who is regulating whom and how. However, the task of modeling dynamic systems with large number of variables can be computationally challenging. In gene regulatory networks, genes, mRNA, or proteins correspond to the network nodes and the links among the nodes stand for the regulatory interactions (activations or inhibitions). In this section, we will describe some of the methods and tools used to reconstruct, visualize, and explore gene networks.

4.1. Gene network reconstruction algorithms

Two main approaches have been used to develop models for gene regulatory networks [35]. One method is based on Bayesian inference theory which seeks to find the most probable network given the observed expression patterns of the genes to be included in the network. The regulatory interactions among genes and their directions are derived from expression data. Several network structures are proposed and scored on the basis of how well they explain the data as it has been successfully implemented in yeast [36]. The second approach is based on "mutual information" as a measure of correlation between gene expression patterns [37]. A regulatory interaction between two genes is established if the mutual information on their expression patterns is significantly larger than a p-threshold value calculated from the mutual information between random permutations of the same patterns. Unlike the Bayesian theory, which tries out whole networks and selects the one that best explains the observed data, the mutual information method constructs a network by selecting or rejecting regulatory interactions between pairs of genes. This method does not provide the direction of regulatory interactions. We present below selected tools that implement either of the aforementioned approaches to reverse-engineer gene regulatory networks.

BNArray (Bayesian Network Array): http://www.cls.zju.edu.cn/binfo/BNArray/

BNArray is a tool developed in R for inferring gene regulatory networks from DNA microarray data by using a Bayesian network [38]. It allows the reconstruction of significant submodules within regulatory networks using an extended subnetwork mining algorithm. BNArray can handle microarray data with missing values.

BANJO (Bayesian Network Inference with Java Objects): http://www.cs.duke.edu/~amink/software/banjo/

Banjo is a tool developed in Java for inferring gene networks [39]. Banjo implements Bayesian and dynamic Bayesian networks to infer networks from both steady-state and time-series expression data. A "proposer" component of Banjo uses heuristic approaches to search the network space for potential network structures. Each network structure is explored and an overall network's score is computed based on the parameters of the conditional probability density distribution. The network with the best overall score is accepted by a "decider" component of the software. The network retained is processed by Banjo to compute influence scores on the edges indicating the direction of the regulation between genes. The software displays the output network.

GNA (Genetic Network Analyzer): http://www-helix.inrialpes.fr/article122.html

GNA is a freely available software used for modeling and simulating genetic regulatory networks from gene expression data and regulatory interaction information [40]. In GNA, the dynamics of a regulatory network is modeled by a class of piecewise-linear differential equations. The biological data are transformed into mathematical formalism. Thus the software uses qualitative constraints in the form of algebraic inequalities instead of numerical values.

PathwayAssist http://www.ariadnegenomics.com/products/pathway-studio

PathwayAssist allows the users to create their own pathways by combining the user-submitted microarray expression data with knowledge from biological databases such as BIND, KEGG, DIP [41]. The software provides a graphical user interface and publication quality figures.

4.2. Network visualization tools

As a result of the explosion and advances in experimental technologies that allow genome-wide characterization of molecular states and interactions among thousands of genes, researchers are often faced with the need for tools for the visualization, display, and evaluation of large structure data. The main aim of these tools is to provide a summarized yet understandable view of large amount of data while integrating additional information regarding the biological processes and functions. Several network visualization tools have been developed of which we will describe some of the most popular.

Cytoscape—http://www.cytoscape.org/

Cytoscape is a general-purpose, open-source software environment for the large scale integration of molecular interaction network data [42]. Dynamic states on molecules and molecular interactions are handled as attributes on nodes and edges, whereas static hierarchical data, such as protein-functional ontologies, are supported by use of annotations. The Cytoscape core handles basic features such as network layout and mapping of data attributes to visual display properties. Many Cytoscape plug-ins extend this core functionality.

CellDesigner http://www.celldesigner.org/

CellDesigner is a structured diagram editor for drawing gene-regulatory and biochemical networks based on standardized technologies and with wide transportability to other systems biology markup language (SBML) compliant applications and systems biology workbench (SBW) [43]. Networks are drawn based on the process diagram, with graphical notation system. The user can browse and modify existing SBML models with references to existing databases, simulate and view the dynamics through an intuitive graphical interface. CellDesigner runs on Windows, MacOS X, and Linux.

VANTED (Visualization and Analysis of Networks with related Experimental Data): http://vanted.ipk-gatersleben.de/

Vanted is a freely available tool for network visualization that allows users to map their own experimental data on networks drawn in the tool, downloaded from KEGG pathway database, or imported using standard imported formats [44]. The software graphically represents the genes in their underlying metabolic context. Statistical methods implemented in VANTED allow the comparison between treatments or groups of genes, the generation of correlation matrix, or the clustering of genes based on expression pattern.

Osprey http://biodata.mshri.on.ca/osprey/servlet/Index

Osprey is a software for visualization and manipulation of complex interaction networks [45]. Osprey allows user defined colors to indicate gene function, experimental systems, and data sources. Genes are colored by their biological process as defined by standardized gene ontology (GO) annotations. As a network complexity increases, Osprey simplifies network layouts through user-implemented node relaxation, which disperses nodes and edges according to anyone of a number of layout options.

VisANT (Integrative Visual Analysis Tool for Biological Networks and Pathways): http://visant.bu.edu/

VisANT is a freely available open-source tool for integrating biomolecular interaction data into a cohesive, graphical interface [45–47]. VisANT offers an online interface for a large range of published datasets on biomolecular interactions, as well as databases for organized annotation, including GenBank, KEGG, and SwissProt.

4.3. Network exploration tools

One of the main focuses in the postgenomic era is to study the network of molecular interactions in order to reveal the complex roles played by genes, gene products, and the cellular environments in different biological processes. The nodes (genes) of a network can be associated with additional information regarding the gene products, gene positions in the chromosome, or the gene functional annotation. The edges in the network symbolize specific interaction that can be associated with a transcription factor-promoter bond for instance. This information can be automatically retrieved in a number of specialized and publicly accessible databases containing data about the nodes and the interactions. Network exploration tools enable the user to perform analysis on single genes, gene families, patterns of molecular interactions, as well as on the global structure of the network. These tools are able to incorporate both microscale and macroscale analysis using heterogenous data. They can connect to a large number of disparate databases. The user usually has an option to construct interaction networks either by curation or by computation and to associate microarray expression data with known metabolic pathways. Here, we describe some of the most popular network exploration tools.

MetNet (Metabolic Networking Database): http://www.metnetdb.org/

MetNet is a publicly available software for analysis of genome-wide mRNA, protein, and metabolite profiling data [48]. The software is designed to enable the biologist to visualize, statistically analyze, and model a metabolic and regulatory network map of Arabidopsis, combined with gene expression profiling data. MetNet provides a framework for the formulation of testable hypotheses regarding the function of specific genes. The tools within MetNet allow the user to map metabolic and regulatory networks; to integrate and visualize data together; to explore and model the metabolic and regulatory flow in the network.

BiologicalNetworks: http://biologicalnetworks.net/

BiologicalNetworks is a bioinformatics and systems biology software platform for visualizing molecular interaction networks, sequence and 3D structure information [49]. The tool performs easy retrieval, construction, and visualization of complex biological networks, including genome-scale integrated networks of protein-protein, protein-DNA, and genetic interactions. BiologicalNetworks also allow the analysis and the mapping of expression profiles of genes or proteins onto regulatory, metabolic, and cellular networks.

PaVESy (Pathway Visualization Editing System): http://pavesy.mpimp-golm.mpg.de/PaVESy.htm

PaVESy is a data managing system for editing and visualization of biological pathways [50]. The main component of PaVESy is a relational SQL database system that stores biological objects, such as metabolites, proteins, genes, and their interrelationships. The user can annotate the biological objects with specific attributes that are integrated in the database. The specific roles of the objects are derived from these attributes in the context of user-defined interactions. PaVESy can display an individualized view on the database content that facilitates user customization.

Genevestigator: https://www.genevestigator.ethz.ch

Genevestigator provides a detailed analysis and navigation through biochemical and/or regulatory pathways. It combines automatically produced or user-created graphical representations of networks (e.g., gene modules or pathways) for the exploratory analysis of a large compendium of gene expression profiles. Effects on gene expression can be projected onto these networks for the following ontologies: anatomy, development, stimulus, and mutation, in form of comparison sets.

5. BIOLOGICAL PATHWAY RESOURCES

One of the downstream applications of the reconstruction of a gene regulatory networks or the identification of clusters of functionally related genes is to associate the genes and their interconnections with known metabolic pathways. Biochemists summarized the sequence of enzyme-catalyzed metabolic reactions between biomolecules as a network of interactions that results from the conversion of one organic substance (substrate) to another (product). Depending on the type of interactions analyzed, several types of biochemical networks are identified. These biochemical networks represent the potential mechanistic associations between genes and gene products that are involved in specific biological processes [52]. Because of the curse of dimensionality that sometimes hampers the whole network analysis, investigators often focus on "pathway" rather than "network" when they are investigated a small number of gene interactions. Many specialized databases are available that store and summarize large amount of information on metabolic reactions. Increasingly, identifying and searching the right database is a critical and necessary step in most biological researches. This task can be tedious due to the large number of databases available. For a more comprehensive list of biological pathways resources on the web, the reader is referred to pathguide (http://www.pathguide.org). Following is the list of the most popular pathways resources on the web.

KEGG (Kyoto Encyclopedia of Genes and Genomes): http://www.genome.jp/kegg

KEGG aims to link lower-level information (genes, proteins, enzymes, reaction molecules, etc.) with higher-level information (interactions, enzymatic reactions, pathways, etc.). Pathways are included for over 100 species.

MetaCyc: http://MetaCyc.org/

MetaCyc is a database of metabolic pathways and enzymes [53]. Its goal is to serve as a metabolic encyclopedia, containing a collection of nonredundant pathways, enzymatic reactions, enzymes, chemical compounds, genes and review-level comments. Enzyme information includes substrate specificity, kinetic properties, activators, inhibitors, cofactor requirements and links to sequence and structure databases. AracCyc (http://www.arabidopsis.org/biocyc/index.jsp) uses MetaCyc as reference database for visualization of *Arabidopsis thaliana* biochemical pathways. Table 2 indicates web links to more online pathways databases.

BioCarta: http://www.biocarta.com/genes/index.asp

BioCarta is a web-based resource for exploring biological pathways. BioCarta catalogs pathways, regulation and interaction information for over 120,000 genes covering most model organisms. Data in BioCarta are constantly updated, and new pathways are suggested by the life science research community.

GeneNeT: http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/

The GeneNet system is designed for formalized description and automated visualization of gene networks [54]. The GeneNet system includes database on gene network components, Java program for the data visualization. GeneNet allows the users to select entities that are involved in the functioning of a particular gene network, to describe the regulatory relations for a particular gene network, and to search for potential transcription factors.

6. TRANSCRIPTION REGULATION ANALYSIS TOOLS

Most organisms encode a large number of DNA-binding proteins that act as transcription factors. In *Arabidopsis*, more than 5% of the genes have been estimated to encode transcription factors [55]. Transcription factors bind to short conserved DNA motifs (cis-acting regulatory elements CARE) located at the 5'end of the gene (in a region called promoter) to initiate mRNA transcription. Thus DNA-binding proteins play a key role in all aspects of genetic activity within an organism. They participate in promoting or repressing the transcription of specific genes. Elucidating the mechanisms that underlie the expression of genomes is one of the major challenges in bioinformatics. An interesting hypothesis one might formulate after a successful microarray study is that the genes that are coexpressed may also be coregulated at the transcriptional level. One way to test this hypothesis is to identify overrepresented oligonucleotides sequences as potential binding sites for transcriptions factors in promoter regions of genes clustered in the same group. The statistical test for overrepresentation of regulatory motifs in intergenic regions is the general principle implemented in

Table 2: Additional links for pathways databases on the internet.

Database name	Description	URL
PathDB	Biochemical pathways, compounds and metabolism	http://www.ncgr.org/pathdb
UM-BBD	University of Minnesota biocatalysis and biodegradation database	http://umbbd.ahc.umn.edu/
BIND	Biomolecular interaction network database	http://www.bind.ca/
BRITE	Biomolecular relations in information transmission and expression, part of KEGG	http://www.genome.ad.jp/brite/
PAJEK	Program for large network analysis	http://vlado.fmf.uni-lj.si/pub/networks/pajek/
DDIB	Database of domain interactions and binding	http://www.ddib.org/
DIP	Database of interacting proteins: experimentally determined protein-protein interactions	http://dip.doe-mbi.ucla.edu/
IntAct project	Protein-protein interaction data	http://www.ebi.ac.uk/intact/
InterDom	Putative protein domain interactions	http://interdom.i2r.a-star.edu.sg/
PSIbase	Interaction of proteins with known 3D structures	
Reactome	A knowledgebase of biological pathways	http://www.reactome.org/
STRING	Predicted functional associations between proteins	http://string.embl.de/
TRANSPATH	Gene regulatory networks and microarray analysis	http://www.biobase-international.com/pages/index.php?id=transpathdatabases

most algorithms for regulatory motif detection [55]. CAREs can also be predicted through phylogenetic footprinting that is based on sequence similarity between orthologous promoters [56]. Some other approaches have been proposed that integrates comparative, structural, and functional genomics to identify conserved motifs in coregulated genes. The detailed description of these approaches is beyond the scope of this chapter. Following is a list of transcription factors database and tools (Table 3).

Plant Promoter Database (PlantProm DB): http://mendel.cs.rhul.ac.uk or http://www.softberry.com/

PlantProm is a plant promoter database. The database represents a collection of annotated, nonredundant proximal promoter sequences for RNA polymerase II with experimentally determined transcription start site from various plant species [57].

The Arabidopsis information resource (TAIR) motif analysis software: http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp

The motif analysis tool of the TAIR compares the frequency of 6-mer motif in promoter regions of query set of genes with the frequency of the 6-mer motif in the whole *A. thaliana* genome. A binomial distribution p-value is computed for each motif identified. The user can specify the size of the genes 5'upstream region to 500 bp or 1 kb. The tool does not account for multiple testing.

TRANSFAC: http://www.biobase-international.com/pages/index.php?id=transfacdatabases

TRANSFAC is an international unique database on eukaryotic transcriptional regulation [58]. The database contains data on transcription factors, their target genes and their experimental-proven binding sites in genes. Tools within TRANSFAC allow the users to automatically visualize gene-regulatory networks based on interlinked factor and gene entries in the database.

AthaMap: http://www.athamap.de/index.php

AthaMap is a database that organizes a genome-wide map of potential transcription factor binding sites in *Arabidopsis thaliana* [59]. AthaMap allows the user to test for the overrepresentation of transcription factors in a set of query genes. A colocalization tool performs combinatorial analysis to identify synchronized binding of pairs of transcription factors.

Table 3: Databases for transcription factors available on the internet.

Databse name	Description	URL
ACTIVITY	Functional DNA/RNA site activity	http://wwwmgs.bionet.nsc.ru/mgs/systems/activity/
DoOP	Database of orthologous promoters: chordates and plants	http://doop.abc.hu/
EPD	Eukaryotic promoter database	http://www.epd.isb-sib.ch/
JASPAR	PSSMs for transcription factor DNA-binding sites	http://jaspar.cgb.ki.se/
MAPPER	Putative transcription factor binding sites in various genomes	http://bio.chip.org/mapper
TESS	Transcription element search system	http://www.cbil.upenn.edu/tess/
TRANSCompel	Composite regulatory elements affecting gene transcription in eukaryotes	http://www.gene-regulation.com/pub/databases.html#transcompel
TRED	Transcriptional regulatory element database	http://rulai.cshl.edu/tred/
TRRD	Transcription regulatory regions of eukaryotic genes	http://www.bionet.nsc.ru/trrd/
AthaMap	Genome-wide map of putative transcription factor binding sites in *Arabidopsis thaliana*	http://www.athamap.de/
DATF	Database of *Arabidopsis* transcription factors	http://datf.cbi.pku.edu.cn/

PlantCARE (Plant Cis-Acting Regulatory Elements): http://bioinformatics.psb.ugent.be/webtools/plantcare/html/

PlantCARE is a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences [60]. The database can be queried on names of TF binding sites, function, species, cell type, genes, and reference literatures. The program returns a list of entries with links to other information within the database or beyond through accession to TRANSFAC, EMBL, GenBANK, or MEDLINE.

PLACE (Plant Cis-acting regulatory DNA Elements): http://www.dna.affrc.go.jp/PLACE/

PLACE is a database of motifs found in plant cis-acting regulatory DNA elements, all from previously published reports [61]. In addition to the motifs originally reported their variations in other genes or in other plant species reported later are also compiled. The PLACE database also contains a brief description of each motif and relevant literature with PubMed ID numbers.

Athena: http://www.bioinformatics2.wsu.edu/cgi-bin/Athena/cgi/home.pl

Athena is a database which contains over 30 000 predicted Arabidopsis promoters sequences and consensus sequences for 105 previously characterized TF binding sites [62]. Athena enables the user to visualize and rapidly inspect key regulatory elements in multiple promoters. The software includes tools for testing the overrepresentation of TF sites among subset of promoters. A data-mining tool allows the selection of promoter sequences containing specific combination of TF binding sites. Athena does not adjust for multiple testing.

AGRIS (Arabidopsis Gene Regulatory Information Server): http://arabidopsis.med.ohio-state.edu/

AGRIS is an information resource for retrieving Arabidopsis promoter sequences, transcription factors and their target genes [63]. AGRIS integrates transcriptional regulatory information from multiple sources. Users can query the database with a gene name, gene symbol to retrieve its promoter along with other genes regulated by the same transcription factor.

7. 'OMICS DATA INTEGRATION TOOLS

Various innovative and advanced technologies have allowed scientists to rapidly generate genome-scale or "omics" datasets at virtually every cellular level. These individual omics provide a wealth of information about living cells and organisms. However, it is only by integrating genomics, transcriptomics proteomics, metabolomics, and other recent omics types of data such as "interactomics," "localizomics," "lipidomics," and "phenomics" that biologists can gain access to a more complete picture of living organisms and unexplored areas of biology. This challenging task requires a systems level approach to perform systematic data mining, cross-knowledge validation, and cross-species interpolation. Some investigators attempted the integration of genomic data and transcriptomic data [64], and the integration of

Table 4: Proteomics databases available on the internet.

Databse name	Description	URL
RPD	Rice proteome database	http://gene64.dna.affrc.go.jp/RPD/
ANPD	Arabidopsis nucleolar protein database	http://bioinf.scri.sari.ac.uk/cgi-bin/atnopdb/home/
AMPD	Arabidopsis mitochondrial protein database	http://www.plantenergy.uwa.edu.au/applications/ampdb/index.html/
PA-GOSUB	Protein sequences from model organisma, GO assignement and subcellular localization	http://www.cs.ualberta.ca/~bioinfo/PA/GOSUB/
Swiss-Prot	A curated protein sequence database which strives to provide a high level of annotation	http://expasy.org/sprot/
AAindex	Database of various physicochemical and biochemical properties of amino acids and pairs of amino acids	http://www.genome.ad.jp/aaindex/
Prosite	Database of protein domains, families and functional sites, as well as associated patterns and profiles	http://www.expasy.ch/prosite/
PLANT-PIs	Database of information on the distribution and functional properties of protease inhibitors in higher plants	http://www.ba.itb.cnr.it/PLANT-PIs/
GeneFarm	Annotation of Arabidopsis genes and proteins	http://urgi.versailles.inra.fr/Genefarm/

protein-protein interation data and transcriptomic data [65] to analyze the dynamics of biological networks in yeast. The approach commonly used comprises three steps: (1) identification of the network that describes all interactions between cellular components from integrating various genome scale data; (2) decomposing the network into its constituent parts or network modules; (3) building a mathematical model that simulates biological systems for the purpose of simulation or prediction [66]. We describe below proteomics and metabolomics, and the potential of their integration with transcritomic data.

7.1. Proteomics

Gene mRNA expression profiling on a global scale in response to specific conditions is not sufficient to render the complexities and dynamics of systems biology. The ultimate products of genes are proteins. Furthermore, mRNA levels are not always well correlated with the levels of the corresponding protein [67] and one gene can produce several protein species. Indeed, proteins undergo a series of post-translational molecular modifications such as glycosylation, phosphorylation, cleavage or complex formation may also occur that overall influence their function. Proteomics is the systematic large-scale study of proteins of an organism or a specific type of tissue, particularly their structure, function, and spatiotemporal distribution. Thus proteomics is an essential component of any functional genomics study aiming at understanding biological processes. The integration of transcriptome and proteome data has not always resulted in consistent results [68]. The methods and techniques used to measure the transcript level and the protein level may affect the results concordance. Nonetheless, the interpretation of the data in terms of biological pathways or functional groups gives better correlation of transcriptome with proteome in yeast [69].

Many plant proteomics databases have been constructed in recent years. As the plant model organism of choice, Arabidopsis proteome database contains more data compared to other species. Protein amino acid sequence databases and repositories for two-dimensional polyacrylamide gel electrophoresis as reference maps of proteomes are becoming popular as tools for analyzing and comparing the plant proteome. SWISS-2DPAGE is a two-dimensional polyacrylamide gel electrophoresis database (http://expasy.org/ch2d). PhytoProt (http://urgi.versailles.inra.fr/phytoprot) is a database of clusters of all the plants full-length protein sequences retrieved from SwissProt/TrEMBL. Proteins are grouped into clusters based on their peptide sequence similarity in order to track erroneous annotations made at the genome level. The database can be searched for any protein or group of proteins using protein ID or words appearing in protein description. Additional plant proteomics databases are provided in Table 4.

7.2. Metabolomics

Metabolomics is the study of all low molecular weight chemicals in a plant as the end products of the cellular processes. The metabolome represents the collection of all metabolites in an organism. Metabolic profiling provides an instantaneous snapshot of the chemistry of a sample

Table 5: Main features of the types of bioinformatics tools used for the analysis of DNA microarray data.

Tools and resources	Goal	Methods
Class level functional Annotation	Determine a biological meaning to groups of related genes identified by microarray analysis	Overrepresentation test of gene ontology (GO) terms
Gene coexpression	Identify common expression patterns between genes in order to infer biological function	Correlation tests of gene expression
Gene network Analysis	Capture the interconnectedness of cellular components in order to explain biological phenomena	Systems biology approach
Gene network reconstruction	Develop models for gene regulatory networks	Bayesian inference theory Mutual information theory
Network visualization	Display a simplified view of large amount biological components and their interactions	Graph theory
Network exploration	Associate network nodes and edges with biological information	Incorporate heterogeneous data from various databases
Biological pathway resources	Map biological pathways information into inferred network	Collect and process information from pathway databases
Transcriptional regulation analysis	Identify transcription factors that regulate gene expression	Overrepresentation test of regulatory motifs in promoter regions of related genes

and defines the biochemical phenotype of a cell or a tissue [70]. Similar to transcript level and protein level, the level of metabolites in an organism or a tissue is influenced by the biological context [71]. Thus measure of mRNA gene expression and protein content of a sample do not tell the whole story of biological phenomena unfolding in that sample. Although plant metabolomics is still in its infancy, recent advances in mass spectrometry have enabled the accumulation of metabolites data on a large scale for some species. Applications of metabolomics data to functional genomics are numerous. Metabolomics provide scientist with the ability (1) to characterize genotypes, ecotypes, or phenotypes with metabolites levels; (2) to identify sites within a genetic network where metabolites levels are regulated; (3) to analyze genes functions at the light of metabolites levels [70]. Currently, one of the most pressing needs in the fields of metabolomics for bioinformatics application is the creation of specific databases and biochemical ontologies. Such tools would help clearly describe the function, localization, and interaction of metabolites. However, databases imbedded in KEGG and AraCyc can be useful at least in part for the purpose of metabolites referencing.

8. CONCLUSION

The deluge of large-scale biological data in the recent years has made the development of computational tools critical to biological investigation. Microarray studies enables scientist to simultaneously interrogate thousands of genes throughout the genome. A great variety of tools have been developed for the specific task of drawing biological meaning from microarray data. Most of the tools available exploit prior biological knowledge accumulated in numerous publicly available databases in an attempt to provide a comprehensive view of biological phenomena. Table 5 summarizes the main features of each class of bioinformatics tool described. These tools differ in many respects and the guidance provided in this review will help biologists with little knowledge in statistics understand some of the key concepts. The integration of transcriptomics data with all other omics data is a challenging task that can be addressed by a systems-levelapproach.

REFERENCES

[1] M. A. Harris, J. Clark, A. Ireland, et al., "The Gene Ontology (GO) database and informatics resource," *Nucleic Acids Research*, vol. 32, database issue, pp. D258–D261, 2004.

[2] J. I. Clark, C. Brooksbank, and J. Lomax, "It's all GO for plant scientists," *Plant Physiology*, vol. 138, no. 3, pp. 1268–1279, 2005.

[3] P. Khatri and S. Drăghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, no. 18, pp. 3587–3595, 2005.

[4] J. J. Goeman and P. Bühlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, 2007.

[5] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, "Enrichment or depletion of a GO category within a class of genes: which test?" *Bioinformatics*, vol. 23, no. 4, pp. 401–407, 2007.

[6] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.

[7] U. Mansmann and R. Meister, "Testing differential gene expression in functional groups: Goeman's global test versus an ANCOVA approach," *Methods of Information in Medicine*, vol. 44, no. 3, pp. 449–453, 2005.

[8] V. K. Mootha, C. M. Lindgren, K.-F. Eriksson, et al., "PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genetics*, vol. 34, no. 3, pp. 267–273, 2003.

[9] J. Tomfohr, J. Lu, and T. B. Kepler, "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, vol. 6, article 225, pp. 1–11, 2005.

[10] J. J. Goeman, S. A. van de Geer, F. de Kort, and H. C. van Houwellingen, "A global test for groups fo genes: testing association with a clinical outcome," *Bioinformatics*, vol. 20, no. 1, pp. 93–99, 2004.

[11] J. J. Goeman, J. Oosting, A.-M. Cleton-Jansen, J. K. Anninga, and H. C. van Houwelingen, "Testing association of a pathway with survival using gene expression data," *Bioinformatics*, vol. 21, no. 9, pp. 1950–1957, 2005.

[12] S. Draghici, P. Khatri, P. Bhavsar, A. Shah, S. A. Krawetz, and M. A. Tainsky, "Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3775–3781, 2003.

[13] P. Khatri, P. Bhavsar, G. Bawa, and S. Draghici, "Onto-Tools: an ensemble of web-accessible ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments," *Nucleic Acids Research*, vol. 32, pp. W449–W456, 2004.

[14] G. F. Berriz, O. D. King, B. Bryant, C. Sander, and F. P. Roth, "Characterizing gene sets with FuncAssociate," *Bioinformatics*, vol. 19, no. 18, pp. 2502–2504, 2003.

[15] W. T. Barry, A. B. Nobel, and F. A. Wright, "Significance analysis of functional categories in gene expression studies: a structured permutation approach," *Bioinformatics*, vol. 21, no. 9, pp. 1943–1949, 2005.

[16] F. Al-Shahrour, P. Minguez, J. Tárraga, et al., "BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments," *Nucleic Acids Research*, vol. 34, pp. W472–W476, 2006.

[17] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society*, vol. 57, no. 1, pp. 289–300, 1995.

[18] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq, "GOToolBox: functional analysis of gene datasets based on Gene Ontology," *Genome Biology*, vol. 5, no. 12, article R101, pp. 1–8, 2004.

[19] N. H. Shah and N. V. Fedoroff, "CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology," *Bioinformatics*, vol. 20, no. 7, pp. 1196–1197, 2004.

[20] S. Maere, K. Heymans, and M. Kuiper, "*BiNGO*: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks," *Bioinformatics*, vol. 21, no. 16, pp. 3448–3449, 2005.

[21] S. Zhong, K.-F. Storch, O. Lipan, M.-C. J. Kao, C. J. Weitz, and W. H. Wong, "GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology™ space," *Applied Bioinformatics*, vol. 3, no. 4, pp. 261–264, 2004.

[22] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[23] T. R. Hughes, M. J. Marton, A. R. Jones, et al., "Functional discovery via a compendium of expression profiles," *Cell*, vol. 102, no. 1, pp. 109–126, 2000.

[24] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.

[25] H. Wei, S. Persson, T. Mehta, et al., "Transcriptional coordination of the metabolic network in Arabidopsis," *Plant Physiology*, vol. 142, no. 2, pp. 762–774, 2006.

[26] P. Zimmermann, M. Hirsch-Hoffmann, L. Hennig, and W. Gruissem, "GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox," *Plant Physiology*, vol. 136, no. 1, pp. 2621–2632, 2004.

[27] P. Zimmermann, L. Hennig, and W. Gruissem, "Gene-expression analysis and network discovery using Genevestigator," *Trends in Plant Science*, vol. 10, no. 9, pp. 407–409, 2005.

[28] K. Toufighi, S. M. Brady, R. Austin, E. Ly, and N. J. Provart, "The botany array resource: e-Northerns, expression angling, and promoter analyses," *The Plant Journal*, vol. 43, no. 1, pp. 153–163, 2005.

[29] D. Steinhauser, B. Usadel, A. Luedemann, O. Thimm, and J. Kopka, "CSB.DB: a comprehensive systems-biology database," *Bioinformatics*, vol. 20, no. 18, pp. 3647–3651, 2004.

[30] L. Shen, J. Gong, R. A. Caldo, et al., "BarleyBase—an expression profiling database for plant genomics," *Nucleic Acids Research*, vol. 33, database issue, pp. D614–D618, 2005.

[31] I. W. Manfield, C.-H. Jen, J. W. Pinney, et al., "*Arabidopsis* Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis," *Nucleic Acids Research*, vol. 34, pp. W504–W509, 2006.

[32] M. I. Arnone and E. H. Davidson, "The hardwiring of development: organization and function of genomic regulatory systems," *Development*, vol. 124, no. 10, pp. 1851–1864, 1997.

[33] G. L. G. Miklos and G. M. Rubin, "The role of the genome project in determining gene function: insights from model organisms," *Cell*, vol. 86, no. 4, pp. 521–529, 1996.

[34] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, no. 2, pp. 101–113, 2004.

[35] E. R. Alvarez-Buylla, M. Benítez, E. B. Dávila, A. Chaos, C. Espinosa-Soto, and P. Padilla-Longoria, "Gene regulatory network models for plant development," *Current Opinion in Plant Biology*, vol. 10, no. 1, pp. 83–91, 2007.

[36] E. Segal, M. Shapira, A. Regev, et al., "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data," *Nature Genetics*, vol. 34, no. 2, pp. 166–176, 2003.

[37] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, supplement 2, pp. S231–S240, 2002.

[38] X. Chen, M. Chen, and K. Ning, "*BNArray*: an R package for constructing gene regulatory networks from microarray data by using Bayesian network," *Bioinformatics*, vol. 22, no. 23, pp. 2952–2954, 2006.

[39] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo, "How to infer gene networks from expression profiles," *Molecular Systems Biology*, vol. 3, article 78, pp. 1–10, 2007.

[40] H. de Jong, J. Geiselmann, C. Hernandez, and M. Page, "Genetic network analyzer: qualitative simulation of genetic regulatory networks," *Bioinformatics*, vol. 19, no. 3, pp. 336–344, 2003.

[41] A. Nikitin, S. Egorov, N. Daraselia, and I. Mazo, "Pathway studio—the analysis and navigation of molecular networks," *Bioinformatics*, vol. 19, no. 16, pp. 2155–2157, 2003.

[42] P. Shannon, A. Markiel, O. Ozier, et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[43] A. Funahashi, M. Morohashi, H. Kitano, and N. Tanimura, "CellDesigner: a process diagram editor for gene-regulatory and biochemical networks," *BIOSILICO* , vol. 1, no. 5, pp. 159–162, 2003.

[44] B. H. Junker, C. Klukas, and F. Schreiber, "Vanted: a system for advanced data analysis and visualization in the context of biological networks," *BMC Bioinformatics*, vol. 7, article 109, pp. 1–13, 2006.

[45] B.-J. Breitkreutz, C. Stark, and M. Tyers, "Osprey: a network visualization system," *Genome Biology*, vol. 4, no. 3, article R22, pp. 1–4, 2003.

[46] Z. Hu, J. Mellor, J. Wu, and C. DeLisi, "VisANT: an online visualization and analysis tool for biological interaction data," *BMC Bioinformatics*, vol. 5, article 17, pp. 1–8, 2004.

[47] Z. Hu, D. M. Ng, T. Yamada, et al., "VisANT 3.0: new modules for pathway visualization, editing, prediction and construction," *Nucleic Acids Research*, vol. 35, pp. W625–632, 2007.

[48] E. S. Wurtele, J. Li, L. Diao, et al., "MetNet: software to build and model the biogenetic lattice of *Arabidopsis*," *Comparative and Functional Genomics*, vol. 4, no. 2, pp. 239–245, 2003.

[49] M. Baitaluk, M. Sedova, A. Ray, and A. Gupta, "BiologicalNetworks: visualization and analysis tool for systems biology," *Nucleic Acids Research*, vol. 34, pp. W466–W471, 2006.

[50] A. Lüdemann, D. Weicht, J. Selbig, and J. Kopka, "PaVESy: pathway visualization and editing system," *Bioinformatics*, vol. 20, no. 16, pp. 2841–2844, 2004.

[51] T. Toyoda and A. Konagaya, "KnowledgeEditor: a new tool for interactive modeling and analyzing biological pathways based on microarray data," *Bioinformatics*, vol. 19, no. 3, pp. 433–434, 2003.

[52] L. J. Lu, A. Sboner, Y. J. Huang, et al., "Comparing classical pathways and modern networks: towards the development of an edge ontology," *Trends in Biochemical Sciences*, vol. 32, no. 7, pp. 320–331, 2007.

[53] C. J. Krieger, P. Zhang, L. A. Mueller, et al., "MetaCyc: a multiorganism database of metabolic pathways and enzymes," *Nucleic Acids Research*, vol. 32, database issue, pp. D438–D442, 2004.

[54] E. A. Ananko, N. L. Podkolodny, I. L. Stepanenko, et al., "GeneNet in 2005," *Nucleic Acids Research*, vol. 33, database issue, pp. D425–D427, 2005.

[55] S. Rombauts, K. Florquin, M. Lescot, K. Marchal, P. Rouzé, and Y. van de Peer, "Computational approaches to identify promoters and *cis*-regulatory elements in plant genomes," *Plant Physiology*, vol. 132, no. 3, pp. 1162–1176, 2003.

[56] W. W. Wasserman, M. Palumbo, W. Thompson, J. W. Fickett, and C. E. Lawrence, "Human-mouse genome comparisons to locate regulatory sites," *Nature Genetics*, vol. 26, no. 2, pp. 225–228, 2000.

[57] I. A. Shahmuradov, A. J. Gammerman, J. M. Hancock, P. M. Bramley, and V. V. Solovyev, "PlantProm: a database of plant promoter sequences," *Nucleic Acids Research*, vol. 31, no. 1, pp. 114–117, 2003.

[58] V. Matys, E. Fricke, R. Geffers, et al., "TRANSFAC®: transcriptional regulation, from patterns to profiles," *Nucleic Acids Research*, vol. 31, no. 1, pp. 374–378, 2003.

[59] C. Galuschka, M. Schindler, L. Bülow, and R. Hehl, "AthaMap web tools for the analysis and identification of co-regulated genes," *Nucleic Acids Research*, vol. 35, database issue, pp. D857–D862, 2007.

[60] M. Lescot, P. Déhais, G. Thijs, et al., "PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 325–327, 2002.

[61] K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga, "Plant cis-acting regulatory DNA elements (PLACE) database: 1999," *Nucleic Acids Research*, vol. 27, no. 1, pp. 297–300, 1999.

[62] T. R. O'Connor, C. Dyreson, and J. J. Wyrick, "Athena: a resource for rapid visualization and systematic analysis of *Arabidopsis* promoter sequences," *Bioinformatics*, vol. 21, no. 24, pp. 4411–4413, 2005.

[63] R. V. Davuluri, H. Sun, S. K. Palaniswamy, et al., "AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis *cis*-regulatory elements and transcription factors," *BMC Bioinformatics*, vol. 4, article 25, pp. 1–11, 2003.

[64] N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, "Genomic analysis of regulatory network dynamics reveals large topological changes," *Nature*, vol. 431, no. 7006, pp. 308–312, 2004.

[65] J.-D. J. Han, N. Bertin, T. Hao, et al., "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, no. 6995, pp. 88–93, 2004.

[66] A. R. Joyce and B. Ø. Palsson, "The model organism as a system: integrating 'omics' data sets," *Nature Reviews Molecular Cell Biology*, vol. 7, no. 3, pp. 198–210, 2006.

[67] S. P. Gygi, Y. Rochon, B. R. Franza, and R. Aebersold, "Correlation between protein and mRNA abundance in yeast," *Molecular and Cellular Biology*, vol. 19, no. 3, pp. 1720–1730, 1999.

[68] K. M. Waters, J. G. Pounds, and B. D. Thrall, "Data merging for integrated microarray and proteomic analysis," *Briefings in Functional Genomics and Proteomics*, vol. 5, no. 4, pp. 261–272, 2006.

[69] M. P. Washburn, A. Koller, G. Oshiro, et al., "Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 6, pp. 3107–3112, 2003.

[70] L. W. Sumner, P. Mendes, and R. A. Dixon, "Plant metabolomics: large-scale phytochemistry in the functional genomics era," *Phytochemistry*, vol. 62, no. 6, pp. 817–836, 2003.

[71] D. B. Kell, "Metabolomics and systems biology: making sense of the soup," *Current Opinion in Microbiology*, vol. 7, no. 3, pp. 296–307, 2004.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 619832, 12 pages
doi:10.1155/2008/619832

Resource Review

Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics

Ana Conesa and Stefan Götz

Bioinformatics Department, Centro de Investigación Príncipe Felipe, 4012 Valencia, Spain

Correspondence should be addressed to Ana Conesa, aconesa@cipf.es

Received 5 October 2007; Accepted 26 November 2007

Recommended by Chunguang Du

Functional annotation of novel sequence data is a primary requirement for the utilization of functional genomics approaches in plant research. In this paper, we describe the Blast2GO suite as a comprehensive bioinformatics tool for functional annotation of sequences and data mining on the resulting annotations, primarily based on the gene ontology (GO) vocabulary. Blast2GO optimizes function transfer from homologous sequences through an elaborate algorithm that considers similarity, the extension of the homology, the database of choice, the GO hierarchy, and the quality of the original annotations. The tool includes numerous functions for the visualization, management, and statistical analysis of annotation results, including gene set enrichment analysis. The application supports InterPro, enzyme codes, KEGG pathways, GO direct acyclic graphs (DAGs), and GOSlim. Blast2GO is a suitable tool for plant genomics research because of its versatility, easy installation, and friendly use.

1. INTRODUCTION

Functional genomics research has expanded enormously in the last decade and particularly the plant biology research community has extensively included functional genomics approaches in their recent research proposals. The number of Affymetrix plant GeneChips, for example, has doubled in the last two years [1] and extensive international genomics consortia exist for major crops (see last PAG Conference reports for an updated impression on current plant genomics, http://www.intl-pag.org). Not less importantly, many middle-sized research groups are also setting up plant EST projects and producing custom microarray platforms [2]. This massive generation of plant sequence data and rapid spread of functional genomics technologies among plant research labs has created a strong demand for bioinformatics resources adapted to vegetative species. Functional annotation of novel plant DNA sequences is probably one of the top requirements in plant functional genomics as this holds, to a great extent, the key to the biological interpretation of experimental results. Controlled vocabularies have imposed along the way as the strategy of choice for the effective annotation of the function of gene products. The use of controlled vocabularies greatly facilitates the exchange of biological knowledge and the benefit from computational resources that manage this knowledge. The gene ontology (GO, http://www.geneontology.org) [3] is probably the most extensive scheme today for the description of gene product functions but also other systems such as enzyme codes [4], KEGG pathways [5], FunCat [6], or COG [7] are widely used within molecular databases. Many bioinformatics tools and methods have been developed to assist in the assignment of functional terms to gene products (reviewed in [8]). Fewer resources, however, are available when it comes to the large-scale functional annotation of novel sequence data of nonmodel species, as would be specifically required in many plant functional genomics projects. Web-based tools for the functional annotation of new sequences include AutoFact [9], GOanna/AgBase [10], GOAnno [11], Goblet [12], GoFigure + GoDel [13], GoPET [14], Gotcha [15], HT-GO-FAT (liru.ars.usda.gov/ht-go-fat.htm), InterProScan [16], JAFA [17], OntoBlast [18], and PFP [19]. Additionally, functional annotation capabilities are usually incorporated in EST analysis pipelines. A few relevant examples are ESTExplorer, ESTIMA, ESTree. or JUICE (see [2] for a survey in EST analysis). These resources are valuable tools

for the assignment of functional terms to uncharacterized sequences but usually lack high-throughput and data mining capabilities, in the first case, or provide automatic solutions without much user interactivity, in the second. In this paper, we describe the Blast2GO (B2G, www.blast2go.org) application for the functional annotation, management, and data mining of novel sequence data through the use of common controlled vocabulary schemas. The philosophy behind B2G development was the creation of an extensive, user-friendly, and research-oriented framework for large-scale function assignments. The main application domain of the tool is the functional genomics of nonmodel organisms and it is primarily intended to support research in experimental labs where bioinformatics support may not be strong. Since its release in September 2005 [20], more than 100 labs worldwide have become B2G users and the application has been referenced in over thirty peer-reviewed publications (www.blast2go.org/citations). Although B2G has a broad species application scope, the project originated in a crop genomics research environment and there is quite some accumulated experience in the use of B2G in plants, which includes maize, tobacco, citrus, Soybean, grape, or tomato. Projects range from functional assignments of ESTs [21–24] to GO term annotation of custom or commercial plant microarrays [25, 26], functional profiling studies [27–29], and functional characterization of specific plant gene families [30, 31].

In the following sections we will explain more extensively the concepts behind Blast2GO. We will describe in detail main functionalities of the application and show a use case that illustrates the applicability of B2G to plant functional genomics research.

2. BLAST2GO HIGHLIGHTS

Four main driving concepts form the foundation of the Blast2GO software: biology orientation, high-throughput, annotation flexibility,and data-mining capability.

Biology orientation. The target users of Blast2GO are biology researchers working on functional genomics projects in labs where strong bioinformatics support is not necessarily present. Therefore, the application has been conceived to be easy to install, to have minimal setup and maintenance requirements, and to offer an intuitive user interface. B2G has been implemented as a multiplatform Java desktop application made accessible by Java Webstart technology. This solution employs the higher versatility of a locally running application while assuring automatic updates provided that an internet connection is available. This implementation has proven to work very efficiently in the fast transfer to users of new functionalities and for bug fixes. Furthermore, access to data in B2G is reinforced by graphical parameters that on one hand allow the easy identification and selection of sequences at various stages of the annotation process and, on the other hand, permit the joint visualization of annotation results and highlighting of most relevant features.

High-throughput while interactive. Blast2GO strives to be the application of choice for the annotation of novel sequences in functional genomics projects where thousands of fragments need to be characterized. In principle, B2G accepts any amount of records within the memory resources of the user's work station. Typical data files of 20 to 30 thousand sequences can be easily annotated on a 2 Giga RAM PC (larger projects may use the graphical interface free version of Blast2GO). During the annotation process, intermediate results can be accessed and modified by the user if desired.

Flexible annotation. Functional annotation in Blast2GO is based on homology transfer. Within this framework, the actual annotation procedure is configurable and permits the design of different annotation strategies. Blast2GO annotation parameters include the choice of search database, the strength and number of blast results, the extension of the query-hit match, the quality of the transferred annotations, and the inclusion of motif annotation. Vocabularies supported by B2G are gene ontology terms, enzyme codes (EC), InterPro IDs, and KEGG pathways.

Data mining on annotation results. Blast2GO is not a mere generator of functional annotations. The application includes a wide range of statistical and graphical functions for the evaluation of the annotation procedure and the final results. Especially, (relative) abundance of functional terms can be easily assessed and visualized.

The first release of B2G covered basic application functionalities: high-throughput blast against NCBI or local databases, mapping, annotation, and gene set enrichment analysis; scalar vector graphics (SVG) combined graphs and basic distributions charts. Enhanced modules for massive blast, modification of annotation intensity, curation, additional vocabularies, high-performing customizable graphs and pathway charts, data mining and sequence handling, as well as a wide array of input and output formats have been incorporated into the Blast2GO suite.

3. THE BLAST2GO APPLICATION

Figure 1 shows the basic components of the Blast2GO suite. Functional assignments proceed through an elaborate annotation procedure that comprises a central strategy plus refinement functions. Next, visualization and data mining engines permit exploiting the annotation results to gain functional knowledge.

3.1. The annotation procedure

The Blast2GO annotation procedure consists of three main steps: blast to find homologous sequences, mapping to collect GO terms associated to blast hits, and annotation to assign trustworthy information to query sequences. Once GO terms have been gathered, additional functionalities enable processing and modification of annotation results.

Blast step. The first step in B2G is to find sequences similar to a query set by blast [32]. B2G accepts nucleotide and protein sequences in FASTA format and supports the four basic blast programs (blastx, blastp, blastn, and tblastx). Homology searches can be launched against public databases

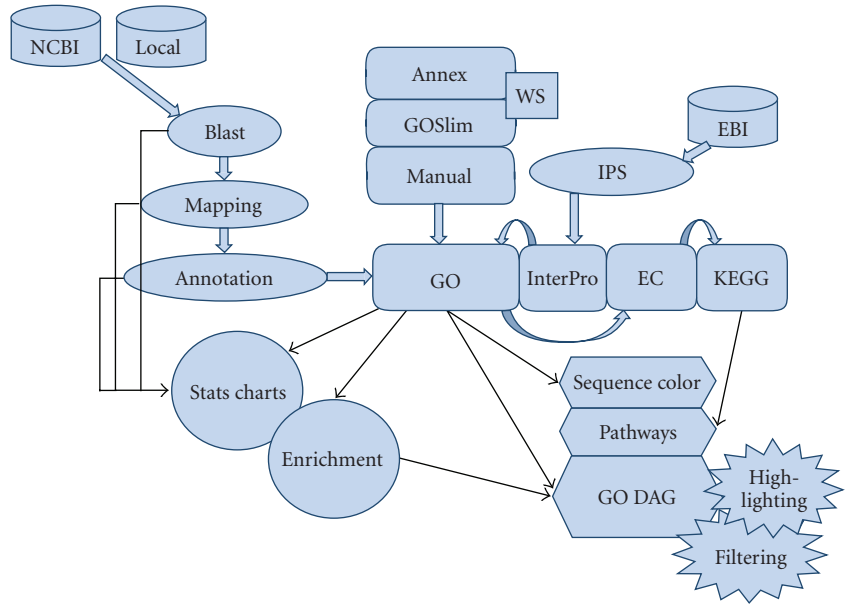


Figure 1: Schematic representation of Blast2GO application. GO annotations are generated through a 3-step process: blast, mapping, annotation. InterPro terms are obtained from InterProScan at EBI, converted and merged to GOs. GO annotation can be modulated from Annex, GOSlim web services and manual editing. EC and KEGG annotations are generated from GO. Visual tools include sequence color code, KEGG pathways, and GO graphs with node highlighting and filtering options. Additional annotation data-mining tools include statistical charts and gene set enrichment analysis functions.

$$\begin{aligned} \mathrm{DT} &= \max\left(\text{similarity} \times \mathrm{EC}_{\text{weight}}\right) \\ \mathrm{AT} &= (\#\mathrm{GO} - 1) \times \mathrm{GO}_{\text{weight}} \\ \mathrm{AR} &: \text{lowest.node}(\mathrm{AS}(\mathrm{DT} + \mathrm{AT}) \geq \text{threshold}) \end{aligned}$$

Figure 2: Blast2GO annotation rule.

such as (the) NCBI nr using a query-friendly version of blast (QBlast). This is the default option and in this case, no additional installations are needed. Alternatively, blast can be run locally against a proprietary FASTA-formatted database, which requires a working www-blast installation. The Make Filtered Blast-GO-BD function in the Tools menu allows the creation of customized databases containing only GO-annotated entries, which can be used in combination with the local blast option. Other configurable parameters at the blast step are the expectation value (*e*-value) threshold, the number of retrieved hits, and the minimal alignment length (hsp length) which permits the exclusion of hits with short, low *e*-value matches from the sources of functional terms. Annotation, however, will ultimately be based on sequence similarity levels as similarity percentages are independent of database size and more intuitive than *e*-values. Blast2GO parses blast results and presents the information for each sequence in table format. Query sequence descriptions are obtained by applying a language processing algorithm to hit descriptions, which extracts informative names and avoids low-content terms such as "hypothetical protein" or "expressed protein".

Mapping step. Mapping is the process of retrieving GO terms associated to the hits obtained after a blast search. B2G performs three different mappings as follows. (1) Blast result accessions are used to retrieve gene names (symbols) making use of two mapping files provided by NCBI (gene-info, gene2accession). Identified gene names are searched in the species-specific entries of the gene product table of the GO database. (2) Blast result GI identifiers are used to retrieve UniProt IDs making use of a mapping file from PIR (Non-redundant Reference Protein database) including PSD, UniProt, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB. (3) Blast result accessions are searched directly in the DBXRef Table of the GO database.

Annotation step. This is the process of assigning functional terms to query sequences from the pool of GO terms gathered in the mapping step. Function assignment is based on the gene ontology vocabulary. Mapping from GO terms to enzyme codes permits the subsequent recovery of enzyme codes and KEGG pathway annotations. The B2G annotation algorithm takes into consideration the similarity between query and hit sequences, the quality of the source of GO assignments, and the structure of the GO DAG. For each query sequence and each candidate GO term, an annotation score (AS) is computed (see Figure 2). The AS is composed of two terms. The first, direct term (DT), represents the highest similarity value among the hit sequences bearing this GO term, weighted by a factor corresponding to its evidence code (EC). A GO term EC is present for every annotation in the GO database to indicate the procedure of functional assignment. ECs vary from experimental evidence, such as inferred by direct assay (IDA) to unsupervised assignments such as

inferred by electronic annotation (IEA). The second term (AT) of the annotation rule introduces the possibility of abstraction into the annotation algorithm. Abstraction is defined as the annotation to a parent node when several child nodes are present in the GO candidate pool. This term multiplies the number of total GOs unified at the node by a user-defined factor or GO weight (GOw) that controls the possibility and strength of abstraction. When all ECw's are set to 1 (no EC control) and the GOw is set to 0 (no abstraction is possible), the annotation score of a given GO term equals the highest similarity value among the blast hits annotated with that term. If the ECw is smaller than one, the DT decreases and higher query-hit similarities are required to surpass the annotation threshold. If the GOw is not equal to zero, the AT becomes contributing and the annotation of a parent node is possible if multiple child nodes coexist that do not reach the annotation cutoff. Default values of B2G annotation parameters were chosen to optimize the ratio between annotation coverage and annotation accuracy [20]. Finally, the AR selects the lowest terms per branch that exceed a user-defined threshold.

The annotation step in B2G can be further adjusted by setting additional filters to the hit sequences considered as annotation source. A lower limit can be set at the e-value parameter to ensure a minimum confidence at the level of homology. Similarly, % "hit" filter has been implemented to assure that a given percentage of the hit sequence is actually spanned by the query. This parameter is of importance to prevent potential function transfer from nonmatching sequence regions of modular proteins. Additionally, the minimal hsp length required at the blast step permits control of the length of the matching region.

3.2. Modulation of annotation

Blast2GO includes different functionalities to complete and modify the annotations obtained through the above-defined procedure.

Additional vocabularies. Enzyme codes and KEGG pathway annotations are generated from the direct mapping of GO terms to their enzyme code equivalents. Additionally, Blast2GO offers InterPro searches directly from the B2G interface. The user, identified by his/her email address, has the possibility of selecting different databases available at the InterProEBI web server [33]. B2G launches sequence queries in batch, and recovers, parses, and uploads InterPro results. Furthermore, InterPro IDs can be mapped to GO terms and merged with blast-derived GO annotations to provide one integrated annotation result. In this process, B2G ensures that only the lowest term per branch remains in the final annotation set, removing possible parent-child relationships originating from the merging action.

Annotation fine-tuning. Blast2GO incorporates three additional functionalities for the refinement of annotation results. Firstly, the Annex function allows annotation augmentation through the Second Layer concept developed by The Norwegian University of Science and Technology (http://www.goat.no, [34]). Basically, the Second Layer database is a collection of manually curated univocal relationships between GO terms from the different GO categories that permits the inference of biological process and cellular component terms from molecular function annotations. Up to 15% of annotation increase and around 30% of GO term confirmations are obtained through the Annex dataset [20]. Secondly, annotation results can be summarized through GOSlim mapping. GOSlim consists of a subset of the gene ontology vocabulary encompassing key ontological terms and a mapping function between the full GO and the GOSlim. Different GOSlim mappings are available, adapted to specific biological domains. At present, GOSlim mappings for plant, yeast, from GOA and Tair, as well as a generic one are available from the GO through Blast2GO. Thirdly, the manual curation function means that the user has the possibility of editing annotation results and manually modifying GO terms and sequence descriptors.

3.3. Visualization and data mining

One aspect of the uniqueness of the Blast2GO software is the availability of a wide array of functions to monitor, evaluate, and visualize the annotation process and results. The purpose of these functions is to help understand how functional annotation proceeds and to optimize performance.

Statistical charts. Summary statistics charts are generated after each of the annotation steps. Distribution plots for e-value and similarity within blast results give an idea of the degree of homology that query sequences have in the searched database. Once mapping has been completed, the user can check the distribution of evidence codes in the recovered GO terms and the original database sources of annotations. These charts give an indication of suitable values for B2G annotation parameters. For example, when a good overall level of sequence similarity is obtained for the dataset, the default annotation cutoff value could be raised to improve annotation accuracy. Similarly, if evidence code charts indicate a low representation of experimentally derived GOs, the user might choose to increase the weight given to electronic annotations. After the final annotation step, new charts show the distribution of annotated sequences, the number of GOs per sequence, the number of sequences per GO, and the distribution of annotations per GO level, which jointly provide a general overview of the performance of the annotation procedure.

Sequence coloring. The visual approach of B2G is further represented by the color code given to annotated sequences. During the annotation process, the background color of active sequences changes according to their analysis status. Nonblasted sequences are displayed in white and change to light red once a positive blast result is obtained. If the result was negative, they will stay dark red. Mapped sequences are depicted in green while annotated sequences become blue. Finally, manually curated sequences can be labeled and colored purple (see Figure 3(A)). Sequence coloring is a simple and effective way of identifying sequences that have reached differential stages during the annotation process. Furthermore, sequences can be selected by their color. This is a very

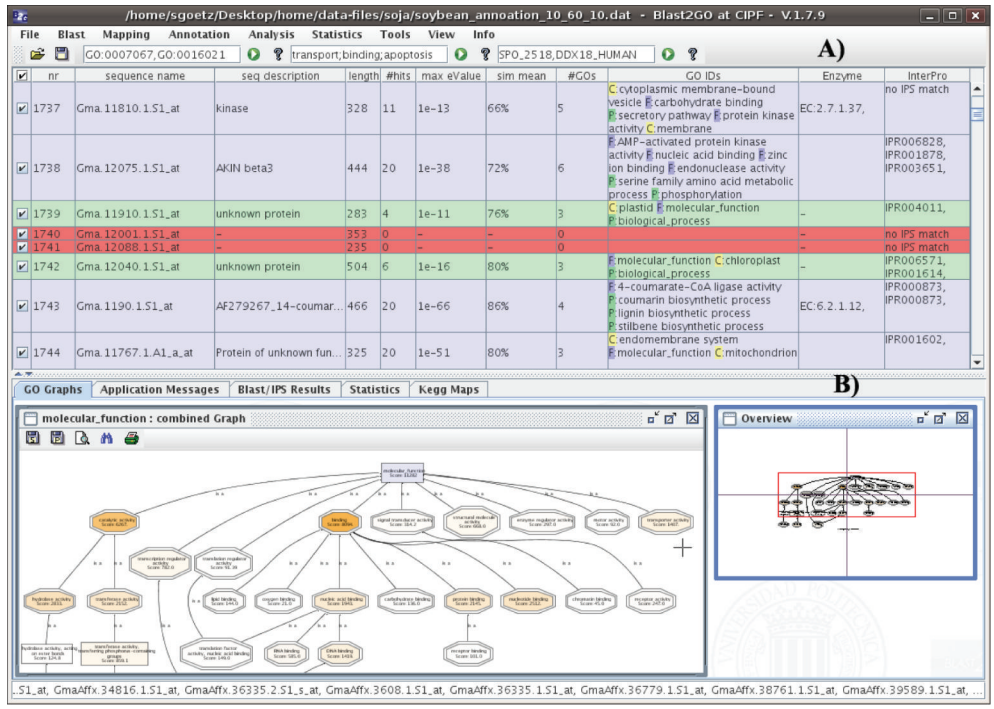


FIGURE 3: Blast2GO user interface. (A) Main sequence table showing sequence color codes. (B) Graphical tab showing a combined graph with score highlighting.

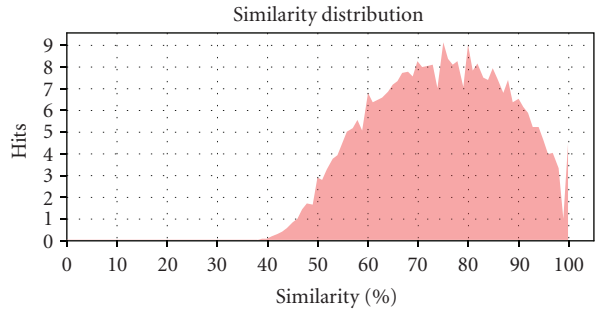


FIGURE 4: Similarity distribution of Soybean GeneChip. Similarity is computed of each query-hot pair as the sum of similarity values for all matching hsps.

useful function for the interactive use of the application. For example, sequences that stayed dark red after blast (no positive result) can be selected to be launched to InterProScan. Sequences that remained green (mapping code) after the annotation step can be selected and reannotated with more permissive parameters.

Combined graph. A core functionality of Blast2GO is the joint visualization of groups of GO terms within the structure of the GO DAG. The combined graph function is typically used to study the collective biological meaning of a set of sequences. Combined graphs are a good alternative to enrichment analysis (see below) where no reference set is to be considered or the number of involved sequences is low. B2G includes several parameters to make these combined graphs easy to analyze and navigate. Firstly, the ZWF format [35], a powerful scalable vector graphics engine, has been adopted to make zooming and browsing through the DAG fast and light. Secondly, annotation-rich areas of the generated DAG can be readily spotted by a node-coloring function. B2G colors nodes either by the number of sequences gathered at that term (additive function) or by a node information score (exponential function, $\sum_{GOs} seq \cdot \alpha^{dist}$) that considers the places of direct annotation. This B2G score takes into account the amount of sequences collected at a given term but penalizes by the distance to the node of actual annotation [20]. The B2G score has shown to be a useful parameter for the identification of "hot" terms within a specific DAG (see Figure 3(B); Conesa, unpublished). Thirdly, the extension and density of the plotted DAG can be modulated by a node filter function. When the number of sequences involved in the combined graph is large, the resulting DAG can be too big to be practical. B2G permits filtering out of low informative terms by imposing a threshold on the number of annotated sequences or B2G scores for a node to be displayed. In this case, the number of omitted nodes is given for each branch, which is an indication of the level of local compression applied to that branch.

Enrichment analysis. A typical data mining approach applied in functional genomics research is the identification of functional classes that statistically differ between two lists of terms. For example, one might want to know the functional categories that are over- or underrepresented in the set of differentially expressed genes of a microarray experiment, or

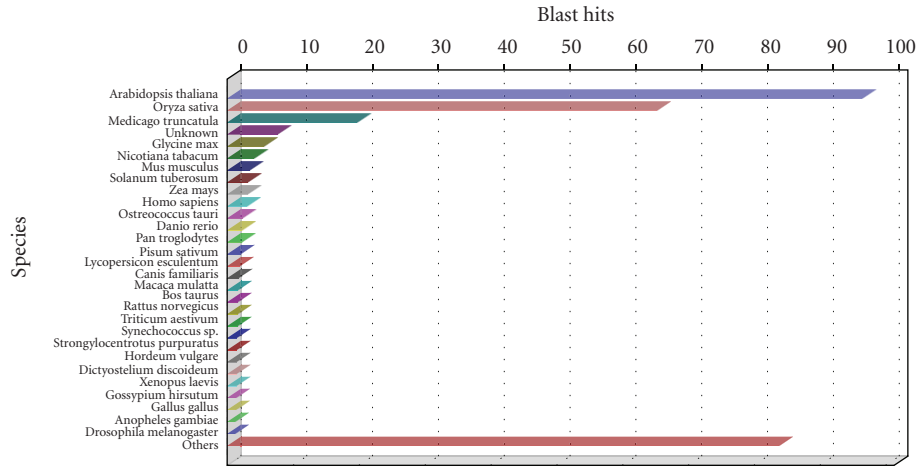


Figure 5: Species distribution chart of Soybean GeneChip after blastx to NCBI nr.

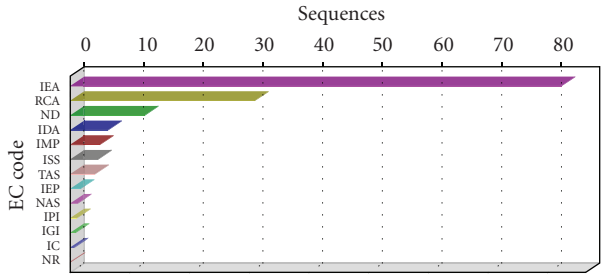


Figure 6: Evidence code distribution chart of Soybean GeneChip after mapping to B2G database.

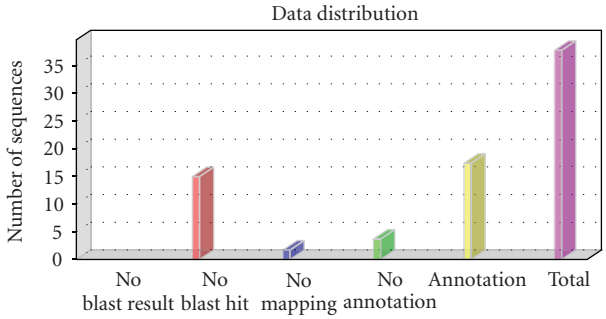


Figure 7: Annotation process results for Soybean Affymetrix GeneChip.

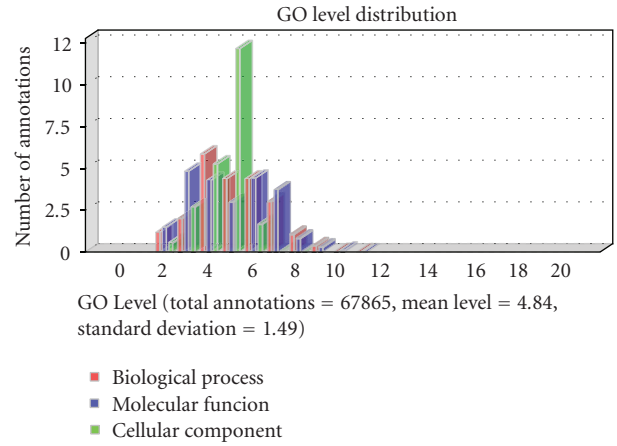


Figure 8: GO level distribution chart for Soybean Affymetrix GeneChip. Most sequences have between 3 and 6 GO terms annotated.

it could be of interest to find which functions are distinctly represented between different libraries of an EST collection. Blast2GO has integrated the Gossip [36] package for statistical assessment of differences in GO term abundance between two sets of sequences. This package employs the Fisher's exact test and corrects for multiple testing. For this analysis, the involved sequences with their annotations must be loaded in the application. B2G returns the GO terms under- or over-represented at a specified significance value. Results are given as a plain table and graphically as a bar chart and as a DAG with nodes colored by their significance value. Also in this case, graph pruning and summarizing functions are available.

3.4. Other functionalities

Next to the annotation and data mining functions, Blast2GO comprises a number of additional functionalities to handle data. In this section, we briefly comment on some of them.

Import and export. B2G provides different formats for the exchange of data. Typically, B2G inputs are FASTA-formatted sequences and returns a tab-delimited file with GO annotations. Other supported output formats are GOstats and GOSpring. Furthermore, B2G also accepts blast results in xml format. This option permits skipping the first step of the B2G annotation procedure when a blast result is already present. Similarly, when accession IDs or gene symbols are known for the query sequences, these can be directly uploaded in B2G and the application will query the B2G

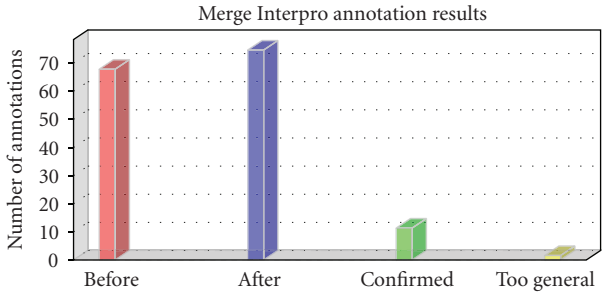


Figure 9: InterPro merging statistics. Red column: total number of GO annotations before adding InterPro-based GO terms. Blue: total number of GO annotations after adding InterPro-based GO terms. Green: number of blast-based GO terms confirmed after InterPro merging. Yellow: too general terms removed after InterPro merging.

database for their annotations. Moreover, the Main Sequence table (see Figure 3) can be saved to a file at any moment to store intermediate results. Finally, graph and enrichment analysis results are presented both graphically and as text files.

Validation. The "true path rule" defined by the gene ontology consortium for the GO DAG assures that all the terms in the pathway from a term up to the root must always be true for a given gene product. The B2G annotation validation function applies this property to annotation results by removing any parent term that has a child within the sequence annotation set. B2G always executes validation after any modification has been made to the existing annotation, for example, after InterPro merging, Annex augmentation, or manual curation.

Comparison of two sets of GO terms. Given two annotation results, Blast2GO can compare their implicit DAG structures. B2G computes the number of identical nodes, more general and more specific terms within the same branch, and terms located to different branches or different GO main categories. Comparison is directional; this means that the active annotation file is contrasted to a reference or external one. Each GO term is compared to all terms in the reference set and the best matching comparison result is recorded. Once a term is matched, it is removed from the query set.

3.5. Some performance figures

The annotation accuracy of Blast2GO has been evaluated by comparing B2G GO annotation results to the existing annotation in a set of manually annotated Arabidopsis proteins that had been previously removed from the nr database. This evaluation indicated that using B2G default parameters, nearly 70% of identical branch recovery was achievable, which is at the top end of the methods that are based on homology search [20]. More recent evaluations have shown that Blast2GO annotation behavior is consistent across species and datasets. In general, the blast step has shown to be decisive in the annotation coverage. For a great deal of sequences with a positive blast result, functional information is available in the GO database and the final annotation success is related to the length and quality of the query sequence and the strictness of annotation parameters. Typically and using default parameters, around 50–60% of annotation success is common for EST datasets and slightly higher values are obtained for full-length proteins (Table 1).

