BIOINFORMATICS RESOURCES FOR *MEDICAGO TRUNCATULA* – 2006

Contributed by:

Klaus Mayer: **UrMeLDB: The European Medicago genome database**
with Manuel Spannagl, Heiko Schoof

Jérôme Gouzy: ***Medicagoo: Searching Medicago resources***
with Thomas Faraut, Thomas Schiex, Philippe Bardou, Céline Noirot, Yoann Beausse, Sébastien Carrere, Emmanuel Courcelle, Marie-Josée Cros, Sylvain Foissac, Annick Moisan, Christine Gaspin

Helge Küster: ***EST analysis and expression profiling software from Bielefeld University***
with Thomas Bekel, Kolja Henckel, Michael Dondrup, Alexander Goesmann

Gregory D. May: ***The Legume Information System (LIS): An integrated, dynamic comparative legume information resource***
with Michael Gonzales, Andrew Farmer, Kamal Gajendran, William D. Beavis

Ernie Retzel: ***MtDB and medicago.org***
with John Crow, James E. Johnson, Timothy Kunau

Nevin D. Young: **The *Medicago truncatula* Genome Sequencing Website**
with Steven Cannon, Shelley Wang, Jay Vasdewani, Ethalinda Cannon, Ben Chacko, Joann Mudge, Atif Ahmed, Min Wang, Roxanne Denny, Bing-Bing Wang

Table of contents:

1. Introduction

The *Medicago* genomics community is extremely fortunate to have such a large and growing body of bioinformatics tools available. For the most part, these resources are represented as websites of major *Medicago* and legume research groups. They consist of database and software tools for organizing, visualizing, and manipulating *Medicago* genomics information. Enabling fundamental *Medicago* genomics research and extending that knowledge to crop legumes are the principle goals of these efforts. There is little doubt that the many powerful genomics resources developed for *Medicago* over the past few years – genome sequence, EST libraries, microarrays and DNA chips, growing bodies of proteomic and metabolomic results — all require extensive, logical, and easy to use tools for bioinformatics.

This brief chapter reviews six of the most exciting *Medicago* websites available today (Summer 2006). Although non-exhaustive, the chapter does give a flavor of the many database resources in place for *Medicago* (and legume) research, including examples of analysis, query, and visualization tools that can be applied to *Medicago* and comparative legume genomics data.

With the *Medicago* genome sequence underway (a draft expected in late 2006 and the complete euchromatin sequence two years later), informatic tools to monitor and coordinate the sequencing effort, to "walk" along the minimum tiling path and sequence assembly, and intelligently annotate the underlying genes are essential. For EST libraries, query tools that enable "electronic" Northerns are frequently the focus, and for microarrays and DNA chips, databases that track array features and integrate that information with the massive image files generated during hybridization experiments are of principle importance. Comparative genomics tools are also required in order for scientists to expand beyond *Medicago* and utilize genomics resources in the context of other legume species and their own EST libraries, genetic and physical maps, and (eventually) genome sequences.

It is essential for anyone venturing into the bioinformatic analysis of *Medicago* genomics data to remember that as extensive as the resources may appear today, both genomic and bioinformatic tools are highly dynamic and rapidly evolving. Millions of new base pairs in genome sequence data are generated each month, functional roles for a growing number of genes are being described, and knowledge about proteomics and metabolomics is expanding rapidly. At the same time, strategies to integrate informatics resources more effectively among multiple sites are evolving. In this chapter, for example, the growing role of shared web services, including BioMoby models for sharing data and tools among websites, plays a prominent role. In some cases, development of new models for informatics analysis and resource sharing started in *Medicago*. Clearly, things look very promising for research in *Medicago* today, in large part, due to the impressive body of informatics tools that are now available.

2. UrMeLDB: The European Medicago Genome Database

Klaus Mayer, Manuel Spannagl, Heiko Schoof
MIPS/IBI Inst. for Bioinformatics, GSF Research Center for Environment and Health,
Ingolstädter Landstr. 1, 85764 Neuherberg; Germany

The European *Medicago truncatula* and legumes database, UrMeLDB, (http://www.urmeldb.net) serves as a comprehensive genome information resource for *Medicago truncatula* and legumes.

UrMeLDB integrates data produced within the European Grain Legumes Integrated Project (GLIP) as well as publicly available *Medicago* genome sequences from the international *Medicago* sequencing initiative. Besides regularly updated *Medicago* genome sequence information, UrMeLDB provides access to associated analysis data, generated both by project partners as well as via the integration of other external annotation and information resources. The UrMeLDB graphical web interface supports various search, download and data analysis options and enables browing of database contents. Its flexible architecture is designed to facilitate integration of new data sources and displays.

Data Content and Sources

UrMeLDB integrates data from both the international *Medicago* sequencing project and the European *Medicago* project. All publicly available *Medicago* genomic clone data, completed as well as unfinished, are continiously integrated into UrMeLDB. Annotation of the genomic sequences involves gene prediction, detection of noncoding genetic elements like transposons, repeats and RNA genes. Gene prediction is undertaken in tight collaboration with the International *Medicago* Genome Annotation Group (IMGAG, http://www.*Medicago*.org/genome/IMGAG.php) (Town, 2006). UrMeLDB aims to provide comprehensive, state-of-the-art analysis of *Medicago* and legume genomic sequences sequences combined with tools that have been tested and selected for legume-centric analysis. A special emphasis is put on providing supporting evidence for all features depicted.

