Databases and Information Integration for the *Medicago truncatula* Genome and Transcriptome[1]

Steven B. Cannon, John A. Crow, Michael L. Heuer, Xiaohong Wang, Ethalinda K.S. Cannon, Christopher Dwan[2], Anne-Francoise Lamblin[3], Jayprakash Vasdewani, Joann Mudge, Andrew Cook, John Gish, Foo Cheung, Steve Kenton, Timothy M. Kunau, Douglas Brown, Gregory D. May, Dongjin Kim, Douglas R. Cook, Bruce A. Roe, Chris D. Town, Nevin D. Young, and Ernest F. Retzel*

Department of Plant Pathology, University of Minnesota, St. Paul, Minnesota 55108 (S.B.C., X.W., E.K.S.C., J.V., J.M., N.D.Y.); Center for Computational Genomics and Bioinformatics, University of Minnesota, Minneapolis, Minnesota 55455 (J.A.C., M.L.H., C.D., A.F.L., T.M.K., E.F.R.); Department of Plant Pathology, University of California, Davis, California 95616 (A.C., J.G., D.J.K., D.R.C.); The Institute for Genomic Research, Rockville, Maryland 20850 (F.C., C.D.T.); Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019 (B.A.R., S.K.); Plant Biology Division, The Samuel Roberts Noble Foundation, Ardmore, Oklahoma 73401 (G.D.M.); and North Carolina State University, Fungal Genomics Laboratory, Department of Plant Pathology, Raleigh, North Carolina 27695 (D.B.)

An international consortium is sequencing the euchromatic genespace of *Medicago truncatula*. Extensive bioinformatic and database resources support the marker-anchored bacterial artificial chromosome (BAC) sequencing strategy. Existing physical and genetic maps and deep BAC-end sequencing help to guide the sequencing effort, while EST databases provide essential resources for genome annotation as well as transcriptome characterization and microarray design. Finished BAC sequences are joined into overlapping sequence assemblies and undergo an automated annotation process that integrates ab initio predictions with EST, protein, and other recognizable features. Because of the sequencing project's international and collaborative nature, data production, storage, and visualization tools are broadly distributed. This paper describes databases and Web resources for the project, which provide support for physical and genetic maps, genome sequence assembly, gene prediction, and integration of EST data. A central project Web site at medicago.org/genome provides access to genome viewers and other resources project-wide, including an Ensembl implementation at medicago.org, physical map and marker resources at mtgenome.ucdavis.edu, and genome viewers at the University of Oklahoma (www.genome.ou.edu), the Institute for Genomic Research (www.tigr.org), and Munich Information for Protein Sequences Center (mips.gsf.de).

Legumes are the third largest plant family in the world and the second most important crop family. *Medicago truncatula* has been chosen as a reference legume for genome sequencing because of its importance as a model for rhizobial and mycorrhizal relationships (Limpens and Bisseling, 2003; Cook, 2004), its relatively small genome (Arumuganathan and Earle, 1991; Blondon et al., 1994), tractable genetics, excellent genetic and physical map resources, and extensive synteny with larger legume crop and forage genomes. Previously, Medicago was the target of survey bacterial artificial chromosome (BAC) sequencing at the University of Oklahoma and funded by the Samuel Roberts Noble Foundation (May and Dixon, 2004). In 2003, the U.S. National Science Foundation and European Union 6th Framework Program initiated support for large-scale genome sequencing, coordinated by an international steering committee.

Substantial new bioinformatics resources now are needed to sequence and annotate the Medicago genespace. They include data management and project coordination for the sequencing effort, as well as convenient access, query, and visualization for the broader user community. Delivering genomic resources to the user community while simultaneously coordinating the sequencing project requires the implementation of existing open-source tools, and the rapid development of mechanisms to process and display a rapidly growing body of genomic information.

A BAC-by-BAC strategy was selected for Medicago by the international steering committee because previous research indicated that most genes are found in relatively gene-rich euchromatic arms, with few found in or around centromeres (Kulikova et al., 2001; Pedrosa et al., 2002; Kulikova et al., 2004; Young et al., 2005). Thus, a BAC-by-BAC strategy targeting euchromatic arms is expected to efficiently capture most of the

[1] This work was supported by the National Science Foundation (awards DBI–0321460, DBI–0196197, DBI–0110206, DBI–9975806, and DBI–9872565), by the U.S. Department of Agriculture Cooperative State Research, Education, and Extension Service/National Research Initiative Program, and by the Samuel Noble Roberts Foundation.