On average, between 3 and 6 GO terms are assigned per sequence at a mean GO level very close to 5. InterPro, Annex, and GOw annotation parameters significantly increase annotation intensity—around 15%—and validate annotation results. Furthermore, default annotation options tend to provide coherent results and resemble the functional assignment obtained by a human computational reviewed analysis [37].

3.6. Use case

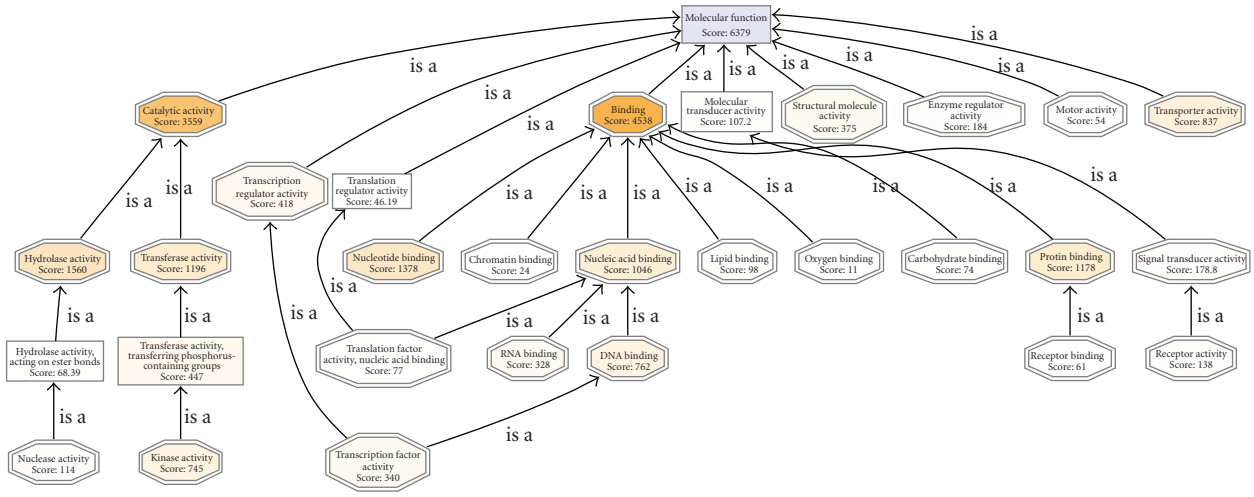
In this section, we present a typical use case of Blast2GO to illustrate the major application features described in the previous sections. We will address the functional annotation of the Soybean Affymetrix GeneChip. The GeneChip Soybean Genome Array targets over 37,500 Soybean transcripts (www.affymetrix.com). The array also contains transcripts for studying two pathogens important for Soybean research. Sequence data and a detailed annotation sheet for the Soybean Genome Array are provided at the Blast2GO site (http://blast2go.bioinfo.cipf.es/b2gdata/soybean).

Blast

Sequence data in FASTA format were uploaded into the application from the menu File → Open File. After selecting the Blast menu, a dialog opens where we can indicate the parameters for the blast step. In our case, the easiest option is to select the nr protein database and perform blast remotely on the NCBI server through Qblast. Additional blast parameters are kept at default values: *e*-value threshold of 1*e*-3 and a recovery of 20 hits per sequence. These permissive values are chosen to retrieve a large amount of information at this first time-consuming step. Annotation stringency will be decided later in the annotation procedure. Furthermore, we set the hsp filter to 33 to avoid hits where the length of the matching region is smaller than 100 nucleotides. After launching, blast sequences turn red as results arrive, up to a total of 22,788. Once blast is completed, we can visualize different charts (similarity, *e*-value, and species distributions, see supplementary material available online at doi:10.1155/2008/619832) to get an impression of the quality of the query sequences and the blast procedure. For example, Statistics → Blast statistics → Similarity distribution chart (see Figure 4) shows that most sequences have blast similarity values of 50–60% or higher. This information is useful for choosing the annotation cutoff parameter at the annotation step, and suggests that taking a value of 60 would be adequate. Furthermore, the Species distribution chart (see Figure 5) shows a great majority of Arabidopsis sequences

Table 1: Blast2GO performance figures of seven cDNA datasets. FE: percentage of sequences with some functional evidence (Mapping or InterProScan positive). BA: percentage of blast-based annotated sequences, #GO: number of GOs per sequence. GO *L*: mean GO level. IP: percentage of annotation increase by InterProScan. Ann: percentage of annotation increase by Annex. TA: total percentage of annotated sequences (including blast and InterPro). Datasets are described in [37].

DataSet	FE	BA	no. GOs	GO *L*	IP	Ann	TA
C. clementina	70.2	58.2	4.4	5.10	7.9	11.8	62.3
M. incog	70.7	55.7	5.6	4.95	11.8	9.9	63.9
T. harzianum	61.1	47.7	3.6	5.27	14.4	16.2	53.4
G. max	61.8	51.1	4.3	5.11	6.1	11.8	53.5
P. flesus	50.1	34.4	5.2	5.07	21.9	10.6	45.1
A. phagocytophilum	56.6	42.5	3.0	4.91	35.4	20.9	49.1
Whale metagenome	69.5	50.7	3.0	4.45	17.6	18	58.8



Molecular function-combined graph

Figure 10: Molecular function combined graph of GOSlim annotation of the Soybean Affymetrix GeneChip. Nodes are colored by score value.

within the blast hits, followed by Cotton, Medicago, Glycine, and Nicotiana.

Mapping

Mapping is a nonconfigurable option launched from the menu Mapping → Make Mapping. GO terms could be found for 21,079 sequences (56%). Mapping charts (menu Statistics → Mapping Statistics) permit the evaluation of mapping results. The evidence code distribution chart (see Figure 6) shows an overrepresentation of electronic annotations, although other nonautomatic codes, such as review by computational analysis (RCA), inferred by mutant phenotype (IMP), or inferred by direct assay (IDA) are also well represented. This suggests that an annotation strategy that promotes nonelectronic ECs would be meaningful as it would benefit from the high-quality GO terms without totally excluding electronic annotations. Therefore, the default EC weights (menu Annotation → Set Evidence Code Weights) that adjust proportionally to the reliability of the source annotation will be maintained at the annotation step.

Annotation

Taking into consideration the charts generated by the previous steps, we have chosen an annotation configuration with an *e*-value filter of 1*e*-6, default gradual EC weights, a GO weight of 15, and an annotation cutoff of 60. This implies that only sequences with a blast *e*-value lower than 1*e*-6 will be considered in the annotation formula, that the query-hit similarity value adjusted by the EC weight of the GO term should be at least 60, and that abstraction is strongly promoted. This annotation configuration resulted in 17,778 successfully GO annotated sequences with a total of 70,035 GO terms at a mean GO level (distance of the GO term to the ontology root term) of 4.72. Furthermore, 6,345 enzyme codes were mapped to a total of 5,390 sequences. Once annotation has been completed, we can visualize the results at each step of the annotation process (see Figure 7). Reannotation is possible by selecting green or red sequences

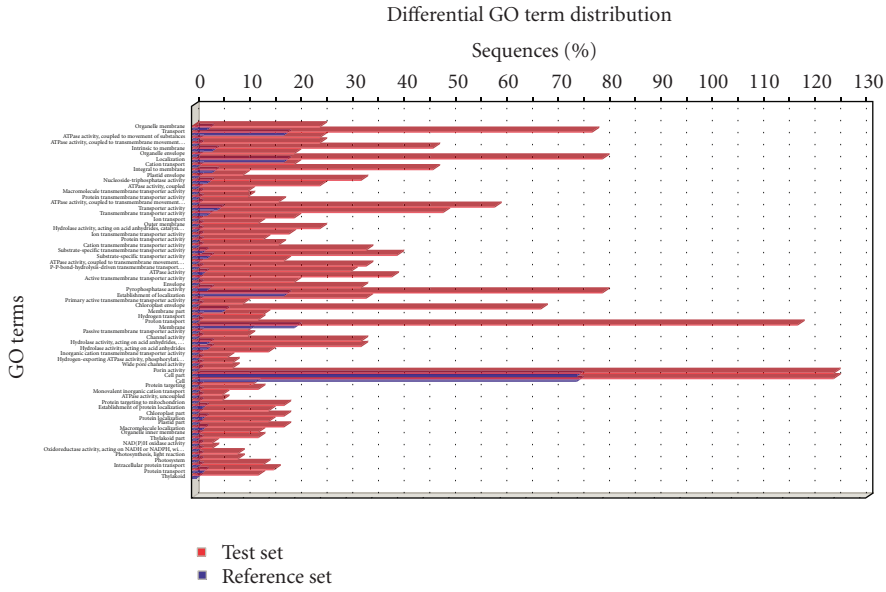


FIGURE 11: Bar chart for functional category enrichment analysis of Soybean membrane proteins. The *Y*-axis shows significantly enriched GO terms and the *X*-axis give the relative frequency of the term. Red bars correspond to test set (membrane) and blue bars correspond to the whole Soybean genome array.

(Tools → Select Sequences by Color) and rerunning blast, mapping, and annotation with different, more permissive parameters. In this way, we obtain a trustworthy annotation for most sequences and behave more permissively only for those sequences which are hard to annotate. Other charts available at the Annotation Statistics menu show the distribution of GO levels (see Figure 8), the length of annotated sequences, and the histogram of GO term abundance.

Annotation augmentation

Blast-based GO annotations can be increased by means of the integrated InterProScan function available under Annotation → Run InterProScan. The user must provide his/her email address and select the motif databases of interest. An InterProScan search against all EBI databases resulted in the recovery of motif functional information for 11,347 sequences and a total of 8,046 GO terms. Once merged to the already existing annotation (Annotation → Add InterProScan GOs to Annotation), 1,189 additional sequences were annotated (see Figure 9). Once Blast plus InterProScan annotations have been gathered, a useful step is to complete implicit annotations through the Annex function (Annotation → Augment Annotation by Annex). After this step, it is recommended to run the function to remove first-level annotations (under Annotation menu). In our use case, the Annex function resulted in the addition of 8,125 new GO terms and a confirmation of 3,892 annotations, which is an average contribution of the Annex function [37].

Manual curation

The manual annotation tool is a useful functionality when information on the automatically generated annotation needs to be changed. For example, the target of GmaAffx.69219.1.S1_at probe was found to be the UDP-glycosyltransferase. The automatic procedure assigned GO terms metabolic process (GO:0008152) and transferase activity, transferring hexosyl groups (GO:0016758) to this sequence. However, as we are aware of the ER localization of this enzyme and its involvement in protein maturation, we would like to add this information to the existing annotation. The manual curation function is available at the Sequence Menu which is displayed by mouse right button click on the selected sequence. From this Menu, the blast and annotation results for this particular sequence can be visualized. Selection of Change Annotations and Description edits the annotation record of GmaAffx.69219.1.S1_at. We can now type in the Annotations box the terms GO:0005783 (endoplasmic reticulum) and GO:0006464 (protein modification process) and mark the manual annotation box. The new annotations are then added and the sequence turns purple (manual annotation color code).

GOSlim

As the number of sequences and different GO terms in the Soybean array is quite large, we are interested in a simpler representation of the functional content of the data. An appropriate option is to map annotations into a GOSlim. At Annotation → Change to GOSlim View, we can select an appropriate GOSlim (generic_plant) for this dataset. Upon completion of slimming sequences acquire the yellow GOSlim coding. The original annotations are stored and can be recovered at any moment. GOSlim mapping generated a set of 105 different annotating GO terms on 18,820 sequences with a mean GO level of 3.41. This means around 40 times less functional diversity than in the original annotation

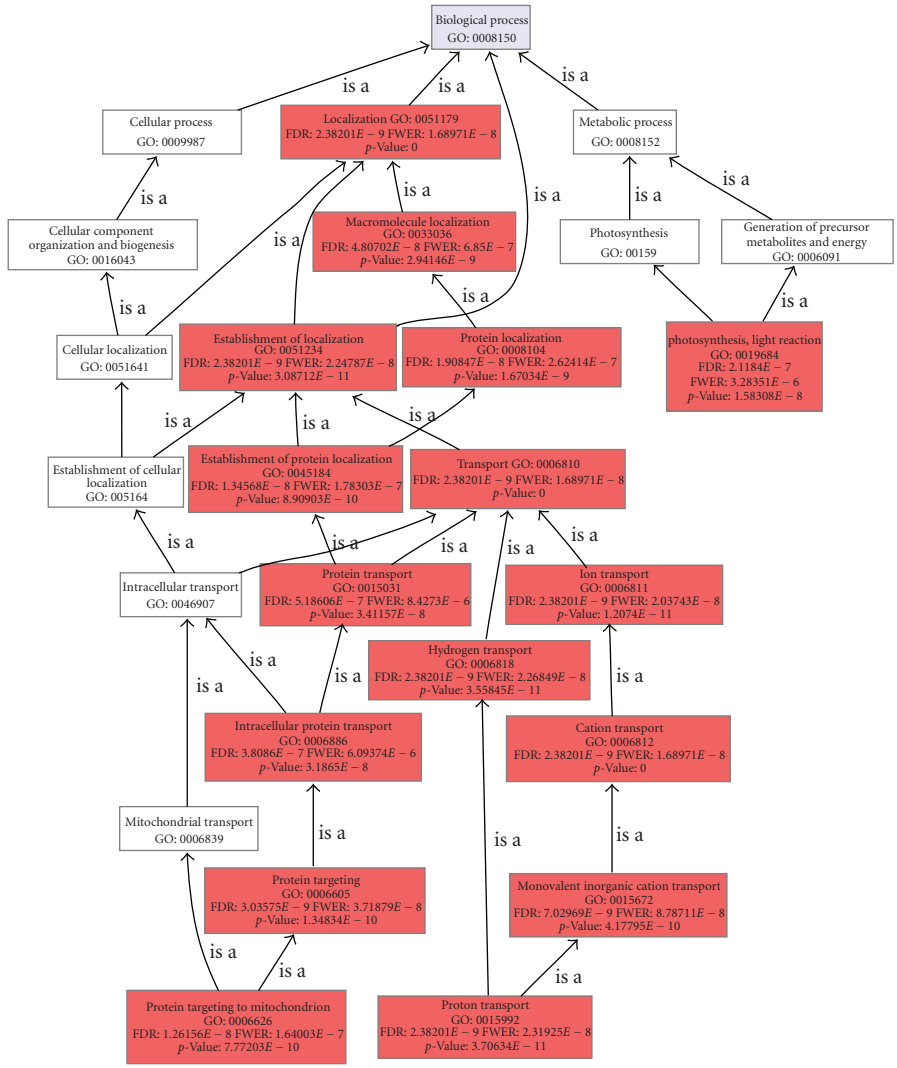


Figure 12: Enriched graph (biological process) of the Soybean membrane subset of sequences. Node filter has been set at FDR < 1*e*-6. Nodes are colored accordingly to their FDR value in the Fisher exact's test against the whole Soybean genome array.

(4533 different terms) and an increase of almost 2 levels of the mean annotation depth.

Combined graph

Once the slimmed annotation is obtained, we can visualize the functional information of the Soybean Genome Array on the GO DAG. This functionality is available under Analysis → Combined Graph. At the Dialog we must indicate the GO category to display (e.g., biological process). To obtain a compact representation of the information, two filters can be applied. For example, by setting the sequence filter to 20, only those nodes with at least 20 sequence assignments will be displayed. By setting the score filter to 20, additionally, parent nodes that do not annotate more sequences than their children terms will be omitted from the graph. Node coloring by score value highlights the areas in the resulting DAG where sequence annotations are most concentrated. Figure 10 shows the Combined Graph for the Molecular Function Category. The two most intensively colored terms at the second GO level indicate the two most abundant functional categories in the Soybean Chip: catalytic activity and binding. Highlighting at lower levels reveals other, most informative, highly represented functional terms, such as hydrolase activity (level 3), kinase activity (level 4), transcription factor activity (level 3), protein binding (level 3), nucleotide binding (level 3), and transporter activity (level 2). The reader is referred to the annotation sheet URL (http://blast2go.bioinfo.cipf.es/b2gdata/soybean) for figure navigation.

Enrichment analysis

The enrichment analysis function in B2G executes a statistical assessment of differences in functional classes between two groups of sequences. To illustrate this function, we have selected all sequences in the Soybean chip which contain the word "membrane" within their description—132

sequences—and compared their annotations to the whole chip. We go to Analysis → Enrichment Analysis → Make Fisher's Exact Test and browse for a text file containing the test set with the names of membrane sequences. As the comparison is made against the complete microarray dataset loaded into the application, no file needs to be selected as Reference. We uncheck the two-tail box to perform only positive enrichment analysis. Upon completion a table with test statistical results is presented in the Statistics tab. This table contains significant GO terms which are ranked according to their significance. Three different significance parameters are given for false-positive control: false discovery rate (FDR), family-wise error rate (FWER), and single test *P*-value (Fisher *P*-value) (see [36] for details). By taking a FDR significance threshold of 0.05, we obtain those functionalities that are strongly significant for membrane proteins in the Soybean Chip. These refer to processes related to transport, protein targeting, and photosynthesis as might be expected for a plant species. Graphical representations of these results can be generated at Analysis → Enrichment Analysis → Bar Chart and Analysis → Enrichment Analysis → Make Enriched Graph. The Bar Chart shows, for each significant GO term, frequency differences between the membrane and the whole chip datasets (see Figure 11). The Enriched Graph shows the DAG of significant terms with a node-coloring proportional to the significance value. This representation helps in understanding the biological context of functional differences and to find pseudoredundancies in the results—parent-child relationships within significant terms—(see Figure 12).

Export results

Once different analyses have been completed the data can be exported in many different ways. The annotation format (menu File → Export → Export Annotations) is the default format for export/import in B2G and simply consists of a tab-delimited file with two columns, one for sequences and other for annotation IDs. Another useful export format is GeneSpring, for communication with this interesting application, which consists of one row per sequence and three different columns showing the descriptions of the GO terms at the three main GO categories. Graphs can be saved in png format. Additionally, all information contained in the Combined Graph can be generated as table (including sequences, GO IDs, levels, and scores) and exported (Analysis → Export Graph Information).

The analysis presented in this use case took about 15 days to complete. Four days were necessary to obtain the totality of 37,500 blast results from the NCBI while twelve days were required for the InterProScan at the EBI web server. Mapping and Annotation were ready within a few hours and one day was necessary to collect and evaluate charts. This shows that with the adequate tools and some training, functional annotation of a plant genome-wide sequence collection is in reach within a couple of weeks.

4. CONCLUSIONS

Functional annotation of novel sequence data is a key requirement for the successful generation of functional genomics in biological research. The Blast2GO suite has been developed to be a useful support to these approaches, especially (but not exclusively) in nonmodel species. This bioinformatics tool is ideal for plant functional genomics research because of the following: (1) it is suitable for any species but can be also customized for specific needs, (2) it combines high throughput with interactivity and curation, and (3) it is user-friendly and requires low bioinformatics efforts to get it running. In our opinion, the major B2G strength is the combination of functional annotation and data mining on annotation results, which means that, within one tool, researchers can generate functional annotation and assess the functional meaning of their experimental results. Further developments of Blast2GO will reinforce this second aspect thought the integration of the tool with the Babelomics (www.babelomics.org, [38]) and GEPAS suites (www.gepas.org, [39]) for the statistical analysis of functional profiling data.

ACKNOWLEDGMENTS

This work has been funded by the Spanish Ramon y Cajal Program and the National Institute of Bioinformatics (a platform of Genoma Espana).

REFERENCES

[1] A. Conesa, J. Forment, J. Gadea, and J. van Dijk, "Microarray technology in agricultural research," in *Microarray Technology Through Applications*, F. Falciani, Ed., pp. 173–209, Taylor & Francis, New York, NY, USA, 2007.

[2] S. H. Nagaraj, R. B. Gasser, and S. Ranganathan, "A hitchhiker's guide to expressed sequence tag (EST) analysis," *Briefings in Bioinformatics*, vol. 8, no. 1, pp. 6–21, 2007.

[3] M. Ashburner, C. A. Ball, J. A. Blake, et al., "Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[4] I. Schomburg, A. Chang, C. Ebeling, et al., "BRENDA, the enzyme database: updates and major new developments," *Nucleic Acids Research*, vol. 32, Database issue, pp. D431–D433, 2004.

[5] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, "KEGG: Kyoto Encyclopedia of Genes and Genomes," *Nucleic Acids Research*, vol. 27, no. 1, pp. 29–34, 1999.

[6] A. Ruepp, A. Zollner, D. Maier, et al., "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes," *Nucleic Acids Research*, vol. 32, no. 18, pp. 5539–5545, 2004.

[7] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, et al., "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 4, p. 41, 2003.

[8] S. Kumar and J. Dudley, "Bioinformatics software for biologists in the genomics era," *Bioinformatics*, vol. 23, no. 14, pp. 1713–1717, 2007.

[9] L. B. Koski, M. W. Gray, B. F. Lang, and G. Burger, "AutoFACT: an automatic functional annotation and classification tool," *BMC Bioinformatics*, vol. 6, p. 151, 2005.

[10] F. M. McCarthy, S. M. Bridges, N. Wang, et al., "AgBase: a unified resource for functional analysis in agriculture," *Nucleic Acids Research*, vol. 35, Database issue, pp. D599–D603, 2007.

[11] F. Chalmel, A. Lardenois, J. D. Thompson, et al., "GOAnno: GO annotation based on multiple alignment," *Bioinformatics*, vol. 21, no. 9, pp. 2095–2096, 2005.

[12] D. Groth, H. Lehrach, and S. Hennig, "GOblet: a platform for Gene Ontology annotation of anonymous sequence data," *Nucleic Acids Research*, vol. 32, pp. W313–W317, 2004.

[13] S. Khan, G. Situ, K. Decker, and C. J. Schmidt, "GoFigure: automated Gene Ontology annotation," *Bioinformatics*, vol. 19, no. 18, pp. 2484–2485, 2003.

[14] A. Vinayagam, C. del Val, F. Schubert, et al., "GOPET: a tool for automated predictions of Gene Ontology terms," *BMC Bioinformatics*, vol. 7, p. 161, 2006.

[15] D. M. A. Martin, M. Berriman, and G. J. Barton, "GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes," *BMC Bioinformatics*, vol. 5, p. 178, 2004.

[16] E. M. Zdobnov and R. Apweiler, "InterProScan—an integration platform for the signature-recognition methods in InterPro," *Bioinformatics*, vol. 17, no. 9, pp. 847–848, 2001.

[17] I. Friedberg, T. Harder, and A. Godzik, "JAFA: a protein function annotation meta-server," *Nucleic Acids Research*, vol. 34, Web Server issue, pp. W379–W381, 2006.

[18] G. Zehetner, "OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3799–3803, 2003.

[19] T. Hawkins, S. Luban, and D. Kihara, "Enhanced automated function prediction using distantly related sequences and contextual association by PFP," *Protein Science*, vol. 15, no. 6, pp. 1550–1556, 2006.

[20] A. Conesa, S. Götz, J. M. García-Gómez, J. Terol, M. Talón, and M. Robles, "Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research," *Bioinformatics*, vol. 21, no. 18, pp. 3674–3676, 2005.

[21] J. Terol, A. Conesa, J. M. Colmenero, et al., "Analysis of 13000 unique *Citrus* clusters associated with fruit quality, production and salinity tolerance," *BMC Genomics*, vol. 8, p. 31, 2007.

[22] J. A. Vizcaíno, F. J. González, M. B. Suárez, et al., "Generation, annotation and analysis of ESTs from *Trichoderma harzianum* CECT 2413," *BMC Genomics*, vol. 7, p. 193, 2006.

[23] A. C. Faria-Campos, F. S. Moratelli, I. K. Mendes, et al., "Production of full-length cDNA sequences by sequencing and analysis of expressed sequence tags from *Schistosoma mansoni*," *Memorias do Instituto Oswaldo Cruz*, vol. 101, supplement 1, pp. 161–165, 2006.

[24] D. S. Durica, D. Kupfer, F. Najar, et al., "EST library sequencing of genes expressed during early limb regeneration in the fiddler crab and transcriptional responses to ecdysteroid exposure in limb bud explants," *Integrative and Comparative Biology*, vol. 46, no. 6, pp. 948–964, 2006.

[25] T. D. Williams, A. M. Diab, S. G. George, et al., "Development of the GENIPOL European flounder (*Platichthys flesus*) microarray and determination of temporal transcriptional responses to cadmium at low dose," *Environmental Science and Technology*, vol. 40, no. 20, pp. 6479–6488, 2006.

[26] M. Gandía, A. Conesa, G. Ancillo, et al., "Transcriptional response of *Citrus aurantifolia* to infection by *Citrus tristeza virus*," *Virology*, vol. 367, no. 2, pp. 298–306, 2007.

[27] A. Reyes-Prieto, J. D. Hackett, M. B. Soares, M. F. Bonaldo, and D. Bhattacharya, "Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions," *Current Biology*, vol. 16, no. 23, pp. 2320–2325, 2006.

[28] M. J. Nueda, A. Conesa, J. A. Westerhuis, et al., "Discovering gene expression patterns in time course microarray experiments by ANOVA-SCA," *Bioinformatics*, vol. 23, no. 14, pp. 1792–1800, 2007.

[29] T. D. Williams, A. M. Diab, S. G. George, V. Sabine, and J. K. Chipman, "Gene expression responses of European flounder (*Platichthys flesus*) to 17-β estradiol," *Toxicology Letters*, vol. 168, no. 3, pp. 236–248, 2007.

[30] J. Ma, D. J. Morrow, J. Fernandes, and V. Walbot, "Comparative profiling of the sense and antisense transcriptome of maize lines," *Genome Biology*, vol. 7, no. 3, p. R22, 2006.

[31] R. T. Nelson and R. Shoemaker, "Identification and analysis of gene families from the duplicated genome of soybean using EST sequences," *BMC Genomics*, vol. 7, p. 204, 2006.

[32] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[33] A. Labarga, F. Valentin, M. Anderson, and R. Lopez, "Web services at the European bioinformatics institute," *Nucleic Acids Research*, vol. 35, Web Server issue, no. supplement 2, pp. W6–W11, 2007.

[34] S. Myhre, H. Tveit, T. Mollestad, and A. Lægreid, "Additional Gene Ontology structure for improved biological reasoning," *Bioinformatics*, vol. 22, no. 16, pp. 2020–2027, 2006.

[35] E. Pietriga, "A toolkit for addressing HCI issues in visual language environments," in *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC '05)*, pp. 145–152, Dallas, Tex, USA, September 2005.

[36] N. Blüthgen, K. Brand, B. Cajavec, M. Swat, H. Herzel, and D. Beule, "Biological profiling of gene groups utilizing Gene Ontology," *Genome Informatics*, vol. 16, no. 1, pp. 106–115, 2005.

[37] S. Goetz, J. M. García-Gómez, J. Terol, et al., "High throughput functional annotation and data mining with the Blast2GO suite," submitted.

[38] F. Al-Shahrour, P. Minguez, J. Tárraga, et al., "BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments," *Nucleic Acids Research*, vol. 34, Web Server issue, pp. W472–W476, 2006.

[39] D. Montaner, J. Tárraga, J. Huerta-Cepas, et al., "Next station in microarray data analysis: GEPAS," *Nucleic Acids Research*, vol. 34, Web Server issue, pp. W486–W491, 2006.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 369601, 6 pages
doi:10.1155/2008/369601

Research Article

The Generation Challenge Programme Platform: Semantic Standards and Workbench for Crop Science

Richard Bruskiewich,[1] Martin Senger,[1] Guy Davenport,[2] Manuel Ruiz,[3] Mathieu Rouard,[4] Tom Hazekamp,[4] Masaru Takeya,[5] Koji Doi,[5] Kouji Satoh,[5] Marcos Costa,[6] Reinhard Simon,[7] Jayashree Balaji,[8] Akinnola Akintunde,[9] Ramil Mauleon,[1] Samart Wanchana,[1,10] Trushar Shah,[2] Mylah Anacleto,[1] Arllet Portugal,[1] Victor Jun Ulat,[1] Supat Thongjuea,[10] Kyle Braak,[2] Sebastian Ritter,[2] Alexis Dereeper,[3] Milko Skofic,[4] Edwin Rojas,[7] Natalia Martins,[6] Georgios Pappas,[6] Ryan Alamban,[1] Roque Almodiel,[1] Lord Hendrix Barboza,[1] Jeffrey Detras,[1] Kevin Manansala,[1] Michael Jonathan Mendoza,[1] Jeffrey Morales,[1] Barry Peralta,[1] Rowena Valerio,[1] Yi Zhang,[1] Sergio Gregorio,[1,11] Joseph Hermocilla,[1,11] Michael Echavez,[1,12] Jan Michael Yap,[1,12] Andrew Farmer,[13] Gary Schiltz,[13] Jennifer Lee,[14] Terry Casstevens,[15] Pankaj Jaiswal,[15] Ayton Meintjes,[16] Mark Wilkinson,[17] Benjamin Good,[18,19] James Wagner,[18,19] Jane Morris,[16] David Marshall,[14] Anthony Collins,[7] Shoshi Kikuchi,[5] Thomas Metz,[1] Graham McLaren,[1] and Theo van Hintum[20]

[1]*Crop Research Informatics Laboratory, International Rice Research Institute (IRRI), DAPO Box 7777, Metro Manila, Philippines*
[2]*Crop Research Informatics Laboratory, International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, DF, Mexico*
[3]*Centre International de Recherche Agronomique pour le Développement (CIRAD), Avenue Agropolis, 34398 Montpellier, Cedex 5, France*
[4]*Bioversity International, Via dei Tre Denari 472/a, 00057 Maccarese (Fiumicino), Rome, Italy*
[5]*National Institute for Agrobiological Sciences (NIAS), Kannondai 2-1-2, Tsukuba, Ibaraki 305-8602, Japan*
[6]*Empresa Brasileira de Pesquisa Agropecuaria (EMBRAPA), Parque Estação Biologia Final W5 Norte, 70770-900 Brasilia, DF, Brazil*
[7]*Centro Internacional de la Papa (CIP), Avenida La Molina 1895, La Molina, Apartado Postal 1558, Lima 12, Peru*
[8]*International Crops Research Institute for the Semi-Arid Tropics, Patancheru, Andhra Pradesh 502324, India*
[9]*International Center for Agricultural Research in the Dry Areas, P.O. Box 5466, Aleppo, Syria*
[10]*National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Phahonyothin Road, Klong 1, Klong Luang, Pathumthani 12120, Thailand*
[11]*Institute of Computer Science, College of Arts and Sciences, University of the Philippines, Los Baños, Laguna 4031, Philippines*
[12]*Department of Computer Science, University of the Philippines, Room 215, Melchor Hall, Diliman, Quezon City 1101, Philippines*
[13]*National Center for Genome Resources, 2935 Rodeo Park Drive East, Santa Fe, NM 87505, USA*
[14]*Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK*
[15]*Department of Plant Breeding, Cornell University, Ithaca, NY 14853, USA*
[16]*African Centre for Gene Technologies, P.O. Box 75011, Lynnwood Ridge 0040, South Africa*
[17]*Department of Medical Genetics, Faculty of Medicine, The University of British Columbia, Vancouver, BC, Canada V6T 1Z3*
[18]*School of Computing Science, Simon Fraser Universtiy, 8888 University Drive, Burnaby, BC, Canada V5A 1S6*
[19]*Bioinformatics Graduate Program, Genome Sciences Centre, BC Cancer Agency, 100-570 West 7th Avenue, Vancouver, BC, Canada V5Z 4S6*
[20]*Centre for Genetic Resources, The Netherlands (CGN), P.O. Box 16, 6700 AA Wageningen, The Netherlands*

Correspondence should be addressed to Richard Bruskiewich, r.bruskiewich@cgiar.org

Received 22 September 2007; Accepted 14 December 2007

Recommended by Chunguang Du

The Generation Challenge programme (GCP) is a global crop research consortium directed toward crop improvement through the application of comparative biology and genetic resources characterization to plant breeding. A key consortium research activity is the development of a GCP crop bioinformatics platform to support GCP research. This platform includes the following: (i) shared, public platform-independent domain models, ontology, and data formats to enable interoperability of data and analysis flows within the platform; (ii) web service and registry technologies to identify, share, and integrate information across diverse, globally dispersed data sources, as well as to access high-performance computational (HPC) facilities for computationally intensive,

high-throughput analyses of project data; (iii) platform-specific middleware reference implementations of the domain model integrating a suite of public (largely open-access/-source) databases and software tools into a workbench to facilitate biodiversity analysis, comparative analysis of crop genomic data, and plant breeding decision making.

1. INTRODUCTION

The fast-moving fields of comparative genomics, molecular breeding, and bioinformatics have the potential to bring new knowledge to bear on problems encountered by resource-poor farmers. These problems include abiotic stresses (such as drought and soil salinity) and biotic stresses (such as plant diseases and pests). The Generation Challenge Programme (GCP; http://www.generationcp.org/) aims to exploit advances in molecular biology to harness the rich global heritage of plant genetic resources and contribute to a new generation of stress-tolerant varieties that meet the needs of these farmers and the consumers of their crops. The GCP brings together three sets of partners: member agricultural research institutes of the Consultative Group on International Agricultural Research (CGIAR; http://www.cgiar.org/), advanced research institutes in developed countries, and national agricultural research and extension systems in developing countries, to undertake a long-term program of globally integrated scientific research, capacity building, and delivery of products for the above goal.

Central to GCP activities is the development of an integrated platform of molecular biology and bioinformatics tools to be applied to the research objectives of the GCP. The resulting platform is also intended to be a "global public good" to be made freely available to all crop researchers and breeders around the world, thus enabling agricultural scientists, particularly in developing countries, to more readily apply information about elite genetic stocks, genomic knowledge, and new breeding technologies that are becoming available to their local breeding programmes.

The goal of this GCP crop informatics platform is to provide solutions for priority end-user needs for biodiversity analysis, comparative analysis of crop genomic data, and plant breeding decision making. Development of the platform is driven by the following observations:

(i) GCP partners (and the international crop research community in general) are globally distributed, each research team having relatively large datasets to share and integrated datasets that reside in diverse online, but locally curated databases;

(ii) GCP research covers a diversity of crop species;

(iii) GCP research spans a wide range of scientific data types, including germplasm, genomic, phenotypic, as well as crop physiological and geographic information, this constellation of data types is evolving with time as new experimental technologies are created;

(iv) GCP scientists (and crop scientists in general) need to apply a wide range of analytical tools already used by their research communities; they also need new tools to meet new or evolving needs; integration of such tools to interoperate with one another is a nontrivial task.

A GCP crop information platform is being developed to better meet these challenges by managing genetic resources, genomics, and crop information using the following components:

(i) shared public platform-independent set of scientific domain models, ontology, and data templates to cross-link all data types and analysis processes within the platform;

(ii) GCP domain model-constrained web service and registry technologies to identify, share, and manage the analysis of information, as well as to integrate it across a network of diverse globally dispersed data sources connected to the Internet;

(iii) reference implementations of platform-specific middleware using the GCP domain model;

(iv) a suite of open-source software tools (adopted or newly developed) integrated into a workbench and accessing web-connected data sources. Included in this suite is software to provide enhanced access to high-performance computational (HPC) grid facilities enabling computationally intensive and/or high-throughput analyses of project data.

This paper will survey progress on some of the central components of the platform, with a special emphasis on the domain model, a reference Java middleware implementation, and Internet protocol aspects of the project.

2. MATERIALS AND METHODS

2.1. GCP domain model

To cope with the scope, diversity, and dispersion of crop information, GCP researchers formulated a vision to specify a consensus blueprint of a scientific domain model and associated ontology. The resulting models and ontology allow a "model-driven architecture" for the development of GCP software and network protocols [1].

The domain model is documented in Unified Modeling Language (UML). Computable versions of the UML model are archived in the DemeterUML folder of the "Pantheon" project in CropForge (http://cropforge.org/) software project repository. The UML diagrams themselves are indexed and published with supporting narratives on a project website (http://pantheon.generationcp.org/demeter). The bulk

of the models are specified with the UML <<interface>> stereotype.

At the heart of the domain model are generic core model interfaces from which other specific scientific model interfaces are derived. This core model starts with the concept of simple identification of data objects in the system (using the *SimpleIdentifier* interface), which is extended by several more specific interfaces. The core includes a general concept of *Entity*, which serves as the superclass for most other interfaces describing major scientific concepts or data types in the system. The *Entity* interface documents generic metadata about objects in the system, including specific annotation of object characteristics using a rich *Feature* model. Other packages in the core models provide utility models for ontology, publication, and experimental study management.

Additional scientific models are derived as extensions of the core models. For example, the base interface classes of most specific major concepts or experimental objects in the scientific domain of discourse of the GCP, such as *Germplasm*, *Map*, or *GeneProduct*, directly extend the *Entity* model, adding subdomain-specific attributes as required. More lightweight concepts in the system extend simpler interfaces such as *Feature*.

For the elaboration of specific components of the core, as well as scientific domain models, the project generally adapts extant public domain models. For example, the *Germplasm* and *Study* subdomain models are derived from the data models of the open-source International Crop Information System (ICIS, http://www.icis.cgiar.org/; [2–4]). Aspects of the genotype (and associated genetic map and genomic sequence) models are influenced by public initiatives such as the Chado relational database schemata of the Generic Model Organism Database (GMOD) project [5]. The production-release GCP domain model is being validated based on feedback from project scientists and developers, who are striving to validate the model by practical application in data management and platform implementation.

A significant feature of the domain model is the reliance on extensible controlled vocabulary and ontology (CVO) to define the full semantics of specialized types, feature attributes, and annotation values of instances of the model classes. Where possible, the GCP is simply adopting existing CVO standards, such as from the gene ontology [6], plant ontology [7], and Microarray Gene Expression Data Society (MGED) ontology [8] consortia. Where no appropriate ontology has yet been formalized, new dictionaries of terms are being compiled in collaboration with GCP scientists. CVO dictionaries selected for the platform are being catalogued in a dedicated online database (at http://pantheon.generationcp.org/) with web browser and web service access. Each selected dictionary is assigned a GCP ontology index number to facilitate platform management of the ontology. Where an existing public ontology already has its own accession identifiers (e.g., GO identifiers for the GO CVO), these identifiers are propagated into the full GCP identifier for the corresponding CVO terms. However, newly specified CVO lacking such a number space are assigned *de novo* GCP accession identifiers.

2.2. GCP platform middleware

Since a March 2006 public review of the GCP domain model, the GCP informatics team has developed selected technology-specific GCP implementations of the model, primarily focusing on Java-based middleware specifying a Model-View-Controller (MVC) architecture (see Figure 1). Although the primary development stream of the project is focusing on a Java language implementation, the GCP domain model is a "platform-independent model" amenable to implementation with other computing languages and is, indeed, being used to guide some complementary work with languages such as Perl, Javascript, and PHP. The Java-based middleware was given the overall name "Pantheon" to account for the usage of various ancient agricultural gods (mostly agricultural, e.g., Demeter, Ceres, Belenus, Osiris) in the naming of the various layers and component parts of the code base. This code base is open source and managed under the Pantheon project in CropForge.

In addition to a Java implementation of the GCP domain model, a Java application programming interface (API) was specified to assist with and standardize software integration of components within the middleware architecture. These interfaces are collected into a core Java library called "PantheonBase" hosted as a module in the Ceres section of Pantheon (under Ceres/projects/Pantheonbase). PantheonBase includes a simple *DataSource* interface for read-only query retrieval of data from any source (local or distributed); a *DataConsumer* interface to guide integration and synchronization of applications and viewers wishing to use data extracted using the middleware; and finally, a *DataTransformer* interface to provide a framework for analysis and transformations (e.g., reformatting) of data. PantheonBase was deliberately designed to be essentially agnostic about the GCP domain model per se, for maximum flexibility and possible reuse with non-GCP-compliant data.

Additional support libraries are being provided within Ceres to support GCP domain model-driven *DataSource* development. In addition to core and support libraries, the Pantheon project provides a clearinghouse for platform and data-type-specific components. These components include adapters implementing the *DataSource* interface for specific data sources (archived in Osiris) for various crop databases at various GCP partner and external sites. Among others, current *DataSource* implementations include a wrapper for the ICIS and for GMOD schemata (Chado, Gbrowse). Other Pantheon components provide application support, including a search engine, data visualization, and web service provider implementations (in Belenus). Examples of the latter are support for NCGR ISYS [9], support for standalone applications based on Eclipse/RCP [10], and a web-based GCP domain-model-compliant web-based search engine (Koios).

2.3. GCP network protocols

The GCP domain model is also being applied to platform-specific implementation of a GCP network based on Internet bioinformatics data exchange protocols such as BioMOBY

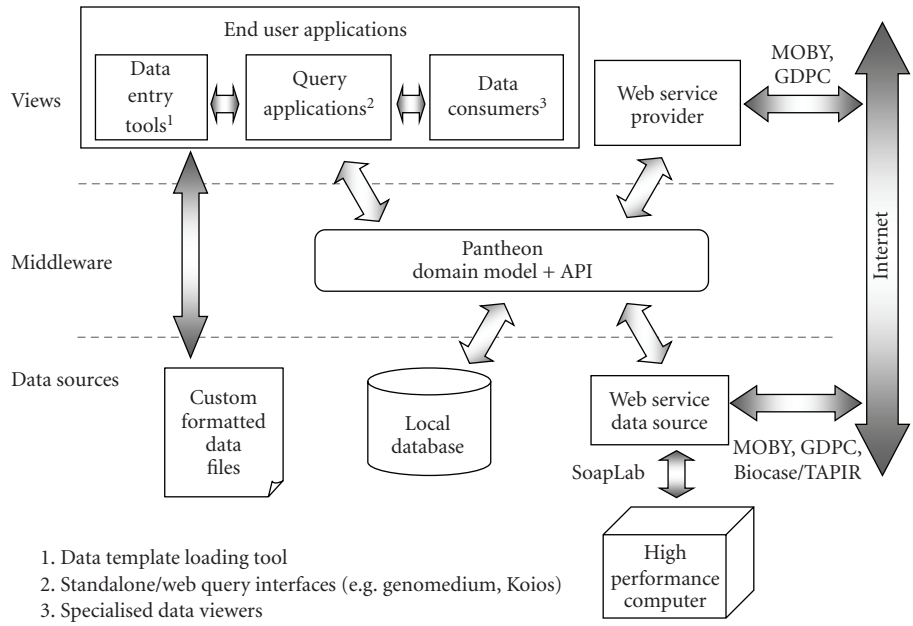


Figure 1

[11], SoapLab [12], SSWAP [13], and Tapir [14]. In this paper, in the interest of brevity, we will discuss only BioMOBY, arepresentative protocol being used in the GCP network.

For BioMOBY, data types were designed using GCP domain model semantics. Although generally faithful in translating the semantics of the Demeter UML specification of the domain model (i.e., the *SimpleIdentifier* interface is represented as a *GCP_SimpleIdentifier* data type), the GCP BioMOBY data types simplify the data representation as a concession to BioMOBY design constraints and to web service performance.

One key example of this is the extensive substitution of *GCP_SimpleIdentifier* objects, instead of fully detailed data objects, at the end of model-to-model association edges found in the Demeter model. The rationale for this is the expectation that, in most cases, web services can apply a concept of "lazy loading" of data-type components, in which one identifies what objects might be embedded in a parent object, but does not necessarily retrieve their details until the user needs them (as a separate web service accepting a *GCP_SimpleIdentifier* of the object but returning the fully populated complex object of the specified type).

UML diagrams with supporting explanatory narration for these GCP-specific BioMOBY data types are published on the Pantheon website (http://pantheon.generationcp.org/moby), which is complemented by a website documenting GCP BioMOBY implementation details (http://moby.generationcp.org/). Supporting the BioMOBY protocol in Pantheon are a series of Pantheon modules for interconversion between GCP MOBY data types and Demeter-compliant Java objects, for web service provider implementation, and for a MOBY client *DataSource* adapter to communicate with GCP-compliant web service providers.

Using GCP model-constrained BioMOBY data types (all prefixed with "GCP_" in their name in the MOBY central registry), various GCP teams are deploying GCP-compliant web services from a common proposed list of documented web service use cases. Concurrently, the MOBY client *DataSource* adapter is being elaborated to communicate with these web services and import remote data into local "workbench" instances of the GCP platform.

2.4. Additional tools integrated into the GCP platform

The GCP domain model and associated platform middleware is not an end in itself. Rather, the goal of these informatics products is to serve as a semantically and operationally rich scaffold for the integration of both local and remote (Internet-connected) bioinformatics data resources and analysis tools.

In addition to data sources and tools already mentioned above, additional open-source third-party analysis tools already coded using Java, but agnostic concerning the GCP framework are being connected to the platform through targeted software engineering. To this end, GCP developers are connecting several public open-source applications by writing suitable *DataSource* adapters, *DataConsumer*, or *DataTransformer* integration code. These include Java software hosted by GMOD such as the Apollo genome browser [15], tools forming part of the Genomic Diversity and Phenotype Connection (GDPC) protocol such as Tassel [16], and tools such as TIGR Multiple Experiment Viewer [17] for microarray analysis, the Comparative Map and Trait Viewer [18] connected to the NCGR ISYS framework [9], the Cytoscape network visualization tool [20], and the MAXD microarray system [19].

3. RESULTS AND DISCUSSION

The GCP consortium was formally established in 2003. The first meeting of the bioinformatics and crop informatics development team of the GCP, designated as Subprogramme 4, was hosted in Rome, in February 2004. The general user needs and project goals were coarsely mapped out at this meeting, with some considerable differences in opinion voiced at how to construct the required informatics framework for the GCP. In May 2004, a smaller team of software experts met in Mexico to discuss project management, identify key user needs and platform requirements, and make some initial progress in the design of the system. Key decisions at this latter meeting were the adoption of the "model-driven architecture" paradigm for system development and to embrace web services as a key technology for global integration of systems. Numerous development meetings have been convened annually since these initial meetings to further refine and advance the design and implementation of the platform.

In particular, a milestone review of the GCP domain model and initial software systems using the model was held in Pretoria, South Africa in March 2006. Since that time, a number of early release versions of software systems based on GCP platform technology have become available, generally documented at http://pantheon.generationcp.org/ and publicly downloadable from various CropForge projects.A special "communications" project for GCP-specific projects is also available on CropForge at the http://cropforge.org/projects/gcpcomm to further inform prospective users on the variety of such GCP software tools now available, and provide a venue for user discussions and feedback about the tools.

3.1. So, what can I do with the GCP platform?

The vision of the platform development team of the bioinformatics and crop informatics subprogramme of the GCP is to establish a truly easy to use but extensible workbench providing interoperability and enhanced data access across all GCP partner sites and, later, across the global crop research community. As indicated above, the GCP domain model has a scope of data type coverage that spans most of the pertinent scientific data types found in crop research from upstream laboratory experiments through germplasm manipulations, in a georeferenced characterized field setting. The diversity of potential data sources and analysis tools is similarly large. What the platform facilitates is transparent data flows between such data sources and tools, whether from locally administered databases or remote Internet-connection resources.

In this light, a number of practical "use cases" may be described in general terms, as a series of data manipulation steps, to highlight some of the anticipated usage of the platform. As an indication of the data retrieval and analysis scope of the GCP platform, we describe a general integrative use case here below, in terms of a series of defined steps.

General GCP platform analysis use case for crop improvement

(1) Retrieve the list of all genetic maps that include a quantitative trait locus (QTL) for a specified trait.
(2) Retrieve selected maps in the list, from a project database or source file containing such maps.
(3) Load this into a suitable mapping tool (e.g., the comparative map and trait visualization tool, CMTV).
(4) Extract the pairs of flanking markers for the QTL.
(5) From a second (crop) database, retrieve the list of all germplasm that have been genotyped with these flanking markers.
(6) Retrieve all the pertinent passport, genotype, and phenotype information about the germplasm in the list.
(7) In parallel to the steps (5) and (6), if available, retrieve any gene locus candidates within (genetic/physical/sequence) map intervals which are defined by flanking markers which are molecular sequence based.
(8) Retrieve gene functional information about the gene loci compiled in step (7).
(9) Retrieve the alleles of "interesting" genes from (8), in the list of germplasm identified in step (5).
(10) Plot germplasm passport, genotype, and phenotype information on geographical information maps.
(11) Retrieve information about the environmental characteristics of the geographical regions identified in step (10).
(12) Identify germplasm, for further detailed evaluation, which appears to be adapted to target environments, which have promising phenotypic values identified in step (6) and which contains target alleles of gene loci identified in step (9).
(13) Identify genotyping (marker) systems potentially available from step (9), for marker assisted selected transfer of target traits from identified germplasm to additional germplasm targets.

4. CONCLUSIONS

The vision of the platform development team of the bioinformatics and crop informatics subprogramme of the GCP is to establish a state-of-art but truly easy-to-use and extensible open-source workbench providing interoperability and enhanced data access across all GCP partner sites and, by extension, the global crop research community.

Although several attempts have been made in the past to build such globally integrative bioinformatics systems, few have the global distribution of partners, scope of crop research, diversity of data types, and magnitude of datasets in comparison to the GCP consortium, nor do they have the long-term project perspective of 10 years. In addition, the GCP platform is specifically targeted to bioinformatics for developing world crop research, in contrast to biomedical research, and also strives to integrate databases from many plants and crops less well represented by well-funded model organisms and crops.

In these respects, the GCP platform effort represents an extremely ambitious but very useful global public good

resource for crop research. It is still conceded to be, in several respects, an incomplete evolving product, one with many rough edges and incompletely met end-user needs; however, the open-source and public nature of the project provides a credible venue for wide participation of interested developers and prospective end users in the future evolution and deployment of the platform.

ACKNOWLEDGMENTS

This work is funded through the Generation Challenge Programme (http://www.generationcp.org/), a consortium funded by several international donors of the Consultative Group on International Agricultural Research (CGIAR; http://www.cgiar.org/). The GCP domain model and platform development team gratefully acknowledges the technical contributions of other scientists at various GCP-funded workshops, in particular, the following individuals: Brigitte Courtois (CIRAD, France); Marco Bink (Wageningen University, The Netherlands); Michel Eduardo Beleza Yamagishi (EMBRAPA, Brazil); Hei Leung and Ken McNally (IRRI, Philippines); and Marilyn Warburton (CIMMYT). Theo van Hintum is the project leader for the GCP Subprogramme 4 on Crop Informatics. The Crop Research Informatics Laboratory is a single operational unit spanning IRRI and CIMMYT, as part of the IRRI-CIMMYT Alliance. Availability: see http://pantheon.generationcp.org/.

REFERENCES

[1] R. Bruskiewich, G. Davenport, T. Hazekamp, et al., "Generation challenge programme (GCP): standards for crop data," *OMICS*, vol. 10, no. 2, pp. 215–219, 2006.

[2] P. N. Fox and B. Skovmand, "The international crop information system (ICIS)—connects genebank to breeder to farmer's field," in *Plant Adaptation and Crop Improvement*, M. Cooper and G. L. Hammer, Eds., pp. 317–326, CAB International, Wallingford, UK, 1996.

[3] R. Bruskiewich, A. B. Cosico, W. Eusebio, et al., "Linking genotype to phenotype: the international rice information system (IRIS)," *Bioinformatics*, vol. 19, supplement 1, pp. i63–i65, 2003.

[4] C. G. McLaren, R. Bruskiewich, A. M. Portugal, and A. B. Cosico, "The international rice information system. A platform for meta-analysis of rice crop data," *Plant Physiology*, vol. 139, no. 2, pp. 637–642, 2005.

[5] http://www.gmod.org/, September 2007.

[6] http://www.geneontology.org/, September 2007.

[7] http://www.plantontology.org/, September 2007.

[8] http://www.mged.org/, September 2007.

[9] A. Siepel, A. Farmer, A. Tolopko, et al., "ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources," *Bioinformatics*, vol. 17, no. 1, pp. 83–94, 2001.

[10] http://wiki.eclipse.org/index.php/Rich_Client_Platform.

[11] M. Wilkinson, H. Schoof, R. Ernst, and D. Haase, "BioMOBY successfully integrates distributed heterogeneous bioinformatics web services. The PlaNet exemplar case," *Plant Physiology*, vol. 138, no. 1, pp. 5–17, 2005.

[12] M. Senger, P. Rice, and T. Oinn, "Soaplab—a unified Sesame door to analysis tools," in *Proceedings of the 2nd UK E-Science, All Hands Meeting*, S. J. Cox, Ed., pp. 509–513, Nottingham, UK, September 2003.

[13] http://www.sswap.info/, September 2007.

[14] http://www.tdwg.org/activities/tapir, September 2007.

[15] S. E. Lewis, S. M. Searle, N. Harris, et al., "Apollo: a sequence annotation editor," *Genome Biology*, vol. 3, no. 12: research0082, 2002.

[16] T. M. Casstevens and E. S. Buckler, "GDPC: connecting researchers with multiple integrated data sources," *Bioinformatics*, vol. 20, no. 16, pp. 2839–2840, 2004.

[17] A. I. Saeed, N. K. Bhagabati, J. C. Braisted, et al., "TM4 microarray software suite," *Methods in Enzymology*, vol. 411, pp. 134–193, 2006.

[18] M. C. Sawkins, A. D. Farmer, D. Hoisington, et al., "Comparative map and trait viewer (CMTV): an integrated bioinformatic tool to construct consensus maps and compare QTL and functional genomics data across genomes and experiments," *Plant Molecular Biology*, vol. 56, no. 3, pp. 465–480, 2004.

[19] D. Hancock, M. Wilson, G. Velarde, et al., "maxdLoad2 and maxdBrowse: standards-compliant tools for microarray experimental annotation, data management and dissemination," *BMC Bioinformatics*, vol. 6, p. 264, 2005.

[20] http://www.cytoscape.org/.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 412696, 9 pages
doi:10.1155/2008/412696

Research Article

SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation

Luciano Carlos da Maia,[1] Dario Abel Palmieri,[2] Velci Queiroz de Souza,[1] Mauricio Marini Kopp,[1] Fernando Irajá Félix de Carvalho,[1] and Antonio Costa de Oliveira[1]

[1] *Plant Genomics and Breeding Laboratory, Eliseu Maciel School of Agronomy, Federal University of Pelotas, Pelotas, RS 96.001-970, Brazil*
[2] *Laboratory for Environmental Studies, Catholic University of Salvador, Salvador, BA, 40.220-140, Brazil*

Correspondence should be addressed to Antonio Costa de Oliveira, acostol@terra.com.br

Received 5 October 2007; Revised 29 January 2008; Accepted 20 May 2008

Recommended by Chunguang Du

Microsatellites or SSRs (*simple sequence repeats*) are ubiquitous short tandem duplications occurring in eukaryotic organisms. These sequences are among the best marker technologies applied in plant genetics and breeding. The abundant genomic, BAC, and EST sequences available in databases allow the survey regarding presence and location of SSR loci. Additional information concerning primer sequences is also the target of plant geneticists and breeders. In this paper, we describe a utility that integrates SSR searches, frequency of occurrence of motifs and arrangements, primer design, and PCR simulation against other databases. This simulation allows the performance of global alignments and identity and homology searches between different amplified sequences, that is, amplicons. In order to validate the tool functions, SSR discovery searches were performed in a database containing 28 469 nonredundant rice cDNA sequences.

1. INTRODUCTION

Microsatellites or SSRs (*simple sequence repeats*) are sequences in which one or few bases are tandemly repeated for varying numbers of times [1]. Variations in SSR regions originate mostly from errors during the replication process, frequently DNA polymerase slippage, generating insertion or deletion of base pairs, resulting, respectively, in larger or smaller regions [2, 3]. SSR assessments in the human genome have shown that many diseases are caused by mutation in these sequences [4].

SSRs can be found in different regions of genes, that is, coding sequences, untranslated sequences (5′-UTR and 3′-UTR), and introns, where the expansions and/or contractions can lead to gene gain or loss of function [5]. Also, there are evidences that genomic distribution of SSRs is related to chromatin organization, recombination, and DNA repair. SSRs are found throughout the genome, in both protein-coding and noncoding regions. Genome fractions as low as 0.85% (*Arabidopsis thaliana*), 0.37% (*Zea mays*), 0.21% (*Caenorhabtis elegans*), 0.30% (*Sacharomyces cerevisae*) and as high as 3.0% (*Homo sapiens*) and 3.21% (*Fugu rubripes*) have been found. Some bias for defined genomic locations has also been reported [6, 7]. This class of markers is broadly applied in genetics and plant breeding, due to its reproducibility, multiallelic, codominant nature, and genomic abundance. Its use for integrating genetic maps, physical mapping, and anchoring gives geneticists and plant breeders a pathway to link genotype and phenotype variations [8].

The protocols for isolating SSR loci for a new species were always very labor-intensive. Currently, with the accumulation of biological data originating from whole genome sequence initiatives, the use of bioinformatics tools helps to maximize the identification of these sequences and consequently, the efficiency in the number of generated markers [9].

The first in silico studies of SSRs were developed using FASTA [10] and BLAST [11] packages. Later, more specific algorithms, such as SPUTINICK [12], REPEATMASKER

[13], TRF-*Tandem Repeat Find* [14], TROLL [15], MISA [16] and SSRIT (*Simple Sequence Repeat Tool*) [17], were obtained [9].

SSR detection is generally followed by the use of another program for primer design, to be anchored on flanking sequences. Also, in some applications, a third step using e-PCR [18] is added, with the goal of verifying primer redundancy. The sequential use of a number of software is often called a pipeline. Building such a pipeline can be a very difficult task for research groups not familiar with programming tools.

In the present work, a computing tool with an interface for Windowsusers was developed, called SSR Locator. The application integrates the following functions: (i) detection and characterization of SSRs and minisatellite motifs between 1 and 10 base pairs; (ii) primer design for each locus found; (iii) simulation of PCR (polymerase chain reaction), amplifying fragments with different primer pairs from a given set of fasta files; (iv) global alignment between amplicons generated by the same primer pair; and (v) estimation of global alignment scores and identities between amplicons, generating information on primer specificity and redundancy. The described tool is publicly available at the site http://www.ufpel.edu.br/~lmaia.faem.

2. MATERIAL AND METHODS

2.1. Algorithms

The algorithms used for the searches, alignment, and homology estimates are described separately.

2.2. SSR search

The algorithm used for perfect and imperfect micro-/minisatellite searches was written in Perl and consists of the generation of a matrix that mixes A(adenine), T(thymine), C(cytosine), and G(guanine) in all possible composite arrangements between 1 and 10 nucleotides. The script instructions perform readings on fasta files, searching all possible arrangements in each database sequence.

Several instructions in the algorithm used in SSRLocator resemble those from MISA [16] and SSRIT [17]. However, additional instructions have been inserted in SSRLocator's code. Instead of allowing the overlap of a few nucleotides when two SSRs are adjacent to each other and one of them is shorter than the minimum size for a given class as found in MISA and SSRIT, a module written in Delphi language records the data and eliminates such overlaps.

The SSR Locator software contains windows focused on the selection and configuration of SSR and minisatellite types (mono- to 10-mers) and a minimum number of repeats for each one of the selected types. The algorithm calls a perfect repeat when one locus is present with adjacent loci at an up or downstream distance higher than 100 bp.

The algorithm calls an imperfect repeat when the same motif is present on both sides of a fragment containing up to 5 base pairs.

The algorithm identifies a composite locus when two or more adjacent loci were found at distances between 6 and 100 bp [16].

In this study, only "Class I" ($\geq$20 bp) repeats are shown. These repeats have been described as the most efficient loci for use as molecular markers [17]. The software SSRLocator was configured to locate a minimum of 20 bp SSRs: monomers(x20), 2-mers(x10), 3-mers(x7), 4-mers(x5), 5-mers(x4), 6-mers(x4), and minissatellites: 7-mers(x3), 8-mers(x3), 9-mers(x3), and 10-mers(x3).

In order to validate the efficiency of SSRLocator in finding SSRs and minisatellites, the same database was analyzed withMISA and SSRIT, using the same parameters for minimum number of repeats.

2.3. Primer design

An algorithm written in Delphi language performs calls to Primer3 [19], which execute primer designs. These results are fed to a module that performs Virtual-PCRs and allocates individual identification, forward and reverse primer sequences, and a sequence fragment corresponding to the region flanked by the primers (original amplicon) to each SSR locus. A window allows the selection of Primer3 parameters, such as range of primer and amplicon sizes, as well as optimum primer size, ranges of melting temperature (TM) (minimum, maximum, and optimum) and GC content (minimum and optimum). For primer searches, the software automatically looks for five base pair distances from both SSR (5′ and 3′) flanking sites. In this study, the following parameters were used: amplicon size between 100 and 280 bp; minimum, optimum, and maximum annealing temperature (TM) of 45, 50, and 55, respectively, minimum, optimum, and maximum primer size of 15, 20, and 25 bp, respectively.

2.4. Virtual-PCR

The module used to simulate a PCR reaction was written in Delphi. The algorithm consists in reading the file generated by the previous module (SSR locus, forward and reverse primers, and original amplicon), followed by a search of sequences containing primer annealing sites. When annealing sites are found for the two primers, the flanked region and the primer sequences are copied to a new variable called "paralog amplicon."

2.5. Global alignment

For the global alignment between paralog and original amplicon sequences and score calculations (match, mismatch, gaps), a routine was written in Delphi language using the algorithms of Needleman and Wunsch (1970) [20] and Smith and Waterman (1981) [21]. Also, in the same module, amplicon identities were calculated according to Waterman (1994) [22] and Vingron and Waterman (1994) [23].

2.6. Implementation

The strategy of creating a two-language hybrid program was established as a function of: (i) the higher speed achieved by

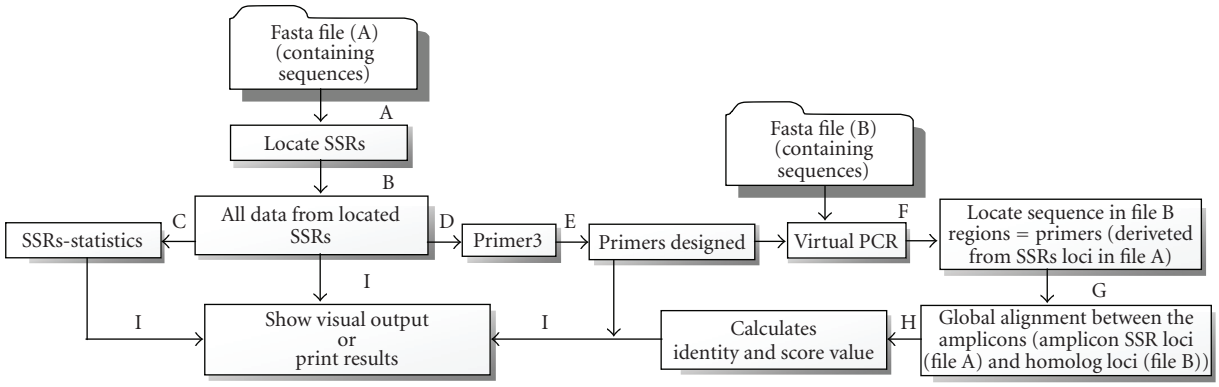


Figure 1: Flow-chart showing the functional structure of SSR Locator. (A) Perl script to search SSRs; (B) text file where information from detected SSRs is stored; (C) module for the statistical calculations for SSR motif occurrence; (D) module that formats text files into standard Primer3 input files; (E) running of Primer3; (F) module for running Virtual-PCR (using a second sequence file as a template); (G) module performing global alignment between homologous amplicons; (H) identity and alignment score calculations between homologous amplicons; and (I) file containing SSR, primer, homologous amplicons, identity, and score information.

handling large text files with Perl as compared to Delphi,and (ii) the better fitness of Perl for generating combinatory strings to be located. The Perl module was transformed into an executable file, making unnecessary to install Perl libraries during program installing. The graphic interface built, integrating input and output windows to the Windows operational system, was obtained using the Suite Turbo Delphi, where a menu system executes calls for each of the previously described modules.

2.7. Sequences for analysis

A total of 28 469 rice (*Oryza sativa* ssp. *japonica*- cv. Nipponbare) nonredundant full length nonredundant cDNA sequences, sequenced by *The Rice Full-Length cDNA Consortium*, mapped on the databases derived from the sequencing of *japonica* (*japonica* draft genome, BAC/PAC clones—IRGSP) and *indica* (*indica* draft genome) subspecies [24] were used for the analyses. These sequences are deposited in NCBI as two groups, the first comprising accesses from AK058203 to AK074028, and the second comprising accesses from AK98843 to AK111488. All these sequences can be also found in KOME (Knowledge-based Oryza Molecular Biological Encyclopedia).

A flow chart representing the different steps performed by the software is shown in Figure 1.

3. RESULTS

3.1. Program validation

A total of 3907 micro- and minisatellites were detected by SSRLocator in the 28 469 analyzed cDNA sequences. The same database searched with MISA and SSRIT presented 3913 and 3917 loci, respectively. The mono-, 4-mer, 6-mer, 7-mer, 8-mer, 9-mer, and 10-mer repeats were identical for the three programs. In the case of 2-mer repeats, 594 elements were detected by SSRLocator and 596 elements were detected by MISA and SSRIT. 3-mer repeats were differently scored by SSRLocator (1990) and the other two (1994) algorithms. For 5-mer repeats, SSRLocator and MISA found the same number of repeats (426), while SSRIT (430) found a different value.

3.2. Overall distribution of SSR types

The results obtained with SSRLocator indicate that out of 28 469 cDNA sequences, 3765 (13.22%) presented one or more micro-/minisatellite loci. In other studies, microsatellites were found in the following proportions in ESTs: 3% in arabidopsis [25], 4% in rosaceae [26], 8.11% in barley [16], 2.9% in sugarcane [27], and values ranging between 6–11% [28] and 1.5–4.7% [29] for cereals in general (maize, barley, rye, sorghum, rice, and wheat).

Considering the 3765 *fl*-cDNA sequences, in 3632 (92.96%) only a single micro-/minisatellitelocus was detected. In 125 sequences, two loci were detected, in seven sequences three lociandonly one sequence had four loci, adding up to 3907 occurrences. Among the types analyzed, SSRs (mono to 6-mer repeats) and minisatellites (7- to 10-mer repeats) comprised 96.98% and 4.12% of detected loci, respectively.

The distribution of occurrences detected by SSRLocator was consisted of 138 monomers, 594 2-mers, 1990 3-mers, 251 4-mers, 426 5-mers, 390 6-mers, 82 7-mers, 6 8-mers, 25 9-mers, and 5 10-mers, corresponding to rates of 3.53%, 15.20%, 50.93%, 6.42%, 10.90%, 9.98%, 2.10%, 0.15%, 0.64%, and 0.13%, respectively (see Table 1).