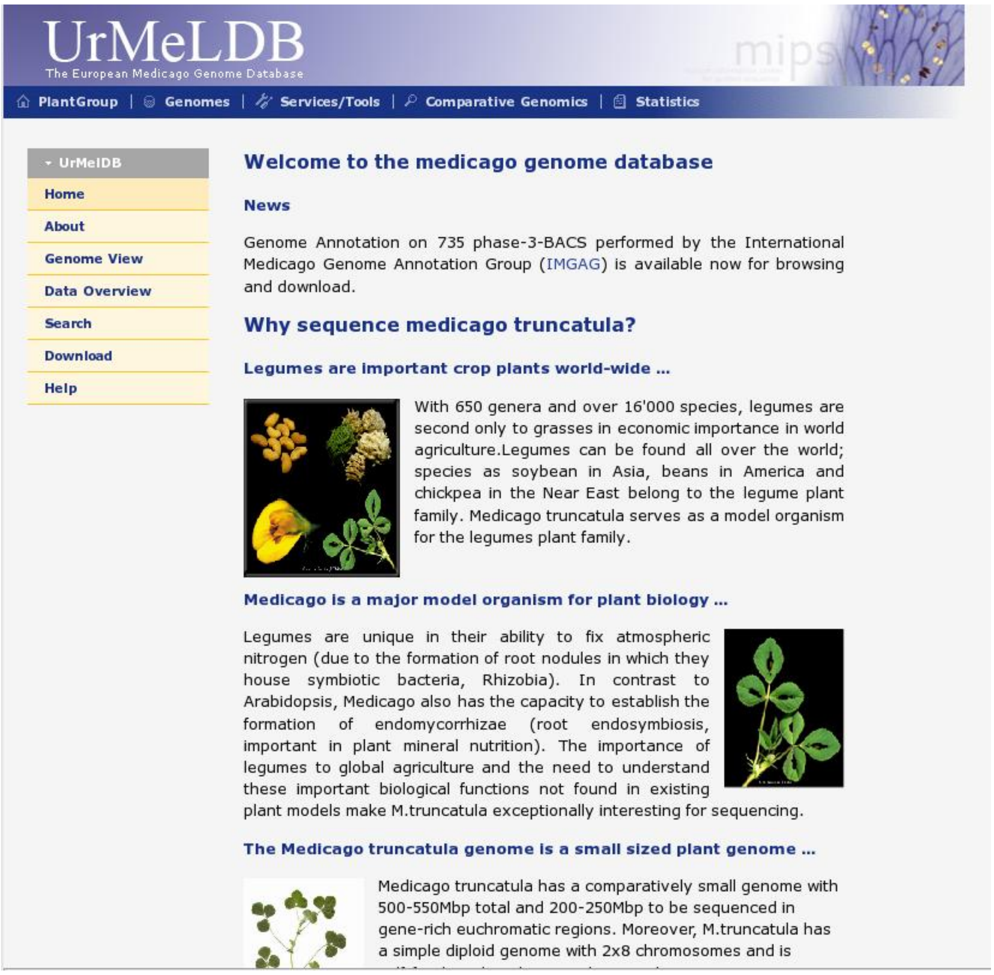
Access and Web Query Interface

The UrMeLDB web interface provides three different routes to access the data: browse, search and download. For genome-oriented visualization and browsing of genetic elements, a graphical interface, Gbrowse[1], has been integrated (Stein *et al*, 2002). GBrowse displays all genetic elements in a given region and allows zoom-in views as well as full-text search of annotations and third party annotation. An important tool for comparative genomics is the prediction of orthologs between genomes. Within protein coding gene reports, possible orthologs in other plant species (as well as in fungi, *etc*) can be extracted from the SIMAP[2] database, a matrix containing precomputed homologies of protein sequences. The UrMeLDB download section (ftp://ftpmips.gsf.de/plants/*Medicago*/) provides ftp access to various data downloads. This includes FASTA-formatted DNA sequence collections for all currently available clone sequences as well as continuously updated data dumps of all current annotated gene and protein sequences within the *Medicago* genome. In addition we developed a sequence export tool that enables downloading of customized, specific sequence datasets, such as all first introns of all protein-coding genes on selected contigs. Data retrieval formats are either multiple FASTA or XML.

[1] http://www.gmod.org/
[2] Similarity Matrix of Proteins, http://mips.gsf.de/proj/simap/

Furthermore, the UrMeLDB download section provides functionality to create and download a GAME-XML file for a specified contig and/or coordinate range. The GAME-XML format is used by the Apollo Genome Browser[3], a detailed graphical viewer for genome data with more flexible interaction possibilities than a browser-based display.

Another route to access UrMeLDB data that is especially useful for machine access, is directed towards the usage of BioMoby web services (Wilkinson & Links, 2002). Through BioMoby web services, data from the plant genome resources at MIPS can be directly accessed from applications, allowing analyses to be performed remotely or included in a web presentation without the need for a local copy. At the time of writing (July 2006) 17 web services are available for accessing *Medicago* data. They can be accessed via the gbrowse_moby client (mips.gsf.de/cgi-bin/proj/planet/gbrowse/gbrowse_moby).


UrMeLDB
The European Medicago Genome Database
mips
PlantGroup | Genomes | Services/Tools | Comparative Genomics | Statistics
UrMelDB
Home
About
Genome View
Data Overview
Search
Download
Help

Welcome to the medicago genome database

News

Genome Annotation on 735 phase-3-BACS performed by the International Medicago Genome Annotation Group (IMGAG) is available now for browsing and download.

Why sequence medicago truncatula?

Legumes are important crop plants world-wide ...

With 650 genera and over 16'000 species, legumes are second only to grasses in economic importance in world agriculture.Legumes can be found all over the world; species as soybean in Asia, beans in America and chickpea in the Near East belong to the legume plant family. Medicago truncatula serves as a model organism for the legumes plant family.

Medicago is a major model organism for plant biology ...

Legumes are unique in their ability to fix atmospheric nitrogen (due to the formation of root nodules in which they house symbiotic bacteria, Rhizobia). In contrast to Arabidopsis, Medicago also has the capacity to establish the formation of endomycorrhizae (root endosymbiosis, important in plant mineral nutrition). The importance of legumes to global agriculture and the need to understand these important biological functions not found in existing plant models make M.truncatula exceptionally interesting for sequencing.

The Medicago truncatula genome is a small sized plant genome ...

Medicago truncatula has a comparatively small genome with 500-550Mbp total and 200-250Mbp to be sequenced in gene-rich euchromatic regions. Moreover, M.truncatula has a simple diploid genome with 2x8 chromosomes and is

Figure 1.
UrMeLDB: The European Medicago Genome Database

[3] http://www.fruitfly.org/annot/apollo/

3. Medicagoo: Searching Medicago resources
http://medicagoo.toulouse.inra.fr/

Jérôme Gouzy[1], Thomas Faraut[2], Thomas Schiex[3]
with contributions from Philippe Bardou[1,2], Céline Noirot[1,3], Yoann Beausse[2], Sébastien Carrere[1], Emmanuel Courcelle[1], Marie-Josée Cros[3], Sylvain Foissac[3], Annick Moisan[3], Christine Gaspin[3]

1. Laboratoire des Interactions Plantes Microorganismes, INRA/CNRS
2. Laboratoire de Génétique Cellulaire, INRA
3. Unité de Biométrie et d'Intelligence Artificielle, INRA, BP 52627, 31326 Castanet Tolosan Cedex, France

The **Medicagoo portal** (*Medicago*.toulouse.inra.fr/) offers an integrated access to *M. truncatula* tools and databases, including databases and analysis tools, locally designed for legumes. The site provides life science researchers with a single web interface facilitating the analysis and exploitation of *Medicago* resources.