[2] Present address: BioTeam Inc., Cambridge, MA.

[3] Present address: University of Minnesota Cancer Center, MMC 806, 420 Delaware St. SE, Minneapolis, MN 55455.

* Corresponding author; e-mail ernest@ccgb.umn.edu; fax 612–626–6069.

www.plantphysiol.org/cgi/doi/10.1104/pp.104.059204.

 © 2005 American Society of Plant Biologists

Medicago genespace. Substantial supporting tools for sequencing also were in place early in the Medicago effort, including a dense genetic map composed of BAC-based markers, a physical map with approximately 11 × coverage, and deep BAC-end sequence data with more than 160,000 sequences from two BAC libraries. Both *Lotus japonicus* and tomato, two other plants with apparently similar genome organization, are also being sequenced BAC-by-BAC. In contrast, since the maize genome is organized with gene-rich islands interspersed among repeat-rich tracts, it likely will be sequenced using a combination of gene-enriched whole-genome shotgun (WGS) and targeted BAC sequencing (Palmer et al., 2003; Whitelaw et al., 2003; Okagaki and Phillips, 2004).

It is expected the Medicago BAC-by-BAC approach will produce contiguous, marker-anchored sequence that will reach chromosome-arm scale in size. These arm-length sequence contigs will be important, as Medicago is intended to serve as a reference for crop legumes with genomes too large or complex to be sequenced efficiently at this time. Contiguous Medicago genome sequence also will be valuable for comparative evolutionary studies. In particular, another model legume, *L. japonicus*, now is being sequenced with a similar strategy at the Kazusa Institute in Japan (Kato et al., 2003). The availability of two reference legume genomes will facilitate important studies of genome evolution, symbiosis, nitrogen fixation, secondary metabolism, and other biological properties in this important plant family (Young et al., 2005).

There are distinct informatics and database requirements for a BAC-based sequencing project that change over the course of the project. The remainder of the paper describes the informatics and computational resources that support genetic and physical mapping, BAC and BAC-end relationships, construction of long sequence assemblies, and, finally, annotation through ab initio predictions and comparisons to EST, gene, and genomic sequences.

PROJECT ORGANIZATION AND STATUS

Sequencing is under way at four centers worldwide. In the United States, the University of Oklahoma Advanced Center for Genome Technology (OU/AGCT) is sequencing chromosomes 1, 4, 6, and 8, while the Institute for Genomic Research (TIGR) is sequencing chromosomes 2 and 7. In the United Kingdom, the Sanger Institute is sequencing chromosome 3 with mapping and organizational support from the John Innes Center, while in France, Genoscope is sequencing chromosome 5 with support from the Institut National de la Recherche Agronomique (INRA). Primary bioinformatics associated with sequencing (such as BAC sequence assembly and selection of new clones) is managed within each sequencing center. Central organizational and informatics management is provided at the University of Minnesota and the Center for Computational Genomics and Bioinformatics (UMN/CCGB) as part of www.medicago.org. A Medicago physical map was developed previously at the University of California at Davis (UCD). Annotation is coordinated by an international consortium known as the International Medicago Genome Annotation Group (IMGAG), with participation from OU, TIGR, the Munich Information for Protein Sequences Center (MIPS), UMN/CCGB, and INRA.

As of February 1, 2004, 1,206 BAC clones had been sequenced, with 695 clones finished (Phase 3). This comprises approximately 143 Mbp with approximately 122 Mbp nonredundant coverage. The sequencing centers will continue to sequence approximately 100 BACs per month through most of 2005. In 2006, the project focus will shift to closing gaps, finishing remaining BAC sequences, constructing chromosome assemblies, annotating assemblies, and conducting initial analyses of the whole genome. Throughout the project, all sequence, marker, contig data, and BAC-level annotations are freely available to all researchers.

DATABASES AND WEB ACCESS TO THE MEDICAGO GENOME SEQUENCE

The participation of multiple centers in mapping and sequencing makes it essential to have common data access and organization, while also allowing for distributed ownership and manipulation of specialized data. Accordingly, the project was designed as a federation, with multiple sequencing and informatics centers participating in data generation, annotation, and display. Web and database resources are distributed, but with a common access point. For example, BAC fingerprinting has been done at a single site (Cook and Kim lab, UC Davis), and these data have been made available to the project via both Web interface and direct database access. This information has been incorporated and expanded in databases at medicago.org/genome to include information on sequencing progress, mapping, and BAC overlap and orientation. Similarly, the acceptance across the project of a single final gene annotation (described below) is one of the hallmarks of cooperation in this project.