For the remaining SSRs, average percentage values have been reported as between 17 and 40% for 2-mer, 54–78% for 3-mer, 2.6–6.6% for 4-mer, 0.4–1.3% for 5-mer, and less than 1% for 6-mer repeats [28] and 26.5% for 2-mer, 65.4% 3-mer, 6.8% 4-mer, 0.77% 5-mer, and 0.45% for 6-mer repeats [30] for barley, maize, wheat, sorghum, rye, and rice, respectively. In nonredundant transcripts from the TIGR database, 15.6% 2-mer, 61.6% 3-mer, 8.5% 4-mer, and 14.4% 5-mer repeats were found in rice [31].

Table 1: Distribution of SSR/minisatellite motifs according to the number of repeats.

Repeats	Mono-	(%)	2-mer	(%)	3-mer	(%)	4-mer	(%)	5-mer	(%)	6-mer	(%)	7-mer	(%)	8-mer	(%)	9-mer	(%)	10-mer	(%)	Total	(%)
3	**0**	—	**0**	—	**0**	—	**0**	—	**0**	—	**0**	—	**78**	95.12	**6**	100	**24**	96	**5**	100	**113**	2.89
4	**0**	—	**0**	—	**0**	—	**0**	—	**348**	81.69	**323**	82.82	**4**	4.88	**0**	0	**1**	4	**0**	0	**676**	17.30
5	**0**	—	**0**	—	**0**	—	**181**	72.11	**69**	16.20	**45**	11.54	**0**	0	**0**	0	**0**	0	**0**	0	**295**	7.55
6	**0**	—	**0**	—	**0**	—	**41**	16.33	**7**	1.64	**13**	3.33	**0**	0	**0**	0	**0**	0	**0**	0	**61**	1.56
7	**0**	—	**0**	—	**1220**	61.31	**9**	3.59	**0**	0	**5**	1.28	**0**	0	**0**	0	**0**	0	**0**	0	**1234**	31.58
8	**0**	—	**0**	—	**441**	22.16	**9**	3.59	**1**	0.23	**1**	0.26	**0**	0	**0**	0	**0**	0	**0**	0	**452**	11.57
9	**0**	—	**0**	—	**173**	8.69	**4**	1.59	**0**	0	**1**	0.26	**0**	0	**0**	0	**0**	0	**0**	0	**178**	4.56
10	**0**	—	**125**	21.04	**68**	3.42	**1**	0.40	**0**	0	**2**	0.51	**0**	0	**0**	0	**0**	0	**0**	0	**196**	5.02
11	**0**	—	**82**	13.80	**32**	1.61	**3**	1.20	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**117**	2.99
12	**0**	—	**76**	12.79	**18**	0.90	**1**	0.40	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**95**	2.43
13	**0**	—	**71**	11.95	**5**	0.25	**1**	0.40	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**77**	1.97
14	**0**	—	**39**	6.57	**2**	0.10	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**41**	1.05
15	**0**	—	**44**	7.41	**5**	0.25	**0**	0	**1**	0.23	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**50**	1.28
16	**0**	—	**30**	5.05	**2**	0.10	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**32**	0.82
17	**0**	—	**33**	5.56	**1**	0.05	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**34**	0.87
18	**0**	—	**15**	2.53	**3**	0.15	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**18**	0.46
19	**0**	—	**17**	2.86	**1**	0.05	**1**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**19**	0.49
20	**21**	15.22	**14**	2.36	**2**	0.10	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**37**	0.95
21	**19**	13.77	**8**	1.35	**2**	0.10	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**29**	0.74
22	**15**	10.87	**6**	1.01	**3**	0.15	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**24**	0.61
23	**8**	5.80	**7**	1.18	**3**	0.15	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**18**	0.46
24	**3**	2.17	**5**	0.84	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**8**	0.20
25	**9**	6.52	**5**	0.84	**1**	0.05	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**15**	0.38
26	**5**	3.62	**4**	0.67	**2**	0.10	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**11**	0.28
27	**3**	2.17	**1**	0.17	**1**	0.05	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**5**	0.13
28	**1**	0.72	**3**	0.51	**3**	0.15	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**7**	0.18
29	**4**	2.90	**0**	0	**1**	0.05	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**5**	0.13
30	**2**	1.45	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**2**	0.05
31	**9**	6.52	**2**	0.34	**1**	0.05	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**12**	0.31
32	**3**	2.17	**3**	0.51	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**6**	0.15
33	**3**	2.17	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**3**	0.08
34	**1**	0.72	**1**	0.17	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**2**	0.05
35	**6**	4.35	**1**	0.17	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**7**	0.18
36	**1**	0.72	**1**	0.17	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**2**	0.05
37	**1**	0.72	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**1**	0.03
38	**4**	2.90	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**4**	0.10
39	**0**	0	**1**	0.17	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**1**	0.03
40	**1**	0.72	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**1**	0.03
41	**1**	0.72	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**1**	0.03
42	**2**	1.45	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**2**	0.05
43	**2**	1.45	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**2**	0.05
44	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0.00
≥45	**14**	10.14	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**0**	0	**14**	0.36
Total	**138**		**594**		**1990**		**251**		**426**		**390**		**82**		**6**		**25**		**5**		**3907**	
(%)	**3.53**		**15.20**		**50.93**		**6.42**		**10.90**		**9.98**		**2.10**		**0.15**		**0.64**		**0.13**		**100.00**	

Table 2: Distribution of SSR/minisatellite repeats in the rice cDNA collection.

Motif		Ocur(1)	(%)(1)	Ocur(2)	(%)(2)	Total	(%) Group	(%) Overall
Mono-	A/T	**111**	88.80	**14**	11.20	**125**	90.58	3.20
	C/G	**10**	76.92	**3**	23.08	**13**	9.42	0.33
2-mer	AG/CT	**97**	36.06	**172**	63.94	**269**	45.29	6.89
	GA/TC	**143**	61.37	**90**	38.63	**233**	39.23	5.96
	CA/TG	**10**	35.71	**18**	64.29	**28**	4.71	0.72
	AT	**24**	100.00	**—**	—	**24**	4.04	0.61
	AC/GT	**6**	31.58	**13**	68.42	**19**	3.20	0.49
	TA	**19**	100.00	**—**	—	**19**	3.20	0.49
	CG	**2**	100.00	**—**	—	**2**	0.34	0.05
3-mer	CCG/CGG	**197**	53.68	**170**	46.32	**367**	18.44	9.39
	CGC/GCG	**218**	61.24	**138**	38.76	**356**	17.89	9.11
	GCC/GGC	**112**	53.08	**99**	46.92	**211**	10.60	5.40
	CTC/GAG	**73**	42.69	**98**	57.31	**171**	8.59	4.38
	AGG/CCT	**34**	30.91	**76**	69.09	**110**	5.53	2.82
	GGA/TCC	**60**	62.50	**36**	37.50	**96**	4.82	2.46
	CAG/CTG	**58**	76.32	**18**	23.68	**76**	3.82	1.95
	AAG/CTT	**34**	50.75	**33**	49.25	**67**	3.37	1.71
	CGA/TCG	**33**	54.10	**28**	45.90	**61**	3.07	1.56
	AGC/GCT	**36**	62.07	**22**	37.93	**58**	2.91	1.48
	GCA/TGC	**47**	83.93	**9**	16.07	**56**	2.81	1.43
	AGA/TCT	**33**	62.26	**20**	37.74	**53**	2.66	1.36
	CCA/TGG	**39**	75.00	**13**	25.00	**52**	2.61	1.33
	ACC/GGT	**22**	48.89	**23**	51.11	**45**	2.26	1.15
	GAA/TTC	**28**	63.64	**16**	36.36	**44**	2.21	1.13
	CAC/GTG	**28**	65.12	**15**	34.88	**43**	2.16	1.10
	GAC/GTC	**18**	54.55	**15**	45.45	**33**	1.66	0.84
	ACG/CGT	**11**	42.31	**15**	57.69	**26**	1.31	0.67
	ATC/GAT	**5**	45.45	**6**	54.55	**11**	0.55	0.28
	TCA/TGA	**5**	50.00	**5**	50.00	**10**	0.50	0.26
	CAA/TTG	**4**	50.00	**4**	50.00	**8**	0.40	0.20
	ACT/AGT	**3**	42.86	**4**	57.14	**7**	0.35	0.18
	TAA/TTA	**1**	14.29	**6**	85.71	**7**	0.35	0.18
	CTA/TAG	**4**	66.67	**2**	33.33	**6**	0.30	0.15
	AAT/ATT	**1**	20.00	**4**	80.00	**5**	0.25	0.13
	CAT/ATG	**4**	100.00	**0**	0	**4**	0.20	0.10
	AAC/GTT	**3**	75.00	**1**	25.00	**4**	0.20	0.10
	ATA/TAT	**1**	50.00	**1**	50.00	**2**	0.10	0.05
	GTA/TAC	**1**	100.00	**0**	0	**1**	0.05	0.03
4-mer	GATC	**18**	100.00	**0**	0	**18**	7.17	0.46
	ATTA/TAAT	**9**	52.94	**8**	47.06	**17**	6.77	0.44
	ATCG/CGAT	**3**	20.00	**12**	80.00	**15**	5.98	0.38
	CATC/GATG	**4**	40.00	**6**	60.00	**10**	3.98	0.26
	AGAA/TTCT	**2**	25.00	**6**	75.00	**8**	3.19	0.20
	GCTA/TAGC	**6**	75.00	**2**	25.00	**8**	3.19	0.20
	GATA/TATC	**1**	14.29	**6**	85.71	**7**	2.79	0.18
	GCGA/TCGC	**3**	42.86	**4**	57.14	**7**	2.79	0.18
	GCAC/GTGC	**2**	33.33	**4**	66.67	**6**	2.39	0.15
	AGGG/CCCT	**2**	33.33	**4**	66.67	**6**	2.39	0.15

Table 2: Continued.

Motif		Ocur(1)	(%)(1)	Ocur(2)	(%)(2)	Total	(%) Group	(%) Overall
5-mer	AGGAG/CTCCT	**3**	15.00	**17**	85.00	**20**	4.69	**0.51**
	CTCTC/GAGAG	**17**	89.47	**2**	10.53	**19**	4.46	**0.49**
	GAGGA/TCCTC	**9**	56.25	**7**	43.75	**16**	3.76	**0.41**
	CCTCC/GGAGG	**12**	80.00	**3**	20.00	**15**	3.52	**0.38**
	AGAGG/CCTCT	**4**	26.67	**11**	73.33	**15**	3.52	**0.38**
	GGAGA/TCTCC	**2**	18.18	**9**	81.82	**11**	2.58	**0.28**
	CTCGC/GCGAG	**7**	77.78	**2**	22.22	**9**	2.11	**0.23**
	AGCTA/TAGCT	**4**	44.44	**5**	55.56	**9**	2.11	**0.23**
	GAAAA/TTTTC	**2**	25.00	**6**	75.00	**8**	1.88	**0.20**
	AGGCG/CGCCT	**2**	25.00	**6**	75.00	**8**	1.88	**0.20**
6-mer	CGCCTC/GAGGCG	**12**	85.71	**2**	14.29	**14**	3.59	0.36
	CGGCGA/TCGCCG	**4**	28.57	**10**	71.43	**14**	3.59	0.36
	CCTCCG/CGGAGG	**9**	81.82	**2**	18.18	**11**	2.82	0.28
	AGGCGG/CCGCCT	**1**	10.00	**9**	90.00	**10**	2.56	0.26
	CCGTCG/CGACGG	**4**	44.44	**5**	55.56	**9**	2.31	0.23
	CGTCGC/GCGACG	**7**	77.78	**2**	22.22	**9**	2.31	0.23
	ACCGCC/GGCGGT	**1**	12.50	**7**	87.50	**8**	2.05	0.20
	CCACCG/CGGTGG	**6**	85.71	**1**	14.29	**7**	1.79	0.18
	GGCGGA/TCCGCC	**5**	71.43	**2**	28.57	**7**	1.79	0.18
	CTCCAT/ATGGAG	**6**	100.00	**0**	0	**6**	1.54	0.15
7-mer	CCGCCGC/GCGGCGG	**4**	66.67	**2**	33.33	**6**	7.32	0.15
	CTCTCTC/GAGAGAG	**4**	80.00	**1**	20.00	**5**	6.10	0.13
	CCTCTCT/AGAGAGG	**4**	100.00	**0**	0	**4**	4.88	0.10
	CTCTCTT/AAGAGAG	**4**	100.00	**0**	0	**4**	4.88	0.10
	CCCAAAT/ATTTGGG	**3**	100.00	**0**	0	**3**	3.66	0.08
	GCCGCCG/CGGCGGC	**3**	100.00	**0**	0	**3**	3.66	0.08
	GCGGCGC/GCGCCGC	**2**	100.00	**0**	0	**2**	2.44	0.05
	AATAAAA/TTTTATT	**2**	100.00	**0**	0	**2**	2.44	0.05
	GTGTGCG/CGCACAC	**2**	100.00	**0**	0	**2**	2.44	0.05
	CGCCGTC/GACGGCG	**2**	100.00	**0**	0	**2**	2.44	0.05
8-mer	TTGGTTTC/GAAACCAA	**2**	100.00	**0**	0	**2**	33.33	0.05
	TGGGCTTG/CAAGCCCA	**1**	100.00	**0**	0	**1**	16.67	0.03
	GCTTCTTG/CAAGAAGC	**1**	100.00	**0**	0	**1**	16.67	0.03
	ACGGGCGA/TCGCCCGT	**1**	100.00	**0**	0	**1**	16.67	0.03
	ATGATGTA/TACATCAT	**1**	100.00	**0**	0	**1**	16.67	0.03
9-mer	TCGGCGGCG/CGCCGCCGA	**2**	100.00	**0**	0	**2**	8.00	0.05
	AGGTGGTGG/CCACCACCT	**2**	100.00	**0**	0	**2**	8.00	0.05
	CCGGTGCGA/TCGCACCGG	**1**	100.00	**0**	0	**1**	4.00	0.03
	ACGAGGAGG/CCTCCTCGT	**1**	100.00	**0**	0	**1**	4.00	0.03
	TCCCTTTTC/GAAAAGGGA	**1**	100.00	**0**	0	**1**	4.00	0.03
	CGGCATGAA/TTCATGCCG	**1**	100.00	**0**	0	**1**	4.00	0.03
	CGGCAGCGA/TCGCTGCCG	**1**	100.00	**0**	0	**1**	4.00	0.03
	ACCATCCCG/CGGGATGGT	**1**	100.00	**0**	0	**1**	4.00	0.03
	ATGGGCGGC/GCCGCCCAT	**1**	100.00	**0**	0	**1**	4.00	0.03
	ATGCAGGGT/ACCCTGCAT	**1**	100.00	**0**	0	**1**	4.00	0.03
10-mer	AGCCCCAACG/CGTTGGGGCT	**1**	50.00	**1**	50.00	**2**	40.00	0.05
	TTTTTTTCTT/AAGAAAAAAA	**1**	100.00	**0**	0	**1**	20.00	0.03
	CCTGCTTTGC/GCAAAGCAGG	1	100	0	0	1	20	0.03
	ATCTCCGCCG/CGGCGGAGAT	1	100	0	0	1	20	0.03

The frequency of micro/minisatellite locus occurrence for each million nucleotides (loci/Mb) [6] in this study was 2.94, 12.64, 42.34, 5.34, 9.06, 8.30, 1.74, 0.13, 0.53, and 0.11 for mono to 10-mer repeats/Mb, respectively. Overall occurrences of 83.13 loci/Mb were found (see Table 1). In other studies, different taxa were described in analyses of EST databases, such as 133 loci/Mb (barley), 161 loci/Mb (wheat, sorghum and rye), and 256 loci/Mb for rice [28]. Also, for nonredundant ESTs in rice, sorghum, barley, wheat, and Arabidopsis, frequencies of 277, 169, 112, 94 and 133 loci/Mb were found, respectively [30]. Frequencies closer to those found in this study were described for CDS regions of Rosaceaespecies, with an average of 40.9–78 loci/Mb for Rose, Almond and Peach, while 39 loci/Mb were found for Arabidopsis [26].

3.3. Occurrence patterns for different SSR and minisatellite types and motifs Monomers, 2-mers, 3-mers, and 4-mers

On Table 2, the contents and percentage values for different micro-/minisatellite motifs are shown. For monomer, 2-mer and 3-mer repeats, all possible arrangements are shown, while for 4-mer to 10-mer repeats, only the ten most frequent motifs are shown.

The A/T monomer repeats were found in 125 loci, with 111 (88.80%) and 14 (11.20%) loci formed by A and T nucleotides, respectively. The C/G motifs were found in 13 loci, with ten (76.92%) and three (23.08%) loci formed by C and G, respectively. A/T containing SSRs were predominant and comprised 90.58% of monomer loci. In the overall distribution, the monomers represent 3.53% of 3907 detected loci. Motifs AG/CT and GA/TC were the most frequent and added up to 8.52% of 2-mer SSRs, and 6.89% and 5.96% of all 3907 detected occurrences. The motifs CT, GA, and TC were the most abundant adding up to 172, 143, and 90 loci, respectively. In maize, barley, rice, sorghum, and wheat ESTs, the motif AG was described as the most frequent [6, 16, 28, 29, 31, 32]. However, in some studies, the most frequent motif was GA [30, 33]. Repeats composed by guanine and cytosine were the most abundant among trimers, with occurrences of 18.44%, 17.89%, and 10.60%, respectively, for the motifs CCG/CGG, CGC/GCG, and GCC/GGC, adding up to 23.9% of the overall frequencies of micro-/minisatellites in the analysis. The motifs CGC, CCG, and CGG were the most frequent comprising 218, 197, and 170 loci, respectively. Many reports indicate the 3-mer CCG as the most frequent in maize, barley, wheat, sorghum and rye [6, 16, 28, 32], sugarcane [27] and rice [29, 31].

Among 4-mers, 100 different arrangements were found, where the motifs GATC (7.17%), ATTA/AAT (6.77%), and ATCG/CGAT (5.98%) were the most frequent. These motifs add up to 19.92% of 4-mer repeats found and represent 1.28% of the overall content of micro-/minisatellites. In barley ESTs, ACGT was reported as the most abundant motif [16, 28]. For other species, AAAG/CTTT and AAGG/CCTT in *Lolium perene* [34], AAAG/CTTT and AAAC/GTTT in arabidopsis UTRs [6, 35], and AAAT and AAAG in citrus [36, 37] were described as most abundant.

3.4. Remaining repeats

Among 5-mers, 188 different arrangements were detected and the most frequent were CTCCT, CTCTC, and CCTCC with 17, 17, and 12 occurrences, respectively. In the analysis of CDS regions, the ACCCG motif was the most frequent in Arabidopsis, AAAAG in *S. cerevisae*, *C. elegans*, and AAAAC in different primates [38]. Also, the motifs AAAAT, AAAAC, and AAAAG were described as the most frequent in eukaryotes [39]. In rice, the motifs AGAGG and AGGGG were the most abundant [31]. Repeats of type 6-mer were detected in 230 different arrangements, where CGCCTC and TCGCCG were the most frequent, occurring in 12 and 10 loci, respectively. Other studies have shown higher frequencies for the motifs AAGATG, AAAAAT in arabidopsis [35], AAAAAG in citrus [36], AACACG in *S. cerevisae*, ACCAGG in *C. elegans* and CCCCGG in primates [38]. For all remaining repeats (minisatellites), the occurrences are widely distributed with low-percentage values for each arrangement. For 7-mer, 8-mer, 9-mer, and 10-mer repeats, the totals of occurrences were 57, 5, 23, and 5, respectively.

3.5. Primer design and PCR simulation

The design of primers for the 3907 detected micro-/minisatellites resulted in 3329 primer pairs, covering 85.20% of loci. The running of "Virtual PCR" generated a total of 4610 amplicons. A module in SSRLocator checks for primer redundancy. A total of 2397 primer pairs amplified only the fragment from its original locus (specific amplicons) and 932 pairs amplified one or more regions besides the original locus. From these, 692 pairs amplified two fragments, one from the original site and a second from another region (paralogous). In this case, 692 specific amplicons plus 692 redundant amplicons, were detected. A total of 143, 90, 2, and 5 primer pairs generated three (two redundancies), four (three redundancies), five (four redundancies), and six (five redundancies) fragments, respectively. The final product of 932 primers with more than one anchoring region resulted in 932 specific amplicons and 1281 redundant amplicons, adding up to 2213 fragments.

To investigate the ability of these primers in amplifying genomic sequences, an extra experiment was performed against the whole rice genomic sequence available at NCBI. The different groups of redundant and nonredundant primer sets, that is, amplifying one, two, three, or more times in the cDNA database, were tested against the genomic sequence. From the 2397 nonredundant primers, only 924 amplified a locus in the genomic sequence. This difference was already expected because of difficulties in amplifying genomic regions, that is, if some primers anneal to a boundary region between two exons in the cDNA, the presence of introns would make this annealing site no more available. It is interesting to note that from the 924 amplicons detected, 914 (99%) did amplify only one locus in the genomic region, agreeing with the cDNA results. When the primer sets that amplified two different cDNAs were run against the genomic sequence, only 294/692 (42.5%) did amplify, having 14.5% been able to amplify two different loci.

Table 3: Distribution of amplicon alignments for specific and redundant amplicons with varying identity levels.

Identity	100	99	98	97	96	95–90	89–80	79–70	69–60	≤59	Total
Amplicons	787	261	151	29	11	8	8	6	5	15	1281
%	61.44	20.37	11.79	2.26	0.86	0.62	0.62	0.47	0.39	1.17	—

Only one primer set did amplify more than two loci. These results indicate that SSR locator performance was consistent between the two databases regarding the nonredundant loci, that is, from those loci that were able to be amplified in both databases, their status of nonredundant was maintained. The changes observed for the redundant loci can be attributable to many causes, including redundancy in the cDNA database, but also to biological reasons due to primer positioning.

3.6. Identity between specific and redundant amplicons

Results of a global alignment between amplicons from original and redundant sites are shown in Table 3. Among the 1281 redundant amplifications, 787 (61.44%) resulted in a perfect alignment between both loci (identity equal to 100). For redundant amplicons with identity levels of 96–99%, and 90–95%, 452 (35.28%) and 8 (0.62%) loci were found, respectively. Alignments with identity levels bellow 90% were found in only 2.65% of cases. The fact that such a high percentage of redundant loci show high identity is probably a consequence of the genome fraction chosen, that is, expressed sequences. This fraction is under tight selection pressure and should not accumulate variations such as substitutions or indels at a high rate. As expected, comparisons to whole genome, generated a great deal of polymorphism, due to the inclusion of intronic regions in the alignments (data not shown).

4. CONCLUSIONS

The software SSRLocator was successfully implemented, adding steps for (1) SSR discovery, (2) primer design, and (3) PCR simulation between the primers obtained from original sequences and other fasta files. Also, the software produces reports for frequency of occurrence, nucleotide arrangement, primer lists with all standard information needed for PCR and global alignments. From the PCR simulation, it was possible to point out which primer pairs were nonredundant, suggesting that these primers are more appropriate for mapping purposes. In this case, however, wet lab experiments should be performed to confirm the advantage of nonredundant over redundant primers for mapping.

It is possible that the results for micro-/minisatellite frequencies (loci/Mb) obtained in this study diverge from the results found in the literature. This can be explained by the different databases used (redundant ESTs, nonredundant ESTs and/or fl-cDNA), different algorithm configurations and minimum requirements set for counting motifs. Another explanation for some contrasting results is the fact that only "Class I" repeats were analyzed in our study.

The results showed that 932 (27.99%) primers presented amplifications in more than one gene sequence. This could be mostly due to the fact that primer pairs derived from a specific gene (cDNA) anchored in similar sites in other duplicated genes, since 5,607/28,469 (19.70%) genes were described as paralogs in the annotation of the database used [24]. Gene duplication along with polyploidy and transposon amplification are the major driving forces in genome evolution [40]. It is therefore not surprising that so many loci have redundancy. Also, a second possibility is that some primers were generated from protein domain regions within the analyzed cDNAs. These domains could be found in protein families with many genome copies, resulting in the observed redundancies. A validation of the redundancies of cDNA results was obtained through a virtual-PCR against the whole rice genome sequence. From the nonredundant primers that generated an amplicon, ca. 99% were nonredundant.

Finally, this tool can be used successfully for data mining strategies to find SSR primers in genomic or expressed sequences (ESTs/cDNAs). Also, this software can be a tool for microsatellite discovery in databanks of related species, anchoring primers in ortholog or paralog regions contained between databases from two different species.

ACKNOWLEDGMENTS

The authors are thankful to the Brazilian Council for Research and Development (CNPq) and the Coordination for Support to Superior Studies (CAPES/Brazil) for grants and fellowships.

REFERENCES

[1] M. Morgante, M. Hanafey, and W. Powell, "Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes," *Nature Genetics*, vol. 30, no. 2, pp. 194–200, 2002.

[2] R. R. Iyer, A. Pluciennik, W. A. Rosche, R. R. Sinden, and R. D. Wells, "DNA polymerase III proofreading mutants enhance the expansion and deletion of triplet repeat sequences in *Escherichia coli*," *Journal of Biological Chemistry*, vol. 275, no. 3, pp. 2174–2184, 2000.

[3] H. Ellegren, "Microsatellites: simple sequences with complex evolution," *Nature Reviews Genetics*, vol. 5, no. 6, pp. 435–445, 2004.

[4] S. M. Mirkin, "DNA structures, repeat expansions and human hereditary disorders," *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 351–358, 2006.

[5] B. Li, Q. Xia, C. Lu, Z. Zhou, and Z. Xiang, "Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes," *Genomics Proteomics & Bioinformatics*, vol. 2, no. 1, pp. 24–31, 2004.

[6] M. Morgante, M. Hanafey, and W. Powell, "Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes," *Nature Genetics*, vol. 30, no. 2, pp. 194–200, 2002.

[7] S. Subramanian, R. K. Mishra, and L. Singh, "Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions," *Genome Biology*, vol. 4, no. 2, p. R13, 2003.

[8] R. K. Varshney, A. Graner, and M. E. Sorrells, "Genic microsatellite markers in plants: features and applications," *Trends in Biotechnology*, vol. 23, no. 1, pp. 48–55, 2005.

[9] M. Bilgen, M. Karaca, A. N. Onus, and A. G. Ince, "A software program combining sequence motif searches with keywords for finding repeats containing DNA sequences," *Bioinformatics*, vol. 20, no. 18, pp. 3379–3386, 2004.

[10] W. R. Pearson and D. J. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 85, no. 8, pp. 2444–2448, 1988.

[11] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[12] C. Abajian, SPUTNIK, 1994, http://www.abajian.com/sputnik.

[13] A. F. A. Smit, R. Hubley, and P. Green, RepeatMasker Open-3.0, 1996, http://www.repeatmasker.org.

[14] G. Benson, "Tandem repeats finder: a program to analyze DNA sequences," *Nucleic Acids Research*, vol. 27, no. 2, pp. 573–580, 1999.

[15] A. T. Castelo, W. Martins, and G. R. Gao, "TROLL—tandem repeat occurence locator," *Bioinformatics*, vol. 18, no. 4, pp. 634–636, 2002.

[16] T. Thiel, W. Michalek, R. K. Varshney, and A. Graner, "Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.)," *Theoretical and Applied Genetics*, vol. 106, no. 3, pp. 411–422, 2003.

[17] S. Temnykh, G. DeClerck, A. Lukashova, L. Lipovich, S. Cartinhour, and S. McCouch, "Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential," *Genome Research*, vol. 11, no. 8, pp. 1441–1452, 2001.

[18] G. D. Schuler, "Sequence mapping by electronic PCR," *Genome Research*, vol. 7, no. 5, pp. 541–550, 1997.

[19] S. Rozen and H. Skaletsky, "Primer3 on the WWW for general users and for biologist programmers," *Methods in Molecular Biology*, vol. 132, part 3, pp. 365–386, 2000.

[20] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[21] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[22] M. Waterman, "Estimating statistical significance of sequence alignments," *Philosophical transactions of the Royal Society of London. Series B*, vol. 344, no. 1310, pp. 383–390, 1994.

[23] M. Vingron and M. S. Waterman, "Sequence alignment and penalty choice. Review of concepts, case studies and implications," *Journal of Molecular Biology*, vol. 235, no. 1, pp. 1–12, 1994.

[24] S. Kikuchi, K. Satoh, T. Nagata, et al., "Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice: the rice full-length cDNA consortium," *Science*, vol. 301, no. 5631, pp. 376–379, 2003.

[25] L. Cardle, L. Ramsay, D. Milbourne, M. Macaulay, D. Marshall, and R. Waugh, "Computational and experimental characterization of physically clustered simple sequence repeats in plants," *Genetics*, vol. 156, no. 2, pp. 847–854, 2000.

[26] S. Jung, A. Abbott, C. Jesudurai, J. Tomkins, and D. Main, "Frequency, type, distribution and annotation of simple sequence repeats in Rosaceae ESTs," *Functional & Integrative Genomics*, vol. 5, no. 3, pp. 136–143, 2005.

[27] G. M. Cordeiro, R. Casu, C. L. McIntyre, J. M. Manners, and R. J. Henry, "Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum," *Plant Science*, vol. 160, no. 6, pp. 1115–1123, 2001.

[28] R. K. Varshney, T. Thiel, N. Stein, P. Langridge, and A. Graner, "In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species," *Cellular & Molecular Biology Letters*, vol. 7, no. 2A, pp. 537–546, 2002.

[29] R. V. Kantety, M. La Rota, D. E. Matthews, and M. E. Sorrells, "Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat," *Plant Molecular Biology*, vol. 48, no. 5-6, pp. 501–510, 2002.

[30] S. K. Parida, K. Anand Raj Kumar, V. Dalal, N. K. Singh, and T. Mohapatra, "Unigene derived microsatellite markers for the cereal genomes," *Theoretical and Applied Genetics*, vol. 112, no. 5, pp. 808–817, 2006.

[31] M. La Rota, R. V. Kantety, J.-K. Yu, and M. E. Sorrells, "Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley," *BMC Genomics*, vol. 6, article 23, 2005.

[32] J.-K. Yu, T. M. Dake, S. Singh, et al., "Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat," *Genome*, vol. 47, no. 5, pp. 805–818, 2004.

[33] N. Nicot, V. Chiquet, B. Gandon, et al., "Study of simple sequence repeat (SSR) markers from wheat expressed sequence tags (ESTs)," *Theoretical and Applied Genetics*, vol. 109, no. 4, pp. 800–805, 2004.

[34] T. Asp, U. K. Frei, T. Didion, K. K. Nielsen, and T. Lübberstedt, "Frequency, type, and distribution of EST-SSRs from three genotypes of *Lolium perenne*, and their conservation across orthologous sequences of *Festuca arundinacea*, *Brachypodium distachyon*, and *Oryza sativa*," *BMC Plant Biology*, vol. 7, article 36, 2007.

[35] L. Zhang, D. Yuan, S. Yu, et al., "Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*," *Bioinformatics*, vol. 20, no. 7, pp. 1081–1086, 2004.

[36] D. Jiang, G.-Y. Zhong, and Q.-B. Hong, "Analysis of microsatellites in citrus unigenes," *Acta Genetica Sinica*, vol. 33, no. 4, pp. 345–353, 2006.

[37] D. A. Palmieri, V. M. Novelli, M. Bastianel, et al., "Frequency and distribution of microsatellites from ESTs of citrus," *Genetics and Molecular Biology*, vol. 30, no. 3, supplement, pp. 1009–1018, 2007.

[38] G. Tóth, Z. Gáspári, and J. Jurka, "Microsatellites in different eukaryotic genomes: surveys and analysis," *Genome Research*, vol. 10, no. 7, pp. 967–981, 2000.

[39] Y.-C. Li, A. B. Korol, T. Fahima, and E. Nevo, "Microsatellites within genes: structure, function, and evolution," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 991–1007, 2004.

[40] E. A. Kellogg and J. L. Bennetzen, "The evolution of nuclear genome structure in seed plants," *American Journal of Botany*, vol. 91, no. 10, pp. 1709–1725, 2004.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 496957, 10 pages
doi:10.1155/2008/496957

Resource Review

MaizeGDB: The Maize Model Organism Database for Basic, Translational, and Applied Research

Carolyn J. Lawrence,[1,2,3] Lisa C. Harper,[4,5] Mary L. Schaeffer,[6,7] Taner Z. Sen,[1,2] Trent E. Seigfried,[1] and Darwin A. Campbell[1]

[1] *USDA-ARS, Corn Insects and Crop Genetics Research Unit, Ames, IA 50011, USA*
[2] *Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011, USA*
[3] *Department of Agronomy, Iowa State University, Ames, IA 50011, USA*
[4] *USDA-ARS, Plant Gene Expression Center, 800 Buchanan Street, Albany, CA 94710, USA*
[5] *Department of Molecular and Biology, University of California Berkeley, Berkeley, CA 94720, USA*
[6] *USDA-ARS, Plant Genetics Research Unit, Columbia, MO 65211, USA*
[7] *Division of Plant Sciences, University of Missouri Columbia, Columbia, MO 65211, USA*

Correspondence should be addressed to Carolyn J. Lawrence, carolyn.lawrence@ars.usda.gov

Received 31 August 2007; Accepted 10 July 2008

Recommended by Chunguang Du

In 2001 maize became the number one production crop in the world with the Food and Agriculture Organization of the United Nations reporting over 614 million tonnes produced. Its success is due to the high productivity per acre in tandem with a wide variety of commercial uses. Not only is maize an excellent source of food, feed, and fuel, but also its by-products are used in the production of various commercial products. Maize's unparalleled success in agriculture stems from basic research, the outcomes of which drive breeding and product development. In order for basic, translational, and applied researchers to benefit from others' investigations, newly generated data must be made freely and easily accessible. MaizeGDB is the maize research community's central repository for genetics and genomics information. The overall goals of MaizeGDB are to facilitate access to the outcomes of maize research by integrating new maize data into the database and to support the maize research community by coordinating group activities.

1. INTRODUCTION

Maize (Zea *mays* L.) is a species that encompasses the subspecies *mays* (commonly called "corn" in the US) as well as the various teosintes that gave rise to modern maize. Maize is an important crop: not only is it one of the most abundant sources of food and feed for people and livestock the world over, it is also an important component of many industrial products. Maize byproducts are present in, for example, glue, paint, insecticides, toothpaste, rubber tires, rayon, and molded plastics, among others. Maize is also currently the nation's major source of ethanol, a major biofuel that is more environmentally friendly than gasoline and that may be a more economical fuel alternative in the long run. Although it is unlikely that ethanol production from maize directly will be sustainable long-term, maize's suitability to serve as a model organism for developing fuelstock grasses is apparent [1]. Indeed, in addition to its value as a commodity, maize has been a premiere model organism for biological research for over 100 years. Many seminal scientific discoveries have first been shown in maize, such as the identification [2] and cloning [3] of transposable elements, the correlation between cytological and genetic crossing over [4], and the discovery of epigenetic phenomena [5]. These exceptional characteristics of maize set this amazing plant apart: no other species serves as both a commodity and a leading model for basic research.

Today, with the accelerated generation of maize genetic and genomic information, the need for a centralized biological data repository is critical. MaizeGDB (the **Maize G**enetics and genomics **D**ata **B**ase [6]) (http://www.maizegdb.org/) is the Model Organism Database (MOD) for maize. Stored at MaizeGDB is comprehensive information on loci (genes

and other genetically defined genomic regions including QTL), variations (alleles and other sorts of polymorphisms), stocks, molecular markers and probes, sequences, gene product information, phenotypic images and descriptions, metabolic pathway information, reference data, and contact information for maize researchers. Described in the results and discussion section are example workflows that could be followed by researchers to utilize the MaizeGDB resource for their research. Other long-term resources serving maize data include Gramene (http://www.gramene.org/) [7], which specializes in grass comparative genomics, and GRIN (the Germplasm Resources Information Network; http://www.ars-grin.gov/npgs/), which provides access to the National Plant Germplasm System's germplasm stocks and related breeding data. MaizeGDB makes an effort to guide researchers to these resources via context-sensitive linkages rather than duplicating data, though some data are shared simply to allow for the context-sensitive linkages to be created. This reduces duplication in effort and allows personnel skilled in comparative genomics and germplasm conservation/plant breeding to interact with maize researchers directly via Gramene and GRIN, respectively.

In addition to storing and making maize data available, the MaizeGDB team also provides services to the community of maize researchers and offers technical support for the Maize Genetics Executive Committee and the Annual Maize Genetics Conference. Also available at the MaizeGDB website, as a service to the maize research community, are bulletin boards for news items, information of interest to cooperators, lists of websites for projects that focus on the scientific study of maize, the Editorial Board's recommended reading list, and educational outreach items.

The genetic and genomic data as well as community-related information maintained by MaizeGDB are highly utilized: MaizeGDB averages 8620 visitors (based on unique Internet Protocol or IP addresses) and over 160 000 page impressions per month (July 2007 to June 2008). In addition, MaizeGDB came in fifth out of 170 in a National Plant Genome Initiative Grantees poll in which lead principal investigators reported most useful websites for their research [8].

2. MATERIALS AND METHODS

2.1. Kinds of data in the database that link genetic and genome sequence information

MaizeGDB is the primary repository for the major genetic and cytogenetic maps and includes details about genes, mutants, QTL (quantitative trait loci), and molecular markers including 2500 RFLPs (restriction fragment length polymorphisms), 4625 SSRs (simple sequence repeats), 363 SNP (single nucleotide polymorphisms), 2500 indels (insertion/deletion sites), and 10 644 overgos (**over**lapping oli**go**nucleotides). These data are described using 1.27 millions synonyms, 42 000 primer sequences, 16 394 raw scores from mapping based upon 16 panels of stocks, and 323 313 links to GenBank [9] accessions. GenBank accessions form the links between the genetic position on a chromosome, the sequence records at MaizeGDB, and the EST (expressed sequence tag) and GSS (genome survey sequence) contig assemblies at PlantGDB [10] and Dana Farber (The Gene Indices at http://compbio.dfci.harvard.edu/tgi/cgi-bin/tgi/gimain.pl?gudb=maize, previously at TIGR [11]). All of the 3 520 247 sequences in MaizeGDB are accessible by BLAST [12] and can be filtered to report only mapped loci, including any SSRs and overgos that may not be mapped genetically, but via BACs (bacterial artificial chromosomes) in anchored contigs.

The inclusion of the public BAC FPC (Finger Print Contig) information [13] adds 439 449 BACs together with associated overgo, SSR, and RFLP markers, which are used to assemble the contigs and to link contigs onto genetic map coordinates. The order of loci on the BAC contigs is represented by over 27 000 sequenced-based loci on the IBM2 FPC057 maps (http://www.maizegdb.org/cgi-bin/displaymapresults.cgi?term=ibm2+fpc0507) in MaizeGDB, by links to contigs at both the Arizona FPC site (http://www.genome.arizona.edu/) and the genome sequencing project (http://www.maizesequence.org/). As the B73 genome sequence progresses, these BAC sequences are added to MaizeGDB along with links to the sequencing project, both from the BAC clones and from genetically mapped loci associated with a BAC.

The newest maps in MaizeGDB, IBM SNP 2007 (http://www.maizegdb.org/cgi-bin/displaymapresults.cgi?term=ibm%20snp%202007), are the first of a new generation of genetic maps from the Maize Diversity Project (http://www.panzea.org/) kindly provided pre-publication by Dr. Mike McMullen. The SNP loci on these maps are associated with allelic sequences from a core set of maize and teosinte germplasm. Because the majority of the anticipated 1128 loci have been previously mapped onto BAC clones [13, 14], these genetic maps tightly link sequence diversity to the B73 genome sequence.

2.2. Methods of access, environments, and the database back end

2.2.1. The production web interface

Maize researchers primarily access MaizeGDB through the series of interconnected Web pages available at http://www.maizegdb.org/ (see Figure 1). These web pages are dynamically generated and are written in PHP (the recursive abbreviation for PHP Hypertext Preprocessor [15]) and Perl [16]. Through this interface, each page shows detailed information on a specific biological entity (such as a gene) as well as basic information about data associated with it (genes are associated with maps, phenotypes, and citations, among others). These additional data types are linked to the gene page, enabling quick access to alternative data views. The site also includes links to related resources at other databases; genes, for example, are linked to Gramene [7].

One may access these individual data pages by using either (1) the search bar located at the top right of every page (Figure 1(A)), or (2) data type-specific advanced querying

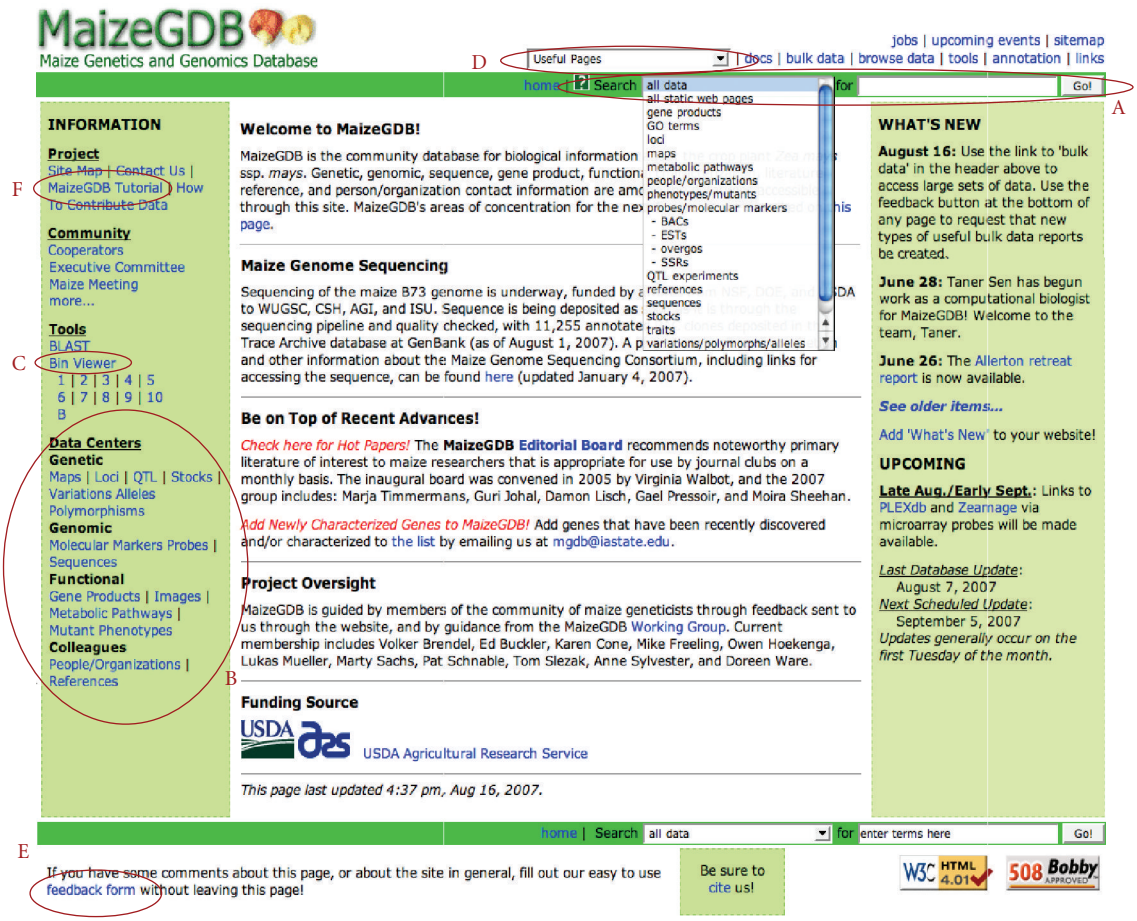


FIGURE 1: The MaizeGDB home page. The most commonly utilized search functionality for MaizeGDB is the search bar (A), which is available within the header of any MaizeGDB page. To browse data and to search specific data types using specific limiters, the Data Centers (B) are also quite useful. Also available is a Bin Viewer (C), which allows for a view of lots of data types within the context of their chromosomal location. To enable access to the Data Centers and other displays of interest from any MaizeGDB page, a pull-down menu for "Useful pages" (D) is accessible on the header of any MaizeGDB page. The footer of all MaizeGDB pages contains a context-sensitive "feedback form" link (E). Researchers use the feedback form to report errors, ask questions, and to contact the MaizeGDB team directly. For newcomers to the site, the MaizeGDB Tutorial (F) can help them to get a jump start on how to use the site.

tools (accessible via the "Data Centers" links; Figure 1(B)) on the left side of the home page, or (3) the Bin Viewer tool (Figure 1(C)), which is located in the left margin of the home page or via a pull down labeled "Useful pages" (Figure 1(D)) accessible at the top of any MaizeGDB page. These tools allow researchers to easily find relevant data displays.

MaizeGDB's method of data delivery has three primary goals: placing information within the framework of its scientific meaning, making this information available to the researcher with minimal input (often only the relevant term), and requiring minimal effort from the researcher to comprehend the data displays. By focusing on biological context and ease of use as the primary focus of this interface (the "production" Web interface), the database is intended to be intuitive to the researcher as their click stream follows a logical path of biological associations. Up-to-date site usage statistics can be accessed online at http://www.maizegdb.org/usage/.

2.2.2. *Structure and relationship of environments: production, staging, and test*

The production Web interface, which most MaizeGDB users interact with, is only one component of the overall MaizeGDB infrastructure (Figure 2). The data accessed by the production Web interface are typically updated on the first Tuesday of each month. Prior to being in that Production Environment, the data are prepared for public accessibility in a Staging Environment. In the Staging Environment, the most up-to-date information is available, new data are added to the database, and existing data are updated with new information. In addition to a Web interface that appears identical to the one in the Production Environment, the Staging Environment offers SQL (Structured Query Language) read-only access to the community so that researchers interested in interacting with the data in a more direct and customized manner can have access to the most

up-to-date information available. In addition, a Disaster Recovery system has been put in place whereby the Curation Database is backed up in a compressed format to a separate machine in Ames, Iowa daily. Once weekly, the Ames file is copied to Columbia, Missouri for off-site storage.

To aid in the modeling of new types of data for inclusion in the MaizeGDB product and to enable programming to be tried out in a safe place, a Test Environment identical to the Staging Environment has been created. Note that three copies of the database exist. While each environment and server has a specific purpose, all are configured such that they could serve a backup to each other. If any one server was to fail, either of the other two could provide full, unrestricted data access and site functionality. The curation database is backed up on a daily basis and is available for download (http://goblin1.zool.iastate.edu/~oracle/) for those who have Oracle Relational Database Management System (RDBMS) installed locally.

2.2.3. Curation

Also available within the Staging Environment are Community Curation Tools to enable researchers to add small datasets to the database directly, as well as a set of Professional Curation Tools developed by Dr. Marty Sachs' group at the Maize Genetics Cooperation-Stock Center in Urbana-Champaign [17]. Whereas the Community Curation Tools have many safeguards to help researchers enter data stepwise and with enforced field requirements, the Professional Curation Tools allow MaizeGDB project members as well as Stock Center personnel to enter datasets in a more stream-lined and powerful fashion with fewer integrity enforcement rules (which slow down the data entry process considerably). It also should be noted that data added to the database via the Community Curation Tools are first marked as "Experimental" that must be "Activated" by professional curators at MaizeGDB. This ensures that only quality information is made publicly accessible. The availability of a Curation Web interface (within the Staging Environment) enables researchers to view the data as they will appear once they are uploaded to Production. Few researchers (about 30 at present) have Community Curation accounts. To increase the use of these tools, training sessions are being organized (see Section 2.3, below). If researchers wish to deposit complex or large datasets, it would not be reasonable to enter the data via the Community Curation Tools because those tools work via a "bottom-up" approach whereby the records are (1) built based upon the most basic information included in the dataset and (2) entered one record at a time (i.e., not in bulk). For complex or large datasets, researchers are encouraged to submit data files to the curators at MaizeGDB. Those data are added to the database directly by curators and the database administrator.

2.2.4. Database back end

Each environment's server has a perpetual license and is supported by Oracle RDBMS powered by 2×2.0 GHz Xeon processors, 4 GB of RAM, 5×73 GB Ultra 320 10 K RPM drives with Red Hat Advanced Server 2.1 operating system installed. The curation database, either partially or in its entirety, can be moved to MySQL, Microsoft Access, and nearly any other portable data format that a researcher would need. Requests to gain read-only SQL access to the Curation database can be made via the feedback link that appears at the bottom of any MaizeGDB page. Data housed at MaizeGDB are in the public domain and are freely available for use without a license.

2.3. Outreach

One of the strengths of MaizeGDB is its responsiveness to community input, received either personally or by the feedback forms accessible at the bottom of each page (Figure 1(E)). To provide outreach and user support as well as to solicit input from researchers in a more active manner, several strategies are employed. The first is tutorials and basic information on MaizeGDB. The MaizeGDB Tutorial (Figure 1(F)) can be reached from the home page at the top of the left margin. A new user can go through this tutorial, and become familiar with how to use the site quickly. In addition, a "Site Tour" with an overview with examples can be found under the "Useful pages" pull down menu at the top of each page. More specific tutorial examples and other educational materials are available via the "Education" link, also within the "Useful pages" pull down menu. Also, on many of the "Data Center" pages (available from the left margin of the front page or via the "Useful pages" pull down) a discussion of the topic of the page that is suitable for the general public appears toward the bottom. Another form of outreach supported by MaizeGDB is assistance at meetings and conferences. Representatives from MaizeGDB attend and help researchers at the Annual Maize Genetics Conference (usually in March), the International Plant and Animal Genome Conference (January), and various other meetings through direct interaction in person. Finally, researchers can request a MaizeGDB site visit. About three times a year, an expert curator travels to various research locations and provides tutorials and support for maize researchers. For these visits, the local maize researchers are asked for a list of specific questions ahead of time. During the one to two day visits, researchers interact in groups and one-on-one with the traveling curator to learn how to utilize MaizeGDB for their research and to deposit data at MaizeGDB.

2.4. Community support services

MaizeGDB provides community support in several ways. Two members of the MaizeGDB team, MLS and TES, serve as ex officio members of the Maize Genetics Conference Steering Committee. They collect electronic abstracts for the Annual Maize Genetics Conference and handle the preparation and printing of the program for the conference. MaizeGDB personnel also manage regular community surveys on behalf of the Maize Genetics Executive Committee. These surveys enable the Executive Committee to summarize the overall research interest of the maize community and to advise funding agencies

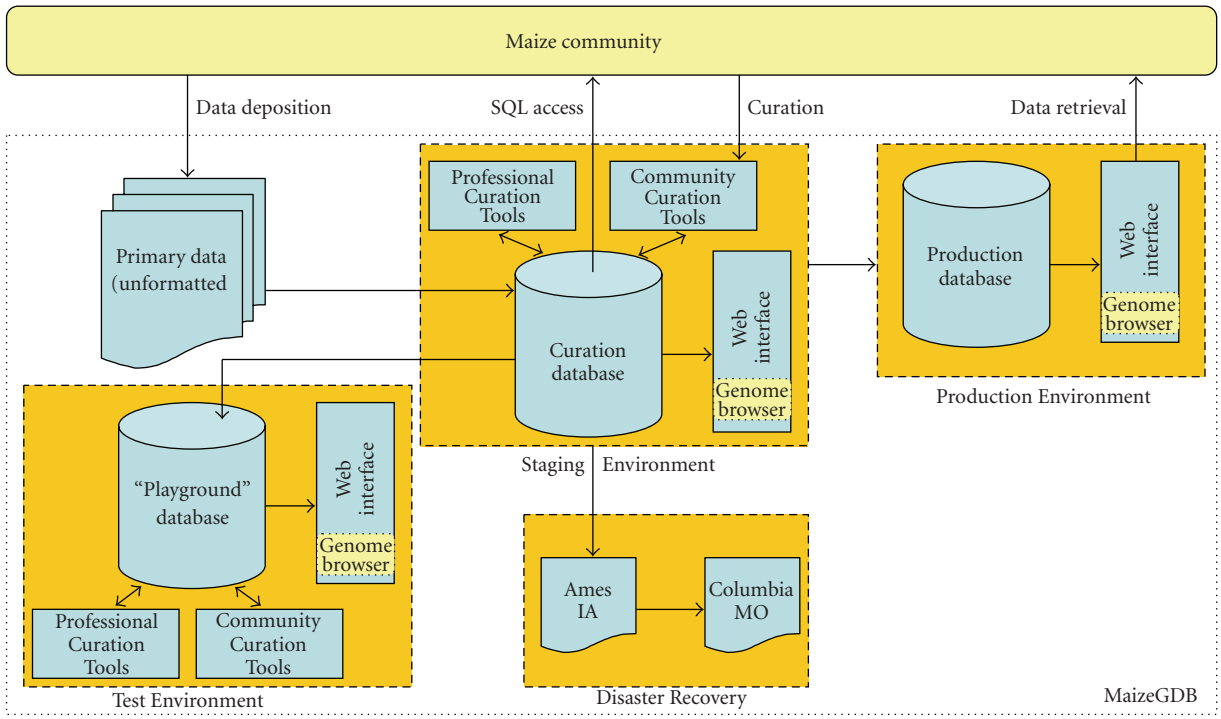


Figure 2: Simplified infrastructure of MaizeGDB. The community of maize researchers can add data to the database (downward-facing arrows from the uppermost yellow box) via direct data deposition (upper left) and via a set of Community Curation Tools that interacts with the Curation Database (upper center). Researchers are also allowed access to maize data (upward-facing arrows from the lower dashed box) via a web interface that can be accessed at http://www.maizegdb.org/ (upper right) and by way of SQL access to the Curation Database, which houses the most up-to-date data available (upper center). These functionalities are supported by two of the three environments: Production and Staging, respectively (upper dashed gold boxes). Available for use by MaizeGDB personnel to facilitate data modeling and trial programming manipulations is a third environment called Test (lower left dashed gold box), which is identical to the Staging Environment. To ensure that the most up-to-date copy of the database is backed up, a Disaster Recovery process has been instituted (lower center dashed gold box) whereby a compressed copy of the database is backed up to a separate machine in Ames, Iowa daily, and to a server in Columbia, Missouri weekly.

on future research directions. MaizeGDB personnel also manage the Executive Committee's website (i.e., http://www.maizegdb.org/mgec.php) and conducts the Executive Committee's elections. MaizeGDB houses the mailing list for the annual Maize Newsletter and project personnel conduct semi-regular mailings to the maize community on behalf of interested researchers by maintaining an electronic list of researchers' contact information. Potential mailings to this list are vetted by the Executive Committee.

3. RESULTS AND DISCUSSION

To demonstrate how researchers utilize MaizeGDB, three example usage cases are presented here. Because researchers with very different goals can all utilize MaizeGDB to advance their work, the usage cases are classified by research type: basic, translational, and applied. See Figure 3 for examples of how these research types fit together. By enabling researchers to carry out workflows that support translational and applied research, MaizeGDB plays a part in influencing crop development directly. Although a single researcher might even include all of these three aspects in his/her research simultaneously, here the researcher types are distinguished as follows: basic researchers investigate the fundamental biology of the organism, translational researchers work to determine the application of basic research outcomes for practical purposes [18], and applied researchers implement proven technologies to improve crops.

3.1. *Basic*

Many basic researchers work with mutants to understand the processes underlying biological phenomena. Once a new mutant is found, there are several standard methods used to elucidate normal gene functions. These efforts include determining whether the mutant represents an allele of a previously described gene, and if not, genetic mapping and cloning of the new gene. Information stored in MaizeGDB is useful in all of these steps.

In a large screen for mutations that change pericarp pigmentation from red to some other color, Researcher 1 has found a plant with a brownish-red pericarp coloration. She first wants search MaizeGDB to find all known mutants that have red pericarp phenotypes

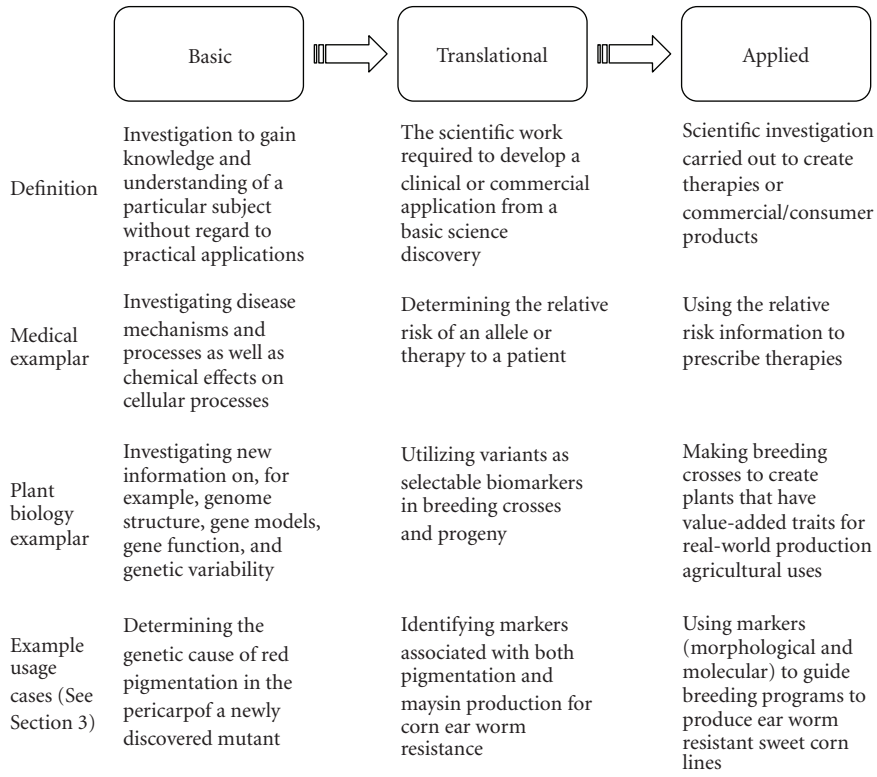


Figure 3: Three types of biological research. Research can be divided into three categories: basic, translational, and applied. Outcomes from basic research feed into translational predictions, and developed uses for these findings constitute the basis for developing real-world applications that benefit humanity and the world. Listed after the flow of research are definitions for each research type as well as medical and plant biological models for how the different divisions are interrelated. Also shown are overviews of the example usage cases presented in Section 3.

to determine whether this mutation represents a newly discovered gene. Because she does not know how others might have described the phenotype, she decides to browse existing phenotype terms and images. From the left margin of the MaizeGDB homepage, she selects "Mutant Phenotypes" under "Data Centers-Functional." On this page (http://www.maizegdb.org/), she selects "pericarp color" from the pull down menu labeled "Show only phenotypes relating to this trait" in the green search bar. A number of possible mutant phenotypes are returned, including "red pericarp." Clicking on the "red pericarp" phenotype link, she finds that the listed mutants are alleles of *p1* (*pericarp color1*). On this page (http://www.maizegdb.org/cgi-bin/displayphenorecord.cgi?id=13818), she scrolls to the bottom and finds that there are many stocks that can be ordered from the Maize Genetics Cooperation-Stock Center that carry *P1-rr* (an allele that causes red pericarp and red cob) or *P1-rw* (red pericarp and white cob). Having these stocks in hand will enable her to test whether the new mutant represents an allele of the *p1* gene, so she decides to order a few for complementation analyses. Clicking on the stock links listed on the variation/allele page allows her access to a shopping cart utility (in the green right hand panel), and she orders seed from the Stock Center directly through the MaizeGDB interface. She then goes back to the results of her "pericarp color" query and repeats the process for "cherry pericarp," ordering stocks for *r1-ch* (*colored1-cherry*), also to be used in her complementation analyses. (Another way she could have found maize stocks that have red pericarp is the following: from the header of any page, select "Useful pages" and click "Stocks." This pulls up the stock search page http://www.maizegdb.org/stock.php. In the green box, select stocks with the phenotype "red pericarp" from the pull down menu of all phenotype names and submit. A long list of stocks that contain alleles of *p1* with red pericarp is returned. Alternatively, the Stock Center Catalog is also available from the Stocks Data Center page.)

Researcher 1 receives several appropriate stocks and performs allelism tests and determines that her mutant (which turns out to be recessive) is not allelic to *p1* or *r1*. She returns to MaizeGDB and again looks through "Mutant Phenotype" results using the "pericarp color" query. Listed there are brown pericarp, orange pericarp, white pericarp, and lacquer red pericarp phenotypes in addition to the red and cherry phenotypes she focused on initially. She finds that there is no stock available for the brown pericarp phenotype (the *brown pericarp1* mutant has been lost), and

all the others are alleles that confer colored pericarp in the dominant condition as a result of the presence of *P1* alleles. To determine whether the new mutation could be an allele of *bp1*, she decides to map it genetically.

MaizeGDB houses the largest collection of publicly available genetic maps of maize (currently over 1,337 maps). These include maps of genes primarily defined by mutants with morphological phenotypes ("Genetic 2005" is the most current), maps based on phenotypic molecular markers, and composite maps where various maps have been integrated. These maps can be easily accessed from the home page, via the left margin link to "Data Centers-Genetic-Maps" (http://www.maizegdb.org/map.php). This page not only allows various map search functions, but also provides information on the most popular maps and a handy reference to explain more about the various composite maps.

The maize genome is divided into genetic bins of approximately 20 centiMorgans each and boundary markers with nearby SSRs can be used for mapping (for further explanation see http://www.maizegdb.org/cgi-bin/bin_viewer.cgi). Researcher 1 decides to utilize SSRs to map her gene to bin resolution. To find the core markers from the home page, she clicks on "Tools-Bin Viewer" in the left margin of the home page. This provides a list of the core bin markers and a link to purchase relevant primers to screen her mapping population. She generates a mapping population, performs PCR experiments using the polymorphic markers, and maps her mutant to bin 9.02.

To see what genes are located in bin 9.02, she goes back to the Bin Viewer (from the homepage), and holds the curser over the image of chromosome 9 until she sees "bin 9.02," then clicks. The result is a long list of genes, other loci, sequences, EST contigs, SSRs, BACs, and other data relating to bin 9.02. Searching through this data, she sees that *bp1* is listed under "other loci" in bin 9.02. This is a "lapsed locus" meaning that the stock has been lost, but perhaps she has found a new allele!

To see more specific genetic mapping data on *bp1*, she goes to the search bar along the top green bar of every page, selects "loci" from the pull down menu, types "bp1" into the field provided, and clicks the button marked "Go!" This brings her to the *bp1* locus page (http://www.maizegdb.org/cgi-bin/displaylocusrecord.cgi?id=61563) where she can see that *bp1* is placed on three genetic maps. Clicking on each map, Researcher 1 learns that in 1935, *bp1* was mapped between *sh1* and *wx1* (*shrunken1* and *waxy1*), two well-studied genes. To search for molecular markers suitable for fine structure mapping, she visits "Data Centers-Genetic-Maps" from the link on the home page. In the green Advanced Search box, she enters *sh1* and *wx1* separately in the "Show only maps containing this locus" lines. This returns only genetic maps that contain both genes. She selects the map with the most markers—IBM2 2005 Neighbors 9 (with 2,488 markers). She finds *sh1* at position 80.30, and *wx1* at 185.00. To choose among several molecular markers, Researcher 1 follows the available links leading her to information about suitable primers, a number of variations (which can help to decide if there may be a polymorphism in her mapping populations), gel patterns, and any available GenBank accession numbers for sequences as well as sequenced BACs. She finally selects markers and performs fine structure mapping. As she finds markers closer and closer to the gene, she can proceed with positional cloning to determine whether the position is consistent with *bp1* (nice examples of how this is done can be found in [19–21]).

3.2. Translational

Research to understand the metabolic pathways that produce pigmentation (like those outlined in Section 3.1) are well studied in maize [22]. One example of a well-characterized gene that confers pigmentation is *p1*, which encodes a transcription factor that regulates synthesis of flavones such as anthocyanins [23]. The *p1* gene, along with its adjacent duplicate *pericarp color2* (*p2*), controls pericarp and cob coloration and causes silks to brown when cut. One flavone produced by the pathway is maysin, a compound which has been shown to be antinutritive to the corn ear worm at concentrations above 0.2% fresh weight if husks limit access to the ear such that feeding on silks is required for the insect to enter [24]. Many QTL for resistance to corn earworm map near loci in the flavone synthesis pathway that are either regulatory genes (such as *p1* and *p2*), or at rate-limiting enzymatic steps, such as *c1* (*chalcone synthase1*) that contribute maysin accumulation in silks [25]. Understanding how maysin functions and how this information could be used for production agriculture is Researcher 2's area of expertise.

Researcher 2 has investigated maysin synthesis for some time, and has decided to clone an uncharacterized maysin QTL near *umc105a*, in the bin 9.02, which is bounded by *bz1* and *wx1* [24]. He believes that the QTL may be a previously described, but lost, *bp1* mutant thought to be involved in maysin synthesis. In the first step, he must first find molecular markers to more finely map the region (his preference would be to use SSRs, since members of the lab are already using them successfully). He plans to follow the strategy of chromosome walking to narrow down the region of interest [19–21] followed by association mapping to identify the actual QTL sequence [26, 27]. Knowing this sequence would enable plant breeders to track the QTL for marker assisted selection.

To find SSR data for mapping to a bin region, Researcher 2 goes to the MaizeGDB home page and clicks on "Data Centers-Genomic-Molecular Markers/Probes" in the left margin, then clicks the "SSR" link at the top of the page (the link is located in "*Specific information is available on BACs, ESTs, overgos, and SSRs.*") Scrolling down to the green "Set Up Criteria" box, he then selects bin 9.02 and submits a search request. A report is returned that lists the available SSRs for bin 9.02, complete with primers, gel patterns for different germplasm, and related maps. By going back to the SSR page, he also downloads tabular reports of map locations of all SSRs on chromosome 9, including those that have been anchored to a BAC contig. Using this information in the laboratory, members of his research group perform mapping experiments using several SSRs in bin 9.02 along with some

others in the more distal part of bin 9.03. They discover that the mid-region peak for the QTL is very near an SSR for *bnlg1372*, which is anchored to a BAC contig.

To find sequenced BACs that may harbor the earworm resistance QTL, Researcher 2 uses the search bar at the top of each MaizeGDB page to find the locus *bnlg1372*. At the top of the *bnlg1372* page, he follows a link to the contig 373 display at the Maize Sequencing Project site (http://www.maizesequence.org/). This is a rather large contig with many sequenced BACs and assigned markers. At the Maize Sequencing Project site, he uses the export function (a button at the left margin) to view a text list of all the markers and sequenced BAC clones that are available on the Finger Print Contig physical map. He finds that *bnlg1372* is assigned to the region "19742100,1974700," encompassed by the sequenced BAC clone, c0324E10. This information provides coordinates for viewing the region on a large contig associated with *bnlg1372*, the sequence of BAC c0324E10, and any other BACs nearby. Researcher 2 sequences candidate regions in diverse germplasm and conducts association analysis using silk maysin levels as a trait. This may require other information about nearby markers, which also are accessible via MaizeGDB [28, 29].