Medicago truncatula genomics

Protein-coding genes annotation*:* the site provides a direct access to the annotations for both *M. truncatula* and *Lotus japonicus* BACs. These annotations are produced by the integrated annotation pipeline based on the **EuGène** gene-finder and used by the International *Medicago* Genome Annotation Group (IMGAG). Users can also submit any DNA sequence for an immediate integrated analysis of their sequence.

RNA gene annotation*:* the **LeARN** resource integrates tools and interfaces covering the three layers of the ncRNA annotation process (i) a detection and clustering pipeline (ii) a web interface allowing for interactive annotation of the RNA molecules (iii) a XML database for storing, analyzing and representing the sequences and their secondary structures. This database is queryable either from a web server or via BioMoby Web-services (Wilkinson & Links, 2002).

Comparative genomics

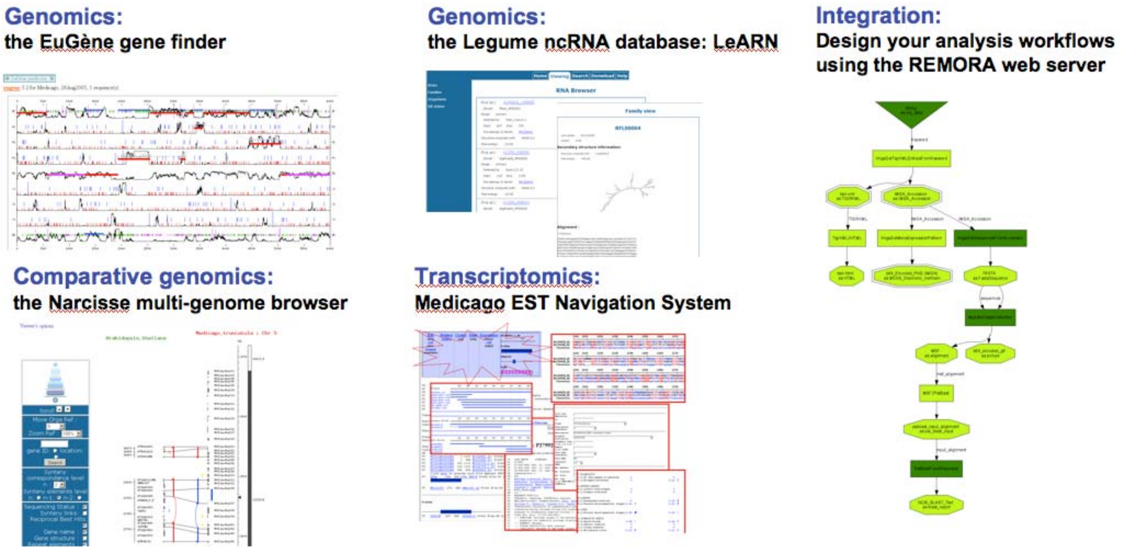
*Medicago*o provides a comparative browser, called **Narcisse**, designed to handle multi-organisms sequence-based maps from whole genomes. In addition to its intrinsic flexibility and capacity to handle very large datasets Narcisse allows for interactive and multi scale analysis of sequence conservations. Starting from "dotplots" giving an overview of the conservation between genomes, it is possible to simply focus on a region to visualize syntenic conservations at the nucleic acid and protein levels. The current list of species integrated in the plant version of Narcisse includes *Arabidopsis thaliana,* rice and *Medicago truncatula*.

Transcriptomics

Analyzing the transcriptome of *M. truncatula* is facilitated by the *Medicago* EST Navigation System (**MENS**), which integrates a database of EST clusters and integrated analysis tools for EST data (Journet *et al*, 2002). MENS provides a variety of pre-computed analyses and tools to quickly and thoroughly explore gene function, gene expression and gene families. Most of these analyses rely on the prediction of transcript-encoded proteins for 28,657 EST clusters (among 37,020 overall). Manual annotations are provided for 6,096 of these clusters. For the remaining ones, a reliable automatic description based on the search of functional or structural motifs (InterPro) is provided.

Analysis workflows

Emerging web-services technology allows interoperability between multiple distributed architectures. *Medicago*o relies on the services of **REMORA** (Carrere & Gouzy, 2006), a web server implemented according to the BioMoby web-service specifications, providing life science researchers with an easy-to-use workflow generator and launcher, a repository of predefined workflows and a survey system. The *Medicago*o portal uses REMORA workflows and underlying BioMoby web-services in order to provide access to the IMGAG annotation.



Genomics: the EuGène gene finder

Genomics: the Legume ncRNA database: LeARN

Integration: Design your analysis workflows using the REMORA web server

Comparative genomics: the Narcisse multi-genome browser

Transcriptomics: Medicago EST Navigation System

Figure 2.
Medicagoo: Searching *Medicago* resources at http://medicagoo.toulouse.inra.fr

4. EST Analysis and Expression Profiling Software from Bielefeld University

Thomas Bekel[1], Kolja Henckel[1], Michael Dondrup[1], Alexander Goesmann[1], Helge Küster[2]

1. Bioinformatics Resource Facility, Center for Biotechnology (CeBiTec), Bielefeld University, D-33594 Bielefeld, Germany
2. Institute for Genome Research, Center for Biotechnology (CeBiTec), Bielefeld University, D 33594 Bielefeld, Germany

The Bioinformatics Resource Facility (Center for Biotechnology, Bielefeld University) has been actively involved in the generation of EST-clustering and analysis along with expression profiling software in the frame of the European Union *MEDICAGO* (http://*Medicago*.toulouse.inra.fr/EU) and GRAIN LEGUMES (http://www.eugrainlegumes.org) projects. All software tools developed in the course of these projects can be accessed via the link http://www.cebitec.uni-bielefeld.de/groups/brf/software and can be launched using standard web browsers. The software tools are connected by a BRIDGE layer that allows cross-referencing of EST and expression profile data (Goesmann *et al*, 2003). Recently, the various tools described here were applied to specify the genetic program activated in arbuscular mycorrhizal roots of *Medicago truncatula* (Hohnjec *et al*, 2006).

SAMS: Sequence Analysis and Management System

The Sequence Analysis and Management System (SAMS) was implemented on the basis of the GenDB genome annotation system (Meyer *et al*., 2003). In SAMS, cDNA libraries, cDNA sequences, bioinformatics tools, tool results, and annotations are modelled in an object-oriented approach. The software was implemented in Perl and a relational MySQL database was chosen as storage component.