The central project portal, http://medicago.org/genome, located at UMN, has an underlying database that mirrors all common-resource data in the project (markers, BACs, contigs, sequence, and relationships). This database provides a Web interface with links from any single data item to the center responsible for that data, including genetic markersand finger print contigs (FPC) at UCD (http://mtgenome.ucdavis.edu), raw BAC sequences at GenBank, and annotation views for each BAC by MIPS, CCGB, TIGR, or OU. Additionally, bulk data sets are available, including all current BAC sequences, all markers and marker sequences, and all BAC-end sequences. To support the genome sequence assembly, the site also offers visualization tools for BAC overlap relationships and se-

quence assemblies (Fig. 1), and specialized queries, for example, to determine the relationships of sequenced BACs via paired BAC-ends.

Two project management features are designed specifically for centers involved in the sequencing effort and require login and password for modification but are viewable by visitors. A BAC registry tracks BACs being sequenced at all sequencing centers, all the way from "intention to sequence" to finished sequence. This is updated twice weekly, and records the sequencing phase, sequencing center, BAC size, and other information. The Registry is associated with a data anomalies system that provides a way for sequencing consortium members to submit, revise, or resolve comments about BACs, markers, or contigs. These comments become part of the database record for the described entity, and are displayed in reports for any such entity. In effect, this serves as bug tracking software for the project.

Marker and physical (FPC-based) map resources, essential to the success of the sequencing effort, are maintained at UC Davis (mtgenome.ucdavis.edu; Fig. 2). This resource includes a database of marker, genotype, primer, marker-BAC, and BAC-contig relationships. Additionally, the site provides Web visualization tools for exploring the genetic map, markers in the context of nearby markers, and FPC contigs. A mapping tool allows a user to input genotypic scoring data and to place new markers in the context of existing markers. The site also provides stored BLAST results between BAC-end and BAC sequences, and between other legume sequences and Medicago BAC sequences.

An example suggests how these sites might be used by a biologist interested in a mapped trait. First, the chromosome of interest can be browsed from the medicago.org/genome home page, using either tabular reports or a chromosome-viewer applet (Fig. 1). A search for a linked marker (using the search field on each page) will locate the marker and associated BAC. From the report or the applet, the user can then link to marker, contig, or BAC reports or genome browsers, or a graphic of the FPC contig (Fig. 2).

GENOME ANNOTATION

The IMGAG consortium has selected a canonical gene prediction and annotation procedure for generating a project-wide protein and coding sequence data set. Annotation involves a multi-institution pipeline, relying on Medicago-trained FGENESH (Salamov and Solovyev, 2000) predictions, the EuGene (Foissac et al., 2003) predictor and gene combiner software, and the TIGR PASA (Haas et al., 2003) procedure for aligning expressed sequence tags (ESTs) to genomic sequence. Eugene combines data from a variety of extrinsic and intrinsic methods. Examples of intrinsic data are results from FGENESH predictions, SpliceMachine (Degroeve et al., 2005), and NetGene2 (Hebsgaard et al., 1996), while examples of extrinsic data are results from BLASTX (Altschul et al., 1997) against predicted peptides from EST clusters, comparisons against ProDom (Bru et al., 2005), and comparisons against ESTs from other legumes. The processing pipeline is described in more detail in a whitepaper at http://medicago.org/genome/about.php.