Although these investigations may require the development of further sequenced-based markers, Researcher 2 hopes that useful markers already exist and decides to explore MaizeGDB for any other sequences or primer-based markers already assigned to the region of interest including SNPs and indels. To do this from the locus page for *bnlg1372*, he clicks on the link to the most current IBM neighbors map listed, then explores the "sequence" and "primer" view versions of the map by clicking on the relevant links at the top of the page just under the map name. The primer view shows primers associated with mapping probes along with the name of the probes—just what he needs to get going with the association mapping work.

3.3. Applied

Interested in breeding plants for organic sweet corn production, Researcher 3 has decided to use molecular markers to select for high maysin content, which would increase resistance to the corn earworm—a cause of significant damage to sweet corn [30]. Although plants could be genetically modified to carry the genes that confer high maysin levels in silks (e.g., see [31]), Researcher 3's farming clients require that their product be certified as both organic and "GMO-free." To meet the producers' needs, he has decided to pursue a marker-assisted selection program to create high maysin sweet inbred lines, which he will use to generate single-cross hybrids. To get started with the work, he searches MaizeGDB to find references, markers, and stocks for the project. Described here are the details on how he could use MaizeGDB to (1) access stocks known to have high maysin content directly and (2) locate relevant stocks based upon associated data with no prior knowledge of which stocks he wants to find. An outline of how he uses MaizeGDB to identify relevant selectable markers for tracking the various QTL associated with maysin accumulation also is described.

In the instance of looking for particular stocks, Researcher 3 has identified GT114 as a high maysin line from [25]. Using the green search bar at the top of any MaizeGDB page, he searches "stocks" for "GT114." At that page, he sees a brief annotation stating that GT114 is a poor pollen producer and makes a note of that observation and plans to cross by IA453 and IA5125, sweet lines that produce pollen well, to ameliorate this potential difficulty. Clicking the link to GT114, he sees that it is an inbred line derived from GT-DDSA (DD Syn A) in Georgia, and it is made available via GRIN. Selecting the link for GRIN, a page opens at that site (http://www.ars-grin.gov/cgi-bin/npgs/html/search.pl?PI+511314). Listed there are the *Crop Science* Registration data, availability (noted as currently unavailable, but a call to Mark Millard, maize curator at the maintenance site indicates that he could access that stock in limited quantities if current resources allow), and an image of bulk kernels among other information. The image of bulked kernels is especially revealing: the kernels are yellow and the cob fragments appear red. Aware that a red cob would be unacceptable for breeding sweet corn (the red pigment could cause quite a mess for those cooking and eating corn on then cob), he decides to search MaizeGDB for other available high maysin stocks.

After a literature search of breeding stocks with a white cob that might still produce maysin in the silks, Researcher 3 starts searching stocks for those known to carry the *P1-wwb* allele, a dominant allele of the *p1* locus that confers white pericarp, white cob, and browning silks. By clicking the "Data Centers-Genetic-Stocks" link from the MaizeGDB homepage, he arrives at the Stocks Data Center page (which is also accessible via the "Useful pages" pull down at the top of every MaizeGDB page). He uses the Advanced Search box to limit the query by variation to those stocks associated with the allele *P1-wwb*. A number of the stocks returned on the results page have been evaluated for silk maysin accumulation (per associated publications) and could be further investigated as potential breeding stocks.

Although the *p1* gene accounts for much of the variability in maysin accumulation [32], association and QTL analyses for candidate genes for maysin accumulation also have identified *anthocyaninless1* (*a1*), *colorless2* (*c2*), and *white pollen1* (*whp1*) as contributing significantly [32, 33]. Researcher 3 can track the dominant *P1-wwb* allele visually by selecting for browning silks given that the sweet lines he will be using in the breeding program have silks that do not brown, but tracking the other factors will require the use of molecular markers. To find molecular markers to select for desirable alleles of, for example, *a1*, Researcher 3 uses the search menu at the top of any page at MaizeGDB to find "loci" using the query "a1." The results page (http://www.maizegdb.org/cgi-bin/displaylocusresults.cgi?term=a1) lists many loci with a1 as a substring, but shows the exact match (the *a1* locus) at the top of the list. Clicking on that link shows the *a1* locus page (http://www.maizegdb.org/cgi-bin/displaylocusrecord.cgi?id=12000), which lists useful information including six probes/molecular markers that could be used for tracking useful *a1* alleles. Using the same process, he also

finds markers for the *c2* and *whp1* loci and sets to work determining which markers to use for his selections.

4. CONCLUSIONS

Because MaizeGDB stores and makes accessible data of use for a variety of applications, it is a resource of interest to maize researchers spanning many disciplines. The fact that basic research outcomes are tied to translational and applied data enables all researcher types to utilize the MaizeGDB resource to further their research goals, and connections to external resources like Gramene, NCBI, and GRIN make it possible for researchers to find relevant resources quickly, irrespective of storage location.

At present, maize geneticists are at the cusp of a milestone: the genome of the maize inbred B73 is being sequenced in the U.S., with anticipated completion in 2008. In addition, scientists working in Mexico at Langebio (the National Genomics for Biodiversity Laboratory) and Cinvestav (Centro de Investigacion y Estudios Avanzados) have announced through a press release (July 12, 2007) that they completely sequenced 95% of the genes with 4X coverage in a native Mexican popcorn called palomero, though the data have not yet been released and the quality of the data is unknown (see http://www.bloomberg.com/apps/news?pid=20601086&sid=aO.Xj8ybAExI&refer=latin_america). At present and as more maize sequence becomes available relating sequences to the *existing* compendium of maize data is the primary need that must be met for maize researchers in the immediate future. Creating and conserving relationships among the data will enable researchers to ask and answer questions about the structure and function of the maize genome that previously could not be addressed. To address this need, MaizeGDB personnel will create a "genome view" by adopting and customizing a Genome Browser that could be used to integrate the outcomes of the Maize Genome Sequencing Project. For genome browser functionality, basic researchers have an interest in visualizing genome structure, gene models, functional data, and genetic variability. Translational researchers would like to be able to assign values to genomic and genetic variants (e.g., the value of a particular allele in a given population) and to view those values within a genomic context. Applied researchers are interested in tagging variants for use as selectable markers and retrieving tags for particular regions of the genome. To best meet these researchers' needs, the "genome view" will allow researchers to visualize a gene within its genomic context and a soon to be created "pathway view" will enable the visualization of a gene product within the context of relevant metabolic pathways annotated with Plant Ontology (http://www.plantontology.org/) [34] and Gene Ontology (http://www.geneontology.org/index.shtml) [35] terms. By making sequence information more easily accessible and fully integrated with other data stored at MaizeGDB, it will become possible for researchers to begin to investigate how sequence relates to the architecture of the maize chromosome complement. How are the chromosomes arranged? Is it possible to relate the genetic and cytological maps to the assembled genome sequence? Are there sequences present at centromeres that signal the cell to construct kinetochores, the machines that ensure proper chromosome segregation to occur, at the correct site? MaizeGDB aims to enable researchers to discover answers to such queries that will enhance the quality of basic maize research and ultimately the value of maize as a crop. It will become possible to interrogate the database to find answers to these and other complex questions, and the content of the genome can better be related to its function, both within the cell and to the plant as a whole. Convergence of traditional biological investigation with the knowledge of genome content and organization is currently lacking, and is a new area of research that will open up once a complete genome sequence and a method for searching through the whole of the data are both in place. It is the ability to investigate and answer such basic research questions that will serve as the basis for devising sound methods to breed better plants. Once the relationships among sequence data and more traditional maize data like genotypes, phenotypes, stocks, and so forth have been captured, it is important that those data be presented to researchers in a way that can be easily understood without requiring that they have any awareness of how the data are actually stored within a database. It is these needs—creating connections between sequence and traditional genetic data, improving the interface to those data, and determining how sequence data relate to the overall architecture of the maize chromosome complement—that the MaizeGDB team seeks to fulfill in the very near future.

ACKNOWLEDGMENTS

We are indebted to the community of maize researchers and the MaizeGDB Working Group (Drs. Volker Brendel, Ed Buckler, Karen Cone, Mike Freeling, Owen Hoekenga, Lukas Mueller, Marty Sachs, Pat Schnable, Tom Slezak, Anne Sylvester, and Doreen Ware) for their continued enthusiasm, help, and guidance. We are grateful to Dr. Bill Beavis for giving us the idea to highlight MaizeGDB's utility for the three user types. We thank Drs. Mike McMullen, Jenelle Meyer, Bill Tracy, and Tom Peterson for helpful discussions concerning *p1* and maysin research as well as Dr. Damon Lisch for suggestions on seminal discoveries in maize and Mark Millard at the USDA-ARS North Central Regional Plant Introduction Station for samples of corn with red cobs.

REFERENCES

[1] C. J. Lawrence and V. Walbot, "Translational genomics for bioenergy production from fuelstock grasses: maize as the model species," *The Plant Cell*, vol. 19, no. 7, pp. 2091–2094, 2007.

[2] B. McClintock, "The origin and behavior of mutable loci in maize," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, no. 6, pp. 344–355, 1950.

[3] N. Fedoroff, S. Wessler, and M. Shure, "Isolation of the transposable maize controlling elements *Ac* and *Ds*," *Cell*, vol. 35, no. 1, pp. 235–242, 1983.

[4] H. B. Creighton and B. McClintock, "A correlation of cytological and genetical crossing-over in Zea mays," *Proceedings of the*

National Academy of Sciences of the United States of America, vol. 17, no. 8, pp. 492–497, 1931.

[5] E. H. Coe Jr., "The properties, origin, and mechanism of conversion-type inheritance at the *B* locus in maize," *Genetics*, vol. 53, no. 6, pp. 1035–1063, 1966.

[6] C. J. Lawrence, M. L. Schaeffer, T. E. Seigfried, D. A. Campbell, and L. C. Harper, "MaizeGDB's new data types, resources and activities," *Nucleic Acids Research*, vol. 35, database issue, pp. D895–D900, 2007.

[7] D. H. Ware, P. Jaiswal, J. Ni, et al., "Gramene, a tool for grass genomics," *Plant Physiology*, vol. 130, no. 4, pp. 1606–1613, 2002.

[8] *Achievements of the National Plant Genome Initiative and New Horizons in Plant Biology*, National Academies Press, Washington, DC, USA, 2008.

[9] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler, "GenBank," *Nucleic Acids Research*, vol. 35, database issue, pp. D21–D25, 2007.

[10] Q. Dong, C. J. Lawrence, S. D. Schlueter, et al., "Comparative plant genomics resources at PlantGDB," *Plant Physiology*, vol. 139, no. 2, pp. 610–618, 2005.

[11] J. Quackenbush, F. Liang, I. Holt, G. Pertea, and J. Upton, "The TIGR gene indices: reconstruction and representation of expressed gene sequences," *Nucleic Acids Research*, vol. 28, no. 1, pp. 141–145, 2000.

[12] S. F. Altschul, T. L. Madden, A. A. Schaffer, et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[13] F. Wei, E. H. Coe Jr., W. Nelson, et al., "Physical and genetic structure of the maize genome reflects its complex evolutionary history," *PLoS Genetics*, vol. 3, no. 7, p. e123, 2007.

[14] J. Gardiner, S. Schroeder, M. L. Polacco, et al., "Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization," *Plant Physiology*, vol. 134, no. 4, pp. 1317–1326, 2004.

[15] R. Lerdorf, P. MacIntyre, and K. Tatroe, *Programming PHP*, O'Reilly, Sebastopol, Calif, USA, 2006.

[16] L. Wall, T. Christiansen, and J. Orwant, *Programming Perl*, O'Reilly, Cambridge, Mass, USA, 2000.

[17] R. Scholl, M. M. Sachs, and D. Ware, "Maintaining collections of mutants for plant functional genomics," *Methods in Molecular Biology*, vol. 236, pp. 311–326, 2003.

[18] S. Carpenter, "Science careers. Carving a career in translational research," *Science*, vol. 317, no. 5840, pp. 966–967, 2007.

[19] E. Bortiri, G. Chuck, E. Vollbrecht, T. Rocheford, R. Martienssen, and S. Hake, "*ramosa2* encodes a LATERAL ORGAN BOUNDARY domain protein that determines the fate of stem cells in branch meristems of maize," *The Plant Cell*, vol. 18, no. 3, pp. 574–585, 2006.

[20] E. Bortiri, D. Jackson, and S. Hake, "Advances in maize genomics: the emergence of positional cloning," *Current Opinion in Plant Biology*, vol. 9, no. 2, pp. 164–171, 2006.

[21] H. Wang, T. Nussbaum-Wagler, B. Li, et al., "The origin of the naked grains of maize," *Nature*, vol. 436, no. 7051, pp. 714–719, 2005.

[22] E. H. Coe Jr., M. G. Neuffer, and D. A. Hosington, "The genetics of corn," in *Corn and Corn Improvement*, G. F. Sprague and J. W. Dudley, Eds., pp. 81–258, American Society of Agronomy, Madison, Wis, USA, 1988.

[23] E. Grotewold, B. J. Drummond, B. Bowen, and T. Peterson, "The *myb*-homologous *P* gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset," *Cell*, vol. 76, no. 3, pp. 543–553, 1994.

[24] B. R. Wiseman, M. E. Snook, and D. J. Isenhour, "Maysin content and growth of corn earworm larvae (Lepidoptera: Noctuidae) on silks from first and second ears of corn," *Journal of Economic Entomology*, vol. 86, no. 3, pp. 939–944, 1993.

[25] P. F. Byrne, M. D. Mcmullen, M. E. Snook, et al., "Quantitative trait loci and metabolic pathways: genetic control of the concentration of maysin, a corn earworm resistance factor, in maize silks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 17, pp. 8820–8825, 1996.

[26] J. M. Thornsberry, M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E. S. Buckler, "*Dwarf8* polymorphisms associate with variation in flowering time," *Nature Genetics*, vol. 28, no. 3, pp. 286–289, 2001.

[27] M. Yano and T. Sasaki, "Genetic and molecular dissection of quantitative traits in rice," *Plant Molecular Biology*, vol. 35, no. 1-2, pp. 145–153, 1997.

[28] S. A. Flint-Garcia, A.-C. Thuillet, J. Yu, et al., "Maize association population: a high-resolution platform for quantitative trait locus dissection," *The Plant Journal*, vol. 44, no. 6, pp. 1054–1064, 2005.

[29] S. Salvi, G. Sponza, M. Morgante, et al., "Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 27, pp. 11376–11381, 2007.

[30] W. F. Tracy, "Sweet corn," in *Specialty Corns*, A. R. Hallauer, Ed., pp. 155–198, CRC Press, Boca Raton, Fla, USA, 2nd edition, 2000.

[31] E. T. Johnson, M. A. Berhow, and P. F. Dowd, "Expression of a maize *Myb* transcription factor driven by a putative silk-specific promoter significantly enhances resistance to *Helicoverpa zea* in transgenic maize," *Journal of Agricultural and Food Chemistry*, vol. 55, no. 8, pp. 2998–3003, 2007.

[32] J. D. F. Meyer, M. E. Snook, K. E. Houchins, B. G. Rector, N. W. Widstrom, and M. D. McMullen, "Quantitative trait loci for maysin synthesis in maize (*Zea mays* L.) lines selected for high silk maysin content," *Theoretical and Applied Genetics*, vol. 115, no. 1, pp. 119–128, 2007.

[33] S. J. Szalma, E. S. Buckler IV, M. E. Snook, and M. D. McMullen, "Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks," *Theoretical and Applied Genetics*, vol. 110, no. 7, pp. 1324–1333, 2005.

[34] K. Ilic, E. A. Kellogg, P. Jaiswal, et al., "The plant structure ontology, a unified vocabulary of anatomy and morphology of a flowering plant," *Plant Physiology*, vol. 143, no. 2, pp. 587–599, 2007.

[35] M. Ashburner, C. A. Ball, J. A. Blake, et al., "Gene ontology: tool for the unification of biology. The gene ontology consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 387812, 5 pages
doi:10.1155/2008/387812

Resource Review

PPNEMA: A Resource of Plant-Parasitic Nematodes Multialigned Ribosomal Cistrons

Francesco Rubino,[1] Amalia Voukelatou,[1] Francesca De Luca,[2] Carla De Giorgi,[1] and Marcella Attimonelli[1]

[1] *Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, Via E. Orabona 4, 70126 Bari, Italy*
[2] *Sezione di Bari, Istituto per la Protezione delle Piante del CNR, Via Amendola 165, 70126 Bari, Italy*

Correspondence should be addressed to Marcella Attimonelli, m.attimonelli@biologia.uniba.it

Received 30 August 2007; Accepted 23 July 2008

Recommended by Chunguang Du

Plant-parasitic nematodes are important pests of crop plants worldwide, and also among the most difficult animals to identify. Their identification based on nuclear ribosomal DNA (rDNA) cistron (18S, 28S, and 5.8S RNA genes, and internal transcribed spacers, ITS1 and ITS2) is becoming a popular tool. Sequences from nuclear ribosomal RNA repeats have been used to demonstrate the identity of isolates from various hosts and to unravel the relationships of cryptic and complex species. In addition, the availability of RNA sequences allows study of phylogenetic relationships between nematodes, also for more complete understanding of their biology as agricultural pests. PPNEMA is a *plant-parasitic nematode* bioinformatic resource. It consists of a database of ribosomal cistron sequences from various species grouped according to nematode genera, and a search system allowing data to be extracted according to both text and pattern searching. PPNEMA offers to the scientific community a preprocessed archive of plant parasitic nematode sequences useful for nematologists. It is a tool to retrieve plant nematode multialigned sequences for phylogenetic studies or to recognize a nematode by comparing its rDNA sequence with the PPNEMA available genus specific multialignments.

1. INTRODUCTION

Plant-parasitic nematodes are devastating parasites of crop plants, reducing the overall yield or lowering the market value of crops [1, 2]. Nematodes are remarkably consistent in their anatomy [3], and their identification is essentially based on morphometric characters. In addition, as variations occur in host responses to attack by various morphologically indistinguishable populations of several parasitic species, correct species identification is fundamental for efficient nematode control. For this reason, direct examination of genetic material has, recently, been used as it represents the most powerful method for nematodes recognition.

Although phytoparasitic nematodes have evolved specific structures for their survival as parasites, these adaptations are essentially built around a basic framework of nematode anatomy. Many biological questions can thus be addressed by placing the nematode *Caenorhabditis elegans,* the best characterised multicellular organism [4], in a phylogenetic and evolutionary context, together with plant-parasitic nematodes.

The nucleotide sequences of fragments of rRNA genes have recently been obtained in various species of plant-parasitic nematodes, yielding a proper platform for both identification and taxonomic approaches [5]. Nematode ribosomal RNA genes are arranged in tandemly repeated clusters (rDNA arrays) containing the genes for 18S, 5.8S, and 26S ribosomal RNA, separated by internal transcribed spacers ITS1 and ITS2 and bordered by IGS intergenic spacers (see Figure 1). Only few sequences available in the primary nucleotide databases span the entire rDNA array, although in several cases phylogenetic relationships within different species of plant-parasitic nematodes have been obtained even when only fragments of ribosomal genes were used [6–8].

This paper describes the PPNEMA database, grouping and analysing rRNA genes sequenced in plant-parasitic nematodes and present in the primary databases. It should be noted that, although specific and important

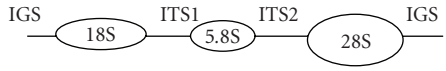


Figure 1: *rDNA cistron scheme*. Representation of nematode ribosomal RNA genes arranged in tandemly repeated clusters: rDNA array.

nematode resources are available on the web, such as WormBase ([9], http://www.wormbase.org/), Nematode.net ([10], http://www.nematode.net/), NemATOL (http://nematol.unh.edu/index.php the Comprehensive Phytopathogen Genome Resource (CPRG) (http://cpgr.tigr.org/index.html), NEMrRNA ([11], http://www.nemamex.ucr.edu/rna/), and NEMBASE ([12], http://www.nematodes.org/nematodeESTs/nembase.html) a database resource for nematode EST datasets. The last three contain sequences from rRNA genes and therefore are likely to be of interest to any reader of this article. However, the innovative aspect of PPNEMA is the availability of the rDNA sequences in groups of multialigned sequences.

2. MATERIALS AND METHODS

2.1. Data source

Sequence data are derived from primary databases (EMBL/GenBank/DDBJ) using the retrieval systems SRS and Entrez. Since a single entry in the primary database can contain more elements of the same cistron, the extraction of the sequences of each element is supported by the information contained in the entry's features table. Moreover, in order to reduce false negatives obtained through the retrieval system, the extracted data are compared to the whole database by applying the Blast database similarity searching system. In this way, sequences of interest for PPNEMA (plant parasitic nematode rDNA sequences) are found, which are not correctly annotated in primary databases and which hence are lost during the text searching retrieval.

2.2. Software

Extracted sequences are analysed by applying (i) the CleanUP software [13] which allows the detection of redundant sequences, and (ii) the ClustalW [14] software which produces the genus/cistron_element specific multialignment. Data so obtained are stored in the PPNEMA database. The database is physically based on MySQL DBMS [15], and the web application is based on an application framework written in Python.

3. RESULTS

3.1. Aim of PPNEMA

The aim of PPNEMA is to offer end-users a ready to use compilation of multialigned plant-parasitic nematode ribosomal cistrons, of which thousands of sequences are available in primary nucleotide databases (EMBL/GenBank/DDBJ). The sequences of several rRNA regions retrieved from primary databases are analysed and stored in the PPNEMA database, grouped by each nematode genus. Thus, PPNEMA is a preprocessed archive of data ready to be used from researchers interested in phylogenetic studies on phytoparasitic nematodes, or to recognize a nematode by comparing its rDNA cistrons with the PPNEMA available genus specific multialigned groups.

3.2. Structure of PPNEMA database

PPNEMA is a well-integrated, web-based, **plant-parasitic nematode** bioinformatics resource, allowing the storage, query, and analysis of phytoparasitic rDNA sequences. PPNEMA consists of a *database* of ribosomal cistron sequences from various species of plant-parasitic nematodes, grouped according to nematode genera and of a search system allowing data to be extracted according to both text and pattern searching. Each entry in the PPNEMA database refers to a complete or partial cistron element of a single isolate within a nematode species; it is identified by a code defining species and function. Sequences derived from the various species are multialigned within each nematode genus. However, since not all sequences span the entire rDNA array, separate multialignments have been produced for single rRNA genes or for portions of the same gene separately, depending on sequence availability. Each multialignment defines a group. The presence within a genus group of perfectly matching sequences (here defined as redundant) is determined by CleanUP software. Redundant sequences are stored in the database, linked to the group containing the group-reference sequence, but they are not enclosed in the multialignment of that group. Thus, each entry in the database is related to a species-specific functional element. Several entries are associated in a group. Several groups are available for the same genus and the same functional element. Figure 2 shows the database structure, and Figure 3 shows an example of a PPNEMA database entry.

3.3. Updating of PPNEMA database

Generally speaking, data in primary databases are organised in such a way that each entry is related to a genomic fragment of DNA related to a genome or one or more genes, complete or partial, so that the extraction of sequences related to the same cistronic element has been so far performed through, very time consuming, a nonautomated procedure. However, we have planned, but not yet implemented, a new updating procedure which will allow the automatic extraction from primary databases of the newly sequenced phytoparasites nematodes rDNA. The automatic procedure will generate one sequence for each entire or partial cistron element of a specific species; this sequence will be analysed through the application of the PPNEMA "characterizing" tool that will guide the automatic procedure in defining its better fitting multialignment group.

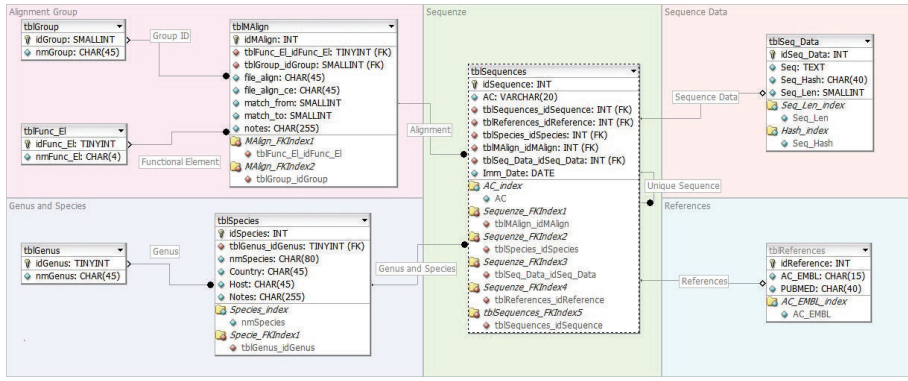
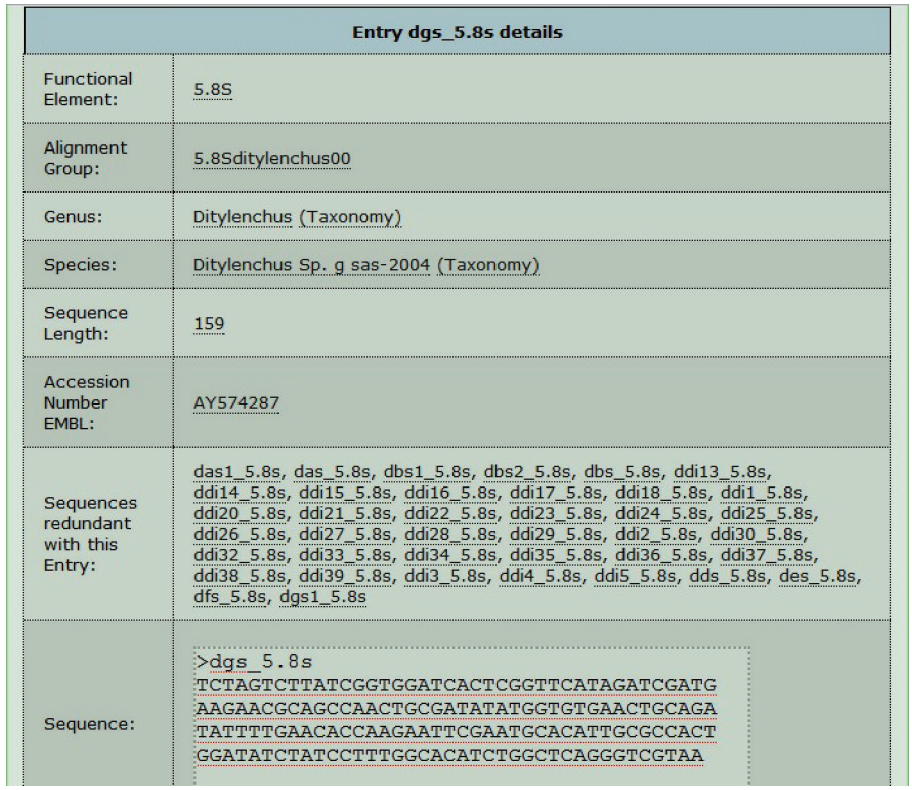


Figure 2: *PPNEMA relational database design*. It includes 8 tables, storing complete set of PPNEMA data linked to each other.



Entry dgs_5.8s details	
Functional Element:	5.8S
Alignment Group:	5.8Sditylenchus00
Genus:	Ditylenchus (Taxonomy)
Species:	Ditylenchus Sp. g sas-2004 (Taxonomy)
Sequence Length:	159
Accession Number EMBL:	AY574287
Sequences redundant with this Entry:	das1_5.8s, das_5.8s, dbs1_5.8s, dbs2_5.8s, dbs_5.8s, ddi13_5.8s, ddi14_5.8s, ddi15_5.8s, ddi16_5.8s, ddi17_5.8s, ddi18_5.8s, ddi1_5.8s, ddi20_5.8s, ddi21_5.8s, ddi22_5.8s, ddi23_5.8s, ddi24_5.8s, ddi25_5.8s, ddi26_5.8s, ddi27_5.8s, ddi28_5.8s, ddi29_5.8s, ddi2_5.8s, ddi30_5.8s, ddi32_5.8s, ddi33_5.8s, ddi34_5.8s, ddi35_5.8s, ddi36_5.8s, ddi37_5.8s, ddi38_5.8s, ddi39_5.8s, ddi3_5.8s, ddi4_5.8s, ddi5_5.8s, dds_5.8s, des_5.8s, dfs_5.8s, dgs1_5.8s
Sequence:	>dgs_5.8s TCTAGTCTTATCGGTGGATCACTCGGTTCATAGATCGATG AAGAACGCAGCCAACTGCGATATATGGTGTGAACTGCAGA TATTTTGAACACCAAGAATTCGAATGCACATTGCGCCACT GGATATCTATCCTTTGGCACATCTGGCTCAGGGTCGTAA

Figure 3: *Example of a PPNEMA entry*. Entry dgs_5.8S (PPNEMA ID) shows functional element, PPNEMA group ID to which entry belongs, Genus and Species names, sequence length, EMBL accession number, list of redundant sequences, and sequence, in FASTA format, which can be downloaded.

3.4. Contents of PPNEMA database

PPNEMA currently contains 2405 sequences, organised in 208 Alignment Groups from 26 genera. Because the plant-parasitic nematode RNA cistrons are not all conserved between and within genera, it is practically impossible to produce one multialignment for each element not only among all species but also among species of the same genera. This means that there are associated multialigned sequences in different groups for the same genera and, in order to have a reference, each multialignment was produced both with and without *Caenhorabditis elegans*, used as outgroup guide. More detailed information about database contents may be obtained through the Statistics option available through the PPNEMA site. Figure 4 shows data obtainable from the statistic option in PPNEMA.

3.5. Functions of PPNEMA

Starting from the PPNEMA home page, two main options are available: search PPNEMA and browse PPNEMA. Both are organised in subsections. Search PPNEMA is used to retrieve specific sequences and/or aligned groups of sequences, through basic search, advanced search, or pattern

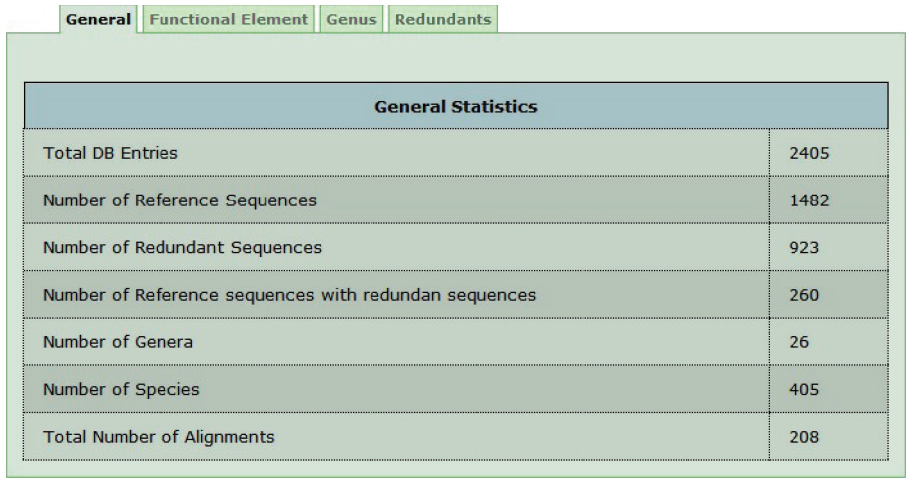


Figure 4: *General statistics* about PPNEMA data. It is also possible to obtain statistical information centred on functional elements or on redundant data content.

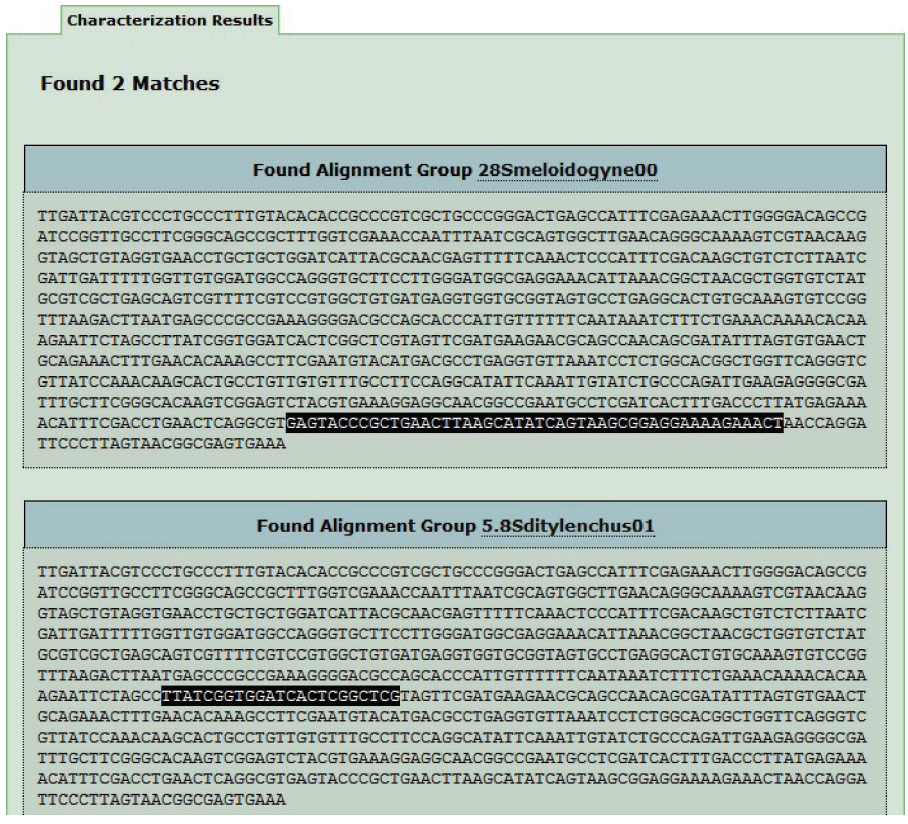


Figure 5: *Result of an anonymous sequence characterization*. The submitted sequence contains 2 fragments matching part of 28smeloidogyne00 and 5.8sditylenchus01 multialignment consensus.

search. Basic search allows retrieval of data according to the following criteria: functional element, genus name, species name, sequence length range. Advanced search allows more elaborate queries combining the various retrieval criteria through the logical operators AND or OR; selection criteria include the possibility to select data through a pattern searching option implemented on the basis of regular expressions. A regular expression is a powerful way of specifying a pattern for a complex search. The primer for the regular expressions used by MySQL is available through the help PPNEMA function. From the advanced search, a pattern search option is implemented within the search menu. The difference between the options "pattern search" and "pattern search through advanced search" is the

output format of the retrieved data. Search results may be grouped by alignment, reference sequences, or redundancy groups. Retrieved sequences grouped by alignment are ready to be analysed with phylogenetic tools. Lastly, the option "characterising a new sequence" can search group/s of multialigned sequences, the consensus sequence of which, defined through regular expressions, matches submitted end-user sequence whose function and/or species paternity is undefined or not completely defined. Figure 5 shows an example of the output obtained by submitting a new sequence for its characterisation. The *browse DB* option allows the list of database species, multialignments, and sequences to be viewed. Starting from any element in the list, related information available in both PPNEMA and cross-referenced databases (e.g., EMBL, GenBank, and Taxonomy) can be obtained. Lastly, the PPNEMA resource contains online help, statistics tables, and an option, designed but not yet implemented, allowing submission of the new sequences on behalf of registered end-users. Registration is already implemented. In progress is the production of the phylogenetic trees, there where data which are variable enough to be informative by the evolutionary point of view. The produced trees will be available in the PPNEMA database.

4. CONCLUSIONS

The PPNEMA database is very helpful in identifying plant parasitic nematodes on a molecular basis, since the availability of multialigned sequences for nematode genera represents a map, on which the sequence of any unidentified nematode species can be located. In addition, the existence of several entries for the same species gives information on the extent of intraspecies variability and can thus help in discriminating between variants or new nematode species. This information is important in view of the expected rapid growth of sequence data from intraspecific studies aimed at both population migration and identification of different pathotypes.

It is important to emphasise that the more sequences obtained, the more informative the PPNEMA database will become. Periodical updating is foreseen, but contribution from sequence producers is welcome.

In conclusion, the perspective is extensive use of the PPNEMA database by plant pathologists who are not specialised in molecular biology.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Annalisa Marsico for her contribution to the database design during her stay in our Department within the Master program in BIOINFORMATICA "Alberto del Lungo" organised at the Siena University (Italy). This work was supported by the University of Bari and by "PROGETTO DI RICERCA MIUR-PNR FIRB" "Laboratorio Internazionale di Bioinformatica."

REFERENCES

[1] J. N. Sasser and D. W. Freckman, "A world perspective on nematology: the role of the society," in *Vistas on Nematology*, D. W. Dickson and J. A. Veech, Eds., pp. 7–14, Society of Nematologists, Hyatttsville, Md, USA, 1987.

[2] K. R. Barker, R. S. Hussey, L. R. Krusberg, et al., "Plant and soil nematodes: societal impact and focus for the future," *Journal of Nematology*, vol. 26, no. 2, pp. 127–137, 1994.

[3] A. F. Bird and J. Bird, *The Structure of Nematodes*, Academic Press, London, UK, 1991.

[4] D. L. Riddle, T. Blumentahal, B. J. Meyer, and J. R. Priess, *C.Elegans II*, Cold Spring Harbor Laboratory Press, Woodbury, NY, USA, 1997.

[5] C. D. Giorgi, P. Veronico, F. De Luca, A. Natilla, C. Lanave, and G. Pesole, "Structural and evolutionary analysis of the ribosomal genes of the parasitic nematode *Meloidogyne artiellia* suggests its ancient origin," *Molecular and Biochemical Parasitology*, vol. 124, no. 1-2, pp. 91–94, 2002.

[6] L. Al-Banna, V. Williamson, and S. L. Gardner, "Phylogenetic analysis of nematodes of the genus *Pratylenchus* using nuclear 26S rDNA," *Molecular Phylogenetics and Evolution*, vol. 7, no. 1, pp. 94–102, 1997.

[7] A. Hugall, J. Stanton, and C. Moritz, "Reticulate evolution and the origins of ribosomal internal transcribed spacer diversity in apomictic Meloidogyne," *Molecular Biology and Evolution*, vol. 16, no. 2, pp. 157–164, 1999.

[8] S. A. Subbotin, A. Vierstraete, P. De Ley, et al., "Phylogenetic relationships within the cyst-forming nematodes (Nematoda, Heteroderidae) based on analysis of sequences from the ITS regions of ribosomal DNA," *Molecular Phylogenetics and Evolution*, vol. 21, no. 1, pp. 1–16, 2001.

[9] T. Bieri, D. Blasiar, P. Ozersky, et al., "WormBase: new content and better access," *Nucleic Acids Research*, vol. 35, database issue, pp. D506–D510, 2007.

[10] T. Wylie, J. C. Martin, M. Dante, et al., "Nematode.net: a tool for navigating sequences from parasitic and free-living nematodes," *Nucleic Acids Research*, vol. 32, database issue, pp. D423–D426, 2004.

[11] S. A. Subbotin, D. Sturhan, N. Vovlas, et al., "Application of the secondary structure model of rRNA for phylogeny: D2-D3 expansion segments of the LSU gene of plant-parasitic nematodes from the family Hoplolaimidae Filipjev, 1934," *Molecular Phylogenetics and Evolution*, vol. 43, no. 3, pp. 881–890, 2007.

[12] J. Parkinson, C. Whitton, R. Schmid, M. Thomson, and M. Blaxter, "NEMBASE: a resource for parasitic nematode ESTs," *Nucleic Acids Research*, vol. 32, database issue, pp. D427–D430, 2004.

[13] G. Grillo, M. Attimonelli, S. Liuni, and G. Pesole, "CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases," *Computer Applications in the Biosciences*, vol. 12, no. 1, pp. 1–8, 1996.

[14] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[15] MySQL 5.0 Manual—Section 12.4.2, http://dev.mysql.com/doc/refman/5.0/en/regexp.html.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 451327, 7 pages
doi:10.1155/2008/451327

Resource Review

Cross-Chip Probe Matching Tool: A Web-Based Tool for Linking Microarray Probes within and across Plant Species

Ruchi Ghanekar,[1,2] Vinodh Srinivasasainagendra,[2] and Grier P. Page[2,3]

[1] *Department of Electrical and Computer Engineering, UAB School of Engineering, University of Alabama at Birmingham, 1530 Third Avenue South, Birmingham, AL 35294-4461, USA*
[2] *Department of Biostatistics, University of Alabama at Birmingham, 1665 University Blvd, Birmingham, Al 35294-0022, USA*
[3] *Statistics and Epidemiology Unit, RTI International, Oxford Building, Suite 119, 2951 Flowers Road South, Atlanta, GA 30341-5533, USA*

Correspondence should be addressed to Grier P. Page, gpage@rti.org

Received 2 November 2007; Accepted 14 August 2008

Recommended by Chunguang Du

The CCPMT is a free, web-based tool that allows plant investigators to rapidly determine if a given gene is present across various microarray platforms, which, of a list of genes, is present on array(s), and which gene a probe or probe set queries and vice versa, and to compare and contrast the gene contents of arrays. The CCPMT also maps a probe or probe sets to a gene or genes within and across species, and permits the mapping of the entire content from one array to another. By using the CCPMT, investigators will have a better understanding of the contents of arrays, a better ability to link data between experiments, ability to conduct meta-analysis and combine datasets, and an increased ability to conduct data mining projects.

1. INTRODUCTION

Microarrays are an incredibly powerful technology that enables the rapid and relatively accurate measurement of thousands of genes in a single sample. Many different microarray platforms have been developed and each has somewhat different content and format. One key difference is the type of probe used to query a gene expression; some platforms use a single probe, and others use many probes. The probes may be short (25 base pairs) oligonucleotides (Affymetrix and NimbleGen arrays), long (50–70 bp) oligonucleotides (Operon, Agilent, CATMA), or cDNA clones (AFGC arrays). Each of the formats has its advantages and disadvantages as well as its proponents and opponents. One thing on which everybody agrees is that arrays will be a part of the experimental techniques of plant biologists for years to come.

Since there are many microarray platforms even within a single species, different investigators may use different platforms to try to address similar or complementary experimental questions or data may be collected across types using different platforms. Also, the large number of datasets that sets in the public domain allow can be used for data mining or meta-analysis if the elements can be connected. However, it is difficult to compare and combine the results due to the difficultly in matching probes across arrays with the genes, or even to determine if a given gene is on a given platform. To make matters worse, while the probe sequences on an array are constant, the genome annotation and gene models are not, and homologous genes may have different names across species. As a result, matching probes across arrays is continually evolving and needs continuing updating.

Investigators have long realized the problem of linking probes across platforms; as a result, several tools have been developed. These include Keck ARray Manager and Annotator (KARMA) [1], RESOURCERER [2], and GeneSeer [3]. Our tool has several advantages over the other tools for several reasons. None of the other tools allows investigators to query for genes within a microarray platform nor do the other tools allow queries by *Arabidopsis* Genome Initiative (AGI) annotation IDs or by TIGR tentative consensus (TC) gene IDs. Furthermore, our tool sends the results to the investigators by email as well as a web-based report making

results' tracking and storage easier. More importantly for plant researchers, only RESOURCERER has any provision for the linking of plant array data, but it has fewer array types.

We developed the CCPMT to enable investigators to rapidly determine (1) if a given gene is present across many types of array platforms within and across species, (2) which, of a list of genes, is present on array(s), and (3) which gene a probe or probe set queries. The CCPMT also maps a probe set or probe sets to a gene or genes within and across species, and permits the mapping of the contents from one array to another, both within and across species.

The CCPMT is the first tool exclusively designed for linking probes from plant microarrays within and across microarray platforms and species. A web-based tool, CCPMT, helps investigators query for annotations at probe level with probe set IDs or even at gene level with gene identifiers such as AGI, EGO [4], and TC IDs. In CCPMT, an investigator can enter either individual or multiple probe set or gene identifiers (separated by commas) in the textbox to query the CCPMT database. Checkboxes for microarray vendors provide the option of selecting multiple arrays while querying the CCPMT database. CCPMT also offers the flexibility to carry out a one-to-one comparison of microarrays. Results are displayed immediately in the web browser and are also sent through email in a *csv file format.

CCPMT has a flexible database design, and in the immediate future additional plant arrays will be added to the database; we will revise the underlying annotation and mapping for the probes based upon new genomic information.

By using the CCPMT, investigators will have a better understanding of the contents of arrays, a better ability to link data between experiments, plus the ability to more easily conduct data mining projects.

2. METHODS

2.1. Arrays selected for initial analysis

Initially we focused upon microarrays with diverse probe types (short and long oligos as well as cDNA) and for both *Poplar* and *Arabidopsis*. *Poplar* and *Arabidopsis* were chosen due to both having completely sequenced genomes and being relatively closely related species. The *Arabidopsis* arrays as tools are the Affymetrix *Arabidopsis* genome (8 K) commonly referred to as AG, Affymetrix *Arabidopsis* genome ATH1-121501 (25 K) commonly referred to as ATH1, Agilent *Arabidopsis* 2 Oligo Microarray (V2) G4136B, *Arabidopsis* Functional Genomics Consortium (AFGC) array, Complete *Arabidopsis* Transcriptome MicroArray (CATMA) array, Operon *Arabidopsis* Genome Oligo Set Version 3.0, and Affymetrix *Poplar* Genome Array. The array that we are calling AFGC actually represents all cDNA clones used in all of the AFGC arrays including the 11 k, 13 k, and 16 k arrays.

2.2. Arabidopsis data preprocessing

We obtained the probe set ID, the vendor's corresponding mapping to AGI ID (for Arabidopsis arrays), and the nucleotide sequences of the probe sets (Table 1) directly from the vendors.

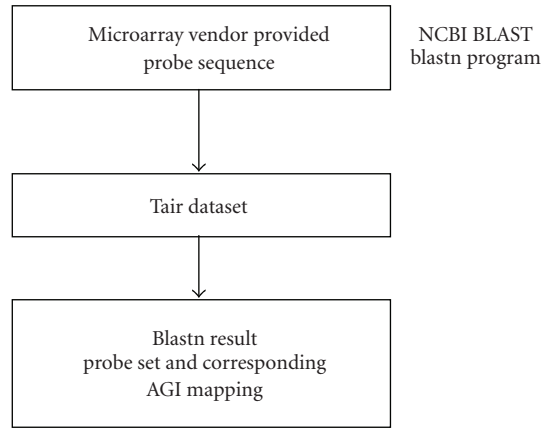


Figure 1: CCPMT *Arabidopsis* BLAST workflow. The workflow in CCPMT to get the probe set to AGI mappings is shown.

In the case of *Arabidopsis*, all vendors provided the mappings between their probe sets and the corresponding AGI gene identifiers. However, due to evolving genome annotation, we derived a new set of mappings between the probe sets and the corresponding AGI IDs. The steps of the process are illustrated in Figure 1. The mapping was accomplished using the NCBI blastn [5] program. Blastn compares a nucleotide query sequence against a nucleotide sequence database. We used two different databases for blastn analysis. For the Affymetrix and Operon probe sequences, which do not contain introns, the AGI CDS database at TAIR was used as the sequence database due to the lack of introns and the UTRs in this database. The AGI CDS dataset is based on the TAIR6.0 release version, and was released in November 2005. For the AFGC and CATMA arrays, which do contain some intronic and UTR sequences, the AGI Transcripts dataset was used. The AGI Transcripts dataset includes all of the coding sequences from *Arabidopsis,* as well as containing the UTRs. Neither database contained intronic sequence. The AGI Transcripts dataset used the TAIR6.0 release version and was released in November 2005. The blastn expected value and percent identity cut-off were 10^{-4} and 98%, respectively.

2.3. Poplar data preprocessing

About 27% of the *Poplar* sequence have significant homology to *Arabidopsis* protein-coding sequences [6] and have been sequenced. Unlike *Arabidopsis*, *Poplar* does not have a universal gene annotation ID; so in CCPMT *Poplar,* probe sets are mapped within the species using the TIGR TC IDs and across plant species using the EGO database. The *Poplar* target sequences were sequence-aligned with the TIGR *Poplar* TC dataset using the blastn program as shown in Figure 2. The blastn expected value and percent identity cut-off were 10^{-4} and 98%, respectively. TIGR also provides a file with a mapping of the EGO ID and the corresponding TCs for all species. From this file, the mappings between EGO

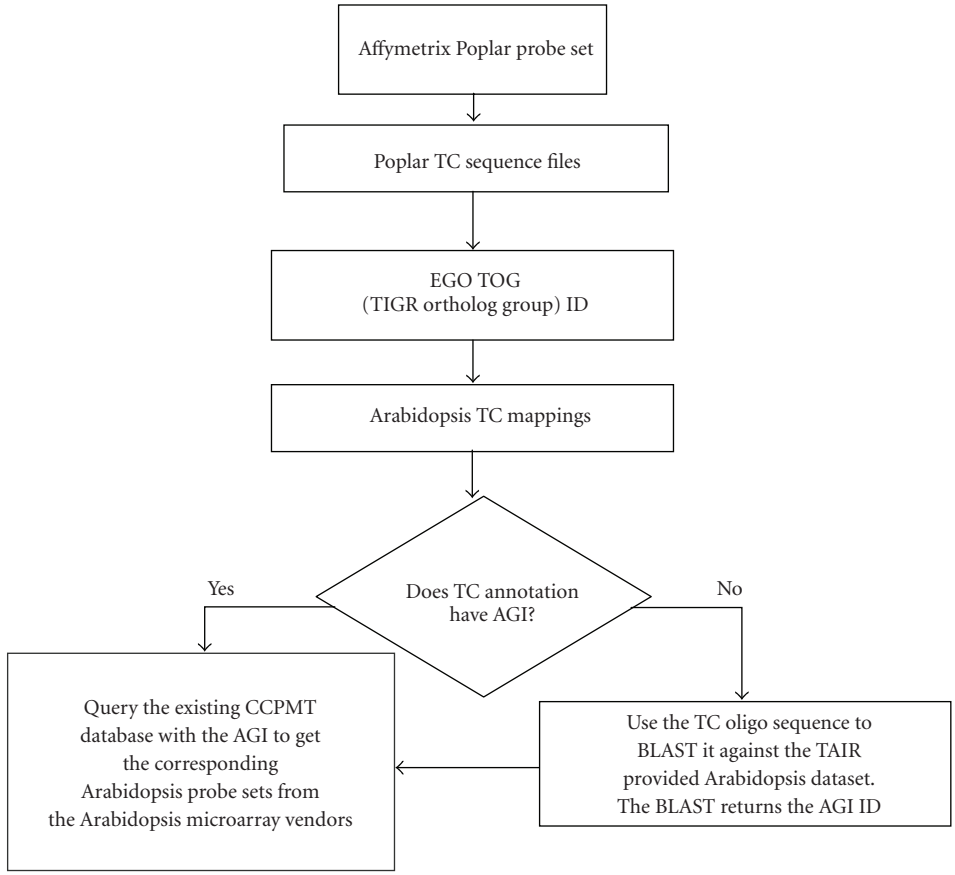


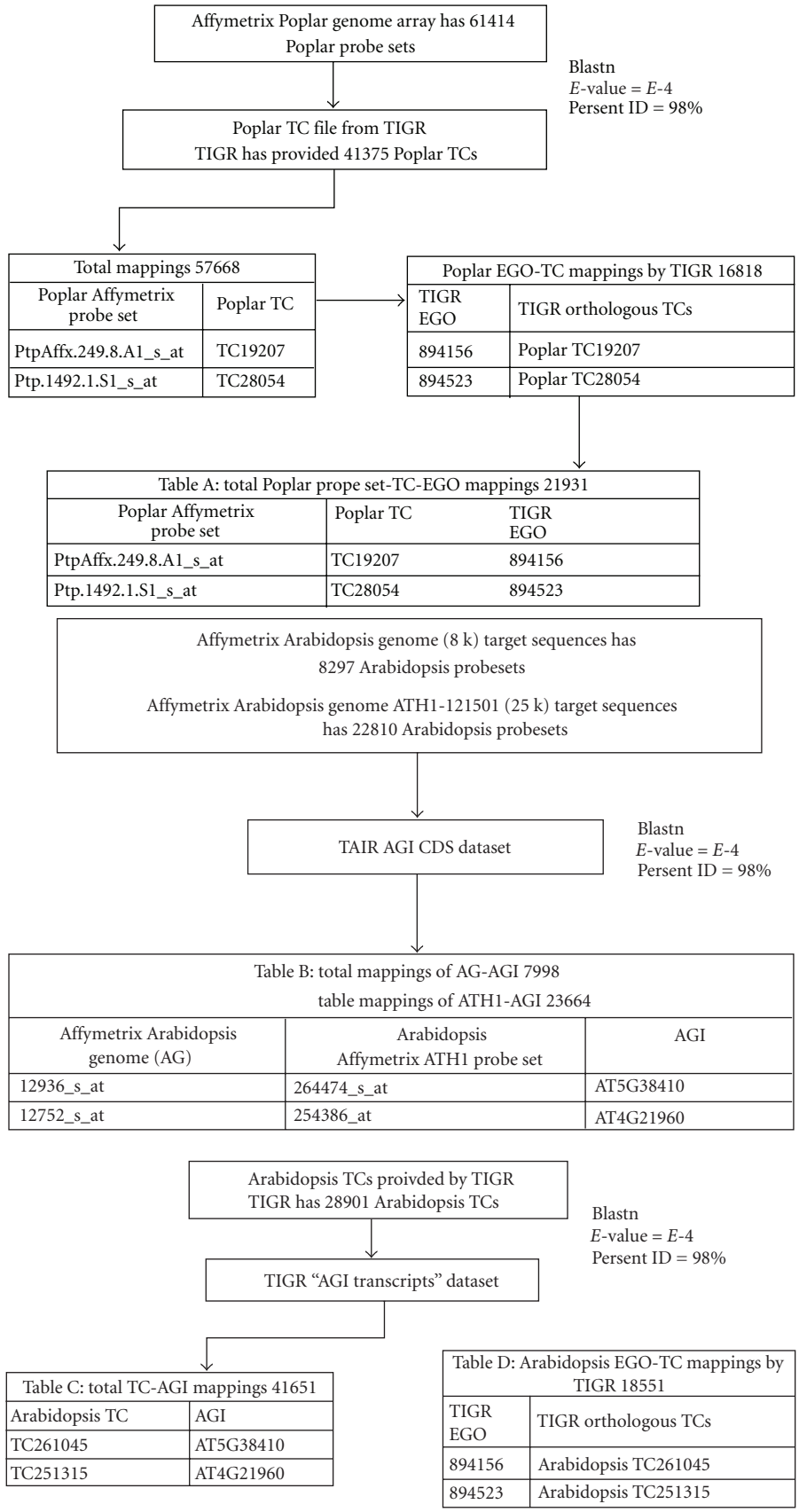
Figure 2: *Poplar-Arabidopsis* mapping. The above workflow explains the steps that were undertaken while mapping the Affymetrix *Poplar* probe set ID with the *Arabidopsis* probe set ID. TIGR EGO ID was used to go across species during mapping.

Table 1: Web pages from where plant microarray data were downloaded.

	Probe sequence file location	Vendor-provided annotation file location
Affymetrix AG	http://www.affymetrix.com/support/technical/byproduct.affx?product=atgenome1	http://www.affymetrix.com/support/technical/byproduct.affx?product=atgenome1
Affymetrix ATH1	http://www.affymetrix.com/support/technical/byproduct.affx?product=arab	http://www.affymetrix.com/support/technical/byproduct.affx?product=arab
Operon	http://omad.operon.com/download/index.php	http://omad.operon.com/download/index.php
CATMA	ftp://ftp.arabidopsis.org/home/tair/Microarrays/CATMA/	ftp://ftp.arabidopsis.org/home/tair/Microarrays/CATMA/
AFGC	ftp://ftp.arabidopsis.org/home/tair/Microarrays/AFGC/	ftp://ftp.arabidopsis.org/home/tair/Microarrays/AFGC/
Agilent	NA (do not provide sequence files)	http://www.chem.agilent.com/Scripts/PDS.asp?lPage=37068
Affymetrix *Poplar* Genome Array	http://www.affymetrix.com/support/technical/byproduct.affx?product=poplar	http://www.affymetrix.com/support/technical/byproduct.affx?product=poplar

IDs and the corresponding *Arabidopsis* and *Poplar* TCs were parsed. The mapping of the TC to EGOs was assumed to be correct. In the future, any plant species with genes mapping to an EGO ID can be easily incorporated into CCPMT. Mapping the *Arabidopsis* TCs to their corresponding AGI IDs was achieved by using the *Arabidopsis* TC sequences (TIGR provides this file) and sequence-aligning with the TAIR "AGI Transcripts" dataset using blastn. Based on the cut-offs used there is the one-to-many mapping at several stages. A probe set can map to multiple genes, and multiple probe sets can map to one gene (Table 2).

As an example, Figure 3 illustrates the mapping of the Affymetrix *Poplar* Genome Array with the Affymetrix AG and Affymetrix ATH1 arrays; similar processes are used for


Affymetrix Poplar genome array has 61414 Poplar probe sets
Blastn
E-value = E-4
Persent ID = 98%
Poplar TC file from TIGR
TIGR has provided 41375 Poplar TCs
Total mappings 57668
Poplar Affymetrix probe set
Poplar TC
PtpAffx.249.8.A1_s_at
TC19207
Ptp.1492.1.S1_s_at
TC28054
Poplar EGO-TC mappings by TIGR 16818
TIGR EGO
TIGR orthologous TCs
894156
Poplar TC19207
894523
Poplar TC28054
Table A: total Poplar prope set-TC-EGO mappings 21931
Poplar Affymetrix probe set
Poplar TC
TIGR EGO
PtpAffx.249.8.A1_s_at
TC19207
894156
Ptp.1492.1.S1_s_at
TC28054
894523
Affymetrix Arabidopsis genome (8 k) target sequences has 8297 Arabidopsis probesets
Affymetrix Arabidopsis genome ATH1-121501 (25 k) target sequences has 22810 Arabidopsis probesets
TAIR AGI CDS dataset
Blastn
E-value = E-4
Persent ID = 98%
Table B: total mappings of AG-AGI 7998
table mappings of ATH1-AGI 23664
Affymetrix Arabidopsis genome (AG)
Arabidopsis Affymetrix ATH1 probe set
AGI
12936_s_at
264474_s_at
AT5G38410
12752_s_at
254386_at
AT4G21960
Arabidopsis TCs proivded by TIGR
TIGR has 28901 Arabidopsis TCs
Blastn
E-value = E-4
Persent ID = 98%
TIGR "AGI transcripts" dataset
Table C: total TC-AGI mappings 41651
Arabidopsis TC
AGI
TC261045
AT5G38410
TC251315
AT4G21960
Table D: Arabidopsis EGO-TC mappings by TIGR 18551
TIGR EGO
TIGR orthologous TCs
894156
Arabidopsis TC261045
894523
Arabidopsis TC251315

Union of table B, table C and table D

Table E: total AG-AGI-TC-EGO mappings 7823 table ATH1-AGI-TC-EGO mappings 20051				
AG probe set	ATH1probe set	AGI	Arabidopsis TC	EGO
12936_s_at	264474_s_at	AT5G38410	TC261045	894156
12752_s_at	254386_at	AT4G21960	TC251315	894523

Union of table A and table E

7744 mappings between Affymetrix Poplar genome array probe sets and Affymetrix AG probe sets 17297 mappings between Affymetrix Poplar genome array probe sets and Affymetrix ATH1 probe sets					
Poplar Affymetrix probe set	Poplar TC	Arabidopsis TC	Arabidopsis AGI	Arabidopsis Affymetrix AG probe set	Arabidopsis Affymetrix ATH1 probe set
PtpAffx.249.8.A1_s_at	TC19207	TC261045	AT5G38410	12936_s_at	264474_s_at
Ptp.1492.1.S1_s_at	TC28054	TC251315	AT4G21960	12752_s_at	254386_at

Figure 3: Workflow for the mapping between Affymetrix *Poplar*, Affymetrix AG, and Affymetrix ATH1 arrays.

Table 2: Comparing microarray vendor and CCPMT mappings.

Type of match	Affymetrix AG	Affymetrix ATH1	Operon	CATMA	AFGC
Number of probes per array type	8297	22810	29954	24576	19108
Nil entries from vendor (no mapping for these probes)	141	250	936	2969	2823
Absent-vendor; present-blast	0	0	0	0	1
Present-vendor; absent-blast	850	930	2335	2990	10952
Many-vendor; one-blast	124	584	0	30	117
One-vendor; many-blast	338	896	480	408	368
Exact match	6932	20193	26138	19551	6413[a]
Percentage of the vendor mapping numbers	84%	89%	87%	80%	34%

the other arrays. Table 3 contains the number of matches that were found between all possible matches among arrays.

2.4. The CCPMT application

The CCPMT (http://www.ssg.uab.edu/ccpmt/) is composed of three pieces, namely, web pages (front end), core methods, and database (back end). The CCPMT web pages are written in JSP. Once the user hits the submit button, all of the data that have been entered are sent to the core code of Java servlets. The servlets act as the core methods that process the information received from the JSP pages and query the database. MySQL is used as the back-end database to store the microarray mappings. The code underlying the CCPMT is available from the corresponding author by request.

2.5. Using the CCPMT

The CCPMT is designed to be flexible and to allow for linking probes across arrays from a variety of starting data. CCPMT can be queried either at the probe set level or with identifiers such as the probe set IDs, AGI IDs, TIGR EGO IDs, or TC IDs, and output can be and is returned in these formats as well. As CCPMT is a web application, users can type or paste their queries in a textbox and, upon submission of the queries, the results are displayed in a browser-friendly format. One can also compare entire arrays by selecting the input array and the output array from the drop-down menu.

2.6. Example of the use of CCPMT

We illustrate the utility of the CCPMT via mapping the probe set 244904_at that is found on the Affymetrix AG array to determine which probe sets on the ATH1 array query the same gene. Step 1 (illustrated in Figure S1 in Supplementary Material available online at doi:10.1155/2008/451327) shows that the user wants to map the input data using Affymetrix probe set IDs. In addition, users' email address is entered so that the results can also be sent as an attachment in comma-separated file format. The next step (see Figure S2) is to enter the probe set(s), 244904_at in this case, and the species of the probe set, and to indicate which arrays to find homologous probe sets (in this example, Affymetrix AG and Affymetrix ATH1 arrays). The results are then displayed in Figure S3 which shows that the probe set 244904_at was mapped to 244922_s_at and 244923_s_at through the respective AGI IDs and that they map to AT2G07674.

Table 3: Summary table of the number of probes that are linked between the various arrays currently in the CCPMT from the array in row to the arrays in columns. The above and below diagonal elements are slightly different for the methods we used such as Blasn, and percent identity is not always reflexive.

	Affymetrix AG	Affymetrix ATH1	AFGC	Agilent	CATMA	Operon	Affymetrix *Poplar* Genome Array
Affymetrix AG	—	7828	12170	7018	7193	8361	7744
Affymetrix ATH1	7827	—	30066	19188	20521	24636	17279
AFGC	12171	30066	—	29622	26070	30509	17793
Agilent	7018	19188	29622	—	18563	21371	16913
CATMA	7192	20521	26070	18561	—	23082	16378
Operon	8362	24636	30509	21371	23081	—	17505
Affymetrix *Poplar* Genome Array	7744	17279	17793	16912	16378	17504	—

3. DISCUSSION

Microarrays are gaining popularity in plant research. In addition, the requirement of many journals to deposit microarray data into public databases has made large amounts of data available for other investigators to use. But because there are a large number of arrays and array types, it can be difficult to compare data across datasets. We developed the CCPMT to allow investigators to identify common elements between databases rapidly and accurately.

While most vendors provide some mapping of probes to genes, in many cases the annotation is out of data or the companies use different standards for mapping. In some cases, there is considerable difference between our mapping and those provided with the arrays. This is due to at least three reasons. The first is that sequence, gene models, and annotation, especially for the incompletely sequenced genomes, can change rapidly. As a result, the provided annotation may be out of date. For example, data for CATMA and AFGC, obtained with TAIR at ftp://ftp.arabidopsis.org/home/tair/Microarrays/, had a timestamp of January 2006, but the FASTA file format has a timestamp of April 2004. The second reason for differences would be the choice of cut-off for mapping. We used >98% and E score of less than 10^{-4} for all but the AFGC arrays. Our choice of >98% is debatable, and somewhat different answers are obtained if other values are used; 98% may identify some paralogous genes, especially across species. It has not been conclusively established what level of sequence similarity is needed between a gene and a probe set for efficient binding. It is known that a single-base-pair difference in a short oligo can (with >50% of the time depending on the position of the SNP) destroy most binding. But since Affymetrix arrays usually have 11 sets of short oligos, the nonbinding of a single probe may or may not affect the overall RNA quantitation [7]. Long oligos bind relatively well with a few (1–3 bp) differences, but there is usually no redundancy of the addition of probes. cDNA clones can be quite long and only a portion of the sequence needs to be homologous for binding. A third source of difference may result from the choice of common genes. We used the TIGR EGO, but the NCBI HomoloGene (http://www.ncbi.nlm.nih.gov/sites/entrez?db=homologene) also identifies homologous genes across species. Unfortunately, these databases give slightly different mapping. We have used TIGR EGO database as it has more plant sequence data and has plant biologists devoted to curating the databases, as opposed to HomoloGene which is mammal-centric. Thus, the choice we made about cut-offs is conservative, but we have probably missed some probes with lower homology that actually do bind certain RNAs, and many others identify paralogous genes. As a result of these issues, our mapping is different from those provided by the vendor. The highest overlap is between the mapping provided by Affymetrix and the CCPMT mapping for the Affymetrix ATH1 array at 89%, while the AFGC has the lowest overlap at about 66%.

We think that the function allowing direct comparison of complete arrays is very useful for several reasons. One of the reasons why we developed the CCPMT was to allow coexpression analysis across arrays and species. This mapping in the CCPMT will be the basis of our next additions to CressExpress (http://www.cressexpress.org/), and others may use this as well for similar projects. Data from experiments that are often collected across time and different array platforms are used, which requires the mapping of probes across array platforms. This ability will be greatly amplified by the ability of the CCPMT to map data across platforms.

The annotation and sequence for genes as well as gene models are continuing to evolve, especially as additional species are sequenced. We have set up the CCPMT to allow for us to rapidly change the various portions of the database and mapping as data change. We plan to revise the CCPMT based upon new genomic information.

CCPMT currently has six *Arabidopsis* microarray arrays and one *Poplar* microarray. The tool was designed in such a way that one can easily incorporate a new microarray vendor for the current plant species as well as for new plant species. In the near future, we will rule out mapping for all Affymetrix-provided arrays for plant species, as well as those long oligo arrays from Operon and Agilent.