The EST analysis pipeline implemented in SAMS provides a web-based sequence import, and user-defined quality values can be set as well as parameters that govern vector-clipping. Before EST sequences are annotated in SAMS, a clustering step is performed that assembles EST reads into TCs (Tentative Consensus sequences). Although SAMS provides user-configurable parameters to adjust the outcome of the clustering process, the standard TIGR (http://www.tigr.org/) parameter set is applied by default.

EST and TC sequences are automatically processed and annotated in SAMS. For annotation, the Metanor automatic function prediction uses a combination of standard tools such as BLAST-based sequence comparisons, Hidden Markov model scans and InterPro-based searches for functional domains. These tools are run automatically on each sequence, leading to a consistent annotation assigning gene names, gene products, descriptions, functional categories, and gene ontology numbers to the TC and EST sequences.

SteN: Statistical electronic Northern Blot

The web-based SteN (Statistical electronic Northern) interface integrated into the SAMS software enables *in silico* predictions of differential gene expression. An easy setup of queries is possible by selecting one of the three states "USE", "DON'T USE" and "IGNORE" for each library in an EST project. "USE" means that a TC must have at least one EST from this library, "DON'T USE" specifies that a TC is not allowed to have an EST from this library, and "IGNORE" means that the library is not considered in the query. Additionally, the logic operators "AND" or "OR" can be used to combine the selected libraries. The result of the query is a list with all TCs matching the selected search criteria. In addition, an R-value (Stekel *et al*, 2000) is calculated for each TC to provide a statistical assessment for the reliability of differential gene expression.

ArrayLIMS and EMMA

Integration of expression data can only be achieved via software tools fulfilling two requirements: (1) the possibility to upload raw data from microarray hybridizations together with a description of experimental conditions into a central data repository and (2) the possibility to combine these data sets in order to relate expression profiles obtained in different experiments. Two databases were set up to fulfill these requirements: an ArrayLIMS (Laboratory Information Management System) for the storage of data on experimental conditions and for storing raw data, and the EMMA (EST Meets MicroArrays) software for data evaluation and data mining.
The ArrayLIMS system requires the user to deliver, in a step-by-step process, all information that is required by others to replicate a microarray experiment. This approach is essential in order to comply to the MIAME (Minimal Information ON A Microarray Experiment; Brazma *et al*, 2001) standard. In ArrayLIMS, the basic information to be entered relates to the biological conditions profiled, the material harvested and the exact growth conditions used. Based on these entries, subsequent steps call for relevant features of the targets isolated from the biological material profiled, the target labeling, and the microarray hybridizations performed. In ArrayLIMS, identifiers for the microarrays to be used in a project are stored and any information entered can be linked to those identifiers. Finally, microarray identifiers are linked to experimental raw data: tiff-files from microarray scanners and data-files derived from image processing. Here, the ArrayLIMS system is compatible with commercial microarray scanners and with the output formats of major commercial image processing software. As a consequence of multiple uploads from different project participants, a central database storing information on all hybridizations performed in the frame of the project is generated.

Subsequent to ArrayLIMS imports, microarray data can be used to set up experiments in EMMA (Dondrup *et al*, 2003), a web-based software designed to facilitate data analysis according to accepted standards for cDNA and 70mer oligonucleotide microarrays. The EMMA system represents an effort to build a microarray analysis software that simplifies complex analysis steps and queries of large datasets, while adhering to standards for data interchange and optimizing the inter-operation with other software. EMMA is written mainly in Perl with some portions programmed in R or Java. The data repository is implemented using O2DBI, an object-oriented code generator that significantly simplifies the creation of complex database applications.

The current release of EMMA includes different pipelines for data normalization, including global mean normalization, lowess (locally weighted scatter plot smooth) normalization, and different print-tip based normalizations. Normalized data can be used to run different significance tests to identify differentially expressed genes, providing one-sided and two-sided pipelines for either query. Higher order analyses are supported by hierarchical clustering pipelines that display results in Java-applets providing several visualisation options to browse clustered datasets. All transformed datasets (normalized data, significance tests, hierarchical clusterings) are stored in a project database and can be downloaded for further analyses. In addition, keyword-based Google-style queries allow the mining of expression data across multiple experimental conditions.

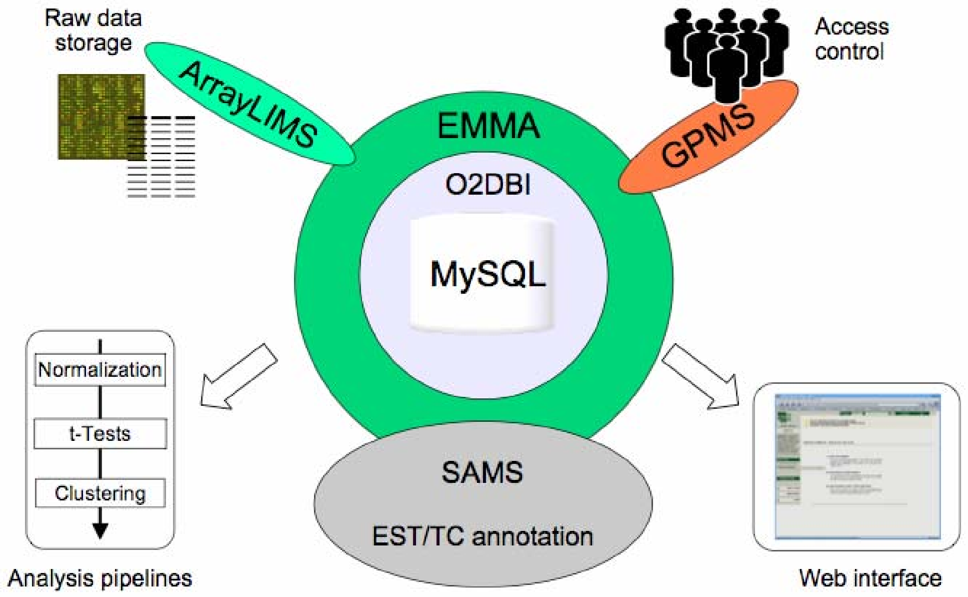


Figure 3:
Architecture of the ArrayLIMS/EMMA database. Raw data from microarray scanners and image processing software are stored in the ArrayLIMS database. From here, data can be imported into the EMMA software which builds on a MySQL database and an O2DBI object to database interface. EMMA is connected to the SAMS software for automated EST and TC annotation. Access to the EMMA databases is controlled by a General Password Management System (GPMS). EMMA provides different pipelines for data analysis. Results are displayed as web front ends, and can thus be queried and downloaded using standard web browsers.

