MtDB: a database for personalized data mining of the model legume Medicago truncatula transcriptome

Anne-Françoise J. Lamblin, John A. Crow, James E. Johnson, Kevin A. T. Silverstein¹, Timothy M. Kunau, Alan Kilian, Diane Benz, Martina Stromvik, Gabriella Endré¹, Kathryn A. VandenBosch¹, Douglas R. Cook², Nevin D. Young¹ and Ernest F. Retzel*

Center for Computational Genomics and Bioinformatics, University of Minnesota, MMC43, 420 Delaware Street S.E., Minneapolis, MN 55455, USA, ¹Department of Plant Biology, University of Minnesota, 220 Bio Sci Center, 1445 Gortner Avenue, St Paul, MN 55408, USA and ²Department of Plant Pathology, University of California, Davis, CA 95616, USA

ABSTRACT
In order to identify the genes and gene functions that underlie key aspects of legume biology, researchers have selected the cool season legume Medicago truncatula (Mt) as a model system for legume research. A set of >170 000 Mt ESTs has been assembled based on in-depth sampling from various developmental stages and pathogen-challenged tissues. MtDB is a relational database that integrates Mt transcriptome data and provides a wide range of user-defined data mining options. The database is interrogated through a series of interfaces with 58 options grouped into two filters. In addition, the user can select and compare unigene sets generated by different assemblers: Phrap, Cap3 and Cap4. Sequence identifiers from all public Mt sites (e.g. IDs from GenBank, CCGB, TIGR, NCGR, INRA) are fully cross-referenced to facilitate comparisons between different sites, and hypertext links to the appropriate database records are provided for all queries’ results. MtDB’s goal is to provide researchers with the means to quickly and independently identify sequences that match specific research interests based on user-defined criteria. The underlying database and query software have been designed for ease of updates and portability to other model organisms. Public access to the database is at http://www.medicago.org/MtDB.

INTRODUCTION
Legumes represent the third largest family of flowering plants. Due to their unique property of symbiotic nitrogen fixation, they are major contributors to the global nitrogen cycle, and thus are key species in many ecosystems. As agricultural species they are important sources of human and animal protein, and of edible and industrial oils. Important economic species are found primarily within two clades, commonly referred to as the tropical (e.g. soybean and bean) and the cool season (e.g. alfalfa, chick pea and peas) legumes. Motivated initially by the need to understand the molecular basis of symbiotic nitrogen fixation, researchers selected the cool season legume Medicago truncatula as a model system for legume biology (1). Its relatively small diploid genome (~450 Mbp) and efficient genetic and molecular properties have made it the system of choice for many studies of basic and applied legume research. A recently-initiated whole genome sequencing effort, combined with deep EST data, has the goal of describing the complete gene repertoire of a legume genome. As part of these efforts it is essential to identify and describe the relationships between the genome of M.truncatula and other legumes, as well as between legumes and other well-characterized genomes (plant, animal and microbial). An appropriately structured database is essential to this goal.

Over the last three years, more than 170 000 expressed sequence tags (ESTs) have been generated world-wide from these groups: the Mt consortium funded by the National Science Foundation, The Samuel Roberts Noble Foundation and a consortium of researchers in Europe with primary funding from the EU programme. In general, EST projects offer a quick source of information for micro- and macroarray based functional genomics, for metabolic reconstructions, a point of reference for proteomics and a quick assessment of the number and diversity of genes being expressed. Mt EST data is accessible from different public repositories worldwide, including NCBI-dbEST database (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map00?taxid=3880), The Institute for Genomic Research (TIGR, http://www.tigr.org/tdb/tgi/mtrcgi/), the National Center for Genome Research (NCGR, https://xgi/ncgr/org/mgi/), the Institut National de la Recherche Agronomique and le Centre National de la Recherche Scientifique (INRA–CNRS, http://medicago.toulouse.inra.fr/Mt/public/Mtruncatula.html) and, the Center for Computational *To whom correspondence should be addressed. Email: ernest@ccgb.umn.edu
Genomics and Bioinformatics at the University of Minnesota (CCGB, http://ccgb.umn.edu/research/bio/medicago). Mt ESTs have been assembled into contig sets or unigenes using programs such as Cap3 (INRA—CNRS) (2), Cap4 (TIGR) (3), and Phrap (CCGB, NCGR) (4).