REFERENCES

[1] K.-H. Cheung, J. Hager, D. Pan, et al., "KARMA: a web server application for comparing and annotating heterogeneous microarray platforms," *Nucleic Acids Research*, vol. 32, web server issue, pp. W441–W444, 2004.

[2] J. Tsai, R. Sultana, Y. Lee, et al., "Resourcerer: a database for annotating and linking microarray resources within and across species," *Genome Biology*, vol. 2, no. 11, pp. 1–4, 2001.
[3] A. J. Olson, T. Tully, and R. Sachidanandam, "GeneSeer: a sage for gene names and genomic resources," *BMC Genomics*, vol. 6, article 134, 2005.
[4] J. Quackenbush, F. Liang, I. Holt, G. Pertea, and J. Upton, "The TIGR Gene Indices: reconstruction and representation of expressed gene sequences," *Nucleic Acids Research*, vol. 28, no. 1, pp. 141–145, 2000.
[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.
[6] B. Stirling, Z. K. Yang, L. E. Gunter, G. A. Tuskan, and H. D. Bradshaw Jr., "Comparative sequence analysis between orthologous regions of the *Arabidopsis* and *Populus* genomes reveals substantial synteny and microcollinearity," *Canadian Journal of Forest Research*, vol. 33, no. 11, pp. 2245–2251, 2003.
[7] J. O. Borevitz, D. Liang, D. Plouffe, et al., "Large-scale identification of single-feature polymorphisms in complex genomes," *Genome Research*, vol. 13, no. 3, pp. 513–523, 2003.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 584360, 16 pages
doi:10.1155/2008/584360

Review Article

Statistical Analysis of Efficient Unbalanced Factorial Designs for Two-Color Microarray Experiments

Robert J. Tempelman

Department of Animal Science, College of Agriculture and Natural Resources, Michigan State University, East Lansing, MI 48824-1225, USA

Correspondence should be addressed to Robert J. Tempelman, tempelma@msu.edu

Received 2 November 2007; Revised 22 January 2008; Accepted 25 April 2008

Recommended by Chunguang Du

Experimental designs that efficiently embed a fixed effects treatment structure within a random effects design structure typically require a mixed-model approach to data analyses. Although mixed model software tailored for the analysis of two-color microarray data is increasingly available, much of this software is generally not capable of correctly analyzing the elaborate incomplete block designs that are being increasingly proposed and used for factorial treatment structures. That is, optimized designs are generally unbalanced as it pertains to various treatment comparisons, with different specifications of experimental variability often required for different treatment factors. This paper uses a publicly available microarray dataset, as based upon an efficient experimental design, to demonstrate a proper mixed model analysis of a typical unbalanced factorial design characterized by incomplete blocks and hierarchical levels of variability.

1. INTRODUCTION

The choice and optimization of experimental designs for two-color microarrays have been receiving increasing attention [1–13]. Interest has been particularly directed towards optimizing experiments that involve a factorial design construction [7, 9, 14] in order to study the joint effects of several factors such as, for example, genotypes, pathogens, and herbicides. It is well known by plant scientists that factorial designs are more efficient than one-factor-at-a-time studies and allow the investigation of potentially interesting interactions between two or more factors. For example, investigators may study how herbicide effects (i.e., mean differences) depend upon plant genotypes or times after application.

Two-color systems such as spotted cDNA or long oligonucleotide microarrays involve hybridizations of two different mRNA samples to the same microarray, each of the two samples being labeled with a different dye (e.g., Cy3 or Cy5; Alexa555 or Alexa647). These microarrays, also simply referred to as arrays or slides, generally contain thousands of probes with generally a few ($\leq$4) spots per probe, and most often just one spot per probe. Each probe specifically hybridizes to a matching mRNA transcript of interest within each sample. After hybridization, microarray images are scanned at two different wavelengths as appropriate for each dye, thereby providing two different fluorescence intensity measurements for each probe. Upon further preprocessing or normalization [15], these dye-specific intensities for each probe are believed to reflect the relative mRNA abundance for the corresponding transcript within the respectively labeled samples. The normalized intensities, or the Cy3/Cy5 ratio thereof, for each spot are typically logarithmically transformed to render data that is generally characterized to be approximately normally distributed.

An increasingly unifying and indisputable message is that the heavily used common reference design is statistically inefficient [1, 9, 10, 12, 13]. Here, the same common reference sample or pool is reused as one of the two samples on every microarray, the other sample deriving from a treatment group of interest. Hence, inferences on differential expression are based only on indirect connections across arrays as samples from different treatments of interest are never directly connected or hybridized together on the

same microarray. In contrast, most of the alternatively proposed efficient designs are incomplete block designs, the most popular being various deviations of the loop design as first proposed for microarrays by Kerr and Churchill [16]. In these designs, direct connections or hybridizations are typically reserved for the most important treatment comparisons with inference on other comparisons being generally as efficient as any based on the common reference design.

The intent of this review is to reemphasize the use of mixed models as the foundation for statistical analysis of efficient factorial designs for microarrays. Mixed model analysis for microarray data was first proposed by Wolfinger et al. [17]. However, this and other previous expositions on the use of mixed model analysis for microarray data have been primarily directed towards the analysis of completely balanced designs [18, 19] whereas many recently proposed designs for microarray studies are unbalanced with respect to, for example, different standard errors on all pairwise comparisons between treatment groups [10, 13]. We will review various aspects of mixed model analysis for unbalanced designs, including a demonstration on publicly available data from a recent plant genomics study [20].

2. THE CONNECTION BETWEEN MIXED MODELS AND EFFICIENT DESIGNS

Efficient experimental designs are typically constructed such that their factors can be broadly partitioned into two categories: *treatment structure* factors and *design structure* factors [21]. The treatment structure naturally includes the factors of greatest interest; for example, herbicides, genotypes, tissues, and so forth, whose effects are deemed to be fixed. In other words, the levels of these *fixed effects* factors are specifically chosen by the investigator such that mean comparisons between such levels, for example, different treatments, are of primary interest. These factors also include any of whose levels are consistently reused over different experiments, such as dye labels, for example, Cy3 versus Cy5, for two-color microarrays. On the other hand, the design structure primarily includes *random effects* factors, whereby the levels of each such factor are considered to be randomly chosen from a conceptually infinite set of such levels [22]. For example, the specific arrays used for a microarray study are considered to be a random sample from a large, perhaps hypothetically infinite, population of arrays; similar claims would be made regarding biological replicates, for example, plants, pools thereof, or even field plots as dependent upon the experimental design [14]. Within each random-effects factor, the effects are typically specified to be normally, independently, and identically distributed (NIID) with variability in effects formally quantified by a variance component (VC).

These design structure or random effects factors are typically further partitioned into two subcategories: *blocking* factors and *experimental error* factors. In two-color microarray experiments, arrays are typically blocking factors as treatments can be directly compared within arrays, although this is not true for the common reference design as previously noted. Blocking represents a longstanding and efficient experimental design strategy for improving precision of inference on treatment comparisons. Experimental error factors, such as plants or pooled samples thereof within treatments, are often necessary to be included as random effects in order to properly specify true experimental replication at the biological level rather than merely at the measurement or technical level. Such specifications are particularly required when multiple aliquots are derived from the same biological specimen for use in multiple arrays [20, 23] or when probes for each gene transcript are spotted more than once on each array. Of course, plants may also alternatively serve as blocking factors in some designs if different tissues are compared within plants.

Currently, there is much software available for microarray data analysis, some of which is only suited for studies having only a treatment structure but no pure design structure. Common examples include the analysis of data generated from single channel systems (e.g., Affymetrix) or of log ratios generated from common reference designs. When no random effects are specified, other than the residuals, the corresponding statistical models are then simply fixed-effects models. Ordinary least squares (OLS) inference is then typically used to infer upon the treatment effects in these studies. OLS is appropriate if the assumption is valid that there is only one composite residual source of variability such that the residuals unique to each observation are NIID.

Conversely, statistical analysis of efficient two-color experiments having a fully integrated treatment and design structure needs to account for fixed and random effects as typical of a *mixed effects* model, more often simply referred to as a mixed model. Generalized least squares (GLS) analysis, also referred to as mixed-model analysis, has been recognized as optimal in terms of minimizing variance of estimates for inference on treatment comparisons. This is true not only for efficient microarray designs [10, 17, 19, 24] but even for general plant science and agronomy research [25–27], including recent applications in plant genomics research [20, 23, 28]. Some of the more recently popular microarray data analysis software has some mixed model analysis capabilities [29, 30].

Recall that some designs may be characterized by different levels of variability thereby requiring particular care in order to properly separate biological from technical replication, for example. Hence, it is imperative for the data analyst to know how to correctly construct the hypothesis test statistics, including the determination or, in some cases, the estimation of the appropriate degrees of freedom. Although, some of these issues have been discussed for balanced designs by Rosa et al. [19], they have not generally been carefully addressed for the analysis of microarray data generated from unbalanced designs. Optimally constructed experimental designs are often unbalanced with respect to inference on all pairwise treatment comparisons, such that even greater care for statistical inference is required than in completely balanced designs. For example, Wit et al. [13] proposed a method for optimizing two-color microarray designs to compare any number of treatment groups. Suppose that 9 different treatment groups are to be compared. Using

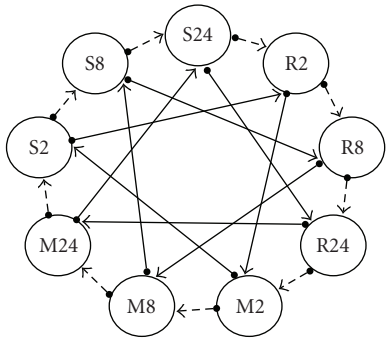


Figure 1: Optimized interwoven loop design for 9 treatments using R package SMIDA (Wit et al., 2005). Each circle represents a different treatment group. Each arrow represents a single array hybridization with circle base representing the Cy3 labeled sample and tail end representing the Cy5 labeled sample.

the methods and software developed by Wit et al. [13], the recommended interwoven loop design that is optimized for A-optimality (lowest average squared standard errors for a particular arrangement of treatment comparisons) is provided in Figure 1. Although Figure 1 appears to be visually symmetric with respect to the treatment labels, including that all treatment groups are dye balanced, not all treatment groups are directly hybridized against each other. Hence, inferences on all pairwise comparisons between treatment groups will not be equally precise. For example, the standard errors for the inference on treatments R2 versus R8 or R8 versus R24 will not be the same as that for treatments R8 versus S24 or R8 versus M2 due to the differences in the number and/or degree of direct and indirect connections for these two sets of comparisons in Figure 1.

Even for some balanced factorial designs, where the standard errors for comparing mean differences for levels of a certain factor are the same for all pairwise comparisons, the experimental error structure can vary substantially for different factors. That is, substantial care is required in deriving the correct test statistics, particularly with split plot arrangements [14]. Of course, even when a completely balanced design is intended, data editing procedures that delete poor quality spots for certain genes would naturally result in unbalanced designs.

3. CASE STUDY

3.1. Design

Zou et al. [20] present an experiment where three different inoculate treatments were applied to soybean (*Glycine max.*) plants 14 days after planting. The three different inoculates included bacteria inoculation along with the avirulence gene *avrB* thereby conferring resistance (R), bacteria inoculation without *avrB* thereby conferring susceptibility (S), and a control group whereby the inoculate simply contained an $MgCl_2$ solution (M). Unfoliated leaves from three to four plants were drawn and pooled for each treatment at each of three different times after postinoculation; 2, 8, and 24 hours. Hence, the treatment structure was comprised of a 3×3 factorial, that is, 3 inoculates $\times 3$ times, for a total of 9 groups. A 10th group involving a fourth null inoculate with leaves harvested at 2 hours postinoculation, N2, was additionally studied by Zou et al. [20]. The complete dataset on gene expression data for all 27 684 genes represented on a set of three microarray platforms as used by Zou et al. [20] is available as accession number GSE 2961 from the NCBI gene expression omnibus (GEO) repository (http://www.ncbi.nlm.nih.gov/geo/). The vast majority of the corresponding probes were spotted only once per array or slide for each platform.

A graphical depiction of the 13 hybridizations that superimposes the design structure upon one replicate of the 3×3 factorial treatment structure plus the additional 14th hybridization involving the 10th group N2 is illustrated in Figure 2. Note that at least two aliquots per each pooled sample are used, each aliquot being labeled with different dyes such that each replicate pool is used in at least two different hybridizations or arrays with opposite dye assignments. In other words, this design is characterized by technical replication such that it is imperative to explicitly model samples within inoculate by time combination as the biological replicates, that is, a set of random effects for modeling experimental error. Failing to do so would confuse pseudoreplication with true replication in the statistical analysis as each of the 2+ aliquots per each pool would then be incorrectly counted as 2+ different experimental replicates. The design in Figure 2 was replicated twice by Zou et al. [20], the second replication being of the exact same dye assignment and hybridization orientation as the first, for a total of 28 hybridizations. Hence, there were 20 samples (pools of leaves) utilized in the experiment, 2 per each of the 9 inoculate by time treatment groups plus 2 samples for the N2 control.

We arbitrarily consider gene expression measurements for just one particular gene based on the GEO submission from Zou et al. [20]: ID_REF #30 located in the metarow-metacolumn-row-column location 1-1-2-14 of each array from microarray platform GPL 1013, one of three different platforms used by Zou et al. [20] and further described in GEO. The statistical analysis of any of the remaining 27 683 genes that were spotted once on each slide across the three different platforms would be exactly the same as that for ID_REF #30, at least for those genes where no observations would be edited out for poor data quality. We use the normalized Cy3 and Cy5 data, provided as fields *S532N* and *S635N* in accession number GSE 2961 for ID_REF #30 from GEO. Hence, for the 28 hybridizations considered for two replications of Figure 2, there were 56 fluorescence intensities (28 Cy3 and 28 Cy5) for each gene. The 56 fluorescence intensities for ID_REF #30, as retrieved from GSE 2961 in GEO, are reproduced in Table 4.

3.2. Statistical model

For the purposes of this review, we concentrate our attention just on the subdesign characterized by the solid arrows in Figure 2 that connect the three primary inoculates (R,S, and

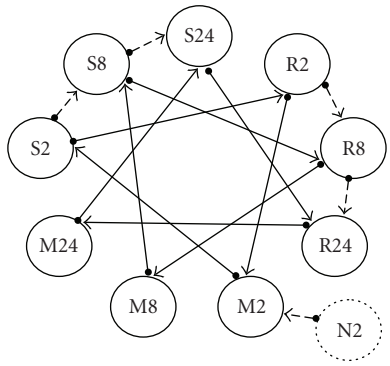


Figure 2: Experimental design for one replicate from Zou et al. (2005). Treatments included a full 3×3 factorial of inoculate and time effects plus a 10th null control group at time 2 (N2). Samples indicated by circles with letters indicating inoculate assignment: bacteria resistant (R), a bacteria susceptible (S), and $MgCl_2$ (M) control inoculate and numbers indicating time (2, 8, or 24 hours) after inoculation. Each arrow represents a single array hybridization with circle base representing the Cy3 labeled sample and tail end representing the Cy5 labeled sample. Solid arrows refer to the A-loop design of Landgrebe et al. (2006).

M) together within each of the 3 different times (2, 8, and 24 hours). The remaining dashed lines in Figure 2 involve either the 10th group (N2) or connect adjacent times (2 with 8 and 8 with 24) within each of two inoculates (R and S); note that no hybridizations connecting any of the three times within inoculate M were provided with GSE 2961 on GEO. Labeling inoculate type as Factor A and time after inoculation as Factor B, the resulting subdesign is an example of the "A-loop" design presented by Landgrebe et al. [9] as illustrated in their Figure 2 (B), albeit for a 3×2 factorial treatment structure in their case. In other words, the only direct connections between the 9 treatment groups within arrays involve comparisons of levels of Factor A within levels of Factor B. Using the log intensities as the response variables for further statistical analysis, an appropriate linear mixed model to specify for this A-loop design would be as follows:

$$y_{ijklm} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \delta_k + r(\alpha\beta)_{l:ij} + s(\beta)_{m:j} + e_{ijklm}, \quad (1)$$

where y_{ijklm} is the log fluorescence intensity pertaining to the lth biological replicate assigned to the ith inoculate ($i = 1, 2, 3$) and jth time ($j = 1, 2, 3$) labeled with the kth dye ($k = 1$ or 2), and hybridized to array $m (m = 1, 2, \ldots, 6)$ within the jth time. Here, μ is the overall mean, α_i is the effect of the ith inoculate, β_j is the effect of the jth time, $\alpha\beta_{ij}$ is the interaction effect between the ith inoculate and jth time, and δ_k is the effect of the kth dye, all of which are defined to be fixed effects. The design structure component of (1) is defined by the random effects of $r(\alpha\beta)_{l:ij}$ for the lth pool or biological replicate ($l = 1, 2$) within the ijth inoculate-time combination, $s(\beta)_{m:j}$ for the mth array ($m = 1, 2, \ldots, 6$) or slide within the jth time, and the residual e_{ijklm} unique to the same subscript identifiers as that for y_{ijklm}. The typical distributional assumptions in mixed models are such that each of the three sets of random effects are NIID with their own VC; that is, $r(\alpha\beta)_{l:ij} \sim \text{NIID}(0, \sigma^2_{R(AB)})$, $s(\beta)_{m:j} \sim \text{NIID}(0, \sigma^2_{S(B)})$, and $e_{ijklm} \sim \text{NIID}(0, \sigma^2_E)$. As clearly demonstrated by Dobbin et al. [31] and based on our experiences, dye effects should be modeled in (1), even after using global normalization procedures such as loess [15], as gene-specific dye effects are common. Nevertheless, one would not normally anticipate interaction effects between dye and other treatment factors (e.g., inoculate or time), and hence these effects are not specified in (1).

It should be somewhat apparent from the A-loop design of Figure 2 why the nesting or hierarchical specifications are specified as such for the random effects. For example, although each pool or replicate is labeled twice, once with each dye, each pool is still part of or nested within the same inoculate by time combination such that samples or replicates are specified to be nested within inoculate by time. Similarly, arrays are nested within times since each array is associated with only one particular level of time; that is, different times are never directly compared or connected within arrays. Hence, one should intuitively recognize from Figure 2 that there would be greater precision for inferring upon inoculate effects than for time effects using the A-loop design. That is, the variability due to arrays is completely confounded with time differences such that it partly defines the experimental unit or replicate for time.

3.3. Classical ANOVA

The complex nature of different levels of replication in the A-loop this design is further confirmed in the classical analysis of variance or ANOVA [21] for this design in Table 1. However, as demonstrated later, classical ANOVA is not necessarily equivalent to a more optimal GLS or mixed model analysis [32]; in fact, estimates of treatment effects based on classical ANOVA are simply equivalent to OLS estimates where all factors are treated as fixed. Nevertheless, the classical ANOVA table, when extended to include expected mean squares (EMS), is instructive in terms of identifying different levels of replication and hence experimental error.

Classical ANOVA is based on equating sums of squares (SS), also called quadratic forms, to their expectations; typically this involves equating mean squares (MS), being SS divided by their degrees of freedom (ν), to their EMS. For completely balanced designs, there is generally one universal manner in which these quadratic forms, and hence the ANOVA table, are constructed [19, 22]. However, for unbalanced designs, such as all of or even just the A-loop component of Figure 2, there are a number of ways of constructing different quadratic forms and hence different ways of constructing ANOVA tables for the same set of data [21, 32]. The most common ANOVA strategy is based on the use of type III quadratic forms as in Table 1 whereby the SS for each factor is adjusted for every other factor in the model. More details on type III and alternative ANOVA quadratic forms for unbalanced data can be found in Milliken and Johnson [21] and Searle [33].

Table 1: Classical ANOVA of log intensities for duplicated A-loop design component of Figure 2 for any particular gene using (1).

Source	SS*	$df^{\dagger}$	Mean square	Expected mean square
Inoculate	SS_A	ν_A	$MS_A = SS_A/\nu_A$	$\sigma^2_E + 1.5\sigma^2_{R(A\cdot B)} + \phi_A$ ‡
Time	SS_B	ν_B	$MS_B = SS_B/\nu_B$	$\sigma^2_E + 2\sigma^2_{R(A\cdot B)} + 2\sigma^2_{S(B)} + \phi_B$
Inoculate*time	SS_{AB}	ν_{AB}	$MS_{AB} = SS_{AB}/\nu_{AB}$	$\sigma^2_E + 1.5\sigma^2_{R(A\cdot B)} + \phi_{AB}$
Dye	SS_D	ν_D	$MS_D = SS_D/\nu_D$	$\sigma^2_E + \phi_D$
Rep(inoculate*time)	$SS_{R(AB)}$	$\nu_{R(AB)}$	$MS_{R(AB)} = SS_{R(AB)}/\nu_{R(AB)}$	$\sigma^2_E + 1.5\sigma^2_{R(A\cdot B)}$
Array(time)	$SS_{S(B)}$	$\nu_{S(B)}$	$MS_{S(B)} = SS_{S(B)}/\nu_{S(B)}$	$\sigma^2_E + 1.5\sigma^2_{S(B)}$
Error	SS_E	ν_E	$MS_E = SS_E/\nu_E$	σ^2_E

* Sums of squares.
† Degrees of freedom.
‡ ϕ_X is the noncentrality parameter for factor X. For example, when $\phi_A = 0$, there are no overall mean inoculate differences such that inoculate and Rep(inoculate*time) have the same expected mean square and $F_A = MS_A/MS_{R(AB)}$ is a random draw from an F distribution with ν_A numerator and $\nu_{R(AB)}$ denominator degrees of freedom.

Table 1 conceptually illustrates the basic components of an ANOVA table; again, for every term, say X, in a statistical model like (1), there is a sum of squares (SS_X), degrees of freedom (ν_X), mean square ($MS_X = SS_X/\nu_X$), and expected mean square (EMS_X). Generally, ANOVA tests on fixed effects are of greatest interest; for example, inoculate, time, and inoculate by time interaction. The correct F ratio test statistic for any fixed effects term in the ANOVA table is constructed such that its MS and a denominator MS have the same EMS if the null hypothesis is true; that is, that there are truly no effects for that particular term. In statistical parlance, no effects for a term X, whether that pertains to the main effects of a factor or the interaction effects between two or more factors, is synonymous with its corresponding *noncentrality parameter* (ϕ_X) being equal to zero; that is, there is no signal due to that model term [32].

Consider, for example, the test for the main effects of inoculate denoted as Factor A in Table 1. If the main effects of inoculate are nonexistent, that is, there are no overall or marginal mean differences between any of the inoculates, then $\phi_A = 0$. It should be clearly noted that when $\phi_A = 0$, the EMS for inoculate matches with the EMS for replicate within inoculate and time, denoted as rep(inoculate*time) in Table 1. In other words, rep(inoculate*time) is said to be the denominator or *error* term for the main effects of inoculate such that rep(inoculate*time) defines the experimental unit or the biological replicate for inoculate effects. Hence, the correct F statistic for testing inoculate effects, as demonstrated from Table 1, is $F_A = MS_A/MS_{R(AB)}$ based on ν_A numerator and $\nu_{R(AB)}$ denominator degrees of freedom. It should also be observed that this same error term or experimental unit would be specified as the denominator MS term for the ANOVA F-test on inoculate by time interaction effects, denoted as inoculate*time in Table 1. That is, when the corresponding noncentrality parameter $\phi_{AB} = 0$, both inoculate*time and rep(inoculate*time) share the same EMS such that the correct F statistic for testing this interaction is $F_{AB} = MS_{AB}/MS_{R(AB)}$ based on ν_{AB} numerator and $\nu_{R(AB)}$ denominator degrees of freedom.

It was previously noted from the A-loop design of Figure 2 that inference on the main effects of time (Factor B) should be less precise than that for the main effects of inoculate. In other words, the size of the experimental unit should be larger for time effects since arrays are nested within levels of time whereas levels of inoculate treatments are directly compared within arrays. This is further demonstrated in Table 1 by the EMS for time with $\phi_B = 0$, being larger than that for inoculate effects with $\phi_A = 0$, under the corresponding true null hypotheses of no main effects for either factor. In fact, the experimental error term for time is composite of both rep(inoculate*time) and arrays(time) such that marginal mean comparisons between the three times, 2, 8, and 24 hours, will be affected by more noise than marginal mean comparisons between the three inoculates which were directly and indirectly connected within arrays.

Note that under the null hypothesis of no time effects ($\phi_B = 0$), there is no one other MS that shares the same EMS $\sigma^2_E + 2\sigma^2_{R(AB)} + 2\sigma^2_{S(B)}$ that would allow one to readily construct an ANOVA F-statistic for the main effects of time. Satterthwaite [34] provided a solution to this problem by proposing the "synthesis" of a denominator *MS*, call it MS*, as being a linear combination of q random effects MS:

$$MS^* = a_1MS_1 + a_2MS_2 + a_3MS_3 + \cdots + a_qMS_q, \quad (2)$$

where $a_1, a_2, \ldots, a_q$ are known coefficients such that MS* has the same expectation as that for a certain model term X having mean square MS_x under the null hypothesis ($\phi_X = 0$). Then $F = MS_X/MS^*$ is approximately distributed as a random variable from a central F distribution with ν_X numerator and ν^* denominator degrees of freedom, where

$$\nu^* = \frac{(MS^*)^2}{\theta}, \quad (3)$$

with θ denoting $(a_1MS_1)^2/\nu_1 + (a_2MS_2)^2/\nu_2 + (a_3MS_3)^2/\nu_3 + \cdots + (a_qMS_q)^2/\nu_q$.

In our example, consider the synthesized $MS^* = 4/3MS_{R(AB)} + 4/3MS_{S(B)} - 5/3MSE$ as being a linear combination of the MS for rep(inoculate*time), array(time), and residual. With reference to (2), MS* is then a linear function of $q = 3$ different MS with $a_1 = 4/3$, $a_2 = 4/3$, and $a_3 = -5/3$. Using the EMS for these three MS provided from

Table 2: Classical ANOVA of log intensities for duplicated A-loop design component of Figure 2 on ID_REF #30 from Zou et al. (2005) using output from SAS PROC MIXED (code in Figure 3).

Type 3 analysis of variance								
Source	DF†	Sum of squares	Mean square	Expected mean square	Error term	Error DF	F value	Pr > F‡
Trt	2	0.7123	0.3561	Var(Residual) + 1.5 Var(sample(inoc*time)) + Q(inoc,inoc*time)	MS(sample(inoc*time))	6	3.13	0.1172
Time	2	3.7737	1.8868	Var(Residual) + 2 Var(sample(inoc*time)) + 2Var(array(time)) + Q(time,inoc*time)	1.3333 MS(array(time)) + 1.3333 MS(sample(inoc*time)) − 1.6667 MS(Residual)	13.969	3.27	0.0683
Inoc*time	4	0.6294	0.1573	Var(Residual) + 1.5 Var(sample(inoc*time)) + Q(inoc*time)	MS(sample(inoc*time))	6	1.38	0.3435
Dye	1	0.0744	0.0744	Var(Residual) + Q(dye)	MS(Residual)	5	2.19	0.1989
Rep(inoc*time)	6	0.6826	0.1137	Var(Residual) + 1.5 Var(sample(inoc*time))	MS(Residual)	5	3.35	0.1030
Array(time)	12	4.3330	0.3610	Var(Residual) + 1.5 Var(array(time))	MS(Residual)	5	10.63	0.0085
Residual	5	0.1699	0.0339	Var(Residual)	.	.	.	.

† Degrees of freedom.
‡ P-value.

Table 1 as $(\sigma_E^2+1.5\sigma_{R(AB)}^2)$, $(\sigma_E^2+1.5\sigma_{S(B)}^2)$, and σ_E^2, respectively, it should be readily seen that the expectation of MS* is then

$$\begin{aligned}\mathrm{EMS}^* &= \frac{4}{3}\left(\sigma_E^2 + 1.5\sigma_{R(AB)}^2\right) + \frac{4}{3}\left(\sigma_E^2 + 1.5\sigma_{S(B)}^2\right) - \frac{5}{3}\sigma_E^2 \\ &= \sigma_E^2 + 2\sigma_{R(AB)}^2 + 2\sigma_{S(B)}^2.\end{aligned} \tag{4}$$

That is, MS* shares the same EMS as that for time in Table 1 when $\phi_B = 0$. Hence, a suitable F statistic for inferring upon the main effects of time would be $F_B = \mathrm{MS}_B/\mathrm{MS}^*$.

To help further illustrate these concepts, let us conduct the ANOVA on the data generated from the A-loop design of Figure 2 for ID_REF #30 from Zou et al. [20]; that is, using data from arrays 1–9 and 15–23 as provided in Table 4. The classical ANOVA table using the *method=type3* option of the popular mixed-model software SAS PROC MIXED [35] for that particular gene is provided in Table 2; SAS code for all statistical analysis presented in this paper is provided in Figure 3 and also available for download, along with the data in Table 4, from http://www.msu.edu/~tempelma/ijpg2008.sas. As noted previously, the correct denominator MS term for testing the main effects of inoculate is replicate within inoculate by time. Hence, the corresponding F statistic = $\mathrm{MS}_A/\mathrm{MS}_{(R(AB))} = F_A = 0.356/0.114 = 3.13$, with $\nu_A = 2$ numerator and $\nu_{R(AB)} = 6$ denominator degrees of freedom leading to a P-value of 0.1172. Similarly, for the inoculate*time interaction, the appropriate F-test statistic is $\mathrm{MS}_{AB}/\mathrm{MS}_{(R(AB))} = F_{AB} = 0.157/0.114 = 1.38$, with $\nu_{AB} = 6$ numerator and $\nu_{R(AB)} = 6$ denominator degrees of freedom leading to a P-value of 0.3435. Even without considering the control of false discovery rates (FDRs) that involve the joint control of type I errors with respect to the remaining 27 683 genes, it seems apparent that neither the main effects of inoculate nor the interaction between inoculate and time would be statistically significant for gene ID_REF #30.

The synthesized denominator MS* for time effects is $\mathrm{MS}^* = 4/3\mathrm{MS}_{R(AB)} + 4/3\mathrm{MS}_{S(B)} - 5/3\mathrm{MSE} = 4/3(0.114) + 4/3(0.361) - 5/3(0.034) = 0.576$. The estimated degrees of freedom for this synthesized MS using (3) is then

$$\begin{aligned}\nu^* &= \frac{(\mathrm{MS}^*)^2}{(a_1\mathrm{MS}_{R(AB)})^2/\nu_{R(AB)} + (a_2\mathrm{MS}_{S(B)})^2/\nu_{S(B)} + (a_3\mathrm{MS}_E)^2/\nu_E} \\ &= \frac{(0.576)^2}{((4/3)\cdot 0.114)^2/6 + ((4/3)\cdot 0.361)^2/12 + (-(5/3)\cdot 0.034)^2/5} \\ &= 13.97.\end{aligned} \tag{5}$$

Hence, the main effects of time, appropriate F-test statistic is $\mathrm{MS}_B/\mathrm{MS}^* = F_B = 1.88/0.576 = 3.27$, with $\nu_B = 2$ numerator and $\nu^* = 13.97$ denominator degrees of freedom leading to a P-value of 0.0683 as also reported in the SAS output provided in Table 2.

3.4. Mixed model analysis

Although the classical ANOVA table is indeed instructive in terms of illustrating the different levels of variability and experimental error, it is not the optimal statistical analysis method for a mixed effects model, especially when the design is unbalanced. A mixed-model or GLS analysis more efficiently uses information on the design structure (i.e., random effects) for inferring upon the fixed treatment structure effects [27, 32].

Table 3: EGLS inference on overall importance of fixed effects for ID_REF #30 based on REML versus ANOVA (type III quadratic forms) for estimation of variance components using output from SAS PROC MIXED (code in Figure 3).

		Type 3 tests of fixed effects using REML			Type 3 tests of fixed effects using ANOVA		
Effect	Num DF*	Den DF*	F value	Pr > F†	Den DF*	F value	Pr > F†
Inoc	2	5.28	3.12	0.1273	6.36	3.48	0.0954
Time	2	17.8	2.81	0.0870	22.8	3.27	0.0563
Inoc*time	4	5.28	1.26	0.3893	6.36	1.38	0.3392
Dye	1	5.43	2.27	0.1879	5.15	2.19	0.1973

*Num Df = numerator degrees of freedom; Den DF = denominator degrees of freedom.
†P-value.

Unfortunately, GLS, in spite of its optimality properties, is generally not attainable with real data because the VC (e.g., $\sigma^2_{R(AB)}$, $\sigma^2_{S(B)}$, and σ^2_E) must be known. Hence, the VC must generally be estimated from the data at hand. There are a number of different methods that are available for estimating VC in mixed models [22]. The classical ANOVA method is based on equating MS to their EMS in the ANOVA table. For example, using the bottom row of Table 1, the EMS of MSE is σ^2_E. So then using the numerical results for ID_REF #30 from Table 2, the type III ANOVA estimate of σ^2_E is simply $\widehat{\sigma}^2_E$ = MSE = 0.034. Now work up one row further in Table 1 to the term array(time). Equating $\text{MS}_{S(B)}$ = 0.361 from the same corresponding row in Table 2 to its EMS of $\sigma^2_E + 1.5\sigma^2_{S(B)}$ using $\widehat{\sigma}^2_E = 0.034$ gives $\widehat{\sigma}^2_{S(B)} = 0.218$. Finally, work up one more (i.e., third to last) row in both tables. Equating $\text{MS}_{R(AB)} = 0.114$ from Table 2 to its EMS of $\sigma^2_E + 1.5\sigma^2_{R(AB)}$ using $\widehat{\sigma}^2_E = 0.034$ leads to $\widehat{\sigma}^2_{R(AB)} = 0.053$. So array variability $\sigma^2_{S(B)}$ is estimated to be roughly four times larger than the biological variability $\sigma^2_{R(AB)}$ which, in turn, is estimated to be somewhat larger than residual variability σ^2_E for ID_REF #30.

Recall that with unbalanced designs, quadratic forms are not unique such that ANOVA estimators of VC will not be unique either. Nevertheless, type III quadratic forms are most commonly chosen as then the SS for each term is adjusted for all other terms, as previously noted. Although ANOVA estimates of VC are unbiased, they are not efficient nor optimal in terms of estimates having minimum standard error [25]. Restricted maximum likelihood (REML) is a generally more preferred method of VC estimation [22, 36, 37] and is believed to have more desirable properties. Nevertheless, the corresponding REML estimates $\widehat{\sigma}^2_E = 0.033$, $\widehat{\sigma}^2_{S(B)} = 0.258$ and $\widehat{\sigma}^2_{R(AB)} = 0.061$ for ID_REF #30 are in some qualitative agreement with the previously provided ANOVA estimates.

Once the VCs are estimated, they are substituted for the true unknown VCs to provide the "estimated" GLS or EGLS of the fixed effects. It is important to note that typically EGLS = GLS for balanced designs, such that knowledge of VC is somewhat irrelevant for point estimation of treatment effects. However, the same is generally not true for unbalanced designs, such as either the A-loop design derived from Figure 2 or even the interwoven loop design from Figure 1. Hence, different methods of VC estimation could lead to different EGLS estimates of treatment effects as we demonstrate later. Suppose that it was of interest to compare the various mean responses of various inoculate by time group combinations in the duplicated A-loop design example. Based on the effects defined in the statistical model for this design in (1), the true mean response for the ith inoculate at the jth time averaged across the two dye effects (δ_1 and δ_2) can be written as

$$\mu_{ij.} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + 0.5\delta_1 + 0.5\delta_2. \tag{6}$$

If the levels are, say, ordered alphanumerically, the mean difference between inoculate $i = 1(M)$ and $i = 2(R)$ at time $j = 1$ (2 hours) is specified as $\mu_{11.} - \mu_{21.}$. Using (6), this difference written as a function of the model effects is then $\mu_{11.} - \mu_{21.} = (\mu + \alpha_1 + \beta_1 + \alpha\beta_{11} + 0.5\delta_1 + 0.5\delta_2) - (\mu + \alpha_2 + \beta_1 + \alpha\beta_{21} + 0.5\delta_1 + 0.5\delta_2) = \alpha_1 - \alpha_2 + \alpha\beta_{11} - \alpha\beta_{21}$. Similarly, the mean difference $\mu_{11.} - \mu_{12.}$ between time $j = 1$ (2 hours) and time $j = 2$ (8 hours) for inoculate $i = 1(M)$ could be derived as $\beta_1 - \beta_2 + \alpha\beta_{11} - \alpha\beta_{12}$. Note that these two comparisons or *contrasts* can be more elegantly written using matrix algebra notation. A better understanding of contrasts is useful to help determine the correct standard errors and statistics used to test these contrasts, including how to write the corresponding SAS code. Hence, a matrix algebra approach to hypothesis testing on contrasts is provided in Appendix 5 that complements the SAS code provided in Figure 3. For now, however, we simply use the "hat" notation ($\widehat{}$) in referring to the EGLS estimates of these two contrasts as $\widehat{\mu}_{11.} - \widehat{\mu}_{21.}$ and $\widehat{\mu}_{11.} - \widehat{\mu}_{12.}$, respectively.

As we already intuitively noted from the A-loop design of Figure 2, inference on $\mu_{11.} - \mu_{21.}$ should be much more precise than that for $\mu_{11.} - \mu_{12.}$ since inoculates are compared within arrays whereas times are not. This distinction should then be reflected in a larger standard error for $\widehat{\mu}_{11.} - \widehat{\mu}_{12.}$ than for $\widehat{\mu}_{11.} - \widehat{\mu}_{21.}$. Indeed, using the REML estimates of VC for EGLS inference, this is demonstrated by $\widehat{se}\,(\widehat{\mu}_{11.} - \widehat{\mu}_{21.}) = 0.2871$ whereas $\widehat{se}\,(\widehat{\mu}_{11.} - \widehat{\mu}_{12.}) = 0.4085$ for ID_REF #30. However, these standard errors are actually slightly understated since they do not take into account the uncertainty of the VC estimates as discussed by Kackar and Harville [38]. Kenward and Roger [39] derive a procedure to take this uncertainty into account which is part of the SAS PROC MIXED implementation using the option *ddfm=kr* [35] as indicated in Figure 3. Invoking this option raises the two standard errors accordingly, albeit very slightly, to $\widehat{se}\,(\widehat{\mu}_{11.} - \widehat{\mu}_{21.}) = 0.2878$ and $\widehat{se}\,(\widehat{\mu}_{11.} - \widehat{\mu}_{12.}) = 0.4088$.

Table 4: Dataset for ID_REF #30 for all hybridizations (14 arrays/loop x2 loops) in Figure 1 for each of two replicates per 10 inoculate by time groups, fluorescence intensities provided as y, log(base 2) intensities provided as ly.

Obs	array	inoculate	time	rep	dye	y	ly
1	1	R	2	1R2	Cy3	16322.67	13.9946
2	1	M	2	1M2	Cy5	20612.48	14.3312
3	2	M	2	1M2	Cy3	10552.21	13.3653
4	2	S	2	1S2	Cy5	10640.89	13.3773
5	3	S	2	1S2	Cy3	24852.98	14.6011
6	3	R	2	1R2	Cy5	21975.92	14.4236
7	4	R	8	1R8	Cy3	30961.96	14.9182
8	4	M	8	1M8	Cy5	13405.08	13.7105
9	5	M	8	1M8	Cy3	13103.51	13.6777
10	5	S	8	1S8	Cy5	15659.44	13.9347
11	6	S	8	1S8	Cy3	20424.47	14.3180
12	6	R	8	1R8	Cy5	34244.92	15.0636
13	7	R	24	1R24	Cy3	15824.29	13.9499
14	7	M	24	1M24	Cy5	13014.05	13.6678
15	8	M	24	1M24	Cy3	17503.11	14.0953
16	8	S	24	1S24	Cy5	27418.99	14.7429
17	9	S	24	1S24	Cy3	37689.16	15.2019
18	9	R	24	1R24	Cy5	55821.64	15.7685
19	10	S	2	1S2	Cy3	28963.28	14.8219
20	10	S	8	1S8	Cy5	38659.44	15.2385
21	11	S	8	1S8	Cy3	41608.78	15.3446
22	11	S	24	1S24	Cy5	41844.79	15.3528
23	12	R	2	1R2	Cy3	12132.41	13.5666
24	12	R	8	1R8	Cy5	19131.53	14.2237
25	13	R	8	1R8	Cy3	31067.04	14.9231
26	13	R	24	1R24	Cy5	26197.03	14.6771
27	14	N	2	1N2	Cy3	18540.91	14.1784
28	14	M	2	1M2	Cy5	24971.88	14.6080
29	15	R	2	2R2	Cy3	9612.25	13.2307
30	15	M	2	2M2	Cy5	9212.11	13.1693
31	16	M	2	2M2	Cy3	10322.23	13.3335
32	16	S	2	2S2	Cy5	10979.19	13.4225
33	17	S	2	2S2	Cy3	8061.40	12.9768
34	17	R	2	2R2	Cy5	6737.37	12.7180
35	18	R	8	2R8	Cy3	8807.09	13.1044
36	18	M	8	2M8	Cy5	8696.95	13.0863
37	19	M	8	2M8	Cy3	15186.20	13.8905
38	19	S	8	2S8	Cy5	23477.49	14.5190
39	20	S	8	2S8	Cy3	19424.30	14.2456
40	20	R	8	2R8	Cy5	18198.99	14.1516
41	21	R	24	2R24	Cy3	19630.00	14.2608
42	21	M	24	2M24	Cy5	15629.14	13.9320
43	22	M	24	2M24	Cy3	10875.49	13.4088
44	22	S	24	2S24	Cy5	20816.21	14.3454
45	23	S	24	2S24	Cy3	24647.70	14.5892
46	23	R	24	2R24	Cy5	22148.96	14.4350
47	24	S	2	2S2	Cy3	17795.09	14.1192
48	24	S	8	2S8	Cy5	34569.11	15.0772
49	25	S	8	2S8	Cy3	44175.28	15.4310
50	25	S	24	2S24	Cy5	38020.46	15.2145

Table 4: Continued.

Obs	array	inoculate	time	rep	dye	y	ly
51	26	R	2	2R2	Cy3	34689.07	15.0822
52	26	R	8	2R8	Cy5	62219.10	15.9251
53	27	R	8	2R8	Cy3	22724.21	14.4719
54	27	R	24	2R24	Cy5	19594.71	14.2582
55	28	N	2	2N2	Cy3	11755.32	13.5210
56	28	M	2	2M2	Cy5	12599.55	13.6211

Now, the denominator degrees of freedom for inference on these two contrasts should also differ given that the nature of experimental error variability somewhat differs for inoculate comparisons as opposed to time comparisons as noted previously from Figure 2. However, with EGLS, there are no SS and hence no corresponding MS or EMS expression for each main effects or interaction term in the model, such that determining the correct test statistic and degrees of freedom is somewhat less obvious than with the previously described classical ANOVA approach [32]. Giesbrecht and Burns [40] introduced a procedure for estimating the denominator degrees of freedom for EGLS inference which, again, is invoked with the *ddfm=kr* option of SAS PROC MIXED. Using this option along with REML estimation of VC for the analysis of ID_REF #30, the estimated degrees of freedom for $\widehat{\mu}_{11.} - \widehat{\mu}_{21.}$ is 5.28 whereas that for $\widehat{\mu}_{11.} - \widehat{\mu}_{12.}$ is 17.0.

Contrasts are also used in EGLS to provide ANOVA-like F tests for the overall importance of various fixed effects; more details based on the specification of contrast matrices to test these effects are provided in Appendix 5. For example, denote the marginal or overall mean of inoculate i averaged across the 3 times and 2 dyes as $\mu_{i..} = (1/3)\sum_{j=1}^{3}\mu_{ij.}$. The $\nu_A = 2$ numerator degrees of freedom hypothesis test for the main effects of inoculates can be written as a combination of two complementary contrasts (A1) $H_0 : \mu_{1..} - \mu_{3..} = 0$ and (A2) $H_0 : \mu_{2..} - \mu_{3..} = 0$; that is, if both contrasts are 0, then obviously $H_0 : \mu_{2..} - \mu_{3..} = 0$ is also true such that then $H_0 : \mu_{1..} = \mu_{2..} = \mu_{3..}$ is true. Similarly, let us suppose that one wished to test the main effects of times (Factor B). Then, it could be readily demonstrated that the corresponding hypothesis test can also be written as a combination of $\nu_B = 2$ complementary contrasts: (B1) $H_0 : \mu_{.1.} - \mu_{.3.} = 0$ and (B2) $H_0 : \mu_{.2.} - \mu_{.3.} = 0$, where $\mu_{.j.} = (1/3)\sum_{i=1}^{3}\mu_{ij.}$ denotes the marginal mean for the jth level of Factor B; that is, the jth time. If both component hypotheses (B1) and (B2) are true, then $H_0 : \mu_{.1.} = \mu_{.2.} = \mu_{.3.} = 0$ is also true thereby defining the composite $\nu_B = 2$ numerator degrees of freedom hypothesis test for the main effects of Factor B.

Now the interaction between inoculate and time is a $\nu_{AB} = \nu_A\nu_B = 2{*}2 = 4$ numerator degrees of freedom test as previously noted from Tables 1 and 2, suggesting that there are 4 complementary contrasts that jointly test for the interaction of the two factors. Of course, it is also well known that the interaction degrees of freedom is typically the product of the main effects degrees of freedom for the two factors considered. Two of the four degrees of freedom for the interaction involve testing whether or not the mean difference between inoculates 1 and 3 is the same within time 1 as it is within time 3, that is, (AB1) $H_0 : \mu_{11.} - \mu_{31.} - (\mu_{13.} - \mu_{33.}) = 0$, and whether or not the mean difference between inoculates 2 and 3 is the same within time 1 as it is within time 3; that is, (AB2) $H_0 : \mu_{21.} - \mu_{31.} - (\mu_{23.} - \mu_{33.}) = 0$. If both hypotheses (AB1) and (AB2) are true then it should be apparent that $H_0 : \mu_{11.} - \mu_{21.} - (\mu_{13.} - \mu_{23.}) = 0$ is also true; that is, the mean difference between inoculates 1 and 2 is the same within time 1 as it is within time 3. The remaining two degrees of freedom for the interaction involve testing whether or not the mean difference between inoculates 1 and 3 is the same within time 2 as it is within time 3; that is, (AB3) $H_0 : \mu_{12.} - \mu_{32.} - (\mu_{13.} - \mu_{33.}) = 0$, and whether or not the mean difference between inoculates 2 and 3 is the same within time 2 as it is within time 3; that is, (AB4) $H_0 : \mu_{22.} - \mu_{32.} - (\mu_{23.} - \mu_{33.}) = 0$. If both hypotheses (AB3) and (AB4) are true then $H_0 : \mu_{12.} - \mu_{22.} - (\mu_{13.} - \mu_{23.}) = 0$ is also true. Hence, contrasts AB1, AB2, AB3, and AB4 completely define the four components or numerator degrees of freedom for the interaction between Factors A and B. That is, the test for determining whether or not the mean differences between all levels of A are the same within each level of B, and vice versa, can be fully characterized by these four complementary contrasts.

The EGLS statistics used for testing the overall importance of these main effects or interactions are approximately distributed as F-random variables with the numerator degrees of freedom defined by the number of complementary components or contrasts as previously described; refer to Appendix 5 and elsewhere [27, 32, 35] for more details. Now, the denominator degrees of freedom for each contrast are dependent upon the design and can be determined based on that using classical ANOVA as in Table 1 or by a multivariate extension of the Satterthwaite-based procedure proposed by Fai and Cornelius [41]; again this option is available as *ddfm=kr* using SAS PROC MIXED (Figure 3).

Unfortunately, much available software used for mixed model analysis of microarray data does not carefully take into consideration that various fixed effects terms of interest may have different denominator degrees of freedom when constructing F test statistics. In fact, a typical strategy of such software is to assume that ν_E (i.e., the residual degrees of freedom) is the denominator degrees of freedom for all tests. This strategy is denoted as the "residual" method for determining denominator degrees of freedom by Spilke et al. [36] who demonstrated using simulation work that the use

```
title "Mixed model analysis of log fluorescence intensity data from gene 30";
proc mixed
    data=gene30 /* name of data as provided in Table 4 */
    method = type3;
    /* Provides classical ANOVA table and EGLS based on ANOVA estimates of VC */
    /* If REML estimates of VC are desired, change above line to method = reml; */
    where ((array <= 9) or (15 <=array <= 23));
       /* Using A-loop component (arrays 1-9, 15-23) of Table 4 data only */
    class rep array inoc time dye;
       /* name of fixed and random classification factors in design */
    model ly = inoc time inoc*time dye
       /* Specify response variable and fixed effects here */
       /ddfm = kr
       /* Use Kenward-Roger's procedure to estimate denominator degrees of freedom */
          e3;
       /* e3 will print the contrast matrices KA, KB and KAB (see (A.8), (A.9) and
          (A.10) of Appendix 5) used to provide the EGLS ANOVA F-test statistics (optional) */
    random array(time) rep(inoc*time) ; /* Specify random effects */
    estimate "k1 contrast"
       int 0 inoc 1 −1 0 time 0 0 0 inoc*time 1 0 0 −1 0 0 0 0 0 dye 0 0;
       /* contrast coefficients as specified for k1 in (A.6) of Appendix 5 */
    estimate "k2 contrast"
       int 0 inoc 0 0 0 time 1 −1 0 inoc*time 1 −1 0 0 0 0 0 0 0 dye 0 0;
       /* contrast coefficients as specified for k2 in (A.7) of Appendix 5 */
run;
```

Figure 3: SAS code for classical ANOVA and EGLS inference. Comments describing purpose immediately provided after corresponding code between /* and */ as with a regular SAS program. EGLS based on REML would simply involve substituting *method = reml* for *method = type3* in the third line of the code.

of the residual method can substantially inflate type I error rate for EGLS inference on fixed effects; in other words, the number of false-positive results or genes incorrectly declared to be differentially expressed between treatments would be unduly increased. Spilke et al. [36] further demonstrated that use of the Kenward-Rogers' method for degrees of freedom estimation and control for uncertainty on VC provided best control of the nominal confidence interval coverage and type I error probabilities.

3.5. Impact of method of variance component estimation on EGLS

It was previously noted that the estimated standard errors for EGLS on two contrasts $\mu_{11.} - \mu_{21.}$ and $\mu_{11.} - \mu_{12.}$ were $\widehat{se}\,(\widehat{\mu}_{11.} - \widehat{\mu}_{21.}) = 0.2878$ and $\widehat{se}\,(\widehat{\mu}_{11.} - \widehat{\mu}_{12.}) = 0.4088$, respectively, when REML was used to estimate the variance components for ID_REF #30. If the VC estimates are computed using type III ANOVA, then these estimated standard errors would differ accordingly; that is, $\widehat{se}\,(\widehat{\mu}_{11.} - \widehat{\mu}_{21.}) = 0.2752$ and $\widehat{se}\,(\widehat{\mu}_{11.} - \widehat{\mu}_{12.}) = 0.3828$, respectively. What perhaps is even more disconcerting is that the estimates of $\mu_{11.} - \mu_{21.}$ and $\mu_{11.} - \mu_{12.}$ also differ between the two EGLS inferences; for example, using REML, $\widehat{\mu}_{11.} - \widehat{\mu}_{21.} = 0.1328$ and $\widehat{\mu}_{11.} - \widehat{\mu}_{12.} = -0.0881$ whereas using ANOVA $\widehat{\mu}_{11.} - \widehat{\mu}_{21.} = 0.1298$ and $\widehat{\mu}_{11.} - \widehat{\mu}_{12.} = -0.0873$.

The overall EGLS tests for ID_REF #30 for testing the main effects of inoculate, time and their interaction as based on the previously characterized complementary contrasts are provided separately for ANOVA versus REML estimates of VC in Table 3; this output is generated as type III tests using the SAS code provided in Figure 3. From here, it should be clearly noted that conclusions upon the overall importance of various fixed effects terms in (1) as derived from EGLS inference subtly depend upon the method of VC estimation; for example, the EGLS *P*-values in Table 3 tend to be several points smaller using ANOVA compared to REML; furthermore, note the differences in the estimated denominator degrees of freedom between the two sets. Naturally, this begs the question as to which method of VC estimation should be used?

In completely balanced designs, ANOVA and REML lead to identical estimates of VC and identical EGLS inference, provided that all ANOVA estimates of VC are positive. ANOVA estimates of VC that are negative are generally constrained by REML to be zero, thereby causing a "ripple" effect on REML estimates of other VC and subsequently on EGLS inference [42]. As noted previously, REML does tend to outperform most other methods for various properties of VC estimation [37]. Furthermore, there is evidence that EGLS based on ANOVA leads to poorer control of type I error rate for inference on fixed effects compared to

EGLS based on REML in unbalanced data structures [36]. However, Stroup and Littell [42] concluded that EGLS using REML may sometimes lead to inference on fixed effects that is too conservative (i.e., actual error rates less than nominal type I error rate) again due to the nonnegative REML restrictions on the VC estimates and associated ripple effects. This issue warrants further study given that it has implications for control of FDR which are most commonly used to control the rate of type I errors in microarray studies [43]. Estimation of FDR inherently depends upon the distribution of *P*-values for treatment effects across all genes such that even mild perturbations on this distribution have potential bias implications for control of false-positive rates.

4. OTHER ISSUES FOR THE DESIGN ANALYSIS INTERFACE

4.1. Log ratio versus log intensity modeling

Recent work on the optimization and comparison of various efficient microarray designs have been based on the assumption of OLS inference; that is, no random sources of variability other than residuals are considered [2, 8, 9, 13]. While this observation may seem to be counterintuitive given that the arguments laid out in this review for the need of (E)GLS to analyze efficient designs, it is important to note at least a couple of things. First, virtually all of the work on design optimization has been based on the assumption that a sample or pool is used only once; the corresponding interwoven loop designs in such cases [13] have been referred to as classical loop designs [10, 19]. However, sometimes two or more aliquots from each sample are used in separate hybridizations [20, 23] such as the A-loop design, example used in this review; the corresponding designs are connected loop designs [10, 19] that require the specification of random biological replicate effects separate from residual effects as previously noted.

Secondly, almost all of the design optimization work has been based on the use of Cy3/Cy5 log ratios as the response variables rather than dye-specific log intensities as used in this review. This data reduction, that is, from two fluorescence intensities to one ratio per spot on an array, certainly eliminates array as a factor to specify in a linear model. However, the use of log ratios can severely limit estimability and inference efficiency of certain comparisons. Suppose that instead of using the 36 log intensities from the duplicated A-loop design from arrays 1–9 and 15–23 of Table 4, we used the derivative 18 Cy3/Cy5 log ratios as the response variables. For example, the two corresponding $\log_2$Cy3 and Cy5 fluorescence intensities for array 1 from Table 4 are 13.9946 and 14.3312. The Cy3/Cy5 log ratio is then the difference or -0.3316 corresponding to a fold change of $2^{-0.3316} = 0.795$. Using log ratios as their response variables, Landgrebe et al. [9] concluded that it was impossible to infer upon the main effects of Factor B (e.g., time) in the A-loop design. However, as we demonstrated earlier, it is possible to infer upon these effects using ANOVA or EGLS analysis on the log intensities. Jin et al. [18] similarly illustrate the utility of log intensity analysis in a split plot design that would not otherwise have been possible using log ratios. Milliken et al. [14] provide much more extensive mixed modeling details on the utility of log intensity analysis in nested or split-plot microarray designs similar to the A-loop design.

The relative efficiency of some designs may be seen to depend upon the relative magnitude of biological to technical variation [10, 44]; sometimes it is only possible to separately estimate these two sources of variability using log intensities rather than log ratios thereby requiring the use of (E)GLS rather than OLS. In fact, analysis of log intensities using mixed effects model appears to be not only more flexible than log-ratio modeling but is statistically more efficient in recovering more data information [1, 45]. That is, as also noted by Milliken et al. [14], treatment effects are more efficiently estimated by combining intraarray and interarray information in a mixed model analysis when an incomplete block design is used, and arrays are explicitly included as random effects by analyzing log intensities rather than log ratios.

4.2. Choosing between efficient experimental designs using mixed models

There are a number of different criteria that might be used to choose between different designs for two-color microarrays. We have already noted that the interwoven loop design in Figure 1 is A-optimal for pairwise comparisons between 9 treatment groups. A-optimality has been criticized for microarray studies because it chooses designs with improved efficiency for certain contrasts at the expense of other perhaps more relevant contrasts and further depends upon the parameterization of the linear model [1, 6, 9]; other commonly considered types of optimality criteria are possible and further discussed by Wit et al. [13] and Landgrebe et al. [9]. At any rate, it is somewhat possible to modify A-optimality to explicitly take into account a particular set of scientific questions [13]; furthermore, optimization with respect to one criterion will generally be nearly optimal for others.

For one particular type of optimality criterion, Landgrebe et al. [9] demonstrated that the A-loop design has the best relative efficiency compared to other designs for inference on the main effects of Factor A and the interaction effects between A and B although the main effects of Factor B could not be inferred upon using an analysis of log ratios as previously noted. How does the A-loop design of Figure 2 generally compare to the interwoven loop design of Figure 1 if a 3×3 factorial treatment structure is imposed on the 9 treatments as implied by the same labels as used in Figure 2? Suppose that Figure 1 is a connected interwoven loop design [10] in the sense that the outer loop of Figure 1 (dashed arrows) connects one biological replicate for each of 9 groups whereas the inner loop of Figure 1 (solid arrows) connects a second biological replicate for each of the 9 groups. Then this design would consume 18 biological replicates and 18 arrays, thereby providing a fair comparison with the duplicated A-loop design of Figure 2.

Recall that Figure 1 is A-optimized for pairwise comparisons between all 9 groups. It is not quite clear what implications this might have for statistical efficiency for the constituent main effects of $A(\nu_A = 2)$, $B(\nu_B = 2)$, and the effects of their interaction $A^*B(\nu_{AB} = 4)$; note, incidentally, that these degrees of freedom independently sum to 8 as required for 9 groups. As duly noted by Altman and Hua [1], pairwise comparisons between all 9 groups may be not as important as various main effects or interaction contrasts with a factorial treatment structure arrangement. Although, as noted earlier, Figure 1 is symmetric with respect to the treatment labels, the classical ANOVA table for this interwoven loop design would be even more complicated (not shown) than that presented for the A-loop design since there is not one single denominator MS that would serve as the experimental error term for inoculate, time or inoculate by time effects!

One should perhaps compare two alternative experimental designs having the same factorial treatment structure, but a different design structure, for contrasts of highest priority, choosing those designs where such contrasts have the smaller standard error. Let us consider the following comparisons: $\mu_{1..} - \mu_{3..}$, $\mu_{.1.} - \mu_{.3.}$, and $\mu_{11.} - \mu_{31.} - (\mu_{13.} - \mu_{33.})$; that is, respectively, the overall mean difference between inoculates 1 and 3, the overall mean difference between times 1 and 3, and the interaction component pertaining to the difference between inoculates 1 and 3 within time 1 versus that same difference within time 3. Recall that these contrasts were components of the EGLS tests on the two sets of main effects and the interaction and previously labeled as (A1), (B1), and (AB1), respectively.

Now the comparison of efficient designs for the relative precision of various contrasts will generally depend upon the relative magnitude of the random effects VC as noted recently by Hedayat et al. [44] and for various microarray design comparisons [10]. Suppose the "true" variance components for σ^2_E, $\sigma^2_{R(AB)}$, and $\sigma^2_{S(B)}$ were 0.03, 0.06, and 0.25, comparable to either set of estimates provided previously on ID_REF #30 from Zou et al. [20]. The linear mixed model for analyzing data generated from Figure 1 would be identical to that in (1) except that arrays would no longer be specified as being nested within times. For the interwoven loop design of Figure 1, the standard errors for each of the three contrasts are $se(\hat{\mu}_{1..} - \hat{\mu}_{3..}) = 0.18$, $se(\hat{\mu}_{.1.} - \hat{\mu}_{.3.}) = 0.21$, and $se(\hat{\mu}_{11.} - \hat{\mu}_{31.} - (\hat{\mu}_{13.} - \hat{\mu}_{33.})) = 0.43$ whereas for the A-loop subdesign of Figure 2, the corresponding standard errors are $se(\hat{\mu}_{1..} - \hat{\mu}_{3..}) = 0.16$, $se(\hat{\mu}_{.1.} - \hat{\mu}_{.3.}) = 0.33$, and $se(\hat{\mu}_{11.} - \hat{\mu}_{31.} - (\hat{\mu}_{13.} - \hat{\mu}_{33.})) = 0.40$. So whereas the optimized design in Figure 1 using Wit et al. [13] provided a substantial improvement for the estimation of overall mean time differences, the A-loop design is indeed more efficient for inferring upon the main effects of inoculate and the interaction between inoculate and time. Hence, the choice between the two designs would reflect a matter of priority for inference on the various main effects and their interactions. It should be carefully noted as demonstrated by Tempelman [10], that designs leading to lower standard errors for certain comparisons do not necessarily translate to greater statistical power as the denominator degrees of freedom for various tests may be substantially different between the two designs.

4.3. Unbalanced designs and shrinkage estimation

Shrinkage or empirical Bayes (EB) estimation is known to improve statistical power for inference on differential gene expression between treatments in microarray experiments [46]. Shrinkage-based estimation is based on the well-established hierarchical modeling concept that more reliable inferences on gene-specific treatment differences are to be attained by borrowing information across all genes [47, 48]. Typically, such strategies have involved improving estimation of standard errors of gene-specific treatment differences by "shrinking" gene-specific variances towards an overall mean or other measure of central tendency. However, most shrinkage estimation procedures have been developed for fixed effects models, that is, for simple experimental designs having a treatment structure but no or very limited design structure, or even treating all design structure factors as fixed [30]. Currently popular shrinkage estimation procedures [49–51] are certainly appropriate for many designs based on one-color Affymetrix systems or for common reference designs. Other proposed shrinkage procedures have facilitated extensions to very special cases of nested designs [47], including some based on rather strong modeling assumptions such as a constant correlation of within-array replicate spots across all genes [52] or a design structure facilitating the use of permutation testing [29]. However, virtually none of the procedures proposed thus far are well adapted to handle unbalanced designs such as the A-loop design where different sizes of experimental units need to be specified for different treatment factors; hence investigators should proceed with caution when using shrinkage estimation for unbalanced mixed-model designs.

5. CONCLUSIONS

We have provided an overview of the use of mixed linear model analysis for the processing of unbalanced microarray designs, given that most efficient incomplete block designs for microarrays are unbalanced with respect to various comparisons. We strongly believe that much mixed-model software currently available for the analysis of microarrays does not adequately address the proper determination of error terms and/or denominator degrees of freedom for various tests. This would be particularly true if we had chosen to analyze all of the data for ID_REF #30 in Table 4 from Zou et al. [20] based on all of the 2×14 hybridizations depicted in Figure 2. Even then, the size of the standard errors and estimated degrees of freedom would still be seen to be somewhat different for estimating the main effects of times compared to estimating the main effects of inoculates given the lower degree of within-array connectivity between the various levels of time as illustrated in Figure 2. If inferences on various comparisons of interest are not conducted correctly in defining a list of differently expressed genes, all subsequent microarray analysis

(e.g., FDR estimates, gene clustering, gene class analysis, etc.) are absolutely futile.

We believe that it is useful to choose proven mixed-model software (e.g., SAS) to properly conduct these tests and, if necessary, to work with an experienced statistician in order to do so. We have concentrated our attention on the analysis of a particular gene. It is, nevertheless, straightforward to use SAS to serially conduct mixed-model analysis for all genes on a microarray [53]; furthermore, SAS JMP GENOMICS (http://www.jmp.com/software/genomics/) provides an even more powerful user interface to the mixed model analysis of microarray data.

APPENDIX

MATRIX REPRESENTATION OF THE MIXED MODEL ANALYSIS OF THE A-LOOP DESIGN OF ZOU ET AL.

Any mixed model, including that specified in (1), can be written in matrix algebra form:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}. \tag{A.1}$$

Here $\mathbf{y} = \{y_{ijklm}\}$ is the vector of all data, $\boldsymbol{\beta}$ is the vector of all fixed effects (e.g., inoculate, time, dye, and inoculate by time interaction effects), $\mathbf{u}$ is the vector of all random effects (e.g., arrays and sample within inoculate by time effects), and $\mathbf{e} = \{e_{ijklm}\}$ is the vector of random residual effects. Furthermore, $\mathbf{X}$ and $\mathbf{Z}$ are corresponding incidence matrices that specify the treatment and design structure of the experiment, thereby linking the treatment and design effects, $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively, to $\mathbf{y}$. Note that $\mathbf{y}$ has a dimension of 36×1 for the duplicated A-loop design of Zou et al. [20]. Now $\boldsymbol{\beta}$ and $\mathbf{u}$ can be further partitioned into the effects as specified in (1); for our example,

$$\begin{aligned}\boldsymbol{\beta} = [&\mu \;\; \alpha_1 \;\; \alpha_2 \;\; \alpha_3 \;\; \beta_1 \;\; \beta_2 \;\; \beta_3 \;\; \alpha\beta_{11} \;\; \alpha\beta_{12} \;\; \alpha\beta_{13},\\ &\alpha\beta_{21} \;\; \alpha\beta_{22} \;\; \alpha\beta_{23} \;\; \alpha\beta_{31} \;\; \alpha\beta_{32} \;\; \alpha\beta_{33} \;\; \delta_1 \;\; \delta_2]',\end{aligned} \tag{A.2}$$

such that $\boldsymbol{\beta}$ is a 18×1 vector of fixed effects; that is, there are 18 elements in (A.2). Furthermore, $\mathbf{u} = [\mathbf{u}'_{R(AB)} \;\; \mathbf{u}'_{S(B)}]'$ can be similarly partitioned into a 18×1 vector of random effects, $\mathbf{u}_{R(AB)}$, for replicates within inoculate by time and another 18×1 vector of random effects, $\mathbf{u}_{S(B)}$, for arrays within time; that is, there are a total of 18 biological replicates and 18 arrays in the study, each characterized by a random effect. Note that it is coincidence that the row dimensions of $\boldsymbol{\beta}$, $\mathbf{u}_{R(AB)}$, and $\mathbf{u}_{S(B)}$ are all 18 for this particular example design.