5. The Legume Information System (LIS): An integrated, dynamic comparative legume information resource

Gregory D. May, Michael Gonzales, Andrew Farmer, Kamal Gajendran, William D. Beavis

National Center for Genomic Resources (NCGR), 2935 Rodeo Park Drive East,
Santa Fe, New Mexico 87505

Comparative genomics is the comparison and analysis of genomes of different species in order to gain a better understanding of how species have evolved and to determine gene function. Clade-oriented information resources such as the Legume Information System (LIS; Gonzales *et al*, 2005) offer data and applications enabling comparative genomics approaches that utilize bioinformatics to leverage genomic information from model and reference organisms for the benefit of legume researchers.

LIS (www.comparative-legumes.org/) is a publicly accessible legume resource that integrates molecular and genetic data from phylogenetically diverse legume species enabling cross-species *transcript*, *genomic* and *map* comparisons. The intent of the LIS is to help researchers leverage data-rich model plants to fill knowledge gaps across crop legume species and provide the ability to traverse between interrelated data types.

Transcript data: The LIS "virtual plant" user interface facilitates intuitive navigation of *M. truncatula*, *Lotus*, soybean and *Arabidopsis* EST and consensus transcript data. Currently, the sequence import functionality at LIS makes use of NCGR's computational pipeline, XGI, which uses a variety of algorithms for sequence pattern recognition, comparison and annotation (www.ncgr.org/xgi) and can handle genomic, EST or ORF sequence data types. Pipeline analyses include: BLASTX searches against NCBI non-redundant protein library; BLASTN and TBLASTX searches against related transcript libraries (Benson *et al*, 2004; Altschul *et al*, 1990; Altschul *et al*, 1997); BLIMPS (Henikoff *et al*, 1994) search against Blocks+ protein motif database (Henikoff *et al*, 2000); searches with the 12 InterProScan algorithms (Zdobnov 2001) against the InterPro database (Mulder *et al*, 2003); identification of signal peptides for extracellular secretion with PexFinder, an algorithm based on SignalP 2.0 (Bendtsen *et al*, 2004); and GenScan (Burge *et al*, 1997) for gene prediction in genomic sequences. The automated post-analysis annotation links BLAST and Blocks+ hits to their cognate Gene Ontology entries (Ashburner *et al*, 2000) and InterPro hits are automatically linked to GO annotations.

Genome data: The XGI genomic pipeline (XGIg) processes LIS genomic data. LIS does not assembly genomics data, but instead uploads data from GenBank as provided by the genome sequencing centers. The Comparative Functional Genomics Browser (CFGB) provides visualization of comparative genomics analysis results including alignment of transcript data within genomic contigs. CFGB also enables dynamic visualization of comparative alignments between genomic contigs through zooming, panning and sorting functions.

Map data: LIS incorporates a CMap-based viewer (www.gmod.org/cmap) that provides users with detailed sequence and annotation viewing through a custom sequence viewing module developed for LIS. CMap provides LIS users access to, where available, the genetic and physical maps of *Medicago*, soybean and *Phaseolus*. Currently, all SoyBase (http://soybase.agron.iastate.edu/) curated linkage map data have been uploaded and incorporated into CMap. For comparative analyses, a map is selected from the database for use as a reference map. Other maps can be subsequently added so that alignments relative to the selected reference map can be compared.

Future implementations of LIS will utilize semantic web services to traverse between data types, allowing users to follow phenotype through to genetic maps onto annotated genome sequence and linkage groups of other species.

The Legume Information System, a component of the Model Plant Initiative (MPI), is developed by the National Center for Genome Resources in cooperation with the USDA Agricultural Research Service. To acknowledge or reference LIS: The Legume Information System (LIS): an integrated information resource for comparative legume biology. Nucleic Acids Research, 2005, Vol. 33, Database issue D660-D665.

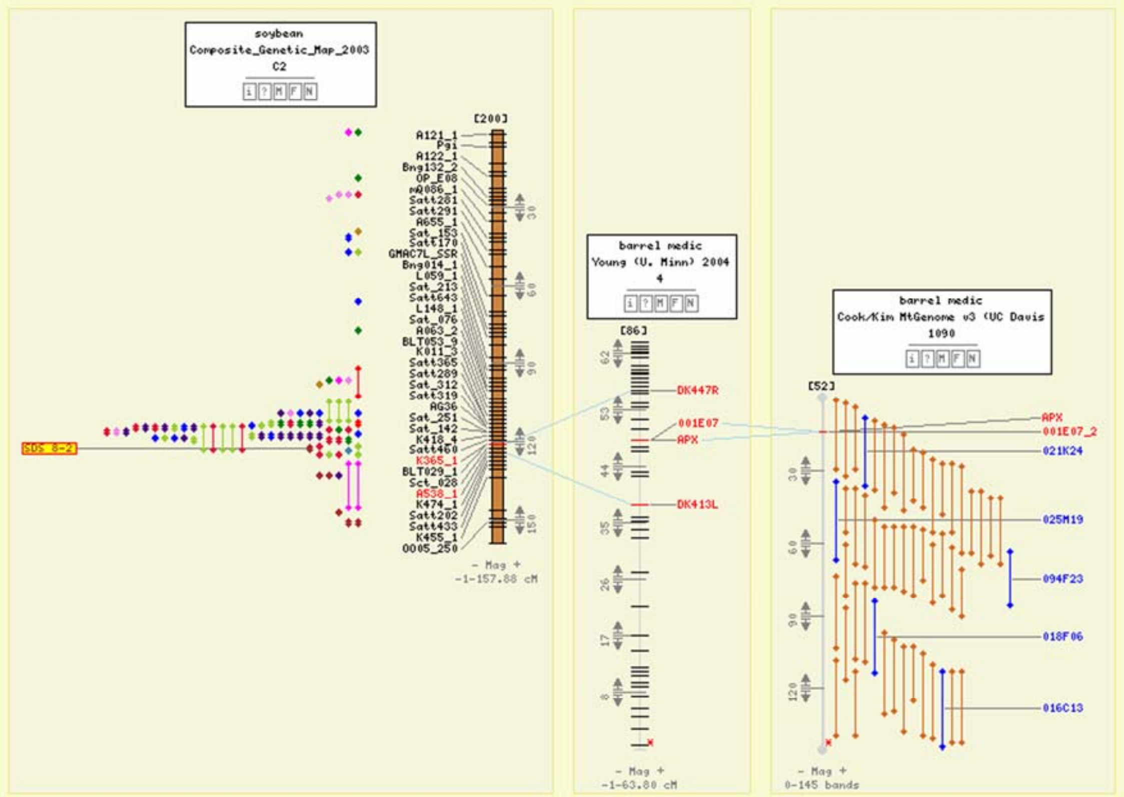
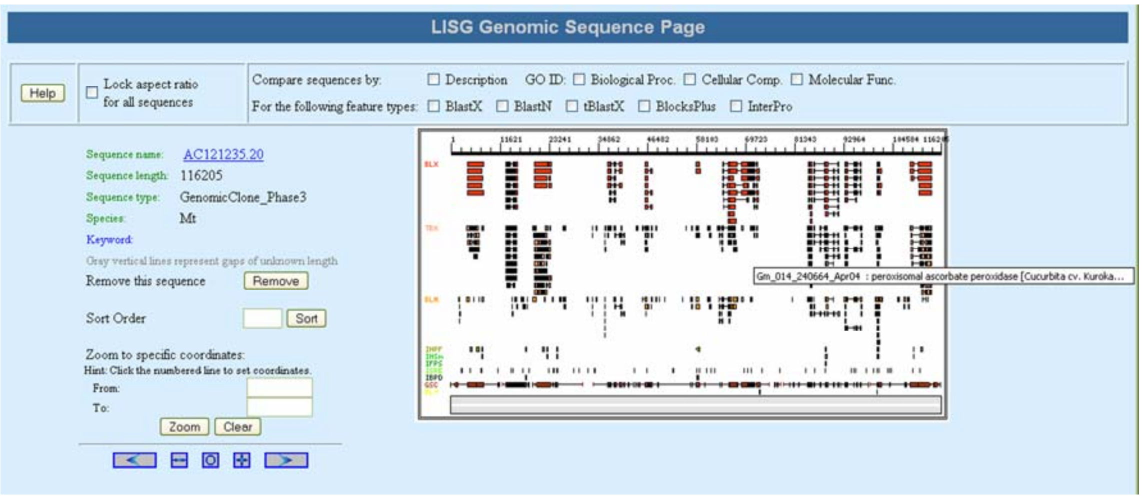