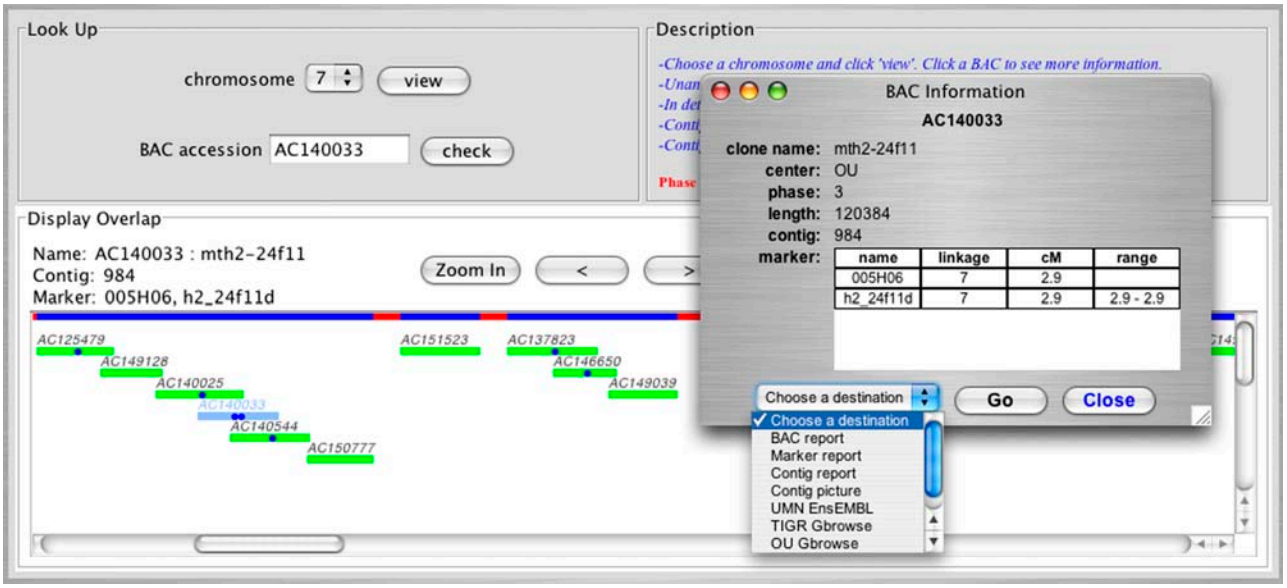

Figure 1. BAC sequence assembly viewer. The database and Web interface at http://medicago.org/genome includes a variety of visualization tools and data formats, including this sequence assembly viewer. The viewer accesses database information for each BAC, shown in the popup menu. Links from the popup menu lead to other database reports and directly to other data sources and views: GenBank records, contig and marker reports, contig images (such as one shown in Fig. 2), and genome viewers.