Again, the distributional assumptions on the random and residual effects are specified the same as in the paper but now written in matrix algebra notation: $\mathbf{u}_{R(AB)} \sim N(\mathbf{0}_{18\times 1}, \mathbf{I}_{18}\sigma^2_{R(AB)})$, $\mathbf{u}_{S(B)} \sim N(\mathbf{0}_{18\times 1}, \mathbf{I}_{18}\sigma^2_{S(B)})$, and $\mathbf{e} \sim N(\mathbf{0}_{36\times 1}, R = \mathbf{I}_{36}\sigma^2_E)$ with $\mathbf{0}_{t\times 1}$ denoting a $t \times 1$ vector of zeros and $\mathbf{I}_t$ denoting an identity matrix of dimension t. Reasonably assuming that $\mathbf{u}_{R(AB)}$ and $\mathbf{u}_{S(B)}$ are pairwise independent of each other (i.e., biological sample effects are not influenced by array effects and vice versa), then the variance-covariance matrix $\mathbf{G}$ of $\mathbf{u}$ is a 36×36 diagonal matrix with the first 18 diagonal elements being $\sigma^2_{R(AB)}$ and the remaining 18 diagonal elements being $\sigma^2_{S(B)}$. The GLS estimator, $\hat{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$ can be written [22, 32] as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \tag{A.3}$$

with its variance-covariance matrix defined by

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}, \tag{A.4}$$

such that $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}$ denotes the generalized inverse of $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})$.

Once the VC are estimated, they are substituted for the true unknown VC in $\mathbf{V}$ to produce $\hat{\mathbf{V}}$ which are then used to provide the "estimated" GLS or EGLS $\tilde{\boldsymbol{\beta}}$, of $\boldsymbol{\beta}$:

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}. \tag{A.5}$$

As noted in the text, typically $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ (i.e., EGLS = GLS) for balanced designs but not necessarily for unbalanced designs, such as those depicted in Figures 1 or 2.

It was previously noted in the paper that the mean difference $\mu_{11.} - \mu_{21.}$ between inoculate $i = 1$ and $i = 2$ at time $j = 1$ as could be written as a function of the model effects in (1) as $\alpha_1 - \alpha_2 + \alpha\beta_{11} - \alpha\beta_{21}$. Similarly, the mean difference $\mu_{12.} - \mu_{12.}$ between time $j = 1$ and time $j = 2$ for inoculate i could be written as $\beta_1 - \beta_2 + \alpha\beta_{11} - \alpha\beta_{12}$. These two contrasts written in matrix notation as $\mathbf{k}'_1\boldsymbol{\beta}$ and $\mathbf{k}'_2\boldsymbol{\beta}$, respectively, where

$$\mathbf{k}'_1 = [\,0 \;\; 1 \;\; -1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; 0 \;\; 0 \;\; -1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0\,], \tag{A.6}$$

$$\mathbf{k}'_2 = [\,0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 1 \;\; -1 \;\; 0 \;\; 1 \;\; -1 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0\,] \tag{A.7}$$

are *contrast vectors* whose coefficients align in order with the elements of $\boldsymbol{\beta}$ in (A.2). For example, note from (A.6) that the nonzero coefficients of 1, −1, 1, and −1 occur within the 2nd, 3rd, 8th, and 11th positions of $\mathbf{k}'_1$, respectively. When these coefficients are multiplied in the same order with the 2nd, 3rd, 8th, and 11th elements of $\boldsymbol{\beta}$ provided in (A.2), one gets $(1)\alpha_1 + (-1)\alpha_2 + (1)\alpha\beta_{11} + (-1)\alpha\beta_{21}$ which is indeed $\mathbf{k}'_1\boldsymbol{\beta} = \alpha_1 - \alpha_2 + \alpha\beta_{11} - \alpha\beta_{21}$ as specified previously. The reader should be able to make a similar observation for $\mathbf{k}'_2\boldsymbol{\beta}$ in considering how the nonzero elements of (A.7) align in position with elements of $\boldsymbol{\beta}$ in (A.2) to produce $\beta_1 - \beta_2 + \alpha\beta_{11} - \alpha\beta_{12}$. In Figure 3, SAS PROC MIXED is used to provide the estimates, standard errors, and test statistics for these two contrasts. That is, note how all of the elements from (A.6) and (A.7) are completely reproduced in the *estimate* statements as "*k1 contrast*" and "*k2 contrast,*" respectively, in Figure 3.

Now, when the VC are known, these two contrasts can be estimated by their GLS, $\mathbf{k}'_1\hat{\boldsymbol{\beta}}$, and $\mathbf{k}'_2\hat{\boldsymbol{\beta}}$. Furthermore, using (A.4), the true standard errors of these two estimates can be determined as $se(\mathbf{k}'_1\hat{\boldsymbol{\beta}}) = \sqrt{\mathbf{k}'_1(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{k}_1}$ and $se(\mathbf{k}'_2\hat{\boldsymbol{\beta}}) = \sqrt{\mathbf{k}'_2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{k}_2}$, respectively. However, as previously noted, the VC are generally not known but must be estimated from the data such that the two contrasts are typically estimated using $\mathbf{k}'_1\tilde{\boldsymbol{\beta}}$ and $\mathbf{k}'_2\tilde{\boldsymbol{\beta}}$ with approximate

standard errors determined by $\widehat{se(\mathbf{k}_1'\tilde{\tilde{\boldsymbol{\beta}}})} = \sqrt{\mathbf{k}_1'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{k}_1}$ and $\widehat{se(\mathbf{k}_2'\tilde{\tilde{\boldsymbol{\beta}}})} = \sqrt{\mathbf{k}_2'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{k}_2}$. Using the REML estimates of VC as provided in the paper, the code from Figure 3 can be executed to provide $\widehat{se(\mathbf{k}_1'\tilde{\tilde{\boldsymbol{\beta}}})} = 0.2871$ whereas $\widehat{se(\mathbf{k}_2'\tilde{\tilde{\boldsymbol{\beta}}})} = 0.4085$ for ID_REF #30 by simply changing *method = type3* to *method = reml* and by deleting *ddfm = kr*. However, these standard errors are actually slightly understated since they do not take into account the uncertainty of the VC or $\hat{\mathbf{V}}$ as an estimate of $\mathbf{V}$ as discussed by Kackar and Harville [38].

Kenward and Roger [39] derive a procedure to take this uncertainty into account and which is part of the SAS PROC MIXED implementation using the *ddfm=kr* option [35] as specified in Figure 3. Invoking this option raises the two standard errors accordingly, albeit very slightly, to $\widehat{se(\mathbf{k}_1'\tilde{\tilde{\boldsymbol{\beta}}})} = 0.2878$ and $\widehat{se(\mathbf{k}_2'\tilde{\tilde{\boldsymbol{\beta}}})} = 0.4088$. Furthermore, the *ddfm=kr* option invokes the procedure of Giesbrecht and Burns [40] to estimate the denominator degrees of freedom for EGLS inference. Using this option and REML, the estimated degrees of freedom for $\mathbf{k}_1'\tilde{\tilde{\boldsymbol{\beta}}}$ is 5.28 whereas that for $\mathbf{k}_2'\tilde{\tilde{\boldsymbol{\beta}}}$ is 17.0 as would be noted from executing the SAS code in Figure 3. The corresponding SAS output will furthermore include the t-test statistics for the two contrasts as $t_1 = \mathbf{k}_1'\tilde{\tilde{\boldsymbol{\beta}}}/\widehat{se(\mathbf{k}_1'\tilde{\tilde{\boldsymbol{\beta}}})} = 0.1328/0.2878 = 0.46$ and $t_2 = \mathbf{k}_2'\tilde{\tilde{\boldsymbol{\beta}}}/\widehat{se(\mathbf{k}_2'\tilde{\tilde{\boldsymbol{\beta}}})} = -0.3799/0.4088 = -0.93$. These statistics when compared to their Student t distributions with their respective estimated degrees of freedom, 5.28 and 17.0, lead to P-values of 0.66 and 0.37, respectively; that is, there is no evidence that either contrast is statistically significant.

Contrast matrices on $\boldsymbol{\beta}$ can be used to derive ANOVA-like F tests for the overall importance of various fixed effects using EGLS. Recall from the paper that the test for the main effects of inoculates can be written as a joint function of $\nu_A = 2$ contrasts $\mu_{1..} - \mu_{3..}$ and $\mu_{2..} - \mu_{3..}$, where $\mu_{i..} = (1/3)\sum_{j=1}^{3}\mu_{ij.}$ with $\mu_{ij.}$ is defined as in (6). These two contrasts, labeled as (A1) and (A2) in the paper, can be jointly written together as a linear function $\mathbf{K}_A'\boldsymbol{\beta}$ of the elements of $\boldsymbol{\beta}$ in (A.2), where

$$\mathbf{K}_A' = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & 0 & 0 \end{bmatrix}. \tag{A.8}$$

For example, the first row of $\mathbf{K}_A'$ specifies the coefficients for testing $\mu_{1..} - \mu_{3..} = (1/3)(\mu_{11.} + \mu_{12.} + \mu_{13.}) - (1/3)(\mu_{31.} + \mu_{32.} + \mu_{33.})$ as a function of the elements of $\boldsymbol{\beta}$ using (6). In other words, matching up, in order, the first row of $\mathbf{K}_A'$ in (A.8) with the elements of $\boldsymbol{\beta}$ in (A.2), the corresponding contrast $\mu_{1..} - \mu_{3..}$ can be rewritten as $\alpha_1 - \alpha_3 + (1/3)\alpha\beta_{11} + (1/3)\alpha\beta_{12} + (1/3)\alpha\beta_{13} - (1/3)\alpha\beta_{31} - (1/3)\alpha\beta_{32} - (1/3)\alpha\beta_{33}$. Similarly, the second row of $\mathbf{K}_A'$ in (A.8) specifies the contrast coefficients for $\mu_{2..} - \mu_{3..}$ as a function of the elements of $\boldsymbol{\beta}$.

Recall that the main effects of times (Factor B) involves a joint test of $\nu_B = 2$ contrasts $\mu_{.1.} - \mu_{.3.}$ and $\mu_{.2.} - \mu_{.3.}$ labeled as (B1) and (B2) in the paper, where $\mu_{.j.} = (1/3)\sum_{i=1}^{3}\mu_{ij.}$. In terms of the elements of $\boldsymbol{\beta}$ in (A.2), these two contrasts are jointly specified as $\mathbf{K}_B'\boldsymbol{\beta}$ with

$$\mathbf{K}_B' = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & -1 & \frac{1}{3} & 0 & -\frac{1}{3} & \frac{1}{3} & 0 & -\frac{1}{3} & \frac{1}{3} & 0 & -\frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & \frac{1}{3} & -\frac{1}{3} & 0 & 0 \end{bmatrix}. \tag{A.9}$$

That is, (A.9) is another 2×18 contrast matrix, just like $\mathbf{K}_A'$, where the two rows of $\mathbf{K}_B'$ specify the coefficients for the contrasts $\mu_{.1.} - \mu_{.3.}$ and $\mu_{.2.} - \mu_{.3.}$, respectively, as a function of the elements of $\boldsymbol{\beta}$ in (6).

Recall that the interaction between the effects of inoculates and times was $\nu_{AB} = 4$ numerator degrees of freedom test based on jointly testing four complementary and independent contrasts, suggesting that there are four rows that determine the corresponding contrast matrix. The complete interaction contrast can then be written as $\mathbf{K}_{AB}'\boldsymbol{\beta}$, where

$$\mathbf{K}_{AB}' = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & -1 & 1 & 0 & 0 \end{bmatrix}. \tag{A.10}$$

Note that the 4 rows in (A.10) specify contrast coefficients on the model effects for each of the 4 constituent component hypotheses, (AB1), (AB2), (AB3), and (AB4) as defined in the paper, when aligned with the coefficients of $\boldsymbol{\beta}$ in (A.2). As a sidenote, the somewhat uninteresting contrast for dye effects could be written using a contrast vector $\mathbf{k}_D'$ (not shown) in order to test the overall mean difference between the two dyes.

The EGLS test statistic for testing the overall importance of any fixed effects term, say X, is specified as $F_X = \tilde{\tilde{\boldsymbol{\beta}}}'\mathbf{K}_X(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{K}_X'\tilde{\tilde{\boldsymbol{\beta}}}$. Here F_X is distributed as an F-random variable under $\mathrm{H}_0 : \mathbf{K}_X'\boldsymbol{\beta} = 0$ with the numerator degrees of freedom being defined by the number of rows of the contrast matrix $\mathbf{K}_X'$ [27, 32, 35]. The denominator degrees of freedom for each contrast is dependent upon the design and can be determined based on that using classical ANOVA as in Table 1 or a multivariate extension of the Satterthwaite-based procedure from Giesbrecht and Burns [40] as proposed by Fai and Cornelius [41]; again this option is available as *ddfm=kr* using SAS PROC MIXED (Figure 3). The corresponding EGLS ANOVA output for ID_REF #30, based on either ANOVA or REML estimation of VC, is provided in Table 3.

ACKNOWLEDGMENT

Support from the Michigan Agricultural Experiment Station (Project MICL 1822) is gratefully acknowledged.

REFERENCES

[1] N. S. Altman and J. Hua, "Extending the loop design for two-channel microarray experiments," *Genetical Research*, vol. 88, no. 3, pp. 153–163, 2006.

[2] F. Bretz, J. Landgrebe, and E. Brunner, "Design and analysis of two-color microarray experiments using linear models," *Methods of Information in Medicine*, vol. 44, no. 3, pp. 423–430, 2005.

[3] J. S. S. Bueno Filho, S. G. Gilmour, and G. J. M. Rosa, "Design of microarray experiments for genetical genomics studies," *Genetics*, vol. 174, no. 2, pp. 945–957, 2006.

[4] F.-S. Chai, C.-T. Liao, and S.-F. Tsai, "Statistical designs for two-color spotted microarray experiments," *Biometrical Journal*, vol. 49, no. 2, pp. 259–271, 2007.

[5] K. Dobbin, J. H. Shih, and R. Simon, "Questions and answers on design of dual-label microarrays for identifying differentially expressed genes," *Journal of the National Cancer Institute*, vol. 95, no. 18, pp. 1362–1369, 2003.

[6] G. F. V. Glonek and P. J. Solomon, "Factorial and time course designs for cDNA microarray experiments," *Biostatistics*, vol. 5, no. 1, pp. 89–111, 2004.

[7] S. Gupta, "Balanced factorial designs for cDNA microarray experiments," *Communications in Statistics: Theory and Methods*, vol. 35, no. 8, pp. 1469–1476, 2006.

[8] K. F. Kerr, "Efficient 2^k factorial designs for blocks of size 2 with microarray applications," *Journal of Quality Technology*, vol. 38, no. 4, pp. 309–318, 2006.

[9] J. Landgrebe, F. Bretz, and E. Brunner, "Efficient design and analysis of two colour factorial microarray experiments," *Computational Statistics & Data Analysis*, vol. 50, no. 2, pp. 499–517, 2006.

[10] R. J. Tempelman, "Assessing statistical precision, power, and robustness of alternative experimental designs for two color microarray platforms based on mixed effects models," *Veterinary Immunology and Immunopathology*, vol. 105, no. 3-4, pp. 175–186, 2005.

[11] S.-F. Tsai, C.-T. Liao, and F.-S. Chai, "Statistical designs for two-color microarray experiments involving technical replication," *Computational Statistics & Data Analysis*, vol. 51, no. 3, pp. 2078–2090, 2006.

[12] V. Vinciotti, R. Khanin, D. D'Alimonte, et al., "An experimental evaluation of a loop versus a reference design for two-channel microarrays," *Bioinformatics*, vol. 21, no. 4, pp. 492–501, 2005.

[13] E. Wit, A. Nobile, and R. Khanin, "Near-optimal designs for dual channel microarray studies," *Journal of the Royal Statistical Society: Series C*, vol. 54, no. 5, pp. 817–830, 2005.

[14] G. A. Milliken, K. A. Garrett, and S. E. Travers, "Experimental design for two-color microarrays applied in a pre-existing split-plot experiment," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, article 20, 2007.

[15] Y. H. Yang, S. Dudoit, P. Luu, et al., "Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation," *Nucleic Acids Research*, vol. 30, no. 4, pp. 1–10, 2002.

[16] M. K. Kerr and G. A. Churchill, "Statistical design and the analysis of gene expression microarray data," *Genetical Research*, vol. 77, no. 2, pp. 123–128, 2001.

[17] R. D. Wolfinger, G. Gibson, E. D. Wolfinger, et al., "Assessing gene significance from cDNA microarray expression data via mixed models," *Journal of Computational Biology*, vol. 8, no. 6, pp. 625–637, 2001.

[18] W. Jin, R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgell, and G. Gibson, "The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*," *Nature Genetics*, vol. 29, no. 4, pp. 389–395, 2001.

[19] G. J. M. Rosa, J. P. Steibel, and R. J. Tempelman, "Reassessing design and analysis of two-colour microarray experiments using mixed effects models," *Comparative and Functional Genomics*, vol. 6, no. 3, pp. 123–131, 2005.

[20] J. Zou, S. Rodriguez-Zas, M. Aldea, et al., "Expression profiling soybean response to *Pseudomonas syringae* reveals new defense-related genes and rapid HR-specific downregulation of photosynthesis," *Molecular Plant-Microbe Interactions*, vol. 18, no. 11, pp. 1161–1174, 2005.

[21] G. A. Milliken and D. E. Johnson, *Analysis of Messy Data, Volume I: Designed Experiments*, Wadsworth, Belmont, Calif, USA, 1984.

[22] S. R. Searle, G. Casella, and C. E. McCulloch, *Variance Components*, John Wiley & Sons, New York, NY, USA, 1992.

[23] M. Vuylsteke, F. van Eeuwijk, P. Van Hummelen, M. Kuiper, and M. Zabeau, "Genetic analysis of variation in gene expression in *Arabidopsis thaliana*," *Genetics*, vol. 171, no. 3, pp. 1267–1275, 2005.

[24] X. Cui and G. A. Churchill, "How many mice and how many arrays? Replication in mouse cDNA microarray experiments," in *Methods of Microarray Data Analysis III*, S. M. Lin and K. F. Johnson, Eds., pp. 139–154, Kluwer Academic Publishers, Norwell, Mass, USA, 2003.

[25] H. P. Piepho, A. Büchse, and K. Emrich, "A hitchhiker's guide to mixed models for randomized experiments," *Journal of Agronomy and Crop Science*, vol. 189, no. 5, pp. 310–322, 2003.

[26] H. P. Piepho, A. Büchse, and C. Richter, "A mixed modelling approach for randomized experiments with repeated measures," *Journal of Agronomy and Crop Science*, vol. 190, no. 4, pp. 230–247, 2004.

[27] J. Spilke, H. P. Piepho, and X. Hu, "Analysis of unbalanced data by mixed linear models using the MIXED procedure of the SAS system," *Journal of Agronomy and Crop Science*, vol. 191, no. 1, pp. 47–54, 2005.

[28] D. Nettleton, "A discussion of statistical methods for design and analysis of microarray experiments for plant scientists," *The Plant Cell*, vol. 18, no. 9, pp. 2112–2121, 2006.

[29] X. Cui, J. T. G. Hwang, J. Qiu, N. J. Blades, and G. A. Churchill, "Improved statistical tests for differential gene expression by shrinking variance components estimates," *Biostatistics*, vol. 6, no. 1, pp. 59–75, 2005.

[30] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.

[31] K. K. Dobbin, E. S. Kawasaki, D. W. Petersen, and R. M. Simon, "Characterizing dye bias in microarray experiments," *Bioinformatics*, vol. 21, no. 10, pp. 2430–2437, 2005.

[32] R. C. Littell, "Analysis of unbalanced mixed model data: a case study comparison of ANOVA versus REML/GLS," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 7, no. 4, pp. 472–490, 2002.

[33] S. R. Searle, *Linear Models for Unbalanced Data*, John Wiley & Sons, New York, NY, USA, 1987.

[34] F. E. Satterthwaite, "An approximate distribution of estimates of variance components," *Biometrics Bulletin*, vol. 2, no. 6, pp. 110–114, 1946.

[35] R. C. Littell, G. A. Milliken, W. W. Stroup, R. D. Wolfinger, and O. Schabenberger, *SAS for Mixed Models*, SAS Institute, Cary, NC, USA, 2nd edition, 2006.

[36] J. Spilke, H.-P. Piepho, and X. Hu, "A simulation study on tests of hypotheses and confidence intervals for fixed effects in mixed models for blocked experiments with missing data," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 10, no. 3, pp. 374–389, 2005.

[37] W. H. Swallow and J. F. Monahan, "Monte Carlo comparison of ANOVA, MIVQUE, REML, and ML estimators of variance components," *Technometrics*, vol. 26, no. 1, pp. 47–57, 1984.

[38] R. N. Kackar and D. A. Harville, "Approximations for standard errors of estimators of fixed and random effects in mixed linear models," *Journal of the American Statistical Association*, vol. 79, no. 388, pp. 853–862, 1984.

[39] M. G. Kenward and J. H. Roger, "Small sample inference for fixed effects from restricted maximum likelihood," *Biometrics*, vol. 53, no. 3, pp. 983–997, 1997.

[40] F. G. Giesbrecht and J. C. Burns, "Two-stage analysis based on a mixed model: large-sample asymptotic theory and small-sample simulation results," *Biometrics*, vol. 41, no. 2, pp. 477–486, 1985.

[41] A. H.-T. Fai and P. L. Cornelius, "Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments," *Journal of Statistical Computation and Simulation*, vol. 54, no. 4, pp. 363–378, 1996.

[42] W. W. Stroup and R. C. Littell, "Impact of variance component estimates on fixed effect inference in unbalanced linear mixed models," in *Proceedings of the 14th Annual Kansas State University Conference on Applied Statistics in Agriculture*, pp. 32–48, Manhattan, Kan, USA, April 2002.

[43] J. D. Storey and R. Tibshirani, "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 16, pp. 9440–9445, 2003.

[44] A. S. Hedayat, J. Stufken, and M. Yang, "Optimal and efficient crossover designs when subject effects are random," *Journal of the American Statistical Association*, vol. 101, no. 475, pp. 1031–1038, 2006.

[45] J. P. Steibel, *Improving experimental design and inference for transcription profiling experiments*, thesis, Department of Animal Science, Michigan State University, East Lansing, Mich, USA, 2007.

[46] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nature Reviews Genetics*, vol. 7, no. 1, pp. 55–65, 2006.

[47] I. Lönnstedt and T. Speed, "Replicated microarray data," *Statistica Sinica*, vol. 12, no. 1, pp. 31–46, 2002.

[48] G. K. Robinson, "That BLUP is a good thing: the estimation of random effects," *Statistical Science*, vol. 6, no. 1, pp. 15–51, 1991.

[49] P. Baldi and A. D. Long, "A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes," *Bioinformatics*, vol. 17, no. 6, pp. 509–519, 2001.

[50] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[51] G. W. Wright and R. M. Simon, "A random variance model for detection of differential gene expression in small microarray experiments," *Bioinformatics*, vol. 19, no. 18, pp. 2448–2455, 2003.

[52] G. K. Smyth, J. Michaud, and H. S. Scott, "Use of within-array replicate spots for assessing differential expression in microarray experiments," *Bioinformatics*, vol. 21, no. 9, pp. 2067–2075, 2005.

[53] G. Gibson and R. D. Wolfinger, "Gene expression profiling using mixed models," in *Genetic Analysis of Complex Traits Using SAS*, A. M. Saxton, Ed., pp. 251–279, SAS Users Press, Cary, NC, USA, 2004.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 574927, 18 pages
doi:10.1155/2008/574927

Review Article

Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources

Ibrokhim Y. Abdurakhmonov and Abdusattor Abdukarimov

Center of Genomic Technologies, Institute of Genetics and Plant Experimental Biology, Academy of Sciences of Uzbekistan, Yuqori Yuz, Qibray region, Tashkent district 702151, Uzbekistan

Correspondence should be addressed to Ibrokhim Y. Abdurakhmonov, genomics@uzsci.net

Received 21 December 2007; Accepted 18 April 2008

Recommended by Chunguang Du

Compared to the conventional linkage mapping, linkage disequilibrium (LD)-mapping, using the nonrandom associations of loci in haplotypes, is a powerful high-resolution mapping tool for complex quantitative traits. The recent advances in the development of unbiased association mapping approaches for plant population with their successful applications in dissecting a number of simple to complex traits in many crop species demonstrate a flourish of the approach as a "powerful gene tagging" tool for crops in the plant genomics era of 21st century. The goal of this review is to provide nonexpert readers of crop breeding community with (1) the basic concept, merits, and simple description of existing methodologies for an association mapping with the recent improvements for plant populations, and (2) the details of some of pioneer and recent studies on association mapping in various crop species to demonstrate the feasibility, success, problems, and future perspectives of the efforts in plants. This should be helpful for interested readers of international plant research community as a guideline for the basic understanding, choosing the appropriate methods, and its application.

1. INTRODUCTION

The level of the genetic diversity is pivotal for world food security and survival of human civilization on earth. Historically, humans exploited plant species for their livelihoods that resulted in domestication of many of them as improved cultivars to produce food for the better supply of the human diet [1]. Presently, out of 150 plant species cultivated in agriculture, twelve provide about 75% of human food and four produce 50% of human diet [2]. According to Food and Health Organization report, ~800 million people in the developing countries are suffering from food deficiency [3] that underlies an attention to improve agricultural production to eliminate or, at least, reduce the feeding problems.

The narrow genetic base of modern crop cultivars is the serious obstacle to sustain and improve crop productivity due to rapid vulnerability of genetically uniform cultivars by potentially new biotic and abiotic stresses [4]. However, plant germplasm resources worldwide, comprising of wild plant species, modern cultivars, and their crop wild relatives, are the important reservoirs of natural genetic variations, originated from a number of historical genetic events as a respond to environmental stresses and selection through crop domestication [1, 5]. The efficient exploiting these ex situ conserved genetic diversities is vital to overcome future problems associated with narrowness of genetic base of modern cultivars. However, many agriculturally important variations such as productivity and quality, tolerance to environmental stresses, and some of forms of disease resistance are controlled by polygenes and "multifactorial" that greatly depends on *genetic* × *environmental* (G × E) interactions [1, 6]. These complex traits are referred to as quantitative trait loci (QTLs), and it is challenging to identify QTLs based on only traditional phenotypic evaluation. Identification of QTLs of agronomic importance and its utilization in a crop improvement further requires mapping of these QTLs in a genome of crop species using molecular markers [1, 6]. This was the major breakthrough and accomplishment in many crops in "genomics era" since the end of the 20th century, and now extended to flourish in the 21st century.

In this review, we provide a brief description for the concept of genetic mapping; then, as a flourish of the crop genomics era, we thoroughly review one of the powerful genetic mapping tools for crops, linkage disequilibrium (LD)-based association study, as a high-resolution, broader allele coverage, and cost effective gene tagging approach in plant germplasm resources. This provides an opportunity to widely dissect and exploit existing natural variations for crop improvement.

2. GENETIC MAPPING OF CAUSATIVE VARIANTS

The main goal of genetic mapping is to detect neutrally inherited markers in close proximity to the genetic causatives or genes controlling the complex quantitative traits. Genetic mapping can be done mostly in two ways [1]: (1) using the experimental populations (also referred to as "biparental" mapping populations) that is called QTL-mapping as well as "genetic mapping" or "gene tagging," and (2) using the diverse lines from the natural populations or germplasm collections that is called LD-mapping or "association mapping." The details of the traditional QTL-mapping approach has recently been reviewed by Collard et al. [6], and further basic description of the approach here would be a redundant. For detailed concept, models and methodologies, problems, and perspectives of linkage analysis, readers are suggested refer to Liu [7] and Wu et al. [8]. Here, we briefly outline linkage mapping procedure for the sake of highlighting the merits of the alternative approach-association mapping.

So that such a linkage analysis can be done [6–8], firstly, the experimental populations such as F_2, back cross (BC), double haploid (DH), recombinant inbred line (RIL), and near isogenic line (NIL) populations, derived from the genetic hybridization of two parental genotypes with an alternative trait of interest, need to be developed. Secondly, these experimental populations including a large number of progenies or lines are measured for the segregation of a trait of interest in the different environmental conditions. Thirdly, a set of polymorphic DNA markers, differentiating the parental genotypes and segregating in a mapping population, need to be identified and genotyped. For that, usual practice is that, first, the parental genotypes are screened, and if markers are polymorphic over the parents, then, all individuals of a mapping population are genotyped with these polymorphic molecular markers. Once genotypic data of a mapping population is ready, marker data is used to construct the framework linkage maps, representing the order (position) and linkage (a relative genetic distance in cM) of used molecular markers along the linkage groups or segments of particular chromosomes. This is accomplished through assessing of recombination rates between the marker loci. Consequently, these markers ordered along the linkage map are statistically correlated with phenotypic characteristics of individuals of a mapping population, and QTL regions affecting a trait of interest, along with closely positioned marker tags to that QTL, are identified.

One can imagine these linkage marker maps as a "road map," marker tags as the labels directing to specific places, and QTLs to a community/neighborhood (with specific function) on the map [6]. The precision of QTL-mapping largely depends on the genetic variation (or genetic background) covered by a mapping population, the size of a mapping population, and a number of marker loci used. Once QTLs affecting a trait of interest accurately tagged using above-outlined approach, marker tags are the most effective tools in a crop improvement that allows the mobilization of the genes of interest from donor lines to the breeding material through marker-assisted selection (MAS). Although traditional QTL-mapping will continue being an important tool in gene tagging of crops, it is a "now classical approach" and overall is very costly [1, 9], and has low resolution with simultaneous evaluation of only a few alleles [10] in a longer research time scale. In linkage mapping, the major limitation, hampering the fine mapping, is associated with the availability of only a few meiotic events to be used that occurred since experimental hybridization in a recent past [11].

3. ASSOCIATION MAPPING AS AN ALTERNATIVE APPROACH

These limitations, however, can be reduced with the use of "association mapping" [1]. Turning the gene-tagging efforts from biparental crosses to natural population of lines (or germplasm collections), and from traditional QTL-mapping to linkage disequilibrium (LD)-based association study became a powerful tool in mapping of the genes of interest [12]. This leads to the most effective utilization of ex situ conserved natural genetic diversity of worldwide crop germplasm resources. LD refers to a historically reduced (nonequilibrium) level of the recombination of specific alleles at different loci controlling particular genetic variations in a population. This LD can be detected statistically, and has been widely applied to map and eventually clone a number of genes underlying the complex genetic traits in humans [13–16].

The advantages of population-based association study, utilizing a sample of individuals from the germplasm collections or a natural population, over traditional QTL-mapping in biparental crosses primarily are due to (1) availability of broader genetic variations with wider background for marker-trait correlations (i.e., many alleles evaluated simultaneously), (2) likelihood for a higher resolution mapping because of the utilization of majority recombination events from a large number of meiosis throughout the germplasm development history, (3) possibility of exploiting historically measured trait data for association, and (4) no need for the development of expensive and tedious biparental populations that makes approach timesaving and cost-effective [17–19].

Although the overall approach of population-based association mapping in plants varies based on the methodology chosen (see below sections), assuming structured population samples, the performance of association mapping includes the following steps (see Figure 1): (1) selection of a group of individuals from a natural population or germplasm collection with wide coverage of genetic diversity; (2) recording or measuring the phenotypic characteristics (yield, quality,

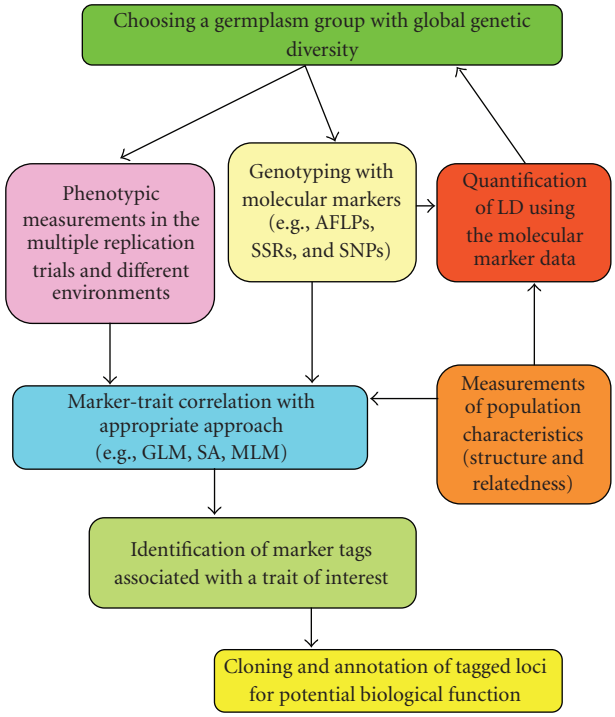


Figure 1: The scheme of association mapping for tagging a gene of interest using germplasm accessions. Note that the outlined scheme may vary based on population characteristics and methodology chosen for association study.

tolerance, or resistance) of selected population groups, preferably, in different environments and multiple replication/trial design; (3) genotyping a mapping population individuals with available molecular markers; (4) quantification of the extent of LD of a chosen population genome using a molecular marker data; (5) assessment of the population structure (the level of genetic differentiation among groups within a sampled population individuals) and kinship (coefficient of relatedness between pairs of each individuals within a sample); and (6) based on information gained through quantification of LD and population structure, correlation of phenotypic and genotypic/haplotypic data with the application of an appropriate statistical approach that reveals "marker tags" positioned within close proximity of targeted trait of interest. Consequently, a specific gene(s) controlling a QTL of interest can be cloned using the marker tags and annotated for an exact biological function (Figure 1). As a starting point for association mapping, it is important to gain knowledge of the patterns of LD for genomic regions of the "target" organisms and the specificity of the extent of LD among different populations or groups to design and conduct unbiased association mapping [20, 21].

4. LINKAGE DISEQUILIBRIUM (LD)

4.1. Concept of LD

Genetic linkage generally refers to coinheritance of different loci within a genetic distance on the chromosome. There are two terms used in population genetics, linkage equilibrium (LE), and linkage disequilibrium (LD) to describe linkage relationships (co-occurrence) of alleles at different loci in a population. LE is a random association of alleles at different loci and equals the product of allele frequencies within haplotypes, meaning that at random combination of alleles at each locus its haplotypes (combination of alleles) frequency has equal value in a population. In contrast, LD is a nonrandom association of alleles at different loci, describing the condition with nonequal (increased or reduced) frequency of the haplotypes in a population at random combination of alleles at different loci. LD is not the same as linkage, although tight linkage may generate high levels of LD between alleles. Usually, there is significant LD between more distant sites or sites located in different chromosomes, caused by some specific genetic factors [9, 22–24] that will be discussed in below sections. Linkage disequilibrium also referred as "gametic phase disequilibrium" (GPD) or "gametic disequilibrium" (GLD) [11, 25] in the literature that describes the same nonrandom association of haplotypes within unrelated populations with a distantly shared ancestry, assuming Hardy-Weinberg equilibrium (HWE).

The concept of LD was first described by Jennings in 1917, and its quantification (D) was developed by Lewtonin in 1964. The simplified explanation of the commonly used LD measure, D or D' (standardized version of D), is the difference between the observed gametic frequencies of haplotypes and the expected gametic haplotype frequencies under linkage equilibrium ($D = P_{AB} - P_A P_B = P_{AB}P_{ab} - P_{Ab}P_{aB}$) [26]. Besides D, a various different measures of LD (D', r^2, D^2, D^*, F, X (2), and δ) have been developed to quantify LD [25, 27–29]. The detail formulae and description of LD quantification was well explained by a number of review papers [10, 25, 26] with a number of hypothetical scenarios for LD and LE. The merits, sensitivity, comparison, appropriate statistical tests, and calculation methodology for these LD measures with the utilization of biallelic or multiallelic loci have been extensively described in the literature in detail [10, 26, 30, 31], and have recently been reviewed by Gupta et al. [25]. Hence here we highlight only some of key utility properties of LD measures to provide a brief understanding the merits of LD in association mapping.

Choosing the appropriate LD measures really depends on the objective of the study, and one performs better than other in particular situations and cases; however, D' and r^2 is the most commonly used measures of LD [25, 26]. D' is informative for the comparisons of different allele frequencies across loci and strongly inflated in a small sample size and low-allele frequencies; therefore, intermediate values of D' is dangerous for comparative analyses of different LD studies and should be verified with the r^2 before using for quantification of the extent of LD [26]. The r^2, the square of the correlation coefficient between the two loci have more reliable sampling properties than D' with the cases of low allele frequencies [26]. The r^2 is affected by both mutation and recombination while D' is affected by more mutational histories (it might indicate minimal historic recombination when high D' values used) [10, 25, 26, 31]. Considering the objective, the most appropriate LD quantification measure

needed for association mapping is r^2 that is also an indicative of marker-trait correlations [25, 26, 32]. The r^2 value varies from 0 to 1, and it will be equal to 1 when only two haplotypes are present. The r^2 value of equal to 0.1 (10%) or above considered the significant threshold for the rough estimates of LD to reveal association between pairs of loci [33].

It is noteworthy to briefly mention here that the estimation of above described GLD (commonly used in association mapping) between different loci ordered within gametes assumes that a targeted population or sampled germplasm is randomly mating and under HWE. Nevertheless, many natural populations violate HWE due to different genetic events (bottleneck, mutation, admixture, artificial selection, population structure, etc.) occurred in history of a population, and are under Hardy-Weinberg disequilibrium (HWD). A concept of "zygotic disequilibrium (ZLD)" was introduced for such a nonequilibrium population [34] that measures LD between different loci of gametes. ZLD, being defined as a deviation of joint zygotic frequencies from the expected values of zero zygotic associations [35, 36], has a power to measure nonrandom associations at both gametic and zygotic level [34, 37]. It shares the most of statistical properties of GLD [36], and the results of GLD and ZLD are mostly in agreement, yet ZLD detects more extensive LD than determined by GLD [37]. The statistical models of ZLD measures for biallelic and multilocus data, its application for natural populations, and inference the genetic and demographic events from the comparisons of GLD and ZLD results as well as implication for whole genome association studies (WGAs) were excellently addressed and described by a number of studies [35–37].

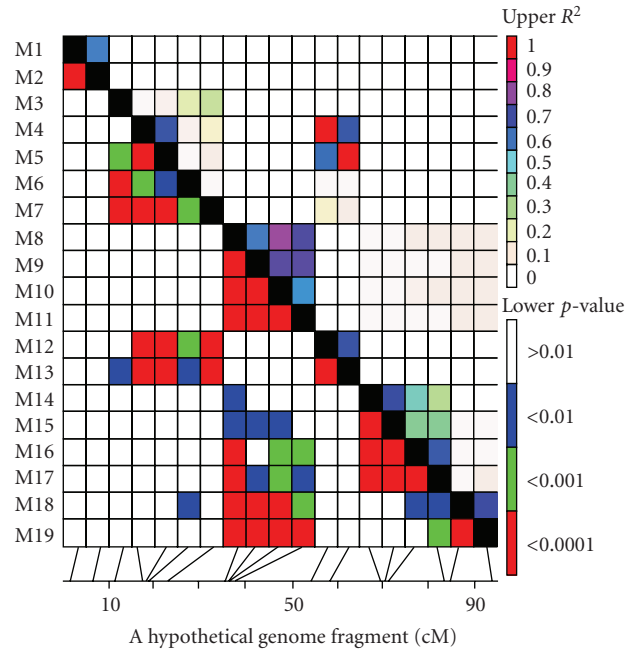


Figure 2: The TASSEL generated triangle plot for pairwise LD between marker sites in a hypothetical genome fragment, where pairwise LD values of polymorphic sites are plotted on both the X- and Y-axis; above the diagonal displays r^2 values and below the diagonal displays the corresponding p-values from rapid 1000 shuffle permutation test. Each cell represents the comparison of two pairs of marker sites with the color codes for the presence of significant LD. Colored bar code for the significance threshold levels in both diagonals is shown. The genetic distance scale for a hypothetical genome fragment was manually drawn. Note: this is for demonstration purposes only and does not have any real impact or correspond to any genomic fragment of an organism.

4.2. Calculation and visualization of LD: LD triangle and decay plots

LD can be calculated using available haplotyping algorithms [26]. One of such efficient methodology is the maximum likelihood estimate (MLE) using an expectation maximization algorithm [38]. Several computer software packages are available and can be utilized for calculation of LD using variety type of molecular markers. These software packages were extensively listed and described in the review by Gupta et al. [25].

Graphical display of pairwise LD between two loci is very useful to estimate the LD patterns measured using a large number of molecular markers. Pairwise LD can be depicted as a color-code triangle plot (Figure 2) based on significant pairwise LD level (r^2, and p-value as well as D') that helps to visualize the block of loci (red blocks) in significant LD. The large red blocks of haplotypes along the diagonal of the triangle plot indicate the high level of LD between the loci in the blocks, meaning that there has been a limited or no recombination since LD block formations. There is freely available specific computer software, "graphical overview of linkage disequilibrium" (GOLD) [39], to depict the structure and pattern of LD. Some other software packages measuring LD such as "Trait Analysis by aSSociation, Evolution and Linkage" (TASSEL) [33, 40] and PowerMarker [41] have also similar graphical display features. The strong block-like LD structures are of a great interest in association mapping which simplifies LD mapping efforts of complex traits [42]. LD blocks are very useful in association mapping when sizes are calculated, which suggest the needs for the minimum number of markers to efficiently cover the genome-wide haplotype blocks in association mapping.

To estimate the size of these LD blocks, the r^2 values (alternatively, D' can also be used) usually plotted against the genetic (cM) or weighted (bp) distance referred to as a "LD decay plot" (Figure 3). One can estimate an average genome-wide decay of LD by plotting LD values obtained from a data set covering an entire genome (i.e., with more or less evenly spaced markers across all chromosomes in a genome) against distance. Alternatively, the extent of LD for particular region (gene or chromosome) can be estimated from an LD decay plot generated using dataset obtained from a region of interest. When such a LD decay plot generated, usual practice is to look for distance point where LD value (r^2) decreases below 0.1 or half strength of D' ($D' = 0.5$) based on curve of nonlinear logarithmic trend line (see, e.g., [33, 43, 44]). This gives the rough estimates of the extent of LD for association study, but for more accurate estimates,

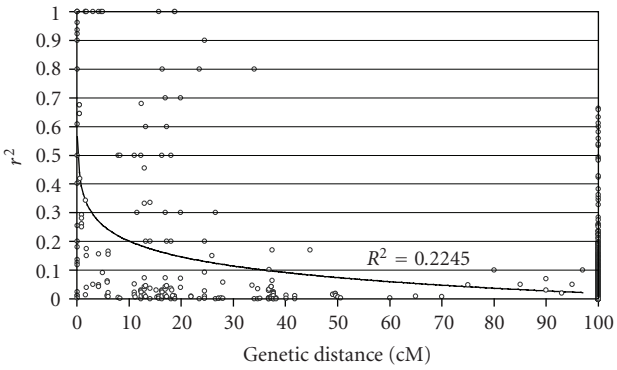


Figure 3: Linkage disequilibrium (LD) decay plot depicted from the LD values of a hypothetical marker data to demonstrate a measure of an average genome-wide LD block sizes. A pairwise LD values (r^2) are plotted against a genetic distance. Inner fitted trend line is a nonlinear logarithmic regression curve of r^2 on genetic distance. LD decay is considered below $r^2 = 0.1$ threshold and based on trend line it is around 38–40 cM in above plot. A pairwise LD between unlinked marker loci is assigned to 100 cM distance point. Note: this is for demonstration purposes only and does not have any real impact or correspond to any genomic fragment of an organism.

highly significant threshold LD values ($r^2 \geq 0.2$) are also used as a cutoff point. The decrease of the LD within the genetic distance indicates that the portion of LD is conserved with linkage and proportional to recombination [22, 25].

4.3. Factors affecting LD and association mapping

There are many genetic and demographic factors that play a role in the shaping of the haplotypic LD blocks in a genome [9, 22, 23, 25, 26]. Although mutation and recombination are one of the strong impact factors influencing LD [24], generally, factors affecting LD can be grouped into two categories: (1) factors that increasing LD, and (2) factors that decreasing LD. The increase of LD is observed with new mutation, mating system (self-pollination), genetic isolation, population structure, relatedness (kinship), small founder population size or genetic drift, admixture, selection (natural, artificial, and balancing), epistasis, and genomic rearrangements [25, 26]. The decrease of LD is observed with high recombination and mutation rate, recurrent mutations, outcrossing, and gene conversions [25, 26].

LD conserved with linkage is very useful for association mapping. However, more often there is a significant LD between pairs of loci located far from each other or even in different chromosomes that might cause spurious correlations in association mapping. These long stretched LD or LD between unlinked loci indicate the existence of other LD generating factors than linkage itself in a genome [9, 22, 23]. One of those factors is selection that generate LD between unlinked loci through "a hitchhiking" effect (high-frequency sweeping and fixation of alleles flanking a favored variant) [45], and epistatic selection or the so-called coadapted genes [46] that is the result of coselection of loci during breeding for multiple traits [26], common in traditional crop breeding programs worldwide.

The population structure (existences of distinctly clustered subdivisions in a population) and population admixture are the main factors to create such an LD between unlinked loci. This primarily happens due to the occurrence of distinct allele frequencies with different ancestry in an admixed or structured population. Theoretically, relatedness generates LD between linked loci, yet it might also generate LD between unlinked loci pairs when predominant parents exist in germplasm groups. There is evidence that relatedness caused LD between linked and unlinked loci in an equal proportion in maize germplasm [22]. The high ratio value of linked to unlinked loci in LD is good indicative to draw conclusion about the role of LD generating factor(s) such as selection or population stratification (cryptic relatedness) [9, 22, 23]. The other factors such as genetic drift or bottlenecks might have also generated LD in a genome [22–24], which is evidenced by nonuniform distribution of LD in chromosomes [24].

Knowing these factors that are increasing or decreasing LD in a genome, obvious question one might ask is whether increased or decreased level of LD is favored in association mapping? Very extensive level of LD (means LD persists within a long distance), theoretically, reduces a number of markers needed for association mapping, but makes resolution lower (coarse mapping). In contrast, less extensive level of LD (means that LD quickly decays within a short distance) requires many markers to tag a gene of interest, but in high resolution (fine mapping). Hence, choosing a population with low or high level of LD depends on the objective of association mapping study. Furthermore, increased LD level due to LD between unlinked loci is not salutary in association mapping since it tends to cause spurious marker-trait associations. LD generated by selection, population structure, relatedness, and genetic drift might be theoretically useful for association mapping in specific situations and population groups that reduces number of markers needed for association mapping [9, 22], but requires serious attention to control factors affecting LD (e.g., population structure and relatedness) to perform unbiased population-based association mapping in plants [41, 47] (see next sections).

There are other factors affecting LD referred to as a whole "ascertainment bias" that are associated with an assayed sample and data characteristics. Some of these factors leading to inaccurate estimate of LD were well reviewed by Gupta et al. [25]. One of such factors largely affecting the LD and leading inaccurate estimates is the presence of minor alleles (also referred as to rare alleles that are present in only 5 to 10% individuals of the sample) in a dataset. Minor alleles are problematic in LD quantification as they largely inflate LD values (in particular the D' and p-values) [43, 48–50]. The r^2 is also very sensitive and has a large variance with rare alleles [43, 51]. Hence in the quantification of LD and association mapping, markers with minor allele frequency of 5–10% (varied from study to study) are (1) removed before analysis (see, e.g., [17, 18, 43, 44, 51]), (2) pooled into common allelic class (see, e.g.,[44, 46]), and (3) replaced with missing values (see, e.g., [52, 53]).

4.4. Estimation of LD using dominant versus codominant markers

The quantification methodology of LD, perfectly suitable for biallelic codominant type of markers (majorly, single nucleotide polymorphisms (SNPs) and now largely extended to multiallelic simple sequence repeats-SSRs), has been well developed and used in human, animal, and plant populations (for reviews see [25, 27–30]). LD quantification using dominant markers (such as random amplified polymorphic DNAs-RAPD, and amplified fragment length polymorphisms-AFLPs) is poorly explored and usually subject to wrong perception and interpretation. However, many underrepresented plant species, like forest trees, or other crops with limited genomic information largely rely on dominant type of markers such as RAPDs and AFLPs [54]. Furthermore, even with codominant, and multiallelic SSR markers, there is a great challenge with assigning correct allelic relationships (identity by decent) of multiple band amplicons when diverse, reticulated, and polyploid germplasm resources, lacking historical pedigree information, are genotyped. Misassignment of allelic relationships of loci is the concern in association analysis [55]. To avoid such a challenging cases, (1) one might select only single band SSR loci and code a dataset as a codominant marker type, yet such a single band SSRs are usually not many in polyploid crop genomes and yield also multiple bands when very diverse germplasm resources are genotyped; (2) alternatively, multiple-band SSRs with unknown allelic relationship may be scored as a dominant marker taking each band as an independent marker locus (uniquely) with a clear size band separation (see, e.g., [52, 56]).

Could a dominant marker data be used for LD quantification? There are some reports where LD level of natural forest tree populations has been measured using dominant markers (AFLPs) and commonly used statistical approach (see, e.g., [57]). There are also a number of reports where dominantly coded (present versus absent) marker data of RAPD, RFLP, AFLP, "candidate gene" (CAPs), and SSRs were successfully used in genome-wide LD analyses and LD-based association mapping in plants (see, e.g., [17–19, 56, 58–60]), demonstrating the feasibility of dominantly coded molecular data in revealing of haplotypic associations. Although a dominant type of coding has limited statistical power compared to codominant markers in population-based analyses because of missing heterozygote information, previous studies suggested that it can be successfully applied to the clustering of individuals and grouping of populations using a Bayesian approach when a large number of loci are genotyped [61, 62]. Dominant-type markers can be a useful tool to estimate the kinship coefficients between individuals [63].

Recently, Li et al. [54] investigated the use of dominant markers in estimation of LD in diploid species and developed appropriate EM algorithm. Based on their conclusion from the comparative data simulation of dominant versus codominant markers, the dominant-type markers could effectively be used in LD analysis with preferentially large number of marker loci and population sample sizes of ≥200 for high heterozygous (proportion of alternative alleles (present versus absent) in a data set, i.e., 0.5 versus 0.5) marker data or with even larger sample size ≥400 for low-heterozygous (i.e., 0.9 versus 0.1) dominant markers. It is also recommended that a mixture of codominant and dominant markers should be used to better characterization of a genetic structure of a population [54].

4.5. LD quantification in plants

LD quantification and LD-based association mapping have been a research objective in plants beginning with the model organism as Arabidopsis, and now extended to crops as maize, barley, durum wheat, spring wheat, rice, sorghum, sugarcane, sugar beet, soybean, and grape, as well as in forest tree species, and forage grasses.

Nordborg et al. [20] sequenced 0.5–1 kb long 13 fragments from a 250 kb region surrounding the flowering time *FRI* gene in a 20 global sample of *A. thaliana*, highly selfing model plant species. They determined that LD decays within a 1 cM distance or 250 kb. Later, investigation of the same authors [21] with markers surrounding the disease resistance locus *RPM1* in a globally-derived set of 96 Arabidopsis accessions revealed that a genome-wide LD extended up to 50–250 kb. LD blocks extended up to 50–100 cM in local Michigan Arabidopsis populations. These long-stretched LDs in local Arabidopsis population were explained as a genetic bottleneck or founder effect through introduction *A. thaliana* into North America in recent past (200 years ago). In contrast, in other study that targeted the region surrounding another disease resistance gene *rps5*, Tian et al. [64] reported much smaller LD block size, extended up to only 10 kb. Likewise, LD quickly decays within 10–50 kb distance around the *CLAVATA* 2 region of Arabidopsis [65]. Recently, Ehrenreich et al. [66] reported the LD decay within ~10 kb in extensive sequence analysis of 600-bp fragments of the regions *MORE AXILLARY GROWTH 2* (*MAX2*) and *MORE AXILLARY GROWTH 3* (*MAX3*) in a panel of 96 accessions from a restricted geographic range in Central Europe. In their genome-wide survey of 1347 fragments of 600-bp lengths, Plagnol et al. [67] reported that LD completely disappears after ~100 kb, which is comparable to that observed in human.

In maize (*Zea mays* L.), a highly outcrossing crop species, very rapid genome-wide LD decay was determined. Tenaillon et al. [68] first reported the extent of LD for maize, genotyping of 21 loci of chromosome 1 over the 25 individuals of the exotic landrace and United States maize germplasm. An average LD decay was determined to occur within 400 bp with $r^2 = 0.2$ and extended up to 1000 bp (~1 kb) in a group of US inbred lines. Later, Remington et al. [43] also reported a very rapid decline of LD in their survey of 6 genes (1.2–10 kb long) in 102 inbred lines, including tropical and semitropical lines with a wide genetic diversity. For these surveyed genes, LD declined generally within 200–2000 bp with $r^2 = 0.1$ except *sugary1* (*su1*) loci, where LD remained significant ($r^2 = 0.3 - 0.4$) for 10 kb distance. This was explained by strong selective episodes in *su1* gene. In the same study, Remington et al. [43] found higher level of LD

with 47 SSR markers compared to those obtained from SNP data. This result was explained by different mutation rate of these two marker systems that tends to capture different historic information.

Long stretches of LD for maize also were reported. Thornsberry et al. [69] measured LD in and around the *Dwarf8* locus. They found "localized LD" (i.e., restricted to particular regions, meaning that high LD stretches interspersed with regions of low LD) extended up to 3 kb. Jung et al. [70] reported the extent of LD within 500 kb in surveying *adh1* locus. Stich et al. [22] examined the genetic diversity and LD in a cross section of 147 European and United States elite inbred material with 100 SSRs. They reported an average significant ($P < .5$) LD block size of 26 cM for flint group, or 41 cM for dent group with nonuniform distribution of LD among 10 chromosomes. They showed a very long stretched LD blocks up to 105 cM in chromosome 2 and up to 103 cM in chromosome 7 in flint and dent groups, respectively. Obtaining of different result from earlier studies [43] was explained due to using (1) much higher marker density, and (2) both related and unrelated inbred lines. In another study, the same authors [9] examined 72 European elite inbred lines with 452 AFLP and 93 SSR markers and reported much shorter average LD block sizes for AFLP (4 cM), but extensive LD for SSR (30-31 cM) in both flint and dent germplasm groups. This suggested a potential for exploiting both markers in association mapping, but with the favor of SSRs over AFLPs because of power of detecting LD. Recently, Andersen et al. [71] reported that LD is persisted over entire 3.5 kb *phenylalanine ammonia lyase* (*PAL*) gene with the $r^2 > 0.2$ in a survey of 32 European maize inbred lines.

In the selfing tetraploid wheat (*Triticum durum* Desf.), Maccaferri et al. [50] quantified LD in a 134 durum wheat accessions that extended up to 10 and 20 cM with $D' = 0.67$ and 0.43, respectively. In hexaploid wheat (*Triticum aestivum* L), almost completely self-pollinating species, strong LD was determined to occur on average within <1 and ~5 cM for region on chromosome 2D and centromeric region 5A that was surveyed with 36 SSR markers in a 95 cultivars of winter wheat [52]. Recently, Chao et al. [72] investigated the genome-wide LD among 43 US wheat elite cultivars and breeding lines representing seven US wheat market classes using 242 SSRs distributed throughout the wheat genome. For this germplasm collection, a genome-wide LD estimates were generally less than 1 cM for the genetically linked loci pairs. Most of the LD regions observed were between loci less than 10 cM apart, suggesting LD is likely to vary widely among wheat populations [72]. Tommasini et al. [56] reported that LD on chromosome 3B extended up to 0.5 cM in 44 varieties or 30 cM in 240 RIL populations of winter wheat, surveyed with 91 SSR and STS markers. This suggested usefulness of cultivar germplasm over biparental mapping population in association mapping.

In rice (*Oryza staiva* L), a selfing species, Garris et al. [73] examined the LD surrounding disease resistance locus *Xa5* using 21 SSRs in a survey of 114 rice accessions. They determined the strong LD within 100 kb with $r^2 = 0.1$. Agrama and Eizenga [74] investigated LD patterns in a worldwide collection of *Oryza staiva*, and its wild relatives using 176 SSR markers. Although it was not specifically indicated, LD decay plot suggests a long range LD declining ~50 cM with $D' = 0.5$ in the "International" and "US" rice collections. Interestingly, LD persisted over an average of 225 cM distance with significant $D' > 0.5$ in a wild accessions. In contrast, many other studies reported a less extent of LD in wild and landrace (broad-based) germplasm and high extent of LD in cultivar (narrow-based) germplasm resources in plants [9, 43]. There is evidence that the LD is remarkably different in other rice species. Rakshit et al. [75] reported that LD in *O. rufipogon* decays within 5 kb, while it declines at 50 kb in *O. sativa* ssp. *indica* accessions. Mather et al. [76] observed that the extent of LD is greatest in temperate *japonica* (>500 kb), followed by tropical *japonica* (~150 kb) and *indica* (~75 kb) that was revealed by using unlinked SNPs. LD extends over a shorter distance in *O. rufipogon* ($\ll$40 kb) than in any of the *O. sativa* groups assayed in their study [76].

LD also has been extensively quantified another highly self-pollinated crop, barley (*Hordeum vulgare* L), where the extent of LD varied from 10 cM to 50 cM range depending on assayed set of a germplasm [17, 77]. Caldwell et al. [51] measured LD in four genes surrounding hardness locus (*Ha*) in three different gene pools and reported a long stretched LD extended up to at least 212 kb in inbred barley and 98 kb in landrace barley germplasm. In contrast to these long range LDs observed in barley germplasm, Morrell et al. [78] reported a rapid decay of LD detected within 300 bp in their study of 18 nuclear genes (average length of 1 361.1 bp) in 25 diverse wild barley accessions. In that, LD completely disappeared within a 1200 bp distance. This demonstrates another example of variability of LD quantification across germplasm resources, breeding material, and regions tested.

Furthermore, genome-wide LD has been quantified for many other plant species that extended up to 10 cM in sugar cane (*Saccharum*) [10], 10–50 cM in soybean (*Glycine max*) [79, 80], 3 cM in sugar beet (*Beta vulgaris* L) [81], 50 cM in sorghum (*Sorghum bicolor*) [44], 5–10 cM in grape (*Vitis vinifera* L) [53], 16–34 kb in poplar (*Populus trichocarpa*) [82], <500 bp in European aspen (*Poplus termula*) [83], 2000 bp in loblolly pine (*Pinus taeda*) [84], 1000 bp in Douglas-fir (*Pseudostuga menziensii*) [85], 100–200 bp in Norway Spruce (*Picea abies*) [86], 200-1, 200 bp in silage maize (*Zea mays L*) [87, 88], and 500–2000 bp in ryegrass (*Lolium perenne*) [89–91]. Also, LD quantification for other important crops, perhaps, is in progress. In this context, recently, we have quantified LD level for improved varieties and landrace stock germplasm of cotton (*Gossypium hirsutum* L) [92]. Survey of 200 microsatellite markers in 335 *G. hirsutum* variety germplasm demonstrated that a genome-wide averages of LD extended up to genetic distance of 25 cM with $r^2 > 0.1$. Likewise, our another companion study using 95 core set microsatellite markers in a total of 286 "exotic" *G. hirsutum* revealed that a genome-wide averages of LD decays within the genetic distance at <10 cM in the landrace stocks germplasm and >30 cM in photoperiodic variety germplasm, providing evidence of the potential for association mapping of agronomically important traits

in cotton (Abdurakhmonov et al. unpublished, submitted elsewhere for publication).

4.6. Implications for association mapping gained from LD quantification studies in plants

Important information and implication for association mapping gained from above studies are that: (1) LD more quickly declines in outcrossing plant species than highly self-pollinating plants, enabling high resolution mapping of a trait of interest in outbreeder plant germplasm. At the same time, LD rapidly declines in crop variety groups (even in selfing species) compared to populations derived from biparental crosses, which provides an advantage of discovery more polymorphisms in the variety germplasms than biparental populations of self-pollinated crops [56]; (2) the extent of LD varies across the genomic regions, among population samples and between species with the examples of "localized LD"; (3) LD measures differ per marker systems used as a reflection of capturing of different historic information in a genome due to different mutation rate (e.g., SNP versus SSR or AFLP versus SSR); (4) an estimate of genome-wide averages for the extent of LD in plant germplasm may not adequately reflect LD patterns of specific regions or specific population groups. Each of these specific regions or population groups should additionally be explored for the extent of LD in order to conduct successful association mapping of variants within regions or populations of interest; (5) LD blocks in narrow-based germplasm groups are longer than broad-based germplasm groups in plants [9, 43]. This suggests an opportunity perform coarse mapping with less number of markers in narrow-based plant germplasm and then fine mapping in broad-based plant germplasm, assuming that genetic causations is sufficiently similar across germplasm groups [12]. This also suggest an opportunity develop a set of mapping populations with the required amount of LD and diversity for high-resolution mapping through directed crossing between selected broad- and narrow-based germplasm groups [86]; and (6) confounding population characteristics and biological behavior have serious impact on pattern and structure of LD in plant germplasm resources that need to be taken into consideration in conducting unbiased association mapping.

5. ASSOCIATION STUDIES IN PLANTS

5.1. The methodology overview

There are many types of different methodologies that have been developed and initially are widely used for association mapping studies in human (comprehensively reviewed by Schulze and McMahon [93]), yet perfectly applicable without change or case-to-case modifications for wide range of organisms, including plants. Lately, some considerably successful achievements have been made to develop powerful, more precise, and unbiased population-based association-mapping methodology for plants. Here, we provide a brief overview for a basic concept and ideology of widely used pioneer methodologies for association mapping, and then highlight the latest developments in the methodology and experimental design of association mapping in plant population with the examples of association mapping of useful traits in crop species.

The classical methodology and design of association mapping is "case and control" (also referred to as "case-control") approach that identifies the causative gene tags in the comparison of allele frequencies in a sample of unrelated affected (referred to as "cases") individuals and a sample of uninfected or healthy individuals (referred to as "controls") [93, 94]. This design requires an equal numbers of unrelated and unstructured "case-control" samples for accurate mapping. The Pearson chi-square test, Fisher's exact test, or Yates continuity correction can be used for a comparison of allele frequencies and detection of an association between a disease phenotype and marker. Although favored, the random sampling individuals from a population do not provide the equal representation of case and controls in the mapping population since cases in the population are usually low, thus requires special efforts to collect the cases. Case and control approach is seriously affected by the existence of population structure and stratification that caught the attention of scientist [93]. Falk and Rubinstein [95] developed a haplotype relative risk (HRR) approach that minimizes, but not eliminates population stratification issues in association mapping [96]. In that, first, a "pseudo-control" group (containing combination of two alleles that are not transmitted to affected offspring) is created; then, the marker allele frequencies in case and "pseudocontrol" groups are correlated [93].

To efficiently eliminate the confounding effects coming from population structure and stratification, Spielman et al. [97] developed transmission disequilibrium test (TDT) method that compares transmission versus nontransmission of marker alleles to affected offspring by using chi-square test [93], assuming a linkage between marker and trait. The TDT design requires genotyping of markers from three individuals: one heterozygous parent, one homozygous parent, and one affected offspring. Although HRR performs better with unstructured sample than TDT because of its power to completely eliminate spurious association with good experimental design, later is widely used as a tool for unbiased fine mapping of traits in the presence of linkage with a biallelic, one marker model that can accommodate pedigree structure [30, 93].

Nonetheless, initial TDT approach had issues with the use of multiallelic markers, multiple markers, missing parental information, extended (larger) pedigrees, and complex quantitative traits [93]. To address these issues, a variety of extensions of TDT approach were developed and applied for multiallelic markers (i.e., GTDT, ETDT, MC-Tm) [98–102], multiple markers [103–105], missing parental information (Curtis-test, S-TDT, SDT, 1-TDT, C-TDT or RC-TDT) [96, 106–110], which were reviewed by Schulze and McMahon [93] in detail. Shortly after publication of various extensions of TDT to multiallelic and multiple markers, the extensions for X-linked genes, such as XS-TDT or XRC-TDT were developed and applied [111]. TDT approach was also extended to pedigrees of any size as a PDT

approach [112, 113] that was demonstrated more powerful than TDT, and S-TDT or SDT under the assumption of high disease prevalence [93, 114].

Further, there were many studies to extend the TDT approaches to QTL and covariates [93]. One of the comprehensive approaches, QTDT was developed with its three different extensions for quantitative traits for any pedigree structure [115, 116]. These family-based association-mapping approaches have their other improvements using more powerful statistical and robust algorithmic procedures, such as likelihood-base statistics and EM algorithm (TDT-LIKE, LRT, EM-LRT) [117–119]. The unified family-based association test package (FBAT) incorporating some of TDT is also developed [120–122] to deal with wide types of experimental designs. The next generation of association mapping approaches in both "case and control" and family-based designs, referred to as identity by decent (IBD) mapping [123], haplotypes-sharing analysis (HSA) [124], and decay of haplotypes sharing (DHS) [125], involves the analysis of haplotypes by testing the length of haplotypes in the data sample, assuming affected individuals will have a longer haplotypes than controls [93].

Although family-based association mapping methodology is effective to control confounding effects of a sample and remove spurious associations, it is less powerful design [126] and have its disadvantageous sides compared to case-control [93] that led to develop the methodologies with better controlling of population structure and stratification. Such an improved methodology for a case and control design or random samples from a population involves the use of additional markers that have neutral effect (null loci) to the trait of interest in the analysis. This approach is referred to as the genomic control (GC) that finds confounding effects of a population and corrects it, thus enabling to remove spurious associations [127, 128]. Although GC is powerful then TDT [128], it will not remove spurious associations in highly structured populations. Zhao et al. [129] put it as

> "*Methods like "genomic control," which simply rescale p-values without changing the ranking of loci are not likely to be useful in genome-wide scans where the existence of true positives is not in doubt.*"

To better deal with highly structured populations, Pritchard et al. [47, 62] developed approach of structured association (SA). SA first searches a population for closely related clusters/subdivisions using Bayesian approach, and then uses the clustering matrices (Q) in association mapping (by a logistic regression) to correct the false associations. Population structure and shared coancestry coefficients between individuals of subdivisions of a population can be effectively estimated with *STRUCTURE* program using several models for linked and unlinked markers [130]. Similar type of methodology measuring and using the population subdivisions (K) in association mapping referred to as "mixture model" was proposed by several other studies [131, 132]. However, SA incorporating only population structure information in the analysis is not good enough itself when highly structured population with some degree of related individuals used in the association mapping.

Hence, recently, Yu et al. [133] developed new methodology, a mixed linear model (MLM) that combines both population structure information (Q-matrix) and level of pairwise relatedness coefficients—"kinship" (K-matrix) in the analysis. To perform MLM: (1) Q-matrix is generated using *STRUCTURE*, (2) the pairwise relatedness coefficients between individuals of a mapping population (K-matrix) [134] measured using *SpaGedi* software [135], and (3) then both Q- and K-matrices are used in association mapping to control spurious associations. Although computationally intensive, MLM approach found to be effective in removing the confounding effects of the population in association mapping [133].