Figure 4a.
Comparative Functional Genomics Browser (see next page for legend)



LISG Genomic Sequence Page

□ Lock aspect ratio for all sequences

Compare sequences by: □ Description GO ID: □ Biological Proc. □ Cellular Comp. □ Molecular Func.

For the following feature types: □ BlastX □ BlastN □ tBlastX □ BlocksPlus □ InterPro

Sequence name: AC121235.20

Sequence length: 116205

Sequence type: GenomicClone_Phase3

Species: Mt

Keyword:

Gray vertical lines represent gaps of unknown length

Remove this sequence

Sort Order

Zoom to specific coordinates:

Hint: Click the numbered line to set coordinates.

From:

To:

Gm_014_240664_Apr04 : peroxisomal ascorbate peroxidase [Cucurbita cv. Kuroka...

Figure 4b.

Comparative Functional Genomics Browser. CFGB facilitates visualization of results of genomics analyses, including comparative transcript data, with gene-sequences aligned to genomic contigs. In this example, regions of the genome where syntenic relationships exist between *Medicago* and soybean, the annotated genomic sequence from *Medicago* can be leveraged to identify candidate genes using the LIS implementation of CMap.

6. *MtDB* and medicago.org

Ernest F. Retzel, John Crow, James E. Johnson, Timothy Kunau

Center for Computational Genomics and Bioinformatics, University of Minnesota▯, 420 Delaware St SE, Mayo MC43▯, Minneapolis, MN 55455

Medicago.org and *MtDB* are two related resources specializing in data from *Medicago* genomics projects. *Medicago*.org was designed to: 1) accommodate *Medicago* genome research projects; 2) present an interface to the data from the legume gene discovery projects, and 3) present a mechanism to distribute core genomic annotation data under the Distributed Annotation System [DAS] (Dowell *et al*, 2001) via an Ensembl database (Birney *et al*, 2003). Other than describing this DAS reference server, the genome browser project will not be discussed here. In addition to the University of Minnesota genome browser, there are multiple instantiations of automatically-annotated *Medicago* genome data both in the United States and in Europe, many of which are discussed elsewhere in this chapter. Beyond general pages and the extensive *medicago*.org/genome site, which documents the genome sequencing effort, an RNAi database is also maintained at *medicago*.org/rnai, displaying results of a *Medicago* root RNAi project.

MtDB (Lamblin *et al*) has defined its mission as providing a biologist-oriented interface to the gene discovery resources available for *Medicago* as well as for the other legume EST projects. In particular, we felt it was important to provide a flexible interface to this data from which biology-based users could easily refine their queries in order to address information gathered from other resources. The initial resource had a fairly flexible web-based interface, but required building custom web pages for each new species that was included. As the EST projects expanded beyond *Medicago*, soybean and *Lotus japonicus* to smaller projects, it became obvious that we would need a more flexible database, and a more extensible interface. The *Nimbus* data model [manuscript in preparation] was designed with an eye to maintainability, ease of use, and the ability to be extended to different species and different features.

MtDB resources. One of the features of *MtDB* is that it is built on nightly-updated data resources. These data sets include GenBank and its included Taxonomy Database (Benson *et al*, 2006; Wheeler *et al*, 2006), UniProt, TAIR peptides (Rhee *et al*, 2003), Pfam (Bateman *et al*, 2004), TigrFams (Haft *et al*, 2003), SMART protein families (Letuni *et al*, 2004; Ponting *et al*, 1999), Gene Ontology (Harris *et al*, 2004), and Kegg data (Kanehisa *et al*, 2004). The inclusion of these databases and their linkage to *MtDB* by similarity searches and HMMs helps create the underpinnings of some of the user-focused resources available to users of *MtDB*.

Interoperability. One of the new directions for *MtDB* and its related resources is the blossuming development of semantic BioMoby services [http://biomoby.org]. This work was begun as a collaboration of the Center for Computational Genomics and Bioinformatics [CCGB] with the Legume Information System (Gonzales *et al*, 2005), and is part of an effort to begin to make distributed data and compute resources inter-operable. The intent of these developments is to provide annotations and computational services to individuals and other annotation services in a web-services-accessible fashion. Remote queries can be fulfilled, operations performed and data returned in a structured format compatible with http protocols. In the future, our intent is that, while the web interface will still be maintained, the ability to integrate information on demand into larger efforts will be the primary utilization of annotation resources.