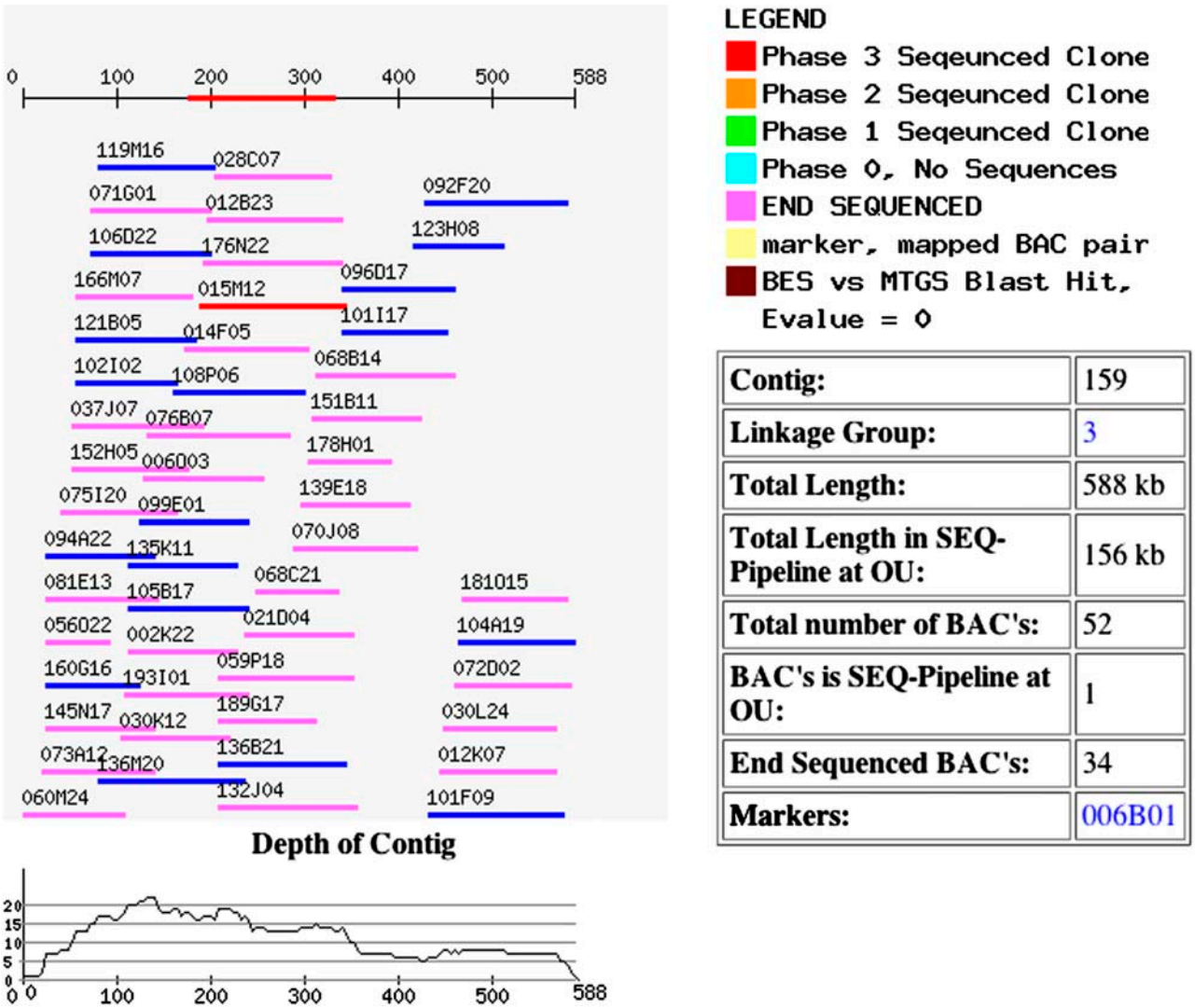

Figure 2. Image of FPC contig. The database and Web interface at http://mtgenome.ucdavis.edu provides access and visualizations for the Medicago physical and genetic map. This image is typical of FPC contigs in the project, showing fingerprint-based BAC overlaps, BACs being sequenced, BACs with and without BAC end sequence, markers, and map position. BES, BAC end sequence; MTGS, *M. truncatula* genome sequence.

In addition to the consortium annotation, independent analysis pipelines and annotation viewers are in place at each sequencing and major bioinformatics center. These resources are integral to the activities at each center and complement the project as a whole, and each browser provides somewhat different data sets and features. For example, the sequencing center at OU maintains a GBrowse (Stein et al., 2002) viewer and analysis pipeline. This gives OU immediate feedback about their newly sequenced BACs, including gene features, overlaps with other BACs, and identification of repetitive DNA. The annotation viewers at all centers will display consensus gene calls from the combined IMGAG annotation effort. In addition, a GBrowse viewer also is available at TIGR, a DBBrowser at MIPS, and an Ensembl at viewer at CCGB (for URLs, see Table I).

THE CCGB ENSEMBL GENOME ANALYSIS ENVIRONMENT

Tightly associated with the medicago.org/genome central portal and BAC Registry database is an Ensembl (Birney et al., 2004) genome analysis pipeline and viewer. In turn, the Ensembl pipeline depends on transcript information from MtDB (*Medicago truncatula* database; described below) and many other data sets. The pipeline is an extended version of Ensembl. Details about the pipeline's computational resources and data targets are available at http://decifr.ccgb.umn.edu/Medicago_truncatula.

Once the BAC sequences within the pipeline are associated with a chromosome and chromosome location using the MySQL BAC Registry database, they are processed through a modified Ensembl pipeline to iden-

Table I. *Selected M. truncatula genomics resources*

M. truncatula Genomics Resources		
UMN/CCGB	Ensembl pipeline, database, genome viewer	http://decifr.ccgb.umn.edu
	MtBR BAC Registry database, Web portal	http://www.medicago.org/genome/
	MtDB2 clustering, query interface	http://www.medicago.org/MtDB2/
UCD	mtgenome physical and genetic map	http://mtgenome.ucdavis.edu
VBI	Medicago functional genomics and bioinformatics	http://medicago.vbi.vt.edu/
NCGR	LIS, Legume Information System	http://www.comparative-legumes.org/
OU	GMOD database, GBrowse genome viewer	http://www.genome.ou.edu/medicago.html
Noble Foundation	Medicago functional genomics	http://www.noble.org/medicago/
TIGR	TIGR database, GBrowse genome viewer	http://www.tigr.org/tdb/e2k1/mta1/
	Gene Index EST clustering, query interface	http://www.tigr.org/tdb/tgi/
	TIGRFAMs gene family HMMs	http://www.tigr.org/TIGRFAMs/
INRA	Medicago EST Navigation System (MENS)	http://medicago.toulouse.inra.fr/Mt/EST/
	Laboratorie Interactions Plantes Microorganismes	http://capoul.toulouse.inra.fr/centre/lipm/
MIPS	UrMeLDB genome browser	http://mips.gsf.de/proj/plant/jsf/medi/index.jsp