Later Zhao et al. [129] extensively tested the MLM approach of Yu et al. [133] in their global set of 95 highly structured Arabidopsis population and came to overall agreement with better performance of $Q + K$ MLM model than any of the other tests that used K- or Q-matrix alone. However, they also noted that (1) K matrix would alone be good enough if a kinship estimated as a proportion of shared haplotypes for each pair of individuals (as denoted K^*); (2) the replacement of Q-matrix (from the computational intensive *structure* analysis) [130] with P-matrix (from more robust principal component analysis) [136] performed similarly to MLM of Yu et al. [133], thus suggesting a potential for future replacements; (3) removing of the confounding effects will also subject to remove true associations with biological effect, which is strongly correlated with population structure that requires a caution; and (4) in a small and highly structured population, the causations with major effect should be expected to be found and, perhaps, larger samples and adequate marker densities are needed for genome-wide dissection of the most traits of interest segregating in an association mapping population [129].

There are other types of mixed models for association mapping that have its own advantages to control population confounding effects and tag a genetic causative of a trait of interest. One of such mixed models utilizes a sample with pedigree information to measure a pedigree-based relatedness and incorporates it directly in QTL-mapping and association mapping [59, 137, 138]. This type of mixed model for known pedigree population combines haplotype effects with pedigree-based structure of variance-covariance relatedness matrix and random polygenic effect that control the population structure [59, 139]. The efficiency of pedigree population for association mapping depends on the population size of pedigree founders (i.e., pedigree population obtained from just two parents will not provide significant level of LD) and the level of relatedness of the founders. Latter is very important and may still lead to spurious association due to initial population structure (mostly unknown) coming from founders that needs to be analyzed also by using *STRUCTURE* [140].

However, as stated by Malosetti et al. [59] and others [140] the pedigree-based mixed model is highly appropriate in association mapping in crops due to (1) plant breeding

programs have already generated many useful pedigree populations that contain LD useful for association studies but cannot be used as an independent LD-mapping population, and (2) many historical trait data sets in plant breeding are unbalanced that have been collected over multiple-years, and multienvironmental trials. At the same time, issues with obtaining the fine-grained pedigree information and difficulty of finding population structure of narrow-based elite cultivars are the concern in pedigree-based mixed model. There is another mixed model that combines the Bayesian variable selection for mapping multiple QTLs and LD mapping method, incorporating estimates of population structure, but not relatedness. This approach was used for association mapping in highly selfing rice germplasm [58]. Authors stated that incorporation of multiple QTL effects and population structure efficiently reduces spurious association and useful for future whole genome associations, with the development of more complex models dealing with differences of LD and effect of QTL alleles between populations.

The other mixed model approach combines QTL and LD analyses of distinct studies. In that, QTLs or candidate genes with already annotated biological function(s) are used as a priori information in association mapping [140, 141]. This is one of the effective alternative strategies in association mapping that allow reducing the total amount of marker genotyping (because of preselecting of markers restricted to QTL region) in less number of individuals. This increases the power and precision of the trait-marker correlations [142].

5.2. Power of association mapping

The power of association mapping is the probability of detecting the true associations within the mapping population size that really depends on (1) the extent and evolution of the LD in a population, (2) the complexity and mode of gene action of the trait of interest, (3) sample size and experimental design. The power can be increased utilizing the better data (knowledgeable experimental design and accurate measurements) and increasing the sample size. In QTL mapping studies, there are specific statistical approaches to estimate the false-positive level of the obtained strong (p-value) associations (control for Type I error) such as a permutation test [143] or false discovery rate (FDR) [144].

A statistical approach within the Bayesian framework is used test the reliability of obtained significance (p-values) in association mapping because of possibility of getting unreliable values due to (1) overestimation of effects (selection bias), (2) association coming from neglecting confounding effects of a sample, (3) poor experimental design, and (4) instability of genetic effects across different environments [142]. Ball [142] developed a methodology, combining the Bayesian and non-Bayesian approaches, that determines the *Bayes* factors guiding to properly design the experiments with given power to detect reliable effects. To detect the reliable effects in association mapping, experiments should be designed at least with the *Bayes* factor of 20 that may require much larger sample sizes. *Bayes* factor provides stronger evidence than conventional p-values [142]. If given *Bayes* factor value (say B = 20) reached with larger sample than the original experimental design, then, the original results indicate a very weak evidence to provide the real effects [142]. At this point, requirement for larger sample size might make association mapping disadvantageous over a traditional QTL-mapping. However, the sample size for association mapping can be decreased keeping the high power with (1) preselecting a priori known QTL regions or candidate genes (from QTL-mapping and expression analyses), (2) using the large populations with samples longer LD block that require a less number of markers to find useful associations, (3) an alternative experimental design (i.e., TDT), and (4) choosing the single marker from the haplotypes of interest that would cut also marker number and so genotyping cost [142]. *Bayes* factor can be calculated using *R* function of *ld.design* from *ldDesign* package [140]

5.3. Examples from reports

The pioneer association studies in plants were performed by Beer et al. [166] in oat, and by Virk et al. [167] in rice. Beer et al. [166] associated 13 QTL with RFLP loci using 64 oat varieties and landraces, yet without considering the population structure that resulted in more increased associations than what were obtained in separate analysis of subpopulations [11]. Virk et al. [167] predicted 6 trait values using RAPD markers in rice germplasm. Later, association mapping was extended to sea beet, barley, maize, wheat, potato, more examples in rice, and Arabidopsis that have utilized population level of LD considering a population structure. Hansen et al. [19] reported association of ALFP markers with bolting gene in sea beat. In barley, various traits such as yield, yield stability, heading date, flowering time, plant height, rachilla length, resistance to mildew and leaf rust were associated with many different types of molecular markers [17, 18, 157, 158]. In maize, flowering time and plant height [43, 69] were associated using SNP and SSRs. Following these pioneer studies of association mapping in maize, several other traits such as phenotypic variation in flowering time, endosperm color, starch production, maysin and chlorogenic acid accumulation, cell wall digestibility, and forage quality were associated using SNP markers of candidate genes [71, 87, 88, 149–153].

In wheat, Breseghello and Sorrells [52] reported first association mapping of kernel size and milling quality in a collection of USA winter wheat using SSRs. Following this work, association mapping of a high molecular-weight glutenin [159] and blotch resistance [56] were reported that utilized SNPs, SSRs, and STS markers. In rice, association mapping has not widely been applied yet due to highly structured population of rice (due to high selfing) [58, 133]. However, Zhang et al. [156] successfully used association mapping for multiple agronomic traits using discriminant analysis (DA) with SSR and AFLP markers. Recently, Iwata et al. [58] associated RFLP markers with width and length of milled rice grains in a set of 332 rice germplasm using their multiple QTL model considering the

Table 1: Linkage disequilibrium and association mapping studies in plants.

Species	Mating system	LD extent	Mapped traits	*Approach used
Arabidopsis	Selfing	10–250 kb and 50–100 cM [20, 21, 64, 66, 67]	Flowering time, growth response, pathogen resistance, and branching architecture [66, 129, 145–148]	One way ANOVA, simple regression, SA, MLM
Maize	Outcrossing	200–2000 bp [43, 68], 3–500 kb [43, 69–71], 4–41 cM [9, 22]	Plant height, flowering time, endosperm color, starch production, maysin and chlorogenic acid accumulation, cell wall digestibility, forage quality, and oleic acid level [43, 69, 71, 87, 88, 149–154]	GLM, SA, MLM, WGA
Rice (*indica, japonica* and *rufipogon*)	Selfing	5–500 kb [73, 75, 76] 50–225 cM [74], 20–30 cM [155]	Multiple agronomic traits such as plant height, heading date, flag leaf length and width, tiller number, steam diameter, panicle length, grain length and width, grain length/width ratio, grain thickness, 1000-grain weight, width and length of milled rice grains [58, 155, 156]	DA, MLM, mixed model with multiple QTL effect
Barley	Selfing	10–50 cM [16, 77], 98–500 kb [51], 300 bp [78]	Yield, yield stability, heading date, flowering time, plant height, rachilla length, resistance to mildew, and leaf rust were associated with many different types of molecular markers [17, 18, 157, 158]	Pearson correlation; regression, ANOVA
Tetraaploid wheat	Selfing	10 and 20 cM [50]	N/A	N/A
Hexaploid wheat	Selfing	<1–10 cM [52, 56, 72]	Kernel size and milling, a high molecular weight glutenin and blotch resistance [52, 56, 159]	GLM-*Q*, LMM
Potato	Selfing	0.3–1 cM [25, 60], 3 cM [160]	Resistance to wilt disease, bacterial blight, *Phytophtora*, and potato quality (tuber shape, flesh color, under water weight, maturity, and etc.)[59, 60, 138, 160]	Nonparametric Mann-Whitney U test, standard two sample *t*-test, GMM
Soybean	Selfing	10–50 cM [79, 80],	Seed protein content [80]	WGA
Sorghum	Outcrossing	50 cM [44]	N/A	N/A
Grape	Vegetative propagation	5–10 cM [53]	N/A	N/A
Sugarcane	Outcrossing/ Vegetative propagation	10 cM [10]	N/A	N/A
Sugar beet	Outcrossing	3 cM [81]	N/A	N/A
Forage grasses (silage maize and ryegrass)	Outcrossing	200–2000 bp [87–91]	Cold tolerance, flowering time and forage quality, water-soluble carbohydrate content [87, 88, 161, 162]	Multiple linear regression; ANOVA
Forest trees (Norway spruce, Loblolly pine, poplar, European aspen, Douglas-fir)	Outcrossing	100–200 bp [86], ~500–2000 bp [83–85]	Early-wood microfibril angle trait, wood density and wood growth rate [141, 163]	ANOVA; combination of LD and QTL mapping

*MLM: mixed linear model [133]; GLM: general linear model without population structure [71]; GLM-*Q*: general linear model using population structure matrix (*Q*) or the least square solution to the fixed effects GLM [56]; DA: discriminant analysis [156]; SA-structured association [47]; LMM: linear mixed model [52]; WGA: whole genome association [154, 164, 165]; GMM: general mixed model [59]; ANOVA: analysis of variance test; N/A—not available (search of known major online library database as of December 2007).

population structure. Association mapping approach was also successfully applied in tetraploid potato where resistance to wilt disease [138], bacterial blight [60], *Phytophtora* [59] that utilized a pedigree-based mixed model.

To date association mapping has also been extended to long lifespan plant species, forest tree populations [163], where associations of polymorphisms in *cinnamoil CoA reductase* (*CCR*) with earlywood microfibril angle trait [141],

and polymorphisms a putative stress response gene with wood density and wood growth rate [163] were reported. There are also the examples of association mapping successes for cold tolerance, flowering time, water-soluble carbohydrate content, and forage quality in forges species that have recently been reviewed by Dobrowolski and Forster (Table 1) [87, 88, 161].

Association mapping of traits in Arabidopsis also has been reported and overall suitability of the approach well documented. Associations of *CRY2* with flowering time were reported [145, 146]. Balasubramanian et al. [147] reported the association of *PHYC* with flowering and growth response in Arabidopsis. Later Zhao et al. [129] revisited to these association results with their mixed model approach and reproved some of previously reported associations (with *PHYC*), but challenged the power of these associations detected by using "standard linear methods without correcting population structure." They put it as "*Clearly, none of these polymorphisms would have been picked up in a genome-wide scan*" while noting the use of different sample and trait measurements in the original studies. They also reported one of the significant flowering time associated polymorphisms in *CLF* gene in their genome-wide analysis using MLM [129]. Flowering time (in *FRI* gene) and pathogen resistance (in *Rpm1*, *Rps5*, and *Rps2* genes) associated polymorphisms were also reported [148]. Recently, Ehrenreich et al. [66] reported polymorphisms of candidate genes (*SPS1*, *MAX2*, and *MAX3*) associated with branching architecture in a survey of 36 genes involved in branch development that were genotyped in a panel of 96 Arabidopsis accessions from Central Europe.

5.4. Choice of the appropriate approach

Table 1 summarizes the LD and association mapping efforts in plants including some of very recent whole genome association mapping studies. As one can see, within the frame of above highlighted association studies in plants, various association mapping methodologies (Table 1), molecular markers (both dominant and co-dominant markers), and plant germplasm resources (including landrace stocks, elite germplasm, and experimental populations—e.g., RILs) have been utilized. Identifying of the most appropriate approach and marker systems, therefore, is challenging and might be irrelevant case-to-case basis.

Choosing the appropriate association mapping depends on (1) the extent and evolution of the linkage disequilibrium in a population, (2) the level of population structure and stratification, (3) availability of pedigree information, (4) complexity of the trait of interest under study, and (5) availability of the genomic information and resources. Based on reported studies, GC is favored approach when population structure is suspected, but failed to be detected [59]; however, MLM considering both relatedness and population structure [133] and pedigree-based mixed model [59] or multiple QTL model [58] performs well in most cases with highly structured and stratified population although one still might argue based on his own experience, knowledge, and type of gemplasm used. According to Stich et al. [23], SA and MLM models do not "explicitly correct" for LD caused by selection and genetic drift, the major factors causing LD in plant germplasm and breeding materials. Hence Stich et al. [23] suggested use of family based association approach [168] with breeding materials. However, again the choice of methodology greatly depends on the germplasm used for mapping. The germplasm materials used for association mapping were comprehensively discussed by Breseghello and Sorrells [169].

6. CONCLUSIONS

Thus the association mapping methodology, initially developed by the human geneticists, has found its successive application in plant germplasm resources, in particular after recent improvements in minimization of spurious associations. The examples of association mapping studies performed in various plant germplasm resources including model plant Arabidopsis and extended to various crop germplasm largely demonstrate the flourish of crop genomics era with the utilization of powerful LD-based association mapping tool. This is also a good indicative of the potential utilization of this technology with the other crops and plant species in the future. Currently, a number of such studies are, perhaps, in progress in many laboratories worldwide. The near-future completion of genome sequencing projects of crop species, powered with more cost-effective sequencing technologies, will certainly create a basis for application of whole genome-association studies [164], accounting for rare and common copy number variants (CNV) (for review see, e.g., [165]) and epigenomics details of the trait of interest in plants, which is widely being applied in human genetics with great success. This will provide with more powerful association mapping tool(s) for crop breeding and genomics programs in tagging true functional associations conditioning genetic diversities, and consequently, its effective utilization.

ACKNOWLEDGMENTS

The authors are grateful to the Academy of Sciences of Uzbekistan and ARS-FSU Scientific Cooperation Program, Office of International Research Programs, USDA-ARS for financial support of their research in Uzbekistan. The authors would like to thank anonymous reviewer(s) of the manuscript for valuable suggestions.

REFERENCES

[1] J. Ross-Ibarra, P. L. Morrell, and B. S. Gaut, "Plant domestication, a unique opportunity to identify the genetic basis of adaptation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, supplement 1, pp. 8641–8648, 2007.

[2] J. E. H. Bermejo and J. Leon, Eds., *Neglected Crops 1492 from a Different Perspective*, Food and Agriculture Organization (FAO) Corporate Document Repository, Botanical Garden of Córdoba, Andalusia, Spain, 1992, http://www.fao.org/docrep/T0646E/T0646E01.htm.

[3] Food and Health Organization's 1999 Report on the State of Food Insecurity in the World, http://www.fao.org/News/1999/img/SOFI99-E.PDF.

[4] G. A. Van Esbroeck, D. T. Bowman, O. L. May, and D. S. Calhoun, "Genetic similarity indices for ancestral cotton cultivars and their impact on genetic diversity estimates of modern cultivars," *Crop Science*, vol. 39, no. 2, pp. 323–328, 1999.

[5] B. A. Meilleur and T. Hodgkin, "*In situ* conservation of crop wild relatives: status and trends," *Biodiversity and Conservation*, vol. 13, no. 4, pp. 663–684, 2004.

[6] B. C. Y. Collard, M. Z. Z. Jahufer, J. B. Brouwer, and E. C. K. Pang, "An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts," *Euphytica*, vol. 142, no. 1-2, pp. 169–196, 2005.

[7] B.-H. Liu, *Statistical Genomics: Linkage, Mapping, and QTL Analysis*, CRC Press, New York, NY, USA, 1998.

[8] R. L. Wu, C.-X. Ma, and G. Casella, *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL*, Springer, New York, NY, USA, 2007.

[9] B. Stich, H. P. Maurer, A. E. Melchinger, et al., "Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers," *Molecular Breeding*, vol. 17, no. 3, pp. 217–226, 2006.

[10] S. A. Flint-Garcia, J. M. Thornsberry, and E. S. Buckler IV, "Structure of linkage disequilibrium in plants," *Annual Review of Plant Biology*, vol. 54, pp. 357–374, 2003.

[11] J. L. Jannink and B. Walsh, "Association mapping in plant populations," in *Quantitative Genetics, Genomics and Plant Breedinging*, M. S. Kang, Ed., pp. 59–68, CAB International, Oxford, UK, 2002.

[12] D. B. Goldstein and M. E. Weale, "Population genomics: linkage disequilibrium holds the key," *Current Biology*, vol. 11, no. 14, pp. R576–R579, 2001.

[13] K. M. Weiss and A. G. Clark, "Linkage disequilibrium and mapping of human traits," *Trends in Genetics*, vol. 18, no. 1, pp. 19–24, 2002.

[14] H. Taniguchi, C. E. Lowe, J. D. Cooper, et al., "Discovery, linkage disequilibrium and association analyses of polymorphisms of the immune complement inhibitor, decay-accelerating factor gene (DAF/CD55) in type 1 diabetes," *BMC Genetics*, vol. 7, article 22, 2006.

[15] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, no. 5281, pp. 1516–1517, 1996.

[16] J. M. Chapman, J. D. Cooper, J. A. Todd, and D. G. Clayton, "Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power," *Human Heredity*, vol. 56, no. 1–3, pp. 18–31, 2003.

[17] A. T. W. Kraakman, R. E. Niks, P. M. M. M. Van den Berg, P. Stam, and F. A. Van Eeuwijk, "Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars," *Genetics*, vol. 168, no. 1, pp. 435–446, 2004.

[18] A. T. W. Kraakman, F. Martínez, B. Mussiraliev, F. A. van Eeuwijk, and R. E. Niks, "Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars," *Molecular Breeding*, vol. 17, no. 1, pp. 41–58, 2006.

[19] M. Hansen, T. Kraft, S. Ganestam, T. Säll, and N.-O. Nilsson, "Linkage disequilibrium mapping of the bolting gene in sea beet using AFLP markers," *Genetical Research*, vol. 77, no. 1, pp. 61–66, 2001.

[20] M. Nordborg, J. O. Borevitz, J. Bergelson, et al., "The extent of linkage disequilibrium in *Arabidopsis thaliana*," *Nature Genetics*, vol. 30, no. 2, pp. 190–193, 2002.

[21] M. Nordborg, T. T. Hu, Y. Ishino, et al., "The pattern of polymorphism in *Arabidopsis thaliana*," *PLoS Biology*, vol. 3, no. 7, p. e196, 2005.

[22] B. Stich, A. E. Melchinger, M. Frisch, H. P. Maurer, M. Heckenberger, and J. C. Reif, "Linkage disequilibrium in European elite maize germplasm investigated with SSRs," *Theoretical and Applied Genetics*, vol. 111, no. 4, pp. 723–730, 2005.

[23] B. Stich, A. E. Melchinger, H.-P. Piepho, et al., "Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations," *Theoretical and Applied Genetics*, vol. 115, no. 4, pp. 529–536, 2007.

[24] G. A. Huttley, M. W. Smith, M. Carrington, and S. J. O'Brien, "A scan for linkage disequilibrium across the human genome," *Genetics*, vol. 152, no. 4, pp. 1711–1722, 1999.

[25] P. K. Gupta, S. Rustgi, and P. L. Kulwal, "Linkage disequilibrium and association studies in higher plants: present status and future prospects," *Plant Molecular Biology*, vol. 57, no. 4, pp. 461–485, 2005.

[26] N. C. Oraguzie, P. L. Wilcox, E. H. A. Rikkerink, and H. N. de Silva, "Linkage disequilibrium," in *Association Mapping in Plants*, N. C. Oraguzie, E. H. A. Rikkerink, S. E. Gardiner, and H. N. de Silva, Eds., pp. 11–39, Springer, New York, NY, USA, 2007.

[27] P. W. Hedrick, "Gametic disequilibrium measures: proceed with caution," *Genetics*, vol. 117, no. 2, pp. 331–341, 1987.

[28] B. Devlin and N. Risch, "A comparison of linkage disequilibrium measures for fine-scale mapping," *Genomics*, vol. 29, no. 2, pp. 311–322, 1995.

[29] L. B. Jorde, "Linkage disequilibrium as a gene-mapping tool," *American Journal of Human Genetics*, vol. 56, no. 1, pp. 11–14, 1995.

[30] J. B. Jorde, "Linkage disequilibrium and the search for complex disease gene," *Genome Research*, vol. 10, no. 10, pp. 1435–1444, 2000.

[31] B. S. Gaut and A. D. Long, "The lowdown on linkage disequilibrium," *Plant Cell*, vol. 15, no. 7, pp. 1502–1506, 2003.

[32] J. M. Abdallah, B. Goffinet, C. Cierco-Ayrolles, and M. Pérez-Enciso, "Linkage disequilibrium fine mapping of quantitative trait loci: a simulation study," *Genetics Selection Evolution*, vol. 35, no. 5, pp. 513–532, 2003.

[33] S. R. Whitt and E. S. Buckler IV, "Using natural allelic diversity to evaluate gene function," in *Plant Functional Genomics: Methods and Protocols*, E. Grotewald, Ed., pp. 123–139, Humana Press, Clifton, NJ, USA, 2003.

[34] B. S. Weir, *Genetic Data Analysis II*, Sinauer Associates, Sunderland, Mass, USA, 1996.

[35] R.-C. Yang, "Zygotic associations and multilocus statistics a nonequilibrium diploid population," *Genetics*, vol. 155, no. 3, pp. 1449–1458, 2000.

[36] R.-C. Yang, "Analysis of mutilocus zygotic associations," *Genetics*, vol. 161, no. 1, pp. 435–445, 2002.

[37] T. Liu, R. J. Todhunter, Q. Lu, et al., "Modeling extent and distribution of zygotic disequilibrium: implications for a multigenerational canine pedigree," *Genetics*, vol. 174, no. 1, pp. 439–453, 2006.

[38] L. Excoffier and M. Slatkin, "Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population," *Molecular Biology and Evolution*, vol. 12, no. 5, pp. 921–927, 1995.

[39] G. R. Abecasis and W. O. C. Cookson, "GOLD—graphical overview of linkage disequilibrium," *Bioinformatics*, vol. 16, no. 2, pp. 182–183, 2000.

[40] Trait Analysis by aSSociation, Evolution and Linkage (TASSEL), http://www.maizegenetics.net/tassel/.

[41] K. Liu and S. V. Muse, "PowerMaker: an integrated analysis environment for genetic maker analysis," *Bioinformatics*, vol. 21, no. 9, pp. 2128–2129, 2005.

[42] K. Zhang, M. Deng, T. Chen, M. S. Waterman, and F. Sun, "A dynamic programming algorithm for haplotype block partitioning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 11, pp. 7335–7339, 2002.

[43] D. L. Remington, J. M. Thornsberry, Y. Matsuoka, et al., "Structure of linkage disequilibrium and phenotypic associations in the maize genome," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 20, pp. 11479–11484, 2001.

[44] M. T. Hamblin, S. E. Mitchell, G. M. White, et al., "Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of Sorghum bicolor," *Genetics*, vol. 167, no. 1, pp. 471–483, 2004.

[45] W. Wang, K. Thornton, A. Berry, and M. Long, "Nucleotide variation along the *Drosophila melanogaster* fourth chromosome," *Science*, vol. 295, no. 5552, pp. 134–137, 2002.

[46] G. B. Cannon, "The effects of natural selection on linkage disequilibrium and relative fitness in experimental populations of *Drosophila melanogaster*," *Genetics*, vol. 48, no. 9, pp. 1201–1216, 1963.

[47] J. K. Pritchard, M. Stephens, N. A. Rosenberg, and P. Donnelly, "Association mapping in structured populations," *American Journal of Human Genetics*, vol. 67, no. 1, pp. 170–181, 2000.

[48] K. L. Mohlke, E. M. Lange, T. T. Valle, et al., "Linkage disequilibrium between microsatellite markers extends beyond 1 cM on chromosome 20 in Finns," *Genome Research*, vol. 11, no. 7, pp. 1221–1226, 2001.

[49] A. F. McRae, J. C. McEwan, K. G. Dodds, T. Wilson, A. M. Crawford, and J. Slate, "Linkage disequilibrium in domestic sheep," *Genetics*, vol. 160, no. 3, pp. 1113–1122, 2002.

[50] M. Maccaferri, M. C. Sanguineti, E. Noli, and R. Tuberosa, "Population structure and long-range linkage disequilibrium in a drum wheat elite collection," *Molecular Breeding*, vol. 15, no. 3, pp. 271–289, 2005.

[51] K. S. Caldwell, J. Russell, P. Langridge, and W. Powell, "Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*," *Genetics*, vol. 172, no. 1, pp. 557–567, 2006.

[52] F. Breseghello and M. E. Sorrells, "Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars," *Genetics*, vol. 172, no. 2, pp. 1165–1177, 2006.

[53] A. Barnaud, T. Lacombe, and A. Doligez, "Linkage disequilibrium in cultivated grapevine, *Vitis vinifera* L," *Theoretical and Applied Genetics*, vol. 112, no. 4, pp. 708–716, 2006.

[54] Y. Li, Y. Li, S. Wu, et al., "Estimation of multilocus linkage disequilibria in diploid populations with dominant markers," *Genetics*, vol. 176, no. 3, pp. 1811–1821, 2007.

[55] P. G. Sand, "A lesson not learned: allele misassignment," *Behavioral and Brain Functions*, vol. 3, article 65, 2007.

[56] L. Tommasini, T. Schnurbusch, D. Fossati, F. Mascher, and B. Keller, "Association mapping of *Stagonospora nodorum* blotch resistance in modern European winter wheat varieties," *Theoretical and Applied Genetics*, vol. 115, no. 5, pp. 697–708, 2007.

[57] Y. Liu, Y. Wang, and H. Huang, "High interpopulation genetic differentiation and unidirectional linear migration patterns in *Myricaria laxiflora* (Tamaricaceae), an endemic riparian plant in the three gorges valley of the Yangtze River," *American Journal of Botany*, vol. 93, no. 2, pp. 206–215, 2006.

[58] H. Iwata, Y. Uga, Y. Yoshioka, K. Ebana, and T. Hayashi, "Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms," *Theoretical and Applied Genetics*, vol. 114, no. 8, pp. 1437–1449, 2007.

[59] M. Malosetti, C. G. van der Linden, B. Vosman, and F. A. van Eeuwijk, "A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato," *Genetics*, vol. 175, no. 2, pp. 879–889, 2007.

[60] C. Gebhardt, A. Ballvora, B. Walkemeier, P. Oberhagemann, and K. Schüler, "Assessing genetic potential in germplasm collections of crop plants by marker-trait association: a case study for potatoes with quantitative variation of resistance to late blight and maturity type," *Molecular Breeding*, vol. 13, no. 1, pp. 93–102, 2004.

[61] P. M. Hollingsworth and R. A. Ennos, "Neighbor joining trees, dominant markers and population genetic structure," *Heredity*, vol. 92, no. 6, pp. 490–498, 2004.

[62] J. K. Pritchard, M. Stephens, and P. Donnelly, "Inference of population structure using multilocus genotype data," *Genetics*, vol. 155, no. 2, pp. 945–959, 2000.

[63] O. J. Hardy, "Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers," *Molecular Ecology*, vol. 12, no. 6, pp. 1577–1588, 2003.

[64] D. Tian, H. Araki, E. Stahl, J. Bergelson, and M. Kreitman, "Signature of balancing selection in *Arabidopsis*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 17, pp. 11525–11530, 2002.

[65] K. A. Shepard and M. D. Purugganan, "Molecular population genetics of the Arabidopsis *CLAVATA2* region: the genomic scale of variation and selection in a selfing species," *Genetics*, vol. 163, no. 3, pp. 1083–1095, 2003.

[66] I. M. Ehrenreich, P. A. Stafford, and M. D. Purugganan, "The genetic architecture of shoot branching in *Arabidopsis thaliana*: a comparative assessment of candidate gene associations *vs.* quantitative trait locus mapping," *Genetics*, vol. 176, no. 2, pp. 1223–1236, 2007.

[67] V. Plagnol, B. Padhukasahasram, J. D. Wall, P. Marjoram, and M. Nordborg, "Relative influences of crossing over and gene conversion on the pattern of linkage disequilibrium in *Arabidopsis thaliana*," *Genetics*, vol. 172, no. 4, pp. 2441–2448, 2006.

[68] M. I. Tenaillon, M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley, and B. S. Gaut, "Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 16, pp. 9161–9166, 2001.

[69] J. M. Thornsberry, M. M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E. S. Buckler, "*Dwarf8* polymorphisms associate with variation in flowering time," *Nature Genetics*, vol. 28, no. 3, pp. 286–289, 2001.

[70] M. Jung, A. Ching, D. Bhattramakki, et al., "Linkage disequilibrium and sequence diversity in a 500-kbp region around the *adh1* locus in elite maize germplasm," *Theoretical and Applied Genetics*, vol. 109, no. 4, pp. 681–689, 2004.

[71] J. R. Andersen, I. Zein, G. Wenzel, et al., "High levels of linkage disequilibrium and associations with forage quality at a *Phenylalanine Ammonia-Lyase* locus in European maize (*Zea mays* L.) inbreds," *Theoretical and Applied Genetics*, vol. 114, no. 2, pp. 307–319, 2007.

[72] S. Chao, W. Zhang, J. Dubcovsky, and M. Sorrells, "Evaluation of genetic diversity and genome-wide linkage disequilibrium among U.S. wheat (*Triticum aestivum* L.) germplasm representing different market classes," *Crop Science*, vol. 47, no. 3, pp. 1018–1030, 2007.

[73] A. J. Garris, S. R. McCouch, and S. Kresovich, "Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.)," *Genetics*, vol. 165, no. 2, pp. 759–769, 2003.

[74] H. A. Agrama and G. C. Eizenga, "Molecular diversity and genome-wide linkage disequilibrium patterns in a worldwide collection of *Oryza sativa* and its wild relatives," *Euphytica*, vol. 160, no. 3, pp. 339–355, 2008.

[75] S. Rakshit, A. Rakshit, H. Matsumura, et al., "Large-scale DNA polymorphism study of *Oryza sativa* and *O. rufipogon* reveals the origin and divergence of Asian rice," *Theoretical and Applied Genetics*, vol. 114, no. 4, pp. 731–743, 2007.

[76] K. A. Mather, A. L. Caicedo, N. R. Polato, K. M. Olsen, S. McCouch, and M. D. Purugganan, "The extent of linkage disequilibrium in rice (*Oryza sativa* L.)," *Genetics*, vol. 177, no. 4, pp. 2223–2232, 2007.

[77] L. V. Malysheva-Otto, M. W. Ganal, and M. S. Röder, "Analysis of molecular diversity, population structure and linkage disequilibrium in a worldwide survey of cultivated barley germplasm (*Hordeum vulgare* L.)," *BMC Genetics*, vol. 7, article 6, 2006.

[78] P. L. Morrell, D. M. Toleno, K. E. Lundy, and M. T. Clegg, "Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 7, pp. 2442–2447, 2005.

[79] Y. L. Zhu, Q. J. Song, D. L. Hyten, et al., "Single-nucleotide polymorphisms in soybean," *Genetics*, vol. 163, no. 3, pp. 1123–1134, 2003.

[80] T.-H. Jun, K. Van, M. Y. Kim, H. S. Lee, and D. R. Walker, "Association analysis using SSR markers to find QTL for seed protein content in soybean," *Euphytica*.

[81] T. Kraft, M. Hansen, and N.-O. Nilsson, "Linkage disequilibrium and fingerprinting in sugar beet," *Theoretical and Applied Genetics*, vol. 101, no. 3, pp. 323–326, 2000.

[82] T.-M. Yin, S. P. DiFazio, L. E. Gunter, S. S. Jawdy, W. Boerjan, and G. A. Tuskan, "Genetic and physical mapping of *Melampsora* rust resistance genes in *Populus* and characterization of linkage disequilibrium and flanking genomic sequence," *New Phytologist*, vol. 164, no. 1, pp. 95–105, 2004.

[83] P. K. Ingvarsson, "Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., salicaceae)," *Genetics*, vol. 169, no. 2, pp. 945–953, 2005.

[84] G. R. Brown, G. P. Gill, R. J. Kuntz, C. H. Langley, and D. B. Neale, "Nucleotide diversity and linkage disequilibrium in loblolly pine," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 42, pp. 15255–15260, 2004.

[85] K. V. Krutovsky and D. B. Neale, "Nucleotide diversity and linkage disequilibrium in cold-hardiness- and wood quality-related candidate genes in douglas fir," *Genetics*, vol. 171, no. 4, pp. 2029–2041, 2005.

[86] A. Rafalski and M. Morgante, "Corn and humans: recombination and linkage disequilibrium in two genomes of similar size," *Trends in Genetics*, vol. 20, no. 2, pp. 103–111, 2004.

[87] C. Guillet-Claude, C. Birolleau-Touchard, D. Manicacci, et al., "Genetic diversity associated with variation in silage corn digestibility for three *O*-methyltransferase genes involved in lignin biosynthesis," *Theoretical and Applied Genetics*, vol. 110, no. 1, pp. 126–135, 2004.

[88] C. Guillet-Claude, C. Birolleau-Touchard, D. Manicacci, et al., "Nucleotide diversity of the *ZmPox3* maize peroxidase gene: relationship between a MITE insertion in exon 2 and variation in forage maize digestibility," *BMC Genetics*, vol. 5, article 19, pp. 1–11, 2004.

[89] Y. Xing, U. Frei, B. Schejbel, T. Asp, and T. Lübberstedt, "Nucleotide diversity and linkage disequilibrium in 11 expressed resistance candidate genes in *Lolium perenne*," *BMC Plant Biology*, vol. 7, article 43, pp. 1–12, 2007.

[90] R. C. Ponting, M. C. Drayton, N. O. I. Cogan, et al., "SNP discovery, validation, haplotype structure and linkage disequilibrium in full-length herbage nutritive quality genes of perennial ryegrass (*Lolium perenne* L.)," *Molecular Genetics and Genomics*, vol. 278, no. 5, pp. 585–597, 2007.

[91] L. Skøt, M. O. Humphreys, I. Armstead, et al., "An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.)," *Molecular Breeding*, vol. 15, no. 3, pp. 233–245, 2005.

[92] I. Y. Abdurakhomonov, R. J. Kohel, S. Saha, et al., "Genome-wide linkage disequilibrium revealed by microsatellite markers and association study of fiber quality traits in cotton," in *Proceedings of the 15th Plant and Animal Genome Conference*, San Diego, Calif, USA, January 2007, W199.

[93] T. G. Schulze and F. J. McMahon, "Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines," *American Journal of Medical Genetics B*, vol. 114, no. 1, pp. 1–11, 2002.

[94] J. Ohashi, S. Yamamoto, N. Tsuchiya, et al., "Comparison of statistical power between 2×2 allele frequency and allele positivity tables in case-control studies of complex disease genes," *Annals of Human Genetics*, vol. 65, no. 2, pp. 197–206, 2001.

[95] C. T. Falk and P. Rubinstein, "Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations," *Annals of Human Genetics*, vol. 51, no. 3, pp. 227–233, 1987.

[96] R. S. Spielman and W. J. Ewens, "The TDT and other family-based tests for linkage disequilibrium and association," *American Journal of Human Genetics*, vol. 59, no. 5, pp. 983–989, 1996.

[97] R. S. Spielman, R. E. McGinnis, and W. J. Ewens, "Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM)," *American Journal of Human Genetics*, vol. 52, no. 3, pp. 506–516, 1993.

[98] H. Bickeböller and F. Clerget-Darpoux, "Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers," *Genetic Epidemiology*, vol. 12, no. 6, pp. 865–870, 1995.

[99] J. P. Rice, R. J. Neuman, S. L. Hoshaw, E. W. Daw, and C. Gu, "TDT with covariates and genomic screens with mod scores: their behavior on simulated data," *Genetic Epidemiology*, vol. 12, no. 6, pp. 659–664, 1995.

[100] P. C. Sham and D. Curtis, "An extended transmission/disequilibrium test (TDT) for multi-allele marker loci," *Annals of Human Genetics*, vol. 59, no. 3, pp. 323–336, 1995.

[101] N. L. Kaplan, E. R. Martin, and B. S. Weir, "Power studies for the transmission/disequilibrium tests with multiple alleles," *American Journal of Human Genetics*, vol. 60, no. 3, pp. 691–702, 1997.

[102] M. A. Cleves, J. M. Olson, and K. B. Jacobs, "Exact transmission-disequilibrium tests with multiallelic markers," *Genetic Epidemiology*, vol. 14, no. 4, pp. 337–347, 1997.

[103] L. C. Lazzeroni and K. Lange, "A conditional inference framework for extending the transmission/disequilibrium test," *Human Heredity*, vol. 48, no. 2, pp. 67–81, 1998.

[104] S. R. Wilson, "On extending the transmission/disequilibrium test (TDT)," *Annals of Human Genetics*, vol. 61, no. 2, pp. 151–161, 1997.

[105] H. Zhao, S. Zhang, K. R. Merikangas, et al., "Transmission/disequilibrium tests using multiple tightly linked markers," *American Journal of Human Genetics*, vol. 67, no. 4, pp. 936–946, 2000.

[106] D. Curtis, "Use of siblings as controls in case-control association studies," *Annals of Human Genetics*, vol. 61, no. 4, pp. 319–333, 1997.

[107] S. Horvath and N. M. Laird, "A discordant-sibship test for disequilibrium and linkage: no need for parental data," *American Journal of Human Genetics*, vol. 63, no. 6, pp. 1886–1897, 1998.

[108] F. Sun, W. D. Flanders, Q. Yang, and M. J. Khoury, "Transmission disequilibrium test (TDT) when only one parent is available: the 1-TDT," *American Journal of Epidemiology*, vol. 150, no. 1, pp. 97–104, 1999.

[109] M. Knapp, "A note on power approximations for the transmission/disequilibrium test," *American Journal of Human Genetics*, vol. 64, no. 4, pp. 1177–1185, 1999.

[110] M. Knapp, "Reconstructing parental genotypes when testing for linkage in the presence of association," *Theoretical Population Biology*, vol. 60, no. 3, pp. 141–148, 2001.

[111] S. Horvath, N. M. Laird, and M. Knapp, "The transmission/disequilibrium test and parental-genotype reconstruction for X-chromosomal markers," *American Journal of Human Genetics*, vol. 66, no. 3, pp. 1161–1167, 2000.

[112] E. R. Martin, S. A. Monks, L. L. Warren, and N. L. Kaplan, "A test for linkage and association in general pedigrees: the pedigree disequilibrium test," *American Journal of Human Genetics*, vol. 67, no. 1, pp. 146–154, 2000.

[113] E. R. Martin, M. P. Bass, and N. L. Kaplan, "Correcting for a potential bias in the pedigree disequilibrium test," *American Journal of Human Genetics*, vol. 68, no. 4, pp. 1065–1067, 2001.

[114] J.-P. Hugot, M. Chamaillard, H. Zouali, et al., "Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease," *Nature*, vol. 411, no. 6837, pp. 599–603, 2001.

[115] G. R. Abecasis, L. R. Cardon, and W. O. C. Cookson, "A general test of association for quantitative traits in nuclear families," *American Journal of Human Genetics*, vol. 66, no. 1, pp. 279–292, 2000.

[116] S. A. Monks and N. L. Kaplan, "Removing the sampling restrictions from family-based tests of association for a quantitative-trait locus," *American Journal of Human Genetics*, vol. 66, no. 2, pp. 576–592, 2000.

[117] J. D. Terwilliger, "A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci," *American Journal of Human Genetics*, vol. 56, no. 3, pp. 777–787, 1995.

[118] C. R. Weinberg, A. J. Wilcox, and R. T. Lie, "A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting," *American Journal of Human Genetics*, vol. 62, no. 4, pp. 969–978, 1998.

[119] C. R. Weinberg, "Allowing for missing parents in genetic studies of case-parent triads," *American Journal of Human Genetics*, vol. 64, no. 4, pp. 1186–1193, 1999.

[120] N. M. Laird, S. Horvath, and X. Xu, "Implementing a unified approach to family-based tests of association," *Genetic Epidemiology*, vol. 19, supplement 1, pp. S36–S42, 2000.

[121] D. Rabinowitz and N. M. Laird, "A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information," *Human Heredity*, vol. 50, no. 4, pp. 211–223, 2000.

[122] S. L. Lake, D. Blacker, and N. M. Laird, "Family-based tests of association in the presence of linkage," *American Journal of Human Genetics*, vol. 67, no. 6, pp. 1515–1525, 2000.

[123] G. J. te Meerman, M. A. Van der Meulen, and L. A. Sandkuijl, "Perspectives of identity by descent (IBD) mapping in founder populations," *Clinical & Experimental Allergy*, vol. 25, supplement 2, pp. 97–102, 1995.

[124] D. F. Levinson, A. Kirby, S. Slepner, I. Nolte, G. T. Spijker, and G. te Meerman, "Simulation studies of detection of a complex disease in a partially isolated population," *American Journal of Medical Genetics B*, vol. 105, no. 1, pp. 65–70, 2001.

[125] M. S. McPeek and A. Strahs, "Assessment of linkage disequilibrium by the decay of haplotype sharing, with application to fine-scale genetic mapping," *American Journal of Human Genetics*, vol. 65, no. 3, pp. 858–875, 1999.

[126] N. Risch and J. Teng, "The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling," *Genome Research*, vol. 8, no. 12, pp. 1273–1288, 1998.

[127] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997–1004, 1999.

[128] S.-A. Bacanu, B. Devlin, and K. Roeder, "The power of genomic control," *American Journal of Human Genetics*, vol. 66, no. 6, pp. 1933–1944, 2000.

[129] K. Zhao, M. J. Aranzana, S. Kim, et al., "An *Arabidopsis* example of association mapping in structured samples," *PLoS Genetics*, vol. 3, no. 1, p. e4, 2007.

[130] K. J. Pritchard and W. Wen, *Documentation for Structure Software*, The University of Chicago Press, Chicago, Ill, USA, 2004.

[131] G. A. Satten, W. D. Flanders, and Q. Yang, "Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model," *American Journal of Human Genetics*, vol. 68, no. 2, pp. 466–477, 2001.

[132] X. Zhu, S. Zhang, H. Zhao, and R. S. Cooper, "Association mapping, using a mixture model for complex traits," *Genetic Epidemiology*, vol. 23, no. 2, pp. 181–196, 2002.

[133] J. Yu, G. Pressoir, W. H. Briggs, et al., "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness," *Nature Genetics*, vol. 38, no. 2, pp. 203–208, 2006.

[134] K. Ritland, "Estimators for pairwise relatedness and inbreeding coefficients," *Genetical Research*, vol. 67, no. 2, pp. 175–186, 1996.

[135] O. J. Hardy and X. Vekemans, "SPAGeDi: a versatile computer program to analyze spatial genetic structure at the individual or population levels," *Molecular Ecology Notes*, vol. 2, no. 4, pp. 618–620, 2002.

[136] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[137] B. Parisseaux and R. Bernardo, "In silico mapping of quantitative trait loci in maize," *Theoretical and Applied Genetics*, vol. 109, no. 3, pp. 508–514, 2004.

[138] I. Simko, S. Costanzo, K. G. Haynes, B. J. Christ, and R. W. Jones, "Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach," *Theoretical and Applied Genetics*, vol. 108, no. 2, pp. 217–224, 2004.

[139] M. J. Sillanpää and M. Bhattacharjee, "Bayesian association-based fine mapping in small chromosomal segments," *Genetics*, vol. 169, no. 1, pp. 427–439, 2005.

[140] R. D. Ball, "Statistical analysis and experimental design," in *Association Mapping in Plants*, N. C. Oraguzie, E. H. A. Rikkerink, S. E. Gardiner, and H. N. de Silva, Eds., pp. 133–196, Springer, New York, NY, USA, 2007.

[141] B. R. Thumma, M. F. Nolan, R. Evans, and G. F. Moran, "Polymorphisms in *cinnamoyl CoA reductase (CCR)* are associated with variation in microfibril angle in *Eucalyptus* spp.," *Genetics*, vol. 171, no. 3, pp. 1257–1265, 2005.

[142] R. D. Ball, "Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies," *Genetics*, vol. 170, no. 2, pp. 859–873, 2005.

[143] G. A. Churchill and R. W. Doerge, "Empirical threshold values for quantitative trait mapping," *Genetics*, vol. 138, no. 3, pp. 963–971, 1994.

[144] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society Series B*, vol. 57, no. 1, pp. 289–300, 1995.

[145] K. M. Olsen, S. S. Halldorsdottir, J. R. Stinchcombe, C. Weinig, J. Schmittt, and M. D. Purugganan, "Linkage disequilibrium mapping of Arabidopsis *CRY2* flowering time alleles," *Genetics*, vol. 167, no. 3, pp. 1361–1369, 2004.

[146] A. L. Caicedo, J. R. Stinchcombe, K. M. Olsen, J. Schmitt, and M. D. Purugganan, "Epistatic interaction between *Arabidopsis FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 44, pp. 15670–15675, 2004.

[147] S. Balasubramanian, S. Sureshkumar, M. Agrawal, et al., "The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*," *Nature Genetics*, vol. 38, no. 6, pp. 711–715, 2006.

[148] M. J. Aranzana, S. Kim, K. Zhao, et al., "Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes," *PLoS Genetics*, vol. 1, no. 5, p. e60, 2005.

[149] K. A. Palaisa, M. Morgante, M. Williams, and A. Rafalski, "Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci," *The Plant Cell*, vol. 15, no. 8, pp. 1795–1806, 2003.

[150] L. M. Wilson, S. R. Whitt, A. M. Ibáñez, T. R. Rocheford, M. M. Goodman, and E. S. Buckler IV, "Dissection of maize kernel composition and starch production by candidate gene association," *The Plant Cell*, vol. 16, no. 10, pp. 2719–2733, 2004.

[151] J. R. Andersen, T. Schrag, A. E. Melchinger, I. Zein, and T. Lübberstedt, "Validation of *Dwarf8* polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.)," *Theoretical and Applied Genetics*, vol. 111, no. 2, pp. 206–217, 2005.

[152] S. J. Szalma, E. S. Buckler IV, M. E. Snook, and M. D. McMullen, "Association analysis of candidate genes for maysin and chlorogenic acid accumulation in maize silks," *Theoretical and Applied Genetics*, vol. 110, no. 7, pp. 1324–1333, 2005.

[153] T. Lübberstedt, I. Zein, J. R. Andersen, et al., "Development and application of functional markers in maize," *Euphytica*, vol. 146, no. 1-2, pp. 101–108, 2005.

[154] A. Beló, P. Zheng, S. Luck, et al., "Whole genome scan detects an allelic variant of *fad2* associated with increased oleic acid levels in maize," *Molecular Genetics and Genomics*, vol. 279, no. 1, pp. 1–10, 2008.

[155] H. A. Agrama, G. C. Eizenga, and W. Yan, "Association mapping of yield and its components in rice cultivars," *Molecular Breeding*, vol. 19, no. 4, pp. 341–356, 2007.

[156] N. Zhang, Y. Xu, M. Akash, S. McCouch, and J. H. Oard, "Identification of candidate markers associated with agronomic traits in rice using discriminant analysis," *Theoretical and Applied Genetics*, vol. 110, no. 4, pp. 721–729, 2005.

[157] E. Igartua, A. M. Casas, F. Ciudad, J. L. Montoya, and I. Romagosa, "RFLP markers associated with major genes controlling heading date evaluated in a barley germ plasm pool," *Heredity*, vol. 83, no. 5, pp. 551–559, 1999.

[158] V. Ivandic, W. T. B. Thomas, E. Nevo, Z. Zhang, and B. P. Forster, "Associations of simple sequence repeats with quantitative trait variation including biotic and abiotic stress tolerance in *Hordeum spontaneum*," *Plant Breeding*, vol. 122, no. 4, pp. 300–304, 2003.

[159] C. Ravel, S. Praud, A. Murigneux, et al., "Identification of *Glu-B1-1* as a candidate gene for the quantity of high-molecular-weight glutenin in bread wheat (*Triticum aestivum* L.) by means of an association study," *Theoretical and Applied Genetics*, vol. 112, no. 4, pp. 738–743, 2006.

[160] B. B. D'hoop, M. J. Paulo, R. A. Mank, H. J. van Eck, and F. A. van Eeuwijk, "Association mapping of quality traits in potato (*Solanum tuberosum* L.)," *Euphytica*, vol. 161, no. 1-2, pp. 47–60, 2008.

[161] M. P. Dobrowolski and J. W. Forster, "Linkage disequilibrium-based association mapping in forage species," in *Association Mapping in Plants*, N. C. Oraguzie, E. H. A. Rikkerink, S. E. Gardiner, and H. N. de Silva, Eds., pp. 197–210, Springer, New York, NY, USA, 2007.

[162] L. Skøt, J. Humphreys, M. O. Humphreys, et al., "Association of candidate genes with flowering time and water-soluble carbohydrate content in *Lolium perenne* (L.)," *Genetics*, vol. 177, no. 1, pp. 535–547, 2007.

[163] P. L. Wilcox, E. C. Echt, and R. D. Burdon, "Gene-assisted selection: applications of association genetics for forest tree breeding," in *Association Mapping in Plants*, N. C. Oraguzie, E. H. A. Rikkerink, S. E. Gardiner, and H. N. de Silva, Eds., pp. 211–247, Springer, New York, NY, USA, 2007.

[164] The Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common

diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.

[165] X. Estivill and L. Armengol, "Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies," *PLoS Genetics*, vol. 3, no. 10, pp. 1787–1799, 2007.

[166] S. C. Beer, W. Siripoonwiwat, L. S. O'donoughue, E. Souza, D. Matthews, and M. E. Sorrels, "Associations between molecular markers and quantitative traits in an oat germplasm pool: can we infer linkages?" *Journal of Agricultural Genomics*, vol. 3, paper 197, 1997.

[167] P. S. Virk, B. V. Ford-Lloyd, M. T. Jackson, H. S. Pooni, T. P. Clemeno, and H. J. Newbury, "Predicting quantitative variation within rice germplasm using molecular markers," *Heredity*, vol. 76, no. 3, pp. 296–304, 1996.

[168] S. Zhang, K. Zhang, J. Li, F. Sun, and H. Zhao, "Test of association for quantitative traits in general pedigrees: the quantitative pedigree disequilibrium test," *Genetic Epidemiology*, vol. 21, supplement 1, pp. S370–S375, 2001.

[169] F. Breseghello and M. E. Sorrells, "Association analysis as a strategy for improvement of quantitative traits in plants," *Crop Science*, vol. 46, no. 3, pp. 1323–1330, 2006.

Hindawi Publishing Corporation
International Journal of Plant Genomics
Volume 2008, Article ID 683509, 16 pages
doi:10.1155/2008/683509

Review Article

Phylogenetic Analyses: A Toolbox Expanding towards Bayesian Methods

Stéphane Aris-Brosou[1,2] and Xuhua Xia[1]

[1] *Department of Biology, Centre for Advanced Research in Environmental Genomics, University of Ottawa, Ontario, Canada K1N 6N5*
[2] *Department of Mathematics and Statistics, University of Ottawa, Ontario, Canada K1N 6N5*

Correspondence should be addressed to Stéphane Aris-Brosou, sarisbro@uottawa.ca

Received 30 November 2007; Accepted 12 February 2008

Recommended by Chunguang Du

The reconstruction of phylogenies is becoming an increasingly simple activity. This is mainly due to two reasons: the democratization of computing power and the increased availability of sophisticated yet user-friendly software. This review describes some of the latest additions to the phylogenetic toolbox, along with some of their theoretical and practical limitations. It is shown that Bayesian methods are under heavy development, as they offer the possibility to solve a number of long-standing issues and to integrate several steps of the phylogenetic analyses into a single framework. Specific topics include not only phylogenetic reconstruction, but also the comparison of phylogenies, the detection of adaptive evolution, and the estimation of divergence times between species.

1. INTRODUCTION

Human cultures have always been fascinated by their origins as a means to define their position in the world, and to justify their hegemony over the rest of the living world. However, scientific (testable) predictions about our origins had to wait for Darwin [1] and his intellectual descendents first to classify [2] and then to reconstruct the natural history of replicating entities, and hereby to kick-start the field of phylogenetics [3, 4]. Rooted in the comparison of morphological characters, phylogenies have for the past four decades focused on the relationships between molecular sequences (e.g., [4]), potentially helped by incorporating morphological information [5], in order to infer ancestor-to-descendent relationships between sequences, populations, or species.

Today, molecular phylogenies are routinely used to infer gene or genome duplication events [6], recombination [7], horizontal gene transfer [8], variation of selective pressures and adaptive evolution [9], divergence times between species [10], the origin of genetic code [11], elucidate the origin of epidemics [12], and host-parasite cospeciation events [13, 14]. As complementary tools for taxonomy (DNA barcoding: [15]), they have also contributed to the formulation of strategies in conservation biology [16]. In addition to untangling the ancestral relationships relating a group of taxa or of a set of molecular sequences, phylogenies have also been used for some time outside of the realm of biological sciences as for instance in linguistics [17, 18] or in forensics [19, 20].

Most of these applications are beyond the scope of plant genomics, but they all suggest that sophisticated phylogenetic methods are required to address most of today's biological questions. While parsimony-based methods are both intuitive and extremely informative, for instance to disentangle genome rearrangements [21], they also have their limitations due to their minimizing the amount of change [22]. These limitations become particularly apparent when analyzing distantly related taxa. A means to overcome, at least partly, some of these difficulties is to adopt a model-based approach, be in a maximum likelihood or in a Bayesian framework. These two frameworks are extremely similar in that they both rely on probabilistic models. Bayesian approaches offer a variety of benefits when compared to traditional maximum likelihood, such as computing speed (although this is not necessarily true, especially under complex models), sophistication of the model, and an appropriate treatment of uncertainty, in particular the one about nuisance parameters.

As a result, Bayesian approaches often make it possible to address more sophisticated biological questions [23], which usually comes at the expense of longer computing times and higher memory requirements than when using simpler models.

Because it is not possible or even appropriate to discuss all the latest developments in a given field of study, this review will focus on a very limited number of key phylogenetic topics. Of notable exceptions, recent developments in phylogenetic hidden Markov models [24] or applications that map ancestral states on phylogenies [25] are not treated. We focus instead on the very first steps involved in *most* phylogenetic analysis, ranging from reconstructing a tree to estimating selective pressures or species divergence times. For each of these steps, some of the most recent theoretical developments are discussed, and pointers to relevant software are provided.

2. RECONSTRUCTING PHYLOGENETIC TREES

2.1. Sequence alignment

The first step in reconstructing a phylogenetic tree from molecular data is to obtain a multiple sequence alignment (MSA) where sequence data are arranged in a matrix that specifies which residues are homologous [26]. A large number of methods and programs exist [27] and most have been evaluated against alignment databases [28], so that it is possible to provide some general guidelines.

The easiest sequences to align are probably those of protein-coding genes: proteins diverge more slowly than DNA sequences and, as a result, proteins are easier to align. The rule-of-thumb is therefore first to translate DNA to amino acid sequences, then perform the alignment at the protein level, before back-translating to the DNA alignment in a final step. This procedure avoids inserting gaps in the final DNA alignment that are not multiple of three and that would disrupt the reading frame. Translation to amino acid sequences can be done directly when downloading sequences, for instance from the National Center for Biotechnology Information (NCBI: www.ncbi.nlm.nih.gov). A number of programs also allow users to perform this translation locally on their computers from an appropriate translation table (e.g., DAMBE [29], MEGA [30, 31]; see Table 1). The second step is to perform the alignment at the protein level. Again, a number of programs exist, but ProbCons [32] appears to be the most accurate *single* method [33]. An alternative for using one single alignment method is to use consensus or meta-methods, that is, to combine several methods [27]. Meta-methods such as M-Coffee can return better MSAs almost twice as often as ProbCons [34]. Finally, when the alignment is obtained at the protein level, back-translation to the DNA sequences can be performed either by using a program such as DAMBE, CodonAlign [35], or by using a dedicated server such as protal2dna (http://bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html) or Pal2Nal (coot.embl.de/pal2nal).

The alignment of rRNA genes with the constraint of secondary structure has now been frequently used in practical research in molecular evolution and phylogenetics [56–60]. The procedure is first to obtain reliable secondary structure and then use the secondary structure to guide the sequence alignment. This has not been automated so far, although both Clustal [61, 62] and DAMBE have some functions to alleviate the difficulties.

What to do with other noncoding genes is still an open question, especially when it comes to aligning a large number (>100) of long (>20,000 residues) and divergent sequences (<25% identity). Some authors have attempted to provide rough guidelines to choose the most accurate program depending on these parameters [28]. However, accuracy figures are typically estimated over a large number of test alignments and may not reflect the accuracy that is expected for any particular alignment [28]. More crucially, most of the alignment programs were developed and benchmarked on protein data, so that their accuracy is generally unknown for noncoding sequences [28]. A very general recommendation is then to use different methods [63] and meta-methods. Those parts of the alignment that are similar across the different methods are probably reliable. The parts that differ extensively are often simply eliminated from the alignment when no external information can be used to decide which positions are homologous. Poorly aligned regions can cause serious problems as, for instance, when analyzing rRNA sequences in which conserved domain and variable domains have different nucleotide frequencies [60]. A simple test of the reliability of an alignment consists in reversing the orientation of the original sequences, and performing the alignment again; because of the symmetry of the problem, reliable MSAs are expected to be identical whichever direction is used to align the sequences [64]. These authors further show that reliability of MSAs decreases with sequence divergence, and that the chance of reconstructing different phylogenies increases with sequence divergence. More sophisticated methods also permit the direct measure of the accuracy of an alignments or the estimation of a distance between two alignments [65]. Applications of Bayesian inference strictly to pairwise [66] and multiple [67, 68] sequence alignment are still in their infancy.

Whichever method is used to obtain an MSA, a final visual inspection is required, and manual editing is often needed. To this end, a number of editors can be used such as JalView [69].

Because an MSA represents a hypothesis about sitewise homology at all the positions, obtaining an accurate MSA presents some circularity; an accurate MSA often necessitates an accurate guide tree, which in turn demands an accurate alignment. The early realization of this "chicken-egg" conundrum led to the idea that both the MSA and the phylogeny should be estimated simultaneously [70]. Although this initial algorithm was parsimony-based, it was already too complex to analyze more than a half-dozen sequences of 100 sites or more. Subsequent parsimony-based algorithms allowed the analysis of larger data sets [71] but still showed some limitations when sequence divergence increases. More recently, a Bayesian procedure was described and implemented in a program, BAli-Phy, where uncertainties with respect to the alignment, the tree, and the parameters of the substitution

Table 1: *List of programs cited in this review*. GUI: graphic user interface; ML: maximum likelihood; PL: penalized likelihood.

Name	Method	Platform	GUI	Inference	Reference
BAMBE	Bayes	DOS, MacOS, Unix	No	Tree	[36]
BayesPhylogenies	Bayes	DOS, MacOS, Unix	No	Tree	[37]
BAli-Phy	Bayes	DOS, MacOS, Unix	No	Simultaneous alignment and tree	[38]
BEAST	Bayes	Windows, MacOS, Unix	Yes	Tree, times	[39]
CONSEL	ML	DOS, MacOS, Unix	No	Tree comparison	[40]
DAMBE	Distances, parsimony, ML	Windows	Yes	Tree	[29]
GARLI	ML (Genetic Algorithm)	Windows, MacOS, Unix	Yes	Tree	[41]
HyPhy	ML	Windows, MacOS, Unix	Yes	Tree, selection, recombination, tree comparison,	[42]
MEGA	Distances, parsimony	Windows	Yes	Tree, times	[30, 31]
MrBayes	Bayes	DOS, MacOS, Unix	No	Tree, selection	[43, 44]
Multidivtime	Bayes	DOS, MacOS, Unix	No	Times	[45–47]
OmegaMap	Bayes	DOS, MacOS, Unix	No	Simultaneous selection and recombination	[48]
PAML	ML	DOS, MacOS, Unix	No	Tree, tree comparison, times, selection	[49, 50]
PAUP$*$	Distances, parsimony, ML	DOS, MacOS, Unix	No	Tree	[51]
PhyloBayes	Bayes	DOS, MacOS, Unix	No	Tree, tree comparison	[52]
PHYML	ML	DOS, MacOS, Unix	No	Tree	[53]
RAxML	ML	DOS, MacOS, Unix	No	Tree	[54]
r8s	PL	DOS, MacOS, Unix	No	Times	[55]

model are all taken into account [38] (see also [72]). Uncertain alignments are a potential problem in large-scale genomic studies [73] or in whole-genome alignments [74]. In these contexts, disregarding alignment uncertainty can lead to systematic biases when estimating gene trees or inferring adaptive evolution [73, 74]. However, these complex Bayesian models [38, 72, 73] still require some nonnegligible computing time and resource, and to date, their performance in terms of accuracy is still unclear.

2.2. Selection of the substitution model

Once a reliable MSA is obtained, the next step in comparing molecular sequences is to choose a metric to quantify divergence. The most intuitive measure of divergence is simply to count the proportion of differences between two aligned sequences (e.g., [75]). This simple measure is known as the p distance. However, because the size of the state space is finite (four letters for DNA, 20 for amino acids, and 61 for sense codons), multiple changes at a position in the alignment will not be observable, and the p distance will underestimate evolutionary distances even for moderately divergent sequences. This phenomenon is generally referred to as saturation. Corrections were devised early to help compensate for saturation. Some of the most famous named nucleotide substitution models are the Jukes-Cantor model or JC [76], the Kimura two-parameter model or K80 [77], the Hasegawa-Kishino-Yano model or HKY85 [78], the Tamura-Nei model or TN93 [79], and the general time-reversible model or GTR [80] (also called REV). Because substitution rates vary along sequences, two components can be added to these substitution models: a "+I" component that models invariable sites [78] and a "+Γ" component that models among-site rate variation either as a continuous [81] or as a discrete [82] mean-one Γdistribution, the latter being more computationally efficient. Amino acid models can also incorporate a "+F" component so that replacement rates are proportional to the frequencies of both the replaced and resulting residues [83].

Given the variety of substitution models, the first step of any model-based phylogenetic analysis is to select the most appropriate model [84, 85]. The rational for doing so is to balance bias and variance: a highly-parameterized model will describe or fit the data much better than a model that contains a smaller number of parameters; in turn however, each parameter of the highly-parameterized model will be estimated with lower accuracy for a given amount of data (e.g., [86]). Besides, both empirical and simulation studies show that the choice of a wrong substitution model can lead not only to less accurate phylogenetic estimation, but also to inconsistent results [87]. The objective of model selection is therefore not to select the "best-fitting" model, as this one will always be the model with the largest number of parameters, but rather to select the most appropriate model that will achieve the optimal tradeoff between

bias and variance. The approach followed by all model selection procedures is therefore to penalize the likelihood of the parameter-rich model for the additional parameters. Because most of the nucleotide substitution models are nested (all can be seen as a special case of GTR +Γ+I), the standard approach to model selection is to perform hierarchical likelihood ratio tests or hLRTs [88]. Note that in all rigor, likelihood ratio tests can also be performed on nonnested models; however, the asymptotic distribution of the test statistic (twice the difference in log-likelihoods) under the null hypothesis (the two models perform equally well) is complicated [89] and quite often impractical. When models are nested, the asymptotic distribution of the test statistic under the null hypothesis is simply a χ^2 distribution whose degree of freedom is the number of additional parameters entering the more complex model (see [90] or [91] for applicability conditions). With the hLRT, then all models are compared in a pairwise manner, by traversing a choice-tree of possible nested models. A number of popular programs allow users to compare pairs of models manually (e.g., PAUP [51], PAML [49, 50]). Readily written scripts that select the most appropriate model among a list of named models also exist, such as ModelTest [92] (which requires PAUP), the *R* package APE [93], or DAMBE. Free web servers are also available; they are either directly based on ModelTest [94] or implement similar ideas (e.g., FindModel, available at hcv.lanl.gov/content/hcv-db/findmodel/findmodel.html). A similar implementation, ProtTest, exists for protein data [95].

However, performing systematic hLRTs is not the optimal strategy for model selection in phylogenetics [96]. This is because the model that is finally selected can depend on the order in which the pairwise comparisons are performed [97]. The Akaike information criterion (AIC) or its variant developed in the context of regression and time-series analysis in small data sets (AIC_c, [98]) is commonly used in phylogenetics (e.g., [96]). One advantage of AIC is that it allows nonnested models to be compared, and it is easily implemented. However, in large data sets, both the hLRT and the AIC tend to favor parameter-rich models [99]. A slightly different approach was proposed to overcome this selection bias, the Bayesian information criterion (BIC: [99]), which penalizes more strongly parameter-rich models. All these model selection approaches (AIC, AIC_c, and BIC) are available in ModelTest and ProtTest. Other procedures exist such as the Decision-Theoretic or DT approach [100]. Although AIC, BIC, and DT are generally based on sound principles, they can in practice select different substitution models [101]. The reason for doing so is not entirely clear, but it is likely due to the data having low-information content. One prediction is that, when these model selection procedures end up with different conclusions, all the selected models will return phylogenies that are not significantly different. It is also possible that applying these different criteria outside of the theoretical context in which they were developed might lead to unexpected behaviors [102]. For instance, $AIC_c$was derived under Gaussian assumptions for linear fixed-effect models [98], and other bias correction terms exist under different assumptions [86].

All the above test procedures compare ratios of likelihood values penalized for an increase in the dimension of one of the models, without directly accounting for uncertainty in the estimates of model parameters. This may be problematic, in particular for small data sets. The Bayesian approach to model selection, called the Bayes factor, directly incorporates this uncertainty. It is also more intuitive as it directly assesses if the data are more probable under a given model than under a different one (e.g., [103]). An extension of this approach makes it possible to select the model not only among the set of named models (JC to GTR) but among all 203 nucleotide substitution models that are possible [104]. An alternative use or interpretation of this approach is to integrate directly over the uncertainty about the substitution model, so that the estimated phylogeny fully accounts for several kinds of uncertainty: about the substitution models, and the parameters entering each of these models. MrBayes (version 3.1.2) [43] implements this feature for amino acid models.

There is an element of circularity in model selection, just as in sequence alignment. In theory, when the hLRT is used for model selection, the topology used for all the computations should be that of the maximum likelihood tree. In practice, model selection is based on an initial topology obtained by a fast algorithm such as neighbor-joining [105, 106] (default setting in ModelTest) or by Weighbor [107] (default setting in FindModel) on JC distances without any correction for among-site rate variation. As mentioned above, it is known that the choice of a wrong model can affect the tree that is estimated, but it is not always clear how the choice of a nonoptimal topology to select the substitution model affects the tree that is finally estimated. Again, this issue with model choice disappears with Bayesian approaches that integrate over all possible time-reversible models as in [104].