The goal of the Mt biology and Mt bioinformatic research communities is to integrate all Mt genomics information, including genome and transcriptome data, genetic markers and physical map information. In this first release, MtDB contain only the transcriptome information (146 000 ESTs and their computed assembly as 26 000 unigenes), its parsed homology search results and a set of unique and elaborate queries for data mining. MtDB contents are updated as new sequences are released. MtDB, the *M. truncatula* database constitutes an important step toward the integration of Mt genomic, genetic and biological information.

**SCOPE OF MtDB**

In creating a new database for Mt EST information, our primary goal was first, to enable researchers to pose complex queries that integrate primary data with downstream analyses and second, to enable them to do it directly, without the need of an intermediary proficient in database languages, as currently done. We used the community web site, medicago.org (http://medicago.org), as a forum to solicit input from *M. truncatula* biologists regarding their intended use of the data. Requests were made for examples of questions of interest, and the responses (http://www.medicago.org/MtDB/Documents/Mt-Queries-examples.html) served to guide database development, from web interface to data modeling and schema design. MtDB is intended to provide a data mining environment that integrates features of NCBI, TIGR, NCGR, INRA-Mt (EST finished sequences, Unigene cluster, BLASTX analysis reports) while incorporating new attributes requested by the user community such as complex Boolean searches and comparative assessment between sequence assembly algorithms’ results. MtDB online query capabilities eliminate waiting periods for the data set to be fully assembled, analyzed, archived and released. Strong consideration in the design of MtDB was given to creating a structure that could be deployed for any organism data set with little modification of the support software required. This emphasis was consequential to the variety of species of the sequences processed at CCGB (animal and vegetal) (http://web.abc.umn.edu/biodata/index.html) and to the staff’s biological expertise in other organisms (*Drosophila*, soybean). MtDB was designed to be maintainable and portable, with the goal of providing a platform for the organization of other organism-specific data sets.

**DATA IN MtDB**

In addition to its analysis software development activities, CCGB is a sequence processing center. Sequence traces are collected from groups worldwide. All sequences are put through the same processing pipeline that consists of base calling with Phred (5) followed by vector and artifact filtering (polyA/T tail, linker sequence, bacterial genomic sequence) using tools developed by CCGB (i.e. Phran, gsvf4, af). Optimal parameters for the filtering tools were set empirically using sequence training sets. This processing pipeline allows us to retain quality scores and other sequence-related information that are useful in downstream visualization and analysis. Via a cooperative effort, we obtain trace data from data producers thus increasing the consistency of analysis for each EST and decreasing the probability of data set contamination with sequences mistakenly deposited in public domains. Although we have taken care to begin curation of MtDB with raw data in the form of trace files, MtDB is designed to integrate data from different sources without a requirement for de novo sequence processing. Processed Mt ESTs were assembled using Phrap to create a unigene set. Both EST and contig sequences were subjected to similarity analysis using BLAST (6) with CCGB’s peptide and dbEST databases. Finished sequences, contigs and analysis reports are archived in BioData (http://ccgb.umn.edu/research/swdev/biodata), a web-based file and visualization system. Customized tools were developed to retrieve, parse and load information from BioData (flat file archival database) into MtDB (Oracle relational database) relational tables. Data extracted from the BioData directories includes basic project bookkeeping information (sources of the data, library descriptions), the base-calling results and quality values of Phred, Phrap assembly results, and computed similarities (BLASTN and BLASTX) for all ESTs and contigs. All relationships with an *E*-value of less than 0.01 are retained in MtDB. The rationale is to allow users to interrogate complete reports from each BLAST analysis, with cut-off values defined by the user during the selection of query parameters. This differs from another Mt
MtDB is currently hosted on an eight processor Sun server (Sun Fire V880, 32GB RAM, Solaris 2.8 OS) with fibre-channel connections to a 4TB Sun StorEdge 3900 storage subsystem. The MtDB schema is implemented on our local instance of Oracle 8i. A simplified view of MtDB is depicted in Figure 1. Supplemental information (i.e. taxonomy id, cDNA clone id, dbEST id, GenPept id, TrEMBL id, definition line) to the MtDB core information (on reads, assemblies, BLAST reports) is extracted from locally produced relational databases containing information parsed from NCBI’s flat-file releases of GenBank, dbEST, taxonomy collections and from SWISS-PROT/TrEMBL, PIR, GenPept databases (http://ccgb.umn.edu/resources/data-collections/). This component-based design facilitates resource sharing with similar databases for other organisms. For improved performance, the results of comparative analyses were separated into four tables based on sequence type (EST, contigs) and BLAST program (BLASTX, BLASTN). Each table was partitioned on E-value and indexed using standard Oracle procedures and performance-training guidelines.