tify repetitive regions that are masked via RepeatMasker (Smit et al., 2004). Then, genic regions and associated transcripts are identified by ab initio Genscan and FGENESH gene prediction (Burge and Karlin, 1997; Salamov and Solovyev, 2000) and BLASTX and BLASTP sequence similarity searches (Altschul et al., 1997). An *M. truncatula*-specific Fgenesh target has been developed and is available upon request. Subsequent similarity searches are then performed against the current versions of the nonredundant UniRef100 set from the Uniprot Consortium (Apweiler et al., 2004), dbEST from National Center for Biotechnology Information (NCBI), the CCGB *M. truncatula* unigene set (Lamblin et al., 2003), and the TIGR gene indices for *M. truncatula* and *Glycine max* (Lee et al., 2005).

Primary input to the analysis pipeline consists of BAC sequences provided by the sequencing consortium (all phase 3 BACs and oriented single-gap phase 2 BACs). This and other sequence data are retrieved by scheduled nightly download processes, subjected to preliminary quality assessments, and fed into the pipeline as appropriate. This automation helps ensure the freshness and completeness of the sequences to be analyzed. Updates to the BLAST targets for UniRef100 and dbEST also are automated on a similar schedule. Releases made at the data source sites are recognized and appropriate files are downloaded. Subsequent processes perform quality assessments of these downloads and trigger the creation of BLAST targets and ancillary metadata. Bioinformatics administrators are notified of these events.

The analysis pipeline is database-centric, relying intrinsically on data storage support from a MySQL database. Two schema types are used: The first is for pipeline processing support and the retention of intermediate results, while the second is for collecting the computed annotations from the pipeline. The latter schema retains genomic sequences, repeats, markers, predicted genes and transcripts, exons, alignments, and protein translations, and provides the BAC annotation information needed for the production of graphical displays at the local Ensembl Web interface.

In addition to the familiar graphical Web interfaces, access to this information is available through two mechanisms: a distributed annotation server (DAS; Dowell et al., 2001), and direct access to the stored information. The reference DAS server for the *M. truncatula* genome can be used to provide reference sequence for a standalone DAS client, for uploading DAS annotation via the Ensembl Web interface, and for DAS annotation servers and clients based at IMGAG member sites allowing local display of consensus gene calls as DAS features locally. Direct network access to the MySQL-stored annotations also is available through standard programming (e.g. database connections via Java JDBC, Perl DBI) and through Ensembl APIs (Birney et al., 2004).

INTEGRATING AND EXPLORING THE MEDICAGO TRANSCRIPTOME

While the genomic sequence is the foundation for annotation, accurate annotation also depends on the extensive Medicago EST resources. EST data will also remain important as an indication of expression patterns, transcript frequencies, and alternate splicing. Four Medicago EST databases have been produced: Medicago EST Navigation System (MENS; Journet et al., 2002), the TIGR Medicago Gene Index (MGI; Lee et al., 2005), which later evolved into the Legume Information System (Bell et al., 2001), and the CCGB MtDB (Lamblin et al., 2003). The MtDB, its successor Nimbus, and supporting analysis pipelines and linkages to Ensembl are described here.

The *M. truncatula* community also has the MtDB investigative database tool, available on-line at http://medicago.org/MtDB2, providing researchers the ability to explore information about the Medicago transcriptome in a variety of ways. EST data from the project is stored along with its associated clone library metadata, and the ESTs themselves are associated with the reports obtained from shotgun sequence assembly. Also stored is information on the unigene sequences

obtained by BLASTX against standard peptide collections, such as UniRef100 from the Uniprot Consortium (Apweiler et al., 2004) and TAIR peptides (Rhee et al., 2003).

These information linkages allow sophisticated queries to be created from a Web-based interface. An example of such a query is ''Identify all unigenes that appear in infected-leaf libraries and show sequence similarity to a known legume peptide.'' The report from this query provides information digests and links to further information. It is important to note the queries available at this site were developed in direct response to surveys of the research community. Feedback from the community was used to design the query pages, the kinds of information presented, and the page layout. In turn, this has allowed the MtDB development team to customize the site for the special interests of this community.