2.3. *Finding the "best" tree and assessing its support*

Once the substitution model is selected, the classical approach proceeds to reconstruct the phylogeny [108]. This is probably one area where phylogenetics has seen mixed progress over the last five years, due to both the combinatorial and the computational complexities of phylogenetic reconstruction.

The combinatorial complexity relates to the extremely large number of tree topologies that are possible with a large number of sequences [109]. For instance, with five sequences, there are 105 rooted topologies, but with ten sequences, this number soars to over 34 million. An exhaustive search for the phylogeny that has the highest probability is therefore not practical even with a moderate number of sequences. Besides, while heuristics exist (e.g., stepwise addition [109]; see [4] for a review), almost none of these is guaranteed to converge on the optimum phylogenetic tree. The common practice is then to use one of these heuristics to find a good starting tree, and then modify repeatedly its topology more or less dramatically to explore its neighborhood for better trees until a stopping rule is satisfied [110]. The art here is in designing efficient tree perturbation methods that adaptively strike a balance between large topological modifications (that almost always lead to a very different tree with

a poor score) and small modifications (that almost always lead to an extremely similar tree with lower score). Some of today's challenges are about choosing between methods that successfully explore large numbers of trees but that can be costly in terms of computing time [110], and methods that are faster but may miss some interesting trees [53]. Several programs such as Leaphy, PhyML, and GARLI[41] are among the best-performing software in a maximum likelihood setting. In a Bayesian framework, the basic perturbation schemes were described early [36] and recently updated [111]. Three popular programs are MrBayes, BAMBE [36], and BEAST [39]. Among all these programs and approaches, PHYML, GARLI, and BEAST are probably among the most efficient programs in terms of computational speed, handling of large data sets and thoroughness of the tree search.

A first aspect of the computational complexity relates to estimating the support of a reconstructed phylogeny. This is more complicated than estimating a confidence interval for a real-valued parameter such as a branch length, because a tree topology is a graph and not a number. The classical approach therefore relies on a nonstandard use of the bootstrap [112]. However, the interpretation of the bootstrap is contentious. Bootstrap proportions P can be perceived as testing the correctness of internal nodes, and failing to do so [113], or $1–P$ can be interpreted as a conservative probability of falsely supporting monophyly [114]. Since bootstrap proportions are either too liberal or too conservative depending on the exact interpretation given to these values [115], it is difficult to adjust the threshold below which monophyly can be confidently ruled out [116]. Alternatively, an intuitive geometric argument was proposed to explain the conservativeness of bootstrap probabilities [117], but the workaround was never actually used in the community or implemented in any popular software. The introduction of Bayesian approaches in the late 1990s [36, 118] suggested a novel approach to estimate phylogenetic support with posterior probabilities. Clade or bipartition posterior probabilities can be relatively fast to compute, even for large data sets analyzed under complicated substitution models [119]. As in model selection, they have a clear interpretation as they measure the probability that a clade is correct, given the data and the model. But as with bootstrap probabilities, some controversies exist. Early empirical studies found that posterior probabilities of highly supported nodes were much larger than bootstrap probabilities [120], and subsequent simulation studies supported this observation (e.g., [121–124]). Some of these differences can be attributed to an artifact of the simulation scheme that was employed [125], but more specific empirical and simulation studies show that prior specifications can dramatically impact posterior probabilities for trees and clades [115, 126, 127]. In the simplest case, the analysis of simulated star trees with four sequences fails to give the expected three unrooted topologies with equal probability (1/3, 1/3, 1/3) but returns large posterior probabilities for an arbitrary topology [115, 126], even when infinitely long sequences are used [128, 129] ([130]). This phenomenon, called the star-tree paradox [126], seems to disappear when polytomies are assigned nonzero prior probabilities and when nonuniform priors force internal branch length towards zero [129]. The second issue surrounding Bayesian phylogenetic methods is about their convergence rate. A theoretical study shows that extremely simple Markov chain Monte Carlo (MCMC) samplers, the technique used to estimate posterior probabilities, could take an extremely long time to converge [131]. In practice, however, MCMC samplers such as those implemented in MrBayes are much more sophisticated. In particular, they include different types of moves [111] and use tempering, where some of the chains of a single run are heated, to improve mixing [43]. As a result, it is unclear whether they suffer from extremely long convergence times. It is also expected that current convergence diagnostic tools such as those implemented in MrBayes would reveal convergence problems [132]. Finally, it is also argued that these controversies such as exaggerated clade support, inconsistently biased priors, and the impossibility of hypothesis testing disappear altogether when posterior probabilities at internal nodes are abandoned in favor of posterior probabilities for topologies [133] (see Section 2.4 below).

The most fundamental aspect of the computational complexity in phylogenetics is due to the structure of the phylogenies: these are trees or binary graphs on which computations are nested and interdependent, which makes these computations intractable or NP-hard [134]. As a result, it is difficult to adopt an efficient "divide and conquer" approach, where a large complicated problem would be split into small simpler tasks, and to take advantage of today's commodity computing by distributing the computation over multicore architectures or heterogeneous computer clusters. Current strategies are limited to distributing the computation of the discrete rate categories (when using a "+Γ" substitution model) and part of the search algorithm [54], or simply to distribute different maximum likelihood bootstrap replicates [53, 54] or different MCMC samplers to available processors [44].

2.4. Comparisons of tree topologies

Science proceeds by testing hypotheses, and it is often necessary to compare phylogenies, for instance to test whether a given data set supports the early divergence of gymnosperms with respect to Gnetales and angiosperms (the anthophyte hypothesis), or whether the Gnetales diverged first (the Gnetales hypothesis) [135, 136]. Because of the importance of comparing phylogenies, a number of tests of molecular phylogenies were developed early. The KH test was first developed to compare two random trees [137]. However, this test is invalid if one of the trees is the maximum likelihood tree [138]. In this case, the SH test should be used [139]. Because the SH test can be very conservative, an approximately unbiased version was developed: the AU test [140]. PAUP and PAML only implement the KH and SH tests; CONSEL [40] also implements the AU test. A Bayesian version of these tests also exists [141], but the computations are more demanding.

Indeed, the Bayesian approach to hypothesis testing relies on computing the probability of the data under a particular model. This quantity is usually not available as a close-form equation, and it must be approximated numerically. The most straightforward approximation is based on the harmonic mean of the likelihood sampled from the posterior

distribution [142]. This approximation was described several times in the context of phylogenies [141, 143] and is available from most Bayesian programs such as MrBayes or BEAST. However, the approximation is extremely sensitive to the behavior of the MCMC sampler [52, 142]: if extremely low-likelihood values happen to be sampled from the posterior distribution, the harmonic mean will be dramatically affected. To date, a couple of more robust approximations have been described and were shown to be preferable to the harmonic mean estimator [52]. The first is based on thermodynamic integration [52] and is available in PhyloBayes (see Table 1). The second approximation [144] is based on a more direct computation [145], but its availability is currently limited to one specific model of evolution.

2.5. More realistic models

While model selection is fully justified on the ground of the bias-variance tradeoff, it should not be forgotten that all these models are simplified representations of the actual substitution process and are all therefore wrong. Stated differently, if AIC selects the GTR +Γ+I to analyze a data set, it should be clear that this conclusion does not imply that the data evolved under this model. All model selection procedures measure a relative model fit. One way to estimate adequacy or absolute model fit is to perform a parametric bootstrap test [146]: *first*, the selected model is compared with a multinomial model by means of a LRT whose test statistic is s (twice the log-likelihood difference); the following steps determine the distribution of s under the null hypothesis that the selected model was the generating model; *second*, the selected model is used to simulate a large number of data sets; *third*, the model selection procedure (LRT) is repeated on each simulated data set, and the corresponding test statistics s^* are recorded; *fourth*, the P-value is estimated as the number of times, the simulated s^* test statistics are more extreme ($>$, for a one-sided test) than the original value of s. The results of such tests suggest that the selected substitution model is generally not an adequate representation of the actual substitution process [85]. Of course, we do not need a model that incorporates all the minute biological features of evolutionary processes. As argued repeatedly (e.g., [147]), we need *useful* models that capture enough of reality of substitution processes to make accurate predictions and avoid systematic biases such as long-branch attraction [148].

More realistic models are obtained by accommodating heterogeneities in the evolutionary process at the level of both sites (space) and lineages (time). The simplest site-heterogeneous model is one, where the aligned data are partitioned, usually based on some prior information. For instance, first and second codon positions are known to evolve slower than third codon positions in protein-coding genes, or exposed residues might evolve faster than buried amino acids in globular proteins. A number of models were suggested to analyze such partitioned data sets (e.g., [149]); these models are implemented in most general-purpose software (e.g., PAML, PAUP, MrBayes) and can be combined with a "+Γ+I" component. A different approach consists in considering that sites can be binned in a number of rate categories; the use of a Dirichlet prior process then makes it possible both to determine the appropriate number of categories and to assign sites to these categories; the application of this method to protein-coding genes was able to recover the underlying codon structure of these genes [150]. However, several studies suggest that evolutionary patterns can be as heterogeneous within a priori partitions as among partitions [37, 151].

Lineage-heterogeneous models or heterotachous models [152] have attracted more attention. In one such approach, different models of evolution are assigned to the different branches of the tree [153], which can make these models extremely parameter-rich. Such a large number of parameters can potentially affect the accuracy of the phylogenetic inference (see the "bias-variance tradeoff" above) and present computational issues (long running times, large memory requirements, and convergence issues). Several simplifications can be made. One assumes that some sets of branches evolve under a particular process [153]. But now these branches must be assigned a priori, and both the determination of the number of sets and their placement on the tree can be difficult (but see Section 4 below for a solution to a similar question). At the other end of the spectrum of heterotachous models lies the simplest model known as the covarion model [154], where sites can either be variable along a branch, or not, and can switch between these two categories across time (e.g., [155], also described in a Bayesian framework [156]).

Between these two extremes are mixture models, which extend the covarion model by allowing more categories of sites. A number of formulations exist, where each site is assumed to have been generated by either several sets of branch lengths [157, 158] or by several rate matrices [37, 96, 151]. One particularity of these models is that they give a semi-parametric perspective to the phylogenetic estimation: if a single simple model cannot approximate a complex substitution process, the hope is that mixing several simple substitution models makes our models more realistic. In some applications, mixture models can also be used to avoid underestimating uncertainty, first when choosing a single model of evolution and then ignoring this uncertainty when estimating the phylogeny. The mixing therefore involves fitting at each site several sets of branch lengths, or several substitution models to the data, and combining these models using a certain weighting scheme. The difference between the numerous mixture models that have been described lies in the choice of the weight factors, and how these are obtained. In one approach, known as model averaging, the weights are determined a priori. A first possibility is to assume that all the models are equally probable, which does not work with an infinite number of models (individual weights are zero in this case). More critically in phylogenetics, this assumption is not coherent for nested models since larger models should be more likely than each submodel. A second possibility is to weight the models with respect to their probability of being the generating model given the data. For practical purposes, this posterior probability can be approximated by Akaike weights [96]. The difficulty here is that model averaging requires analyzing the data even for models that, a posteriori, turn out to have extremely small probabilities or

weights. This may be seen as a waste of resources (computing time and storage space).

2.6. Integrated Bayesian approaches

Mixture models can work within the framework of maximum likelihood, but the treatment of the weight factors is complicated. A sound alternative is to resort to a fully Bayesian approach. A prior distribution is set on the weight factors, and a special form of MCMC sampler whose Markov chain moves across models with different numbers of parameters, a reversible-jump MCMC sampler (RJ-MCMC), is constructed. The advantage of RJ-MCMC samplers is that they allow estimating the phylogeny while integrating over the uncertainty pertaining to the parameters of the substitution model and even integrating over the model itself [104]. Mixture models are available in BayesPhylogenies [37] for nucleotide models. Another Bayesian mixture model, named CAT for CATegories, was developed to analyze amino acid alignments. The CAT model recently proved successful in a number of empirical [159, 160] and simulation [161] studies in avoiding the artifact known as long-branch attraction [148]. This model is freely available in the PhyloBayes software (see Table 1).

All these models assume that each site evolve independently. The independence assumption greatly simplifies the computations, but is also highly unrealistic. Models that describe the evolution of doublets in RNA genes [162], triplets in codon models [163, 164], or other models with local or context dependencies [165–167] exist, but complete dependence models are still in their infancy and, so far, have only been implemented in a Bayesian framework [168, 169]. One particularly interesting feature of this approach is that complete dependence models incorporate information about the three-dimensional (3D) structure of proteins and therefore permit the explicit modeling of structural constraints or of any other site-interdependence pattern [170]. The incorporation of 3D structures also allows the establishment of a direct relationship between evolution at the DNA level and at the phenotypic level. This link between genotype and phenotype is established via a proxy that plays the role of a fitness function which, in retrospect, can be used to predict amino-acid sequences compatible with a given target structure, that is, to help in protein design [171].

3. DETECTING POSITIVE SELECTION

Fitness functions are however difficult to determine at the molecular level. In addition, while examples of adaptive evolution at the morphological level abound, from Darwin's finches in the Galapagos [172] to cichlid fishes in the East African lakes [173], the role of natural selection in shaping the evolution of genomes is much more controversial [147, 174]. First, the neutral theory of molecular evolution asserts that much of the variation at the DNA level is due to the random fixation of mutations with no selective advantage [175]. Second, a compelling body of evidence suggests that most of the genomic complexities have emerged by non-adaptive processes [176]. A number of statistical approaches exist either to test neutrality at the population level or to detect positive Darwinian evolution at the species level [147]. A shortcoming of neutrality tests is their dependence on a demographic model [177] and their sensitivity to processes of molecular evolution such as among-site rate variation [178]. They also do not model alternative hypotheses that would permit distinguishing negative selection from adaptive evolution. The development of demographic models based on Poisson random fields [179] and composite likelihoods [180] makes it possible both to estimate the strength of selection and to assess the impact of a variety of scenarios on allele frequency spectra [9]. But demographic singularities such as bottlenecks can still generate spurious signatures of positive selection [180, 181].

When effective population sizes are no longer a concern, for instance in studies at or above the species level, the detection of positive selection in protein-coding genes usually relies on codon models [163, 164] (see [182] for a review including methods based on amino-acid models). Codon models permit distinguishing between synonymous substitutions, which are likely to be neutral, and nonsynonymous substitutions, which are directly exposed to the action of selection. If synonymous and nonsynonymous substitutions accumulate at the same rate, then the protein-coding gene is likely to evolve neutrally. Alternatively, if nonsynonymous substitutions accumulate slower than synonymous substitutions, it must be because nonsynonymous substitutions are deleterious and this suggests the action of purifying selection. Conversely, the accumulation of nonsynonymous substitutions faster than synonymous substitutions suggests the action of positive selection. The nonsynonymous to synonymous rate ratio, denoted $\omega = d_N/d_S$, is therefore interpreted as a measure of selection at the protein level, with $\omega = 1$, <1 and >1 indicating neutral evolution, negative or positive selection, respectively. This ratio is also denoted K_a/K_s, in particular in studies that rely on counts of nonsynonymous and synonymous sites (e.g., [183]). An extension exists to detect selection in noncoding regions [184], and a promising phylogenetic hidden Markov or phylo-HMM model permits detection of selection in overlapping genes [185].

These rate ratios can be estimated by a number of methods implemented in MEGA, DAMBE, HyPhy [42], and PAML. The most intuitive methods, called counting methods, work in three steps: (i) count synonymous and nonsynonymous sites, (ii) count the observed differences at these sites, and (iii) apply corrections for multiple substitutions [186]. Counting methods are however not optimal in the sense that most work on pairs of sequences and therefore, just like neighbor-joining, fail to account for all the information contained in an alignment. In addition, simulations suggest that counting methods can be sensitive to a variety of biases such as unequal transition and transversion rates, or uneven base, or codon frequencies [187]. Counting methods that incorporate these biases perform generally better than those that do not, but the maximum likelihood method still appears more robust to sever biases [187]. In addition, the maximum likelihood method that accounts for all the information in a data set has good power and good accuracy to detect positive selection [188, 189].

However, the first studies using these methods found little evidence for adaptive evolution essentially because they were averaging ω rate ratios over both lineages and sites [147]. Branch models were then developed [190, 191] quickly followed by site models [192–196] and by branch-site models [189, 197]. All these approaches, as implemented in PAML, rely on likelihood ratio tests to detect adaptive evolution: a model where adaptive evolution is permitted is compared with a null model where ω cannot be greater than one. Simulations show that some of these tests are conservative [189], so that detection of adaptive evolution should be safe as long as convergence of the analyses is carefully checked [198], including in large-scale analyses [199]. If the model allowing adaptive evolution explains the data significantly better than the null model, then an empirical Bayes approach can be used to identify which sites are likely to evolve adaptively [192]. The empirical Bayes approach relies on estimates of the model parameters, which can have large sampling errors in small data sets. Because these sampling errors can cause the empirical Bayes site identification to be unreliable [200], a Bayes empirical Bayes approach was proposed and was shown to have good power and low-false positive rates [201]. Full Bayesian approaches that allow for uncertain parameter estimates were also proposed [202]. Yet, simulations showed that they did not improve further on Bayes empirical Bayes estimates [203], so that the computational overhead incurred by full Bayes methods may not be necessary in this case. One particular case, where a Bayesian approach is however required, is to tell the signature of adaptive evolution from that of recombination, as these two processes can leave similar signals in DNA sequences. Indeed, simulations show that recombination can lead to false positive rates as large as 90% when trying to detect adaptive evolution [204]. The codon model with recombination implemented in OmegaMap [48] can then be used to tease apart these two processes (e.g., see [205]).

4. ESTIMATING DIVERGENCE TIMES BETWEEN SPECIES

The estimation of the dates when species diverged is often perceived to be as important as estimating the phylogeny itself. This explains why so-called "dating methods" were first wished for when molecular phylogenies were first reconstructed [206]. In spite of over four decades of history, molecular dating has only recently seen new developments. One of the reasons for this slow progress is that, unlike the other parts of phylogenetic analysis, divergence times are parameters that cannot be estimated directly. Only sitewise likelihood values and distances between pairs of sequences are identifiable, that is, directly estimable. Distances are expressed as a number of substitutions per site (sub/site) and can be decomposed as the product of two quantities: a rate of evolution (sub/site/unit of time) and a time duration (unit of time). As a result, time durations and, likewise, divergence times cannot be estimated without making an additional assumption on the rates of evolution. The simplest assumption is to posit that rates are constant in time, which is known as the molecular clock hypothesis [207]. This hypothesis can be tested, for instance, with PAUP or PAML, by means of a likelihood ratio test that compares a constrained model (clock) with an unconstrained model (no clock). These two models are nested, so that twice the log-likelihood difference asymptotically follows a χ^2 distribution. If n sequences are analyzed, the constrained model estimates $n - 1$ divergence times, while the unconstrained model estimates $2n - 3$ branch lengths. The degree of freedom of this test is then $(2n - 3) - (n - 1) = n - 2$ [4]. The systematic test of the molecular clock assumption on recent data shows that this hypothesis is too often untenable [208].

The most recent work has then focused on relaxing this assumption, and three different directions have emerged [209]. A first possibility is to relax the clock *globally* on the phylogeny, but to assume that the hypothesis still holds *locally* for closely related species [210–212]. Recent developments of these local clock models now allow the use of multiple calibration points and of multiple genes [213], the automatic placement of the clocks on the tree [214] and the estimation of the number of local clocks [209]. PAML can be used for most of these computations. However, local clock models still tend to underestimate rapid rate change [209]. The second possibility to relax the global clock assumption is to assume that rates of evolution evolve in an autocorrelated manner along lineages and to minimize the amount of rate change over the entire phylogeny. The most popular approach in the plant community is Sanderson's penalized likelihood [215], implemented in r8s [55]. This approach performs well on data sets for which the actual fossil dates are known [216] but still tends to underestimate the actual amount of rate change [209].

Bayesian methods appear today as the emerging approach to estimate divergence times. Taking inspiration from Sanderson's pioneering work [217], Thorne et al. developed a Bayesian framework where rates of evolution change in an autocorrelated manner across lineages [45–47]: the rate of evolution of a branch depends on the rate of evolution of its parental branch; the branches emanating from the root require a special treatment. These Bayesian models work by modeling how rates of evolution change in time (rate prior), and how the speciation/population process shapes the distribution of divergence times (speciation prior). These prior distributions can actually be interpreted as penalty functions [45, 209], and they can have simple or more complicated forms [218]. The Multidivtime program [45–47] is extremely quick to analyze data thanks to the use of a multivariate normal approximation of the likelihood surface. It assumes that rates of evolution change following a stationary lognormal prior distribution. Further work suggested that it might not always be the best performing rate prior [218–220], but these latter studies had two potential shortcomings: (i) they were based on a speciation prior that was so strong that it biased divergence times towards the age of the fossil root [219, 221], and (ii) they used a statistical procedure, the posterior Bayes factor [222], that is potentially inconsistent. One potential limitation of the Bayesian approach described so far is its dependence on one single tree topology, which must be either known ahead of time or estimated by other means. Recently, Drummond et al. found a way to relax this requirement by

positing that rates of evolution are uncorrelated across lineages, while all the branches of the tree are constrained to follow exactly the same rate prior [223]. As a result, their approach is able to estimate the most probable tree (given the data and the substitution model), the divergence times and the position of the root even without any outgroup or without resorting to a nonreversible model of substitution [224]. Drummond et al. further argue that the use of explicit models of rate variation over time might contribute to improved phylogenetic inference [223]. In addition, when the focus is on estimating divergence times, a recent analysis suggests that this uncorrelated model of rate change could outperform the methods described above to accommodate rapid rate change among lineages [209]. Implemented in BEAST, this approach offers a variety of substitution models and prior distributions and presents a graphic user interface that will appeal to numerous researchers [39].

5. CHALLENGES AND PERSPECTIVES

With the advent of high-throughput sequencing technologies such as the whole-genome shotgun approach by pyrosequencing [225], fast, cheap, and accurate genomic information is becoming available for a growing number of species [226]. If low coverage limits the complete assembly of many genome projects, it still allows the quick access to draft genomes for a growing number of species [227]. As a result, phylogenetic inference can now incorporate large numbers of expressed sequence tags (ESTs), genes [228], and occasionally complete genomes [229]. The motivation for developing these so-called phylogenomic approaches is their presumed ability to return fully resolved and well-supported trees by decreasing both sampling errors [230] and misleading signals due for instance to horizontal gene transfer [231] or to hidden paralogy [232]. In practice, these large-scale studies can give the impression that incongruence is resolved [228], but they also can fail to address systematic errors due to the use of too simple models [233]. If the genes incorporated in phylogenomic studies are often concatenated to limit the number of parameters entering the model, it remains important to allow sitewise heterogeneities [234]. If partition models can reduce systematic biases [234], Bayesian mixture models such as CAT [151] appear to be robust to long-branch attraction [159], a rampant issue in phylogenomics [235]. All together, the accumulation of genomic data and these latest methodological developments seem to make the reconstruction of the tree of life finally within reach. In comparison, dating the tree of life is still in its infancy, even if a number of initiatives such as the TimeTree server are being developed [236]. These resources are limited to some vertebrates but will hopefully soon be extended to include other large taxonomic groups such as plants. To achieve this goal, however, phylogenetic studies should systematically incorporate divergence times, as is now routine in some research communities (e.g., [237]). This joint estimation of time and trees is today facilitated by the availability of user-friendly programs such as BEAST. The near future will probably see the development of mixture models for molecular dating and more sophisticated models that integrate most of the topics discussed here from sequence alignment to detection of sites under selection into one single but yet user-friendly [238] toolbox.

ACKNOWLEDGMENTS

Jeff Thorne provided insightful comments and suggestions, and two anonymous reviewers helped in improving the original manuscript. Support was provided by the Natural Sciences Research Council of Canada (DG-311625 to SAB and DG-261252 to XX).

REFERENCES

[1] C. Darwin, *On the Origin of Species by Means of Natural Selection*, J. Murray, London, UK, 1859.

[2] R. R. Sokal and P. H. A. Sneath, *Principles of Numerical Taxonomy*, W. H. Freeman, San Francisco, Calif, USA, 1963.

[3] L. L. Cavalli-Sforza, I. Barrai, and A. W. Edwards, "Analysis of human evolution under random genetic drift," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 29, pp. 9–20, 1964.

[4] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates, Sunderland, Mass, USA, 2004.

[5] H. Glenner, A. J. Hansen, M. V. Sørensen, F. Ronquist, J. P. Huelsenbeck, and E. Willerslev, "Bayesian inference of the metazoan phylogeny: a combined molecular and morphological approach," *Current Biology*, vol. 14, no. 18, pp. 1644–1649, 2004.

[6] B. E. Pfeil, J. A. Schlueter, R. C. Shoemaker, and J. J. Doyle, "Placing paleopolyploidy in relation to taxon divergence: a phylogenetic analysis in legumes using 39 gene families," *Systematic Biology*, vol. 54, no. 3, pp. 441–454, 2005.

[7] E. R. Chare and E. C. Holmes, "A phylogenetic survey of recombination frequency in plant RNA viruses," *Archives of Virology*, vol. 151, no. 5, pp. 933–946, 2006.

[8] H. Philippe and C. J. Douady, "Horizontal gene transfer and phylogenetics," *Current Opinion in Microbiology*, vol. 6, no. 5, pp. 498–505, 2003.

[9] R. Nielsen, C. Bustamante, A. G. Clark, et al., "A scan for positively selected genes in the genomes of humans and chimpanzees," *PLoS Biology*, vol. 3, no. 6, p. e170, 2005.

[10] S. R. Ramírez, B. Gravendeel, R. B. Singer, C. R. Marshall, and N. E. Pierce, "Dating the origin of the Orchidaceae from a fossil orchid with its pollinator," *Nature*, vol. 448, no. 7157, pp. 1042–1045, 2007.

[11] R. D. Knight, S. J. Freeland, and L. F. Landweber, "Rewiring the keyboard: evolvability of the genetic code," *Nature Reviews Genetics*, vol. 2, no. 1, pp. 49–58, 2001.

[12] J. Antonovics, M. E. Hood, and C. H. Baker, "Molecular virology: was the 1918 flu avian in origin?" *Nature*, vol. 440, no. 7088, p. E9, 2006, discussion E9-10.

[13] A. P. Jackson and M. A. Charleston, "A cophylogenetic perspective of RNA-virus evolution," *Molecular Biology and Evolution*, vol. 21, no. 1, pp. 45–57, 2004.

[14] J. P. Huelsenbeck, B. Rannala, and B. Larget, "A Bayesian framework for the analysis of cospeciation," *Evolution*, vol. 54, no. 2, pp. 352–364, 2000.

[15] M. Hajibabaei, G. A. C. Singer, P. D. N. Hebert, and D. A. Hickey, "DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics," *Trends in Genetics*, vol. 23, no. 4, pp. 167–172, 2007.

[16] S.-J. Luo, J.-H. Kim, W. E. Johnson, et al., "Phylogeography and genetic ancestry of tigers (*Panthera tigris*)," *PLoS Biology*, vol. 2, no. 12, p. e442, 2004.

[17] C. J. Howe, A. C. Barbrook, M. Spencer, P. Robinson, B. Bordalejo, and L. R. Mooney, "Manuscript evolution," *Endeavour*, vol. 25, no. 3, pp. 121–126, 2001.

[18] R. D. Gray and Q. D. Atkinson, "Language-tree divergence times support the Anatolian theory of Indo-European origin," *Nature*, vol. 426, no. 6965, pp. 435–439, 2003.

[19] D. M. Hillis and J. P. Huelsenbeck, "Support for dental HIV transmission," *Nature*, vol. 369, no. 6475, pp. 24–25, 1994.

[20] A. Salas, H.-J. Bandelt, V. Macaulay, and M. B. Richards, "Phylogeographic investigations: the role of trees in forensic genetics," *Forensic Science International*, vol. 168, no. 1, pp. 1–13, 2007.

[21] D. Sankoff and J. H. Nadeau, "Chromosome rearrangements in evolution: from gene order to genome sequence and back," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 20, pp. 11188–11189, 2003.

[22] D. L. Swofford, P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers, "Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods," *Systematic Biology*, vol. 50, no. 4, pp. 525–539, 2001.

[23] M. Holder and P. O. Lewis, "Phylogeny estimation: traditional and Bayesian approaches," *Nature Reviews Genetics*, vol. 4, no. 4, pp. 275–284, 2003.

[24] A. Siepel and D. Haussler, "Phylogenetic hidden Markov models," in *Statistical Methods in Molecular Evolution*, R. Nielsen, Ed., pp. 325–351, Springer, New York, NY, USA, 2005.

[25] M. Pagel and A. Meade, "Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo," *American Naturalist*, vol. 167, no. 6, pp. 808–825, 2006.

[26] S. Kumar and A. Filipski, "Multiple sequence alignment: in pursuit of homologous DNA positions," *Genome Research*, vol. 17, no. 2, pp. 127–135, 2007.

[27] C. Notredame, "Recent evolutions of multiple sequence alignment algorithms," *PLoS Computational Biology*, vol. 3, no. 8, p. e123, 2007.

[28] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Current Opinion in Structural Biology*, vol. 16, no. 3, pp. 368–373, 2006.

[29] X. Xia and Z. Xie, "DAMBE: software package for data analysis in molecular biology and evolution," *Journal of Heredity*, vol. 92, no. 4, pp. 371–373, 2001.

[30] S. Kumar, K. Tamura, and M. Nei, "MEGA: molecular evolutionary genetics analysis software for microcomputers," *Computer Applications in the Biosciences*, vol. 10, no. 2, pp. 189–191, 1994.

[31] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1596–1599, 2007.

[32] C. B. Do, M. S. P. Mahabhashyam, M. Brudno, and S. Batzoglou, "ProbCons: probabilistic consistency-based multiple sequence alignment," *Genome Research*, vol. 15, no. 2, pp. 330–340, 2005.

[33] I. M. Wallace, G. Blackshields, and D. G. Higgins, "Multiple sequence alignments," *Current Opinion in Structural Biology*, vol. 15, no. 3, pp. 261–266, 2005.

[34] I. M. Wallace, O. O'Sullivan, D. G. Higgins, and C. Notredame, "M-Coffee: combining multiple sequence alignment methods with T-Coffee," *Nucleic Acids Research*, vol. 34, no. 6, pp. 1692–1699, 2006.

[35] B. G. Hall, *Phylogenetic Trees Made Easy: A How-to Manual*, Sinauer Associates, Sunderland, Mass, USA, 2008.

[36] B. Larget and D. L. Simon, "Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees," *Molecular Biology and Evolution*, vol. 16, no. 6, pp. 750–759, 1999.

[37] M. Pagel and A. Meade, "A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data," *Systematic Biology*, vol. 53, no. 4, pp. 571–581, 2004.

[38] B. D. Redelings and M. A. Suchard, "Joint Bayesian estimation of alignment and phylogeny," *Systematic Biology*, vol. 54, no. 3, pp. 401–418, 2005.

[39] A. J. Drummond and A. Rambaut, "BEAST: Bayesian evolutionary analysis by sampling trees," *BMC Evolutionary Biology*, vol. 7, article 214, pp. 1–8, 2007.

[40] H. Shimodaira and M. Hasegawa, "CONSEL: for assessing the confidence of phylogenetic tree selection," *Bioinformatics*, vol. 17, no. 12, pp. 1246–1247, 2001.

[41] D. Zwickl, "Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion," Ph.D. thesis, University of Texas at Austin, Austin, Tex, USA, 2006.

[42] S. L. Kosakovsky Pond, S. D. W. Frost, and S. V. Muse, "HyPhy: hypothesis testing using phylogenies," *Bioinformatics*, vol. 21, no. 5, pp. 676–679, 2005.

[43] F. Ronquist and J. P. Huelsenbeck, "MrBayes 3: Bayesian phylogenetic inference under mixed models," *Bioinformatics*, vol. 19, no. 12, pp. 1572–1574, 2003.

[44] G. Altekar, S. Dwarkadas, J. P. Huelsenbeck, and F. Ronquist, "Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference," *Bioinformatics*, vol. 20, no. 3, pp. 407–415, 2004.

[45] J. L. Thorne, H. Kishino, and I. S. Painter, "Estimating the rate of evolution of the rate of molecular evolution," *Molecular Biology and Evolution*, vol. 15, no. 12, pp. 1647–1657, 1998.

[46] H. Kishino, J. L. Thorne, and W. J. Bruno, "Performance of a divergence time estimation method under a probabilistic model of rate evolution," *Molecular Biology and Evolution*, vol. 18, no. 3, pp. 352–361, 2001.

[47] J. L. Thorne and H. Kishino, "Divergence time and evolutionary rate estimation with multilocus data," *Systematic Biology*, vol. 51, no. 5, pp. 689–702, 2002.

[48] D. J. Wilson and G. McVean, "Estimating diversifying selection and functional constraint in the presence of recombination," *Genetics*, vol. 172, no. 3, pp. 1411–1425, 2006.

[49] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.

[50] Z. Yang, "PAML 4: phylogenetic analysis by maximum likelihood," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1586–1591, 2007.

[51] D. L. Swofford, *PAUP*: Phylogenetic Analysis Using Parsimony (and other Methods) 4.0 Beta*, Sinauer Associates, Sunderland, Mass, USA, 10th edition, 2002.

[52] N. Lartillot and H. Philippe, "Computing Bayes factors using thermodynamic integration," *Systematic Biology*, vol. 55, no. 2, pp. 195–207, 2006.

[53] S. Guindon and O. Gascuel, "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood," *Systematic Biology*, vol. 52, no. 5, pp. 696–704, 2003.

[54] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 2006.

[55] M. J. Sanderson, "r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock," *Bioinformatics*, vol. 19, no. 2, pp. 301–302, 2003.

[56] K. M. Kjer, "Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs," *Molecular Phylogenetics and Evolution*, vol. 4, no. 3, pp. 314–330, 1995.

[57] C. Notredame, E. A. O'Brien, and D. G. Higgins, "RAGA: RNA sequence alignment by genetic algorithm," *Nucleic Acids Research*, vol. 25, no. 22, pp. 4570–4580, 1997.

[58] R. E. Hickson, C. Simon, and S. W. Perrey, "The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence," *Molecular Biology and Evolution*, vol. 17, no. 4, pp. 530–539, 2000.

[59] X. Xia, "Phylogenetic relationship among horseshoe crab species: effect of substitution models on phylogenetic analyses," *Systematic Biology*, vol. 49, no. 1, pp. 87–100, 2000.

[60] X. Xia, Z. Xie, and K. M. Kjer, "18S ribosomal RNA and tetrapod phylogeny," *Systematic Biology*, vol. 52, no. 3, pp. 283–295, 2003.

[61] J. D. Thompson, D. G. Higgins, and T. J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucleic Acids Research*, vol. 22, no. 22, pp. 4673–4680, 1994.

[62] M. A. Larkin, G. Blackshields, N. P. Brown, et al., "Clustal W and clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[63] T. Golubchik, M. J. Wise, S. Easteal, and L. S. Jermiin, "Mind the gaps: evidence of bias in estimates of multiple sequence alignments," *Molecular Biology and Evolution*, vol. 24, no. 11, pp. 2433–2442, 2007.

[64] G. Landan and D. Graur, "Heads or tails: a simple reliability check for multiple sequence alignments," *Molecular Biology and Evolution*, vol. 24, no. 6, pp. 1380–1383, 2007.

[65] A. S. Schwartz, E. W. Myers, and L. Pachter, "Alignment metric accuracy," http://arxiv.org/abs/q-bio.QM/0510052, 2005.

[66] J. Zhu, J. S. Liu, and C. E. Lawrence, "Bayesian adaptive sequence alignment algorithms," *Bioinformatics*, vol. 14, no. 1, pp. 25–39, 1998.

[67] I. Holmes and W. J. Bruno, "Evolutionary HMMs: a Bayesian approach to multiple alignment," *Bioinformatics*, vol. 17, no. 9, pp. 803–820, 2001.

[68] J. L. Jensen and J. Hein, "Gibbs sampler for statistical multiple alignment," *Statistica Sinica*, vol. 15, no. 4, pp. 889–907, 2005.

[69] M. Clamp, J. Cuff, S. M. Searle, and G. J. Barton, "The Jalview Java alignment editor," *Bioinformatics*, vol. 20, no. 3, pp. 426–427, 2004.

[70] D. Sankoff and R. Cedergren, "Simultaneous comparison of three or more sequences related by a tree," in *Time Wraps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, D. Sankoff and R. Cedergren, Eds., pp. 253–264, Addison-Wesley, Reading, Mass, USA, 1983.

[71] J. Hein, "A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given," *Molecular Biology and Evolution*, vol. 6, no. 6, pp. 649–668, 1989.

[72] G. Lunter, I. Miklós, A. Drummond, J. L. Jensen, and J. Hein, "Bayesian coestimation of phylogeny and sequence alignment," *BMC Bioinformatics*, vol. 6, article 83, pp. 1–10, 2005.

[73] K. M. Wong, M. A. Suchard, and J. P. Huelsenbeck, "Alignment uncertainty and genomic analysis," *Science*, vol. 319, no. 5862, pp. 473–476, 2008.

[74] G. Lunter, A. Rocco, N. Mimouni, A. Heger, A. Caldeira, and J. Hein, "Uncertainty in homology inferences: assessing and improving genomic sequence alignment," *Genome Research*, vol. 18, no. 2, pp. 298–309, 2008.

[75] M. Nei and S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, New York, NY, USA, 2000.

[76] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," in *Mammalian Protein Metabolism*, H. N. Munro, Ed., pp. 21–121, Academic Press, New York, NY, USA, 1969.

[77] M. Kimura, "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences," *Journal of Molecular Evolution*, vol. 16, no. 2, pp. 111–120, 1980.

[78] M. Hasegawa, H. Kishino, and T. Yano, "Dating of the human-ape splitting by a molecular clock of mitochondrial DNA," *Journal of Molecular Evolution*, vol. 22, no. 2, pp. 160–174, 1985.

[79] K. Tamura and M. Nei, "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees," *Molecular Biology and Evolution*, vol. 10, no. 3, pp. 512–526, 1993.

[80] S. Tavare, "Some probabilistic and statistical problems on the analysis of DNA sequences," in *Lectures on Mathematics in the Life Sciences*, vol. 17, pp. 57–86, American Mathematical Society, Providence, RI, USA, 1986.

[81] Z. Yang, "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites," *Molecular Biology and Evolution*, vol. 10, no. 6, pp. 1396–1401, 1993.

[82] Z. Yang, "Estimating the pattern of nucleotide substitution," *Journal of Molecular Evolution*, vol. 39, no. 1, pp. 105–111, 1994.

[83] N. Goldman and S. Whelan, "A novel use of equilibrium frequencies in models of sequence evolution," *Molecular Biology and Evolution*, vol. 19, no. 11, pp. 1821–1831, 2002.

[84] P. Liò and N. Goldman, "Models of molecular evolution and phylogeny," *Genome Research*, vol. 8, no. 12, pp. 1233–1244, 1998.

[85] S. Whelan, P. Liò, and N. Goldman, "Molecular phylogenetics: state-of-the-art methods for looking into the past," *Trends in Genetics*, vol. 17, no. 5, pp. 262–272, 2001.

[86] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference : A Practical Information-Theoretic Approach*, Springer, New York, NY, USA, 2002.

[87] W. J. Bruno and A. L. Halpern, "Topological bias and inconsistency of maximum likelihood using wrong models," *Molecular Biology and Evolution*, vol. 16, no. 4, pp. 564–566, 1999.

[88] D. Posada and K. A. Crandall, "Selecting the best-fit model of nucleotide substitution," *Systematic Biology*, vol. 50, no. 4, pp. 580–601, 2001.

[89] D. R. Cox, "Further results on tests of separate families of hypotheses," *Journal of the Royal Statistical Society. Series B*, vol. 24, no. 2, pp. 406–424, 1962.

[90] N. Goldman and S. Whelan, "Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics," *Molecular Biology and Evolution*, vol. 17, no. 6, pp. 975–978, 2000.

[91] M. Anisimova and O. Gascuel, "Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative," *Systematic Biology*, vol. 55, no. 4, pp. 539–552, 2006.

[92] D. Posada and K. A. Crandall, "MODELTEST: testing the model of DNA substitution," *Bioinformatics*, vol. 14, no. 9, pp. 817–818, 1998.

[93] E. Paradis, J. Claude, and K. Strimmer, "APE: analyses of phylogenetics and evolution in R language," *Bioinformatics*, vol. 20, no. 2, pp. 289–290, 2004.

[94] D. Posada, "ModelTest server: a web-based tool for the statistical selection of models of nucleotide substitution online," *Nucleic Acids Research*, vol. 34, web server issue, pp. W700–W703, 2006.

[95] F. Abascal, R. Zardoya, and D. Posada, "ProtTest: selection of best-fit models of protein evolution," *Bioinformatics*, vol. 21, no. 9, pp. 2104–2105, 2005.

[96] D. Posada and T. R. Buckley, "Model selection and model averaging in phylogenetics: advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests," *Systematic Biology*, vol. 53, no. 5, pp. 793–808, 2004.

[97] D. Pol, "Empirical problems of the hierarchical likelihood ratio test for model selection," *Systematic Biology*, vol. 53, no. 6, pp. 949–962, 2004.

[98] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[99] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[100] V. N. Minin, Z. Abdo, P. Joyce, and J. Sullivan, "Performance-based selection of likelihood models for phylogeny estimation," *Systematic Biology*, vol. 52, no. 5, pp. 674–683, 2003.

[101] Z. Abdo, V. N. Minin, P. Joyce, and J. Sullivan, "Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation," *Molecular Biology and Evolution*, vol. 22, no. 3, pp. 691–703, 2005.

[102] L. Bao, H. Gu, K. A. Dunn, and J. P. Bielawski, "Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on their application to gene and genome data," *BMC Evolutionary Biology*, vol. 7, supplement 1, p. S5, 2007.

[103] M. A. Suchard, R. E. Weiss, and J. S. Sinsheimer, "Bayesian selection of continuous-time Markov chain evolutionary models," *Molecular Biology and Evolution*, vol. 18, no. 6, pp. 1001–1013, 2001.

[104] J. P. Huelsenbeck, B. Larget, and M. E. Alfaro, "Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 1123–1133, 2004.

[105] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, no. 4, pp. 406–425, 1987.

[106] O. Gascuel and M. Steel, "Neighbor-joining revealed," *Molecular Biology and Evolution*, vol. 23, no. 11, pp. 1997–2000, 2006.

[107] W. J. Bruno, N. D. Socci, and A. L. Halpern, "Weighted neighbor-joining: a likelihood-based approach to distance-based phylogeny reconstruction," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 189–197, 2000.

[108] S. L. Baldauf, "Phylogeny for the faint of heart: a tutorial," *Trends in Genetics*, vol. 19, no. 6, pp. 345–351, 2003.

[109] L. L. Cavalli-Sforza and A. W. F. Edwards, "Phylogenetic analysis. Models and estimation procedures," *American Journal of Human Genetics*, vol. 19, no. 3, part 1, pp. 233–257, 1967.

[110] S. Whelan, "New approaches to phylogenetic tree search and their application to large numbers of protein alignments," *Systematic Biology*, vol. 56, no. 5, pp. 727–740, 2007.

[111] M. T. Holder, P. O. Lewis, D. L. Swofford, and B. Larget, "Hastings ratio of the LOCAL proposal used in Bayesian phylogenetics," *Systematic Biology*, vol. 54, no. 6, pp. 961–965, 2005.

[112] J. Felsenstein, "Confidence limits on phylogenies: an approach using the bootstrap," *Evolution*, vol. 39, no. 4, pp. 783–791, 1985.

[113] D. M. Hillis and J. J. Bull, "An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis," *Systematic Biology*, vol. 42, no. 2, pp. 182–192, 1993.

[114] J. Felsenstein and H. Kishino, "Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull," *Systematic Biology*, vol. 42, no. 2, pp. 193–200, 1993.

[115] Z. Yang and B. Rannala, "Branch-length prior influences Bayesian posterior probability of phylogeny," *Systematic Biology*, vol. 54, no. 3, pp. 455–470, 2005.

[116] V. Berry and O. Gascuel, "On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain," *Molecular Biology and Evolution*, vol. 13, no. 7, pp. 999–1011, 1996.

[117] B. Efron, E. Halloran, and S. Holmes, "Bootstrap confidence levels for phylogenetic trees," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 14, pp. 7085–7090, 1996.

[118] B. Mau, M. A. Newton, and B. Larget, "Bayesian phylogenetic inference via Markov chain Monte Carlo methods," *Biometrics*, vol. 55, no. 1, pp. 1–12, 1999.

[119] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback, "Bayesian inference of phylogeny and its impact on evolutionary biology," *Science*, vol. 294, no. 5550, pp. 2310–2314, 2001.

[120] W. J. Murphy, E. Eizirik, S. J. O'Brien, et al., "Resolution of the early placental mammal radiation using Bayesian phylogenetics," *Science*, vol. 294, no. 5550, pp. 2348–2351, 2001.

[121] C. J. Douady, F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. P. Douzery, "Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability," *Molecular Biology and Evolution*, vol. 20, no. 2, pp. 248–254, 2003.

[122] M. P. Cummings, S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka, "Comparing bootstrap and posterior probability values in the four-taxon case," *Systematic Biology*, vol. 52, no. 4, pp. 477–487, 2003.

[123] P. Erixon, B. Svennblad, T. Britton, and B. Oxelman, "Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics," *Systematic Biology*, vol. 52, no. 5, pp. 665–673, 2003.

[124] B. Svennblad, P. Erixon, B. Oxelman, and T. Britton, "Fundamental differences between the methods of maximum likelihood and maximum posterior probability in phylogenetics," *Systematic Biology*, vol. 55, no. 1, pp. 116–121, 2006.

[125] J. P. Huelsenbeck and B. Rannala, "Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models," *Systematic Biology*, vol. 53, no. 6, pp. 904–913, 2004.

[126] P. O. Lewis, M. T. Holder, and K. E. Holsinger, "Polytomies and Bayesian phylogenetic inference," *Systematic Biology*, vol. 54, no. 2, pp. 241–253, 2005.

[127] B. Kolaczkowski and J. W. Thornton, "Effects of branch length uncertainty on Bayesian posterior probabilities for phylogenetic hypotheses," *Molecular Biology and Evolution*, vol. 24, no. 9, pp. 2108–2118, 2007.

[128] M. Steel and F. A. Matsen, "The Bayesian "star paradox" persists for long finite sequences," *Molecular Biology and Evolution*, vol. 24, no. 4, pp. 1075–1079, 2007.

[129] Z. Yang, "Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1639–1655, 2007.

[130] B. Kolaczkowski and J. W. Thornton, "Is there a star tree paradox?" *Molecular Biology and Evolution*, vol. 23, no. 10, pp. 1819–1823, 2006.

[131] E. Mossel and E. Vigoda, "Phylogenetic MCMC algorithms are misleading on mixtures of trees," *Science*, vol. 309, no. 5744, pp. 2207–2209, 2005.

[132] F. Ronquist, B. Larget, J. P. Huelsenbeck, J. B. Kadane, D. Simon, and P. van der Mark, "Comment on "Phylogenetic MCMC algorithms are misleading on mixtures of trees"," *Science*, vol. 312, no. 5772, p. 367, 2006.

[133] W. C. Wheeler and K. M. Pickett, "Topology-Bayes versus clade-Bayes in phylogenetic analysis," *Molecular Biology and Evolution*, vol. 25, no. 2, pp. 447–453, 2008.

[134] B. Chor and T. Tuller, "Maximum likelihood of evolutionary trees: hardness and approximation," *Bioinformatics*, vol. 21, supplement 1, pp. i97–i106, 2005.

[135] M. J. Donoghue, "Progress and prospects in reconstructing plant phylogeny," *Annals of the Missouri Botanical Garden*, vol. 81, no. 3, pp. 405–418, 1994.

[136] S. Aris-Brosou, "Least and most powerful phylogenetic tests to elucidate the origin of the seed plants in the presence of conflicting signals under misspecified models," *Systematic Biology*, vol. 52, no. 6, pp. 781–793, 2003.

[137] H. Kishino and M. Hasegawa, "Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoide," *Journal of Molecular Evolution*, vol. 29, no. 2, pp. 170–179, 1989.

[138] N. Goldman, J. P. Anderson, and A. G. Rodrigo, "Likelihood-based tests of topologies in phylogenetics," *Systematic Biology*, vol. 49, no. 4, pp. 652–670, 2000.

[139] H. Shimodaira and M. Hasegawa, "Multiple comparisons of log-likelihoods with applications to phylogenetic inference," *Molecular Biology and Evolution*, vol. 16, no. 8, pp. 1114–1116, 1999.

[140] H. Shimodaira, "An approximately unbiased test of phylogenetic tree selection," *Systematic Biology*, vol. 51, no. 3, pp. 492–508, 2002.

[141] S. Aris-Brosou, "How Bayes tests of molecular phylogenies compare with frequentist approaches," *Bioinformatics*, vol. 19, no. 5, pp. 618–624, 2003.

[142] A. E. Raftery, "Hypothesis testing and model selection," in *Markov Chain Monte Carlo in Practice*, W. Gilks, S. Richardson, and D. J. Spiegelhalter, Eds., pp. 163–187, Chapman & Hall, Boca Raton, Fla, USA, 1996.

[143] J. A. A. Nylander, F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey, "Bayesian phylogenetic analysis of combined data," *Systematic Biology*, vol. 53, no. 1, pp. 47–67, 2004.

[144] S. C. Choi, A. Hobolth, D. M. Robinson, H. Kishino, and J. L. Thorne, "Quantifying the impact of protein tertiary structure on molecular evolution," *Molecular Biology and Evolution*, vol. 24, no. 8, pp. 1769–1782, 2007.

[145] S. Chib and I. Jeliazkov, "Marginal likelihood from the Metropolis-Hastings output," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 270–281, 2001.

[146] N. Goldman, "Statistical tests of models of DNA substitution," *Journal of Molecular Evolution*, vol. 36, no. 2, pp. 182–198, 1993.

[147] Z. Yang, *Computational Molecular Evolution*, Oxford University Press, Oxford, UK, 2006.

[148] J. Felsenstein, "Cases in which parsimony or compatibility methods will be positively misleading," *Systematic Zoology*, vol. 27, no. 4, pp. 401–410, 1978.

[149] Z. Yang, "Maximum-likelihood models for combined analyses of multiple sequence data," *Journal of Molecular Evolution*, vol. 42, no. 5, pp. 587–596, 1996.

[150] J. P. Huelsenbeck and M. A. Suchard, "A nonparametric method for accommodating and testing across-site rate variation," *Systematic Biology*, vol. 56, no. 6, pp. 975–987, 2007.

[151] N. Lartillot and H. Philippe, "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 1095–1109, 2004.

[152] P. Lopez, D. Casane, and H. Philippe, "Heterotachy, an important process of protein evolution," *Molecular Biology and Evolution*, vol. 19, no. 1, pp. 1–7, 2002.

[153] Z. Yang and D. Roberts, "On the use of nucleic acid sequences to infer early branchings in the tree of life," *Molecular Biology and Evolution*, vol. 12, no. 3, pp. 451–458, 1995.

[154] W. M. Fitch and E. Markowitz, "An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution," *Biochemical Genetics*, vol. 4, no. 5, pp. 579–593, 1970.

[155] C. Tuffley and M. Steel, "Modeling the covarion hypothesis of nucleotide substitution," *Mathematical Biosciences*, vol. 147, no. 1, pp. 63–91, 1998.

[156] J. P. Huelsenbeck, "Testing a covariotide model of DNA substitution," *Molecular Biology and Evolution*, vol. 19, no. 5, pp. 698–707, 2002.

[157] B. Kolaczkowski and J. W. Thornton, "Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogenous," *Nature*, vol. 431, no. 7011, pp. 980–984, 2004.

[158] M. Spencer, E. Susko, and A. J. Roger, "Likelihood, parsimony, and heterogeneous evolution," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1161–1164, 2005.

[159] N. Lartillot, H. Brinkmann, and H. Philippe, "Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model," *BMC Evolutionary Biology*, vol. 7, supplement 1, p. S4, 2007.

[160] E. Jiménez-Guri, H. Philippe, B. Okamura, and P. W. H. Holland, " *Buddenbrockia* is a cnidarian worm," *Science*, vol. 317, no. 5834, pp. 116–118, 2007.

[161] H. Philippe, Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc, "Heterotachy and long-branch attraction in phylogenetics," *BMC Evolutionary Biology*, vol. 5, article 50, pp. 1–8, 2005.

[162] M. Schöniger and A. Von Haeseler, "A stochastic model for the evolution of autocorrelated DNA sequences," *Molecular Phylogenetics and Evolution*, vol. 3, no. 3, pp. 240–247, 1994.

[163] S. V. Muse and B. S. Gaut, "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome," *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 715–724, 1994.

[164] N. Goldman and Z. Yang, "A codon-based model of nucleotide substitution for protein-coding DNA sequences," *Molecular Biology and Evolution*, vol. 11, no. 5, pp. 725–736, 1994.

[165] A. Siepel and D. Haussler, "Phylogenetic estimation of context-dependent substitution rates by maximum likelihood," *Molecular Biology and Evolution*, vol. 21, no. 3, pp. 468–488, 2004.

[166] D. G. Hwang and P. Green, "Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 39, pp. 13994–14001, 2004.

[167] O. F. Christensen, A. Hobolth, and J. L. Jensen, "Pseudo-likelihood analysis of codon substitution models with neighbor-dependent rates," *Journal of Computational Biology*, vol. 12, no. 9, pp. 1166–1182, 2005.

[168] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne, "Protein evolution with dependence among codons due to tertiary structure," *Molecular Biology and Evolution*, vol. 20, no. 10, pp. 1692–1704, 2003.

[169] N. Rodrigue, N. Lartillot, D. Bryant, and H. Philippe, "Site interdependence attributed to tertiary structure in amino acid sequence evolution," *Gene*, vol. 347, no. 2, pp. 207–217, 2005.

[170] N. Rodrigue, H. Philippe, and N. Lartillot, "Assessing site-interdependent phylogenetic models of sequence evolution," *Molecular Biology and Evolution*, vol. 23, no. 9, pp. 1762–1775, 2006.

[171] C. L. Kleinman, N. Rodrigue, C. Bonnard, H. Philippe, and N. Lartillot, "A maximum likelihood framework for protein design," *BMC Bioinformatics*, vol. 7, article 326, pp. 1–17, 2006.

[172] A. Sato, H. Tichy, C. O'Huigin, P. R. Grant B, R. Grant, and J. Klein, "On the origin of Darwin's finches," *Molecular Biology and Evolution*, vol. 18, no. 3, pp. 299–311, 2001.

[173] W. Salzburger, T. Mack, E. Verheyen, and A. Meyer, "Out of Tanganyika: genesis, explosive speciation, key-innovations and phylogeography of the haplochromine cichlid fishes," *BMC Evolutionary Biology*, vol. 5, article 17, pp. 1–15, 2005.

[174] A. L. Hughes, "Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level," *Heredity*, vol. 99, no. 4, pp. 364–373, 2007.

[175] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, New York, NY, USA, 1983.

[176] M. Lynch, *The Origins of Genome Architecture*, Sinauer Associates, Sunderland, Mass, USA, 2007.

[177] R. Nielsen, "Statistical tests of selective neutrality in the age of genomics," *Heredity*, vol. 86, no. 6, pp. 641–647, 2001.

[178] S. Aris-Brosou and L. Excoffier, "The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism," *Molecular Biology and Evolution*, vol. 13, no. 3, pp. 494–504, 1996.

[179] C. D. Bustamante, J. Wakeley, S. Sawyer, and D. L. Hartl, "Directional selection and the site-frequency spectrum," *Genetics*, vol. 159, no. 4, pp. 1779–1788, 2001.

[180] L. Zhu and C. D. Bustamante, "A composite-likelihood approach for detecting directional selection from DNA sequence data," *Genetics*, vol. 170, no. 3, pp. 1411–1421, 2005.

[181] M. Bamshad and S. P. Wooding, "Signatures of natural selection in the human genome," *Nature Reviews Genetics*, vol. 4, no. 2, pp. 99–111, 2003.

[182] M. Anisimova and D. A. Liberles, "The quest for natural selection in the age of comparative genomics," *Heredity*, vol. 99, no. 6, pp. 567–579, 2007.

[183] M. Nei and T. Gojobori, "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions," *Molecular Biology and Evolution*, vol. 3, no. 5, pp. 418–426, 1986.

[184] W. S. W. Wong and R. Nielsen, "Detecting selection in noncoding regions of nucleotide sequences," *Genetics*, vol. 167, no. 2, pp. 949–958, 2004.

[185] S. McCauley, S. de Groot, T. Mailund, and J. Hein, "Annotation of selection strengths in viral genomes," *Bioinformatics*, vol. 23, no. 22, pp. 2978–2986, 2007.

[186] Z. Yang, "Adaptive molecular evolution," in *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop, and C. Cannings, Eds., pp. 229–254, John Wiley & Sons, New York, NY, USA, 2nd edition, 2003.

[187] Z. Yang and R. Nielsen, "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models," *Molecular Biology and Evolution*, vol. 17, no. 1, pp. 32–43, 2000.

[188] W. S. W. Wong, Z. Yang, N. Goldman, and R. Nielsen, "Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites," *Genetics*, vol. 168, no. 2, pp. 1041–1051, 2004.

[189] J. Zhang, R. Nielsen, and Z. Yang, "Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level," *Molecular Biology and Evolution*, vol. 22, no. 12, pp. 2472–2479, 2005.

[190] J. Zhang, S. Kumar, and M. Nei, "Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes," *Molecular Biology and Evolution*, vol. 14, no. 12, pp. 1335–1338, 1997.

[191] Z. Yang, "Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution," *Molecular Biology and Evolution*, vol. 15, no. 5, pp. 568–573, 1998.

[192] R. Nielsen and Z. Yang, "Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene," *Genetics*, vol. 148, no. 3, pp. 929–936, 1998.

[193] Y. Suzuki and T. Gojobori, "A method for detecting positive selection at single amino acid sites," *Molecular Biology and Evolution*, vol. 16, no. 10, pp. 1315–1328, 1999.

[194] Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedersen, "Codon-substitution models for heterogeneous selection pressure at amino acid sites," *Genetics*, vol. 155, no. 1, pp. 431–449, 2000.

[195] T. Massingham and N. Goldman, "Detecting amino acid sites under positive selection and purifying selection," *Genetics*, vol. 169, no. 3, pp. 1753–1762, 2005.

[196] S. L. Kosakovsky Pond and S. D. W. Frost, "Not so different after all: a comparison of methods for detecting amino

acid sites under selection," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1208–1222, 2005.

[197] Z. Yang and R. Nielsen, "Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages," *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 908–917, 2002.

[198] M. Anisimova and Z. Yang, "Molecular evolution of the hepatitis delta virus antigen gene: recombination or positive selection?" *Journal of Molecular Evolution*, vol. 59, no. 6, pp. 815–826, 2004.

[199] S. Aris-Brosou, "Determinants of adaptive evolution at the molecular level: the extended complexity hypothesis," *Molecular Biology and Evolution*, vol. 22, no. 2, pp. 200–209, 2005.

[200] M. Anisimova, J. P. Bielawski, and Z. Yang, "Accuracy and power of Bayes prediction of amino acid sites under positive selection," *Molecular Biology and Evolution*, vol. 19, no. 6, pp. 950–958, 2002.

[201] Z. Yang, W. S. W. Wong, and R. Nielsen, "Bayes empirical Bayes inference of amino acid sites under positive selection," *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 1107–1118, 2005.

[202] J. P. Huelsenbeck and K. A. Dyer, "Bayesian estimation of positively selected sites," *Journal of Molecular Evolution*, vol. 58, no. 6, pp. 661–672, 2004.

[203] S. Aris-Brosou, "Identifying sites under positive selection with uncertain parameter estimates," *Genome*, vol. 49, no. 7, pp. 767–776, 2006.

[204] M. Anisimova, R. Nielsen, and Z. Yang, "Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites," *Genetics*, vol. 164, no. 3, pp. 1229–1236, 2003.

[205] M. Anisimova, J. Bielawski, K. Dunn, and Z. Yang, "Phylogenomic analysis of natural selection pressure in *Streptococcus* genomes," *BMC Evolutionary Biology*, vol. 7, article 154, pp. 1–13, 2007.

[206] E. Zuckerkandl and L. Pauling, "Molecules as documents of evolutionary history," *Journal of Theoretical Biology*, vol. 8, no. 2, pp. 357–366, 1965.

[207] E. Zuckerkandl and L. Pauling, "Evolutionary divergence and convergence in proteins," in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds., Academic Press, New York, NY, USA, 1965.

[208] L. Bromham and D. Penny, "The modern molecular clock," *Nature Reviews Genetics*, vol. 4, no. 3, pp. 216–224, 2003.

[209] S. Aris-Brosou, "Dating phylogenies with hybrid local molecular clocks," *PLoS ONE*, vol. 2, no. 9, p. e879, 2007.

[210] H. Kishino and M. Hasegawa, "Converting distance to time: application to human evolution," *Methods in Enzymology*, vol. 183, pp. 550–570, 1990.

[211] A. Rambaut and L. Bromham, "Estimating divergence dates from molecular sequences," *Molecular Biology and Evolution*, vol. 15, no. 4, pp. 442–448, 1998.

[212] A. D. Yoder and Z. Yang, "Estimation of primate speciation dates using local molecular clocks," *Molecular Biology and Evolution*, vol. 17, no. 7, pp. 1081–1090, 2000.

[213] Z. Yang and A. D. Yoder, "Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse Lemur species," *Systematic Biology*, vol. 52, no. 5, pp. 705–716, 2003.

[214] Z. Yang, "A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times," *Acta Zoologica Sinica*, vol. 50, pp. 645–656, 2004.

[215] M. J. Sanderson, "Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach," *Molecular Biology and Evolution*, vol. 19, no. 1, pp. 101–109, 2002.

[216] A. B. Smith, D. Pisani, J. A. Mackenzie-Dodds, B. Stockley, B. L. Webster, and D. T. J. Littlewood, "Testing the molecular clock: molecular and paleontological estimates of divergence times in the Echinoidea (Echinodermata)," *Molecular Biology and Evolution*, vol. 23, no. 10, pp. 1832–1851, 2006.

[217] M. J. Sanderson, "A nonparametric approach to estimating divergence times in the absence of rate constancy," *Molecular Biology and Evolution*, vol. 14, no. 12, pp. 1218–1231, 1997.

[218] S. Aris-Brosou and Z. Yang, "Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny," *Systematic Biology*, vol. 51, no. 5, pp. 703–714, 2002.

[219] S. Aris-Brosou and Z. Yang, "Bayesian models of episodic evolution support a late Precambrian explosive diversification of the Metazoa," *Molecular Biology and Evolution*, vol. 20, no. 12, pp. 1947–1954, 2003.

[220] S. Y. Ho, M. J. Phillips, A. J. Drummond, and A. Cooper, "Accuracy of rate estimation using relaxed-clock models with a critical focus on the early metazoan radiation," *Molecular Biology and Evolution*, vol. 22, no. 5, pp. 1355–1363, 2005.

[221] J. J. Welch, E. Fontanillas, and L. Bromham, "Molecular dates for the "cambrian explosion": the influence of prior assumptions," *Systematic Biology*, vol. 54, no. 4, pp. 672–678, 2005.

[222] M. Aitkin, "Posterior Bayes factors," *Journal of the Royal Statistical Society B*, vol. 53, no. 1, pp. 111–142, 1991.

[223] A. J. Drummond, S. Y. Ho, M. J. Phillips, and A. Rambaut, "Relaxed phylogenetics and dating with confidence," *PLoS Biology*, vol. 4, no. 5, p. e88, 2006.

[224] J. P. Huelsenbeck, J. P. Bollback, and A. M. Levine, "Inferring the root of a phylogenetic tree," *Systematic Biology*, vol. 51, no. 1, pp. 32–43, 2002.

[225] J. Shendure, R. D. Mitra, C. Varma, and G. M. Church, "Advanced sequencing technologies: methods and goals," *Nature Reviews Genetics*, vol. 5, no. 5, pp. 335–344, 2004.

[226] M. J. Moore, A. Dhingra, P. S. Soltis, et al., "Rapid and accurate pyrosequencing of angiosperm plastid genomes," *BMC Plant Biology*, vol. 6, article 17, pp. 1–13, 2006.

[227] P. Green, "2x genomes—Does depth matter?" *Genome Research*, vol. 17, no. 11, pp. 1547–1549, 2007.

[228] A. Rokas, B. L. Williams, N. King, and S. B. Carroll, "Genome-scale approaches to resolving incongruence in molecular phylogenies," *Nature*, vol. 425, no. 6960, pp. 798–804, 2003.

[229] A. G. Clark, M. B. Eisen, D. R. Smith, et al., "Evolution of genes and genomes on the *Drosophila* phylogeny," *Nature*, vol. 450, no. 7167, pp. 203–218, 2007.

[230] F. Delsuc, H. Brinkmann, and H. Philippe, "Phylogenomics and the reconstruction of the tree of life," *Nature Reviews Genetics*, vol. 6, no. 5, pp. 361–375, 2005.

[231] F. Ge, L. S. Wang, and J. Kim, "The cobweb of life revealed by genome-scale estimates of horizontal gene transfer," *PLoS Biology*, vol. 3, no. 10, p. e316, 2005.

[232] R. D. M. Page, "Extracting species trees from complex gene trees: reconciled trees and vertebrate phylogeny," *Molecular Phylogenetics and Evolution*, vol. 14, no. 1, pp. 89–106, 2000.

[233] M. J. Phillips, F. Delsuc, and D. Penny, "Genome-scale phylogeny and the detection of systematic biases," *Molecular Biology and Evolution*, vol. 21, no. 7, pp. 1455–1458, 2004.

[234] H. Nishihara, N. Okada, and M. Hasegawa, "Rooting the eutherian tree: the power and pitfalls of phylogenomics," *Genome Biology*, vol. 8, no. 9, p. R199, 2007.

[235] N. Rodríguez-Ezpeleta, H. Brinkmann, B. Roure, N. Lartillot, B. F. Lang, and H. Philippe, "Detecting and overcoming systematic errors in genome-scale phylogenies," *Systematic Biology*, vol. 56, no. 3, pp. 389–399, 2007.

[236] S. B. Hedges, J. Dudley, and S. Kumar, "TimeTree: a public knowledge-base of divergence times among organisms," *Bioinformatics*, vol. 22, no. 23, pp. 2971–2972, 2006.

[237] J. E. Janečka, W. Miller, T. H. Pringle, et al., "Molecular and genomic data identify the closest living relative of primates," *Science*, vol. 318, no. 5851, pp. 792–794, 2007.

[238] S. Kumar and J. Dudley, "Bioinformatics software for biologists in the genomics era," *Bioinformatics*, vol. 23, no. 14, pp. 1713–1717, 2007.