DATABASE QUERY AND INTERFACE

The MtDB query pages have been implemented using standard server-side Java technologies (JSP, JDBC) and provide a rich set of options by encapsulating the SQL needed for interaction with the MtDB database itself. Mining of MtDB EST and unigene data sets is accomplished through an integrated suite of eight ‘query pages’. The queries use the cDNA library information and the parsed results of pre-run homology searches of all sequences present in MtDB. A distinguishing feature of MtDB is that it enables users to develop personalized

Figure 2. MtDB query#2 shows the layout of the library and taxonomy filters and the Boolean options Any/All/Only/Contain/Do not contain used in queries 2–7. The library filter counts 32 individual cDNA libraries which have been regrouped in 9 tissue categories based on the tissue used in the library construction. The taxonomy filter counts 28 species options that covers 10 legumes, 5 non-legumes, 4 prokaryotes and 9 pathogens. Some taxonomy options cover the complete taxonomy tree (e.g. bacteria, fungi) while others are species specific (e.g. Medicago truncatula, Arabidopsis thaliana). The Booleans operate the following way: 1. ‘All’, identifies contigs that contain EST from at least all of the libraries selected. 2. ‘Any’, identifies contigs that contain EST from at least one of the selected libraries, i.e. library A OR B. In this case, the identified contigs can be composed of EST from libraries A and B or others, of EST from libraries A and others, of EST from library B and others. 3. ‘Only’, identifies contigs that contain EST from only the selected libraries, i.e. contigs composed of EST from libraries C and D and none others.
Figure 3. Snapshot composite of query result web pages. Many of the features returned in the output tables are hyperlinked to original records in CCGB, TIGR, NCBI or to pre-computed analysis. Numbers in orange indicate the resource to which the information is hyperlinked to. 1. CCGB unique identifier accesses CCGB-BioData sequence record. 2. The type of BLAST program links to the full BLAST report for query sequence in CCGB-BioData file system. 3. The name of the peptide BLAST hit links to Metafam protein family assignment page and outside protein database record. 4. BLAST hit definition line links to NCBI precomputed BLink (BLASTP) record for that hit sequence. 5. Links to NCBI taxonomy page for that taxonomy ID number. 6. Automatically loads the query ID number into MtDB Q#9 ‘CCGB-contigs’ equivalence page (see D). 7. Accesses CCGB-ERGO metabolic reconstruction tentative assignment for that query sequence. 8. Links to NCBI-GenBank record for that GenBank accession number. 9. Links to NCBI-dbEST record for that GenBank gi number. 10. Links to TIGR MtGI record for that TC or EST. 11. Based on the assembly target set, links to either TIGR MtGI or INRA Mt record. 12. Automatically loads the Original clone ID into MtDB Q#8 ‘MtDB-EST aliases’ to search for all identifier aliases. (A) Standard HTML table result display for queries 1–7. A summary of the parameters selected for the query is shown at the top of the page followed by the number of sequences that were identified matching the set parameters. To facilitate browsing of the results, only the top BLAST hit is returned though the search was done on the full BLAST report of individual sequences. (B) Query #8 ‘MtDB EST-aliases’ result display table shows the different identifiers that can be used to either find a sequence in MtDB or learn how a sequence is named in other community databases. (C) Query#9 ‘CCGB-contigs’ ‘Other’s TCs’ equivalence page returns a first result table listing the contigs and TCs that exhibit some equivalence based on EST composition and other parameters. Statistics about the pairwise comparison are displayed. (D) Clicking the ‘Go’ button called for the detailed EST distribution between the two equivalent assembly sequences. A help page is available at http://www.medicago.org/MtDB2/Help/MtDB2-Q9Help.html.
queries that combine library and taxonomy filters with Boolean options (all, any, only) and selectors (‘do contain’, ‘do not contain’) (Fig. 2). The ‘all’, ‘any’, ‘only’ Booleans are the equivalent of ‘and’, ‘inclusive or’, ‘exclusive or’, respectively. The library filter is used to select or identify sequences based on their cDNA library of origin. A total of 30 libraries represent various stages of plant development and exposure to different external agents, such as symbiotic and pathogenic microorganisms. For example, queries can be constructed to extract unigenes that are unique to a particular tissue (or tissues), or are expressed in response to a particular treatment, such as infection by a pathogenic or symbiotic microbe. The taxonomy filter is used to identify sequences that possess taxonomically-restricted homologies. For example, the user might structure a query to identify Mt sequences (peptide or nucleic acid) that have homologies with an Arabidopsis or maize. Such sequences would be candidates for legume-specific genes, perhaps contributing to legume-specific properties such as symbiotic nitrogen fixation.