7. The *Medicago truncatula* Genome Sequencing Website: *medicago.org/genome*

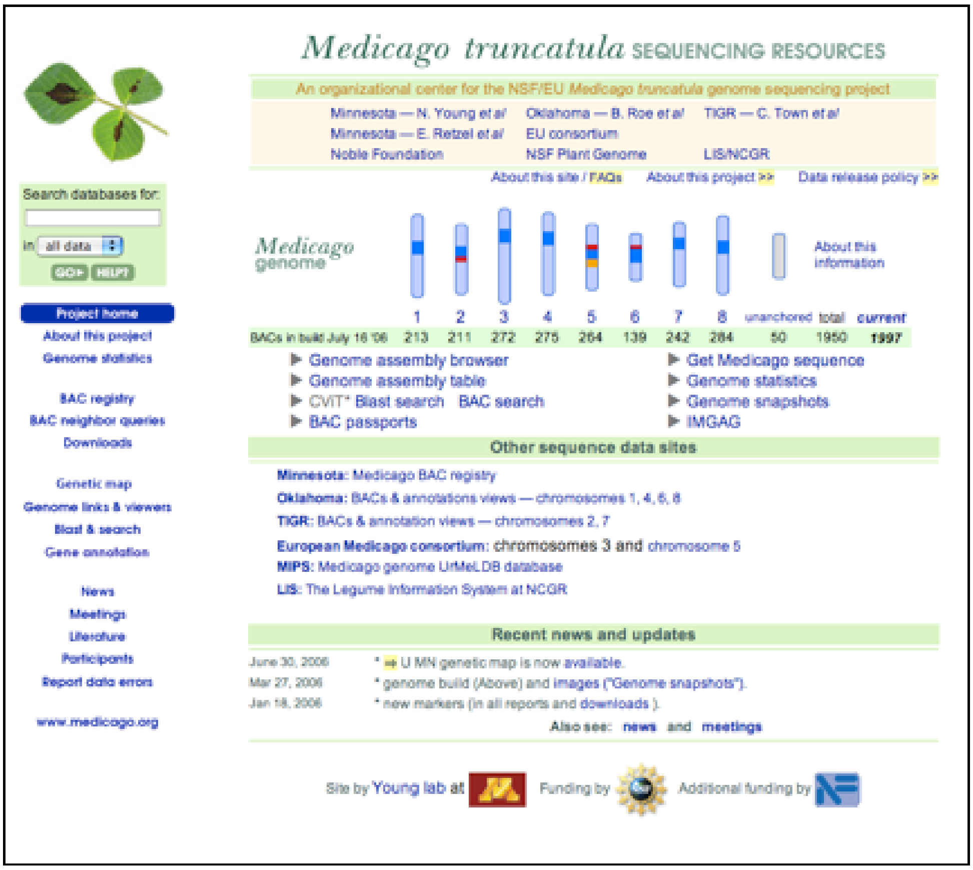
Steven Cannon, Shelley Wang, Jay Vasdewani, Ethalinda Cannon, Ben Chacko, Joann Mudge, Atif Ahmed, Min Wang, Roxanne Denny, Bing-Bing Wang, Nevin Young

Department of Plant Pathology, 495 Borlaug Hall, University of Minnesota, St. Paul, MN 55108

The *medicago.org/genome* website acts as an organizational center for the *Medicago truncatula* genome sequencing initiative. Originally created to assist sequencing centers engaged in the *Medicago* sequencing effort, the site now provides a wide array of tools for biologists interested in utilizing the *Medicago* genome sequence. The site is built upon a comprehensive MySQL relational database that contains nearly all genome sequence-related data generated in the course of the sequencing project. The site is useful to anyone working in the field of *Medicago* genomics and is especially helpful to scientists interested in positional gene cloning in *Medicago*. The site can be accessed directly or through *medicago.org* (E. Retzel *et al*, this chapter), with both sites located together at the University of Minnesota. As a resource for comparative genomics, *medicago.org/genome* also provides strategic links to the Legume Information System (LIS) at the National Center for Genomic Research (NCGR).

The *medicago.org/genome* site maintains the sequencing project's registry where centers register the BAC clones they are sequencing, including a queue of newly chosen BACs selected for future sequencing. This registry keeps track of comments and anomalies associated with individual BACs. The site contains tools for dynamically viewing the most up-to-date genome sequence assembly, summary statistics about the sequencing initiative and the *M. truncatula* genome, graphical access to the composite genetic map of *Medicago*, plus a variety of query and sequence alignment (Blast) tools. The site acts as a convenient portal to many other *Medicago* sequencing and genomics websites through a comprehensive set of context sensitive pull-down menus. Finally, the site provides convenient access to IMGAG (International *Medicago* Genome Annotation Group) gene annotation. Nearly all tabular and sequence data at *Medicago*.org/genome are available in both html and download (text) format.

Blast tools at *medicago.org/genome* include the innovative "CViT" suite of software (E. Cannon *et al*). CViT makes it possible to view the sequence positions of multiple Blast hits against a backdrop of the current genome sequence assembly. The site's Assembly browser tool is a Java applet that graphically displays the minimum tiling path of sequenced BACs, organized into sequence contigs. Where possible, these contigs are joined together into scaffolds/supercontigs through cases of paired BAC ends or other types of physical data. Mousing over the assembly displays information about the underlying BAC, while context sensitive pull-down menus enable drilling down for additional information and relevant links. Still another query tool enables users to identify the next best BAC clone(s) for chromosome walking in regions where the sequencing project has not yet progressed.



Medicago truncatula SEQUENCING RESOURCES

An organizational center for the NSF/EU *Medicago truncatula* genome sequencing project

Minnesota — N. Young *et al* | Oklahoma — B. Roe *et al* | TIGR — C. Town *et al*
Minnesota — E. Retzel *et al* | EU consortium
Noble Foundation | NSF Plant Genome | LIS/NCGR

About this site / FAQs | About this project >> | Data release policy >>

Search databases for:
in all data
GO> HELP?