The success of MtDB and the experience with the system itself led to the creation of a next-generation tool, Nimbus, which extends the capabilities of MtDB while addressing the need for hosting multiple transcriptome projects at a single site. In particular, these exploration entry points quickly lead to further queries based on pathways and functional classifications. This system makes heavy use of established standards, e.g. the UniRef100 peptide set; Pfam, TIGRFAMs, and SMART HMM (hidden Markov model) sets, and controlled vocabularies available from the Gene Ontology Consortium. Reference information such as peptide and motif descriptions are retained and maintained in a shared schema (nimbus_shared), and each individual project contains information on its ESTs, unigenes, and computed annotations, and is maintained in another schema (nimbus_project) by a designated biologist curator.

While the Nimbus system and Ensembl pipelines make no demands upon the freshness of the information it stores and queries, it is still a concern of the project administrator to create new releases periodically. For this reason, we have established a reliable system for downloading target information available at remote sites and automatically creating BLAST, HMM, and other targets (Fig. 3). This system consists of two components: one, which polls remote sites for information changes and downloads new information as appropriate, and the second, which digests this information and creates sequence, HMM, and metadata

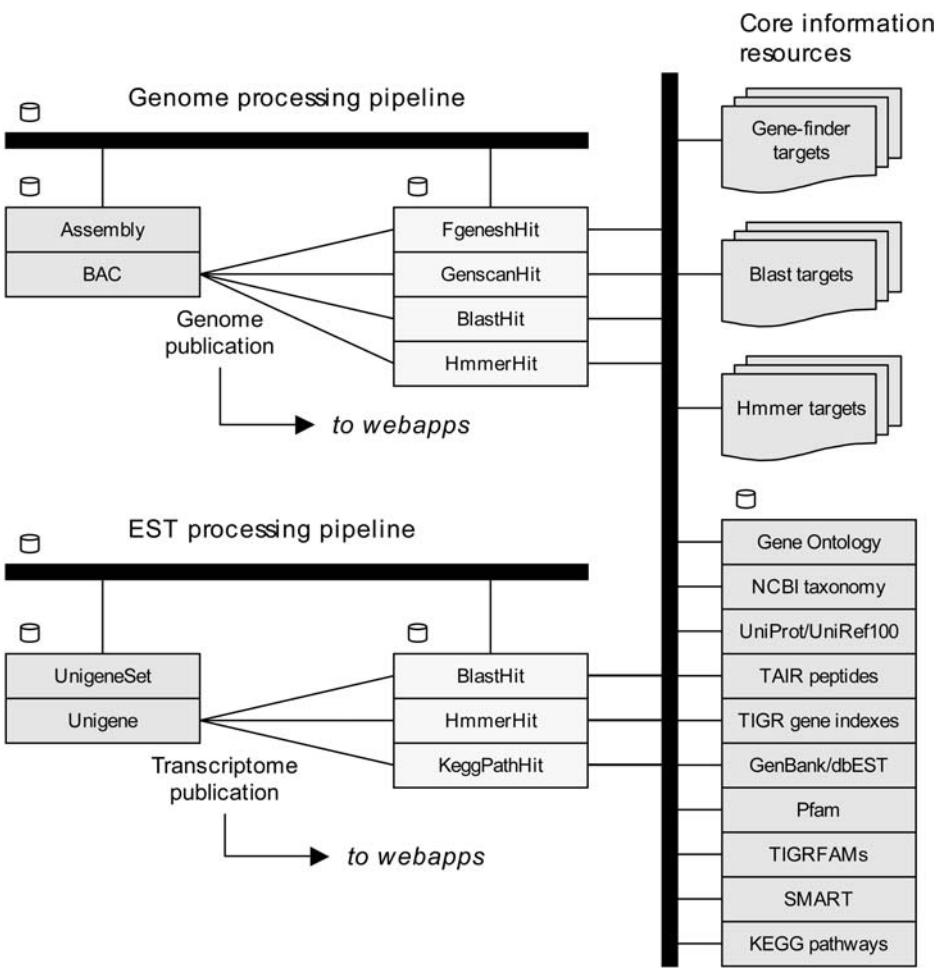

Figure 3. Ensembl and MtDB EST pipelines and information resources. The EnsEBML annotation and the MtDB and Nimbus EST processing pipelines