**QUERY RESULT OUTPUT DISPLAY**

MtDB provides two output formats, each of which can be displayed to the screen or saved as a tab-delimited file. Selecting ‘CCGB ID’ returns the unique CCGB tags of the identified sequences. These tags can be cross-referenced to other identifier types by using the MtDB EST-alias tool page (see Tracking tools below). In the second option, identified sequences are presented in a hypertext linked HTML table that contains statistics related to the top BLAST hit alignment, the definition line of the hit sequence and its species of origin. The information present in the output table is hyperlinked to the original database record (CCGB, NCBI, TIGR, and pre-computed analyses of protein family assignments (CCGBs MetaFam) (7) and BLASTP result against the protein non-redundant (nr) database (NCBI BLink). Hyperlinks are also given to a tentative metabolic pathway assignment (CCGB-ERGO plant metabolic reconstruction), where such relationships are predicted (collaborative effort with Integrated Genomics, Inc.). To facilitate comparison of unigenes predicted by the various Mt data sites (i.e. CCGB, TIGR, INRA or NCGR), the output table is hyperlinked to the most probable assembly equivalence across Unigenes. (Fig. 3A).

**TRACKING TOOLS**

In addition to the database mining query pages, MtDB provides two innovative tools. ‘MtDB EST aliases’ is designed to track sequences across databases by means of a completely cross-referenced list of identifiers used at the various Mt data sites (Fig. 3B). ‘MtDB CCGB-contig(-)Other’s TCs’ enables researchers to quickly identify related contigs within other institutions’ Mt assemblies. Two criteria are currently used to identify related contigs: (i) overlap in EST composition and (ii) consensus sequence similarity. By default, all contigs are displayed that have any common ESTs or a similar subregion (E-value <1.0). These criteria effectively identify all equivalent and homologous contigs between two assemblies. Researchers may sort related contigs preferentially according to multiple criteria [e.g. (i) common-EST-count, (ii) percent-consensus-overlap, (iii) percent-consensus-identity] (Fig. 3C).

If desired, the actual ESTs in common between the analogous contigs, plus those that exclusively belong to only one or the other contigs/assemblies, may be viewed (Fig. 3D). At present, MtDB does not provide a confidence estimate for helping determine which of the analogous contigs are most likely correct. However, the authors are currently working on software to evaluate contigs quality. In the meantime, researchers may use the links to the contigs images to make manual accuracy assessments. A help page is available at http://www.medicago.org/MtDB2/Help/MtDB2-Q9Help.html.

**FUTURE DEVELOPMENT**

Features under development for MtDB include:

- A keyword search tool, initially based on the BLAST hit description line and ultimately improved to reflect standardized ontologies.
- A graphical depiction of relative expression for each Mt EST based on tissue or condition of origin.
- Integration EST data with whole genome sequence information. The derived relationships will include genome context information (i.e. synteny), gene structure verification (e.g. intron-exon junctions), and the location of putative promoter regions.

**CONCLUSION**

MtDB is a high-performance relational database system designed to provide on-line data mining of the integrated Mt transcriptome projects. Using pre-structured queries with user-defined variables, users without knowledge of SQL are able to sort through more than 26 000 contigs and identify sequences of interest. Use of the complete BLAST report for each sequence expands the power of each query and has the potential to identify novel correlations. We anticipate that MtDB will provide a flexible database that can be applied to the management of EST data from many species.

**ACKNOWLEDGEMENTS**

This work supported in part by the National Science Foundation awards DBI-0196197, DBI-0110206, DBI-9975806 and DBI-9872565; by the USDA SCA 58-3625-8-117 funded by the North Central Soybean Research Board and the United Soybean Board, and the USDA SCA 58-1907-0-030. The authors sincerely thank their colleagues who kindly shared data and comments: Jérôme Gouzy, Pascal Gamas and Jean Dénarié (INRA–CNRS-Genoscope Mt project), Gregory May and Angela Scott (Samuel R. Noble Foundation), Christopher Town and Foo Cheung (TIGR), Mark Waugh and William Beavis (NCGR). Special thanks to Suzanne Grindle, Rodney Staggs, Shalini Raghavan and Charles Paule for their help in sequence processing and Chris Dwan for...
improvement and maintenance on processing software pipeline implementation.

REFERENCES