Project home
About this project
Genome statistics

BAC registry
BAC neighbor queries
Downloads

Genetic map
Genome links & viewers
Blast & search
Gene annotation

News
Meetings
Literature
Participants
Report data errors

www.medicago.org

Medicago genome

About this information

	1	2	3	4	5	6	7	8	unanchored	total	current
BACs in build July 16 '06	213	211	272	275	264	139	242	284	50	1950	**1997**

- ▶ Genome assembly browser
- ▶ Genome assembly table
- ▶ CViT* Blast search BAC search
- ▶ BAC passports
- ▶ Get Medicago sequence
- ▶ Genome statistics
- ▶ Genome snapshots
- ▶ IMGAG

Other sequence data sites

Minnesota: Medicago BAC registry
Oklahoma: BACs & annotations views — chromosomes 1, 4, 6, 8
TIGR: BACs & annotation views — chromosomes 2, 7
European Medicago consortium: chromosomes 3 and chromosome 5
MIPS: Medicago genome UrMeLDB database
LIS: The Legume Information System at NCGR

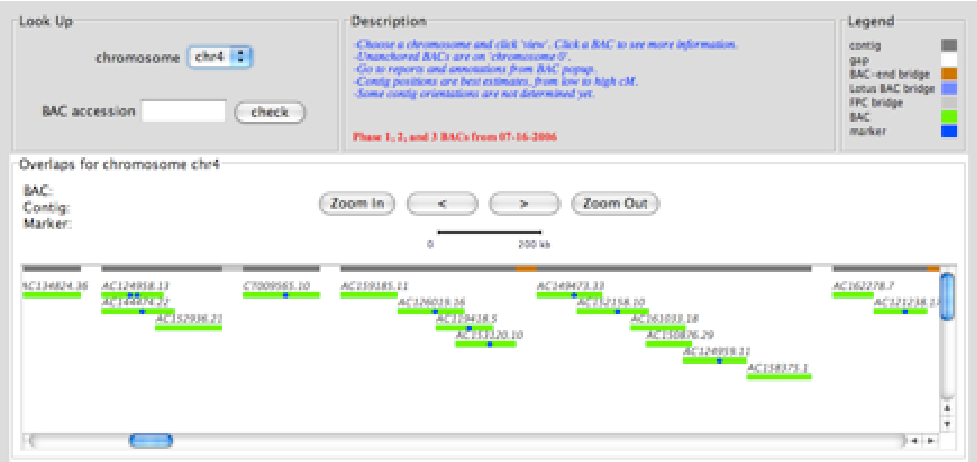
Recent news and updates

June 30, 2006	* ⇒ U MN genetic map is now available.
Mar 27, 2006	* genome build (Above) and images ("Genome snapshots").
Jan 18, 2006	* new markers (in all reports and downloads).

Also see: **news** and **meetings**

Site by Young lab at | Funding by | Additional funding by

medicago.org/genome homepage



Look Up

chromosome chr4

BAC accession [ ] check

Description

-Choose a chromosome and click 'view'. Click a BAC to see more information.
-Unanchored BACs are on 'chromosome 0'.
-Go to reports and annotations from BAC popup.
-Contig positions are best estimates, from low to high cM.
-Some contig orientations are not determined yet.

Phase 1, 2, and 3 BACs from 07-16-2006

Legend

contig	■
gap	□
BAC-end bridge	■
Lotus BAC bridge	■
FPC bridge	■
BAC	🟩
marker	🟦

Overlaps for chromosome chr4

BAC:
Contig:
Marker:

genome assembly browser

Figure 4.
Features of medicago.org/genome website.

8. References

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol*. **215**: 403-410.

Altschul, S.F, Madden, T.L., Schaffer, A.A., Zhang, J., Miller, W., *et al* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. **25**: 3389–3402.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., *et al* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*. **25**: 25-29.

Bateman, A., L. Coin, R. Durbin, R.D. Finn, V. Hollich, *et al* (2004) The Pfam protein families database. *Nucleic Acids Res*. **32**: D138-141.

Bendtsen, J. D., Nielsen, H., von Heijne, G., Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*. **340**: 783-795.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L. (2004) GenBank. *Nucleic Acids Res.* **32**: D23-D26.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., *et al* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet*. **29**: 365-371.

Burge, C., Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol*., **268**: 78-94.

Carrere, S., Gouzy, J. (2006) REMORA: a pilot in the ocean of BioMoby web-services. *Bioinformatics.* **22**: 900-901.

Dondrup, M., Goesmann, A., Bartels, D., Kalinowski, J., Krause, *et al* (2003) EMMA: a platform for consistent storage and efficient analysis of microarray data. *J. Biotechnol*. **106**, 135-146.

Dowell, R.D., R.M. Jokerst, A. Day, S.R. Eddy, L. Stein (2001) The distributed annotation system. *BMC Bioinformatics* **2**: 7.

Gonzales, M. D., Archuleta, E., Farmer, A., Gajendran, K., Grant, D., *et al* (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res*. **33**: D660-D665.

Goesmann, A., Linke, B., Rupp, O., Krause, L., Bartels, D., *et al* (2003) Building a BRIDGE for the integration of heterogeneous data from functional genomics into a platform for systems biology. *J. Biotechnol*. **106**: 157-167.

Haft, D.H., J.D. Selengut, O. White (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res* **31**: 371-373.

Harris, M.A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*. **32**: D258-261.

Henikoff, S., Henikoff, J. G. (1994) Protein family classification based on searching a database of blocks. *Genomics* **19**: 97-107.

Henikoff, S., Henikoff, J. G., Pietrokovski, S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15**: 471-479.

Henikoff, J. G., Greene, E. A., Pietrokovski, S., Henikoff, S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res*. **28**: 228-230.

Hohnjec, N., Henckel, K., Bekel, T., Gouzy, J., Dondrup, *et al*, 2006. Transcriptional snapshots provide insights into the molecular basis of arbuscular mycorrhiza in the model legume *Medicago truncatula*. *Funct. Plant Biol*., **33**: 737-748.

Journet E.-P., van Tuinen D., Gouzy, J., Crespeau, H., Carreau, V., *et al* (2002) Exploring root symbiotic programs in the model legume *Medicago truncatula* using EST analysis. *Nucleic Acids Res.* **30**: 5579–5592.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*. **32**: D277-280.

Letunic, I., R.R. Copley, S. Schmidt, F.D. Ciccarelli, T. Doerks, *et al* (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res*. **32**: 142-144.

Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., *et al* (2003) GenDB: an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*. **31**: 2187-2195.

Mulder N.J., Apweiler R., Attwood T.K., Bairoch A., Barrell D., (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*. **31**: 315-318.

Ponting, C.P., J. Schultz, F. Milpetz, P. Bork (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res*. **27**: 229-232.

Rhee, S.Y., W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, *et al* (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res*. **31**: 224-228.

Stein, L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., *et al* (2002). The generic genome browser: a building block for a model organism system database. *Genome Res*. **12**: 1599-1610.

Stekel, D.J., Git, Y., Falciani, F. (2000) The comparison of gene expression from multiple cDNA libraries. *Genome Res*. **10**: 2055-2061.

Town, C.D. (2006) Annotating the genome of *Medicago truncatula*. *Curr. Opin. Plant Biol*., **9**: 122-127.

Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, *et al* (2006) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. **34**: D173-180.

Wilkinson, M.D., Links, M (2002) BioMOBY: an open source biological web services proposal. *Brief Bioinform*. **3**: 331-341.

Zdobnov E.M., Apweiler R. (2001) InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847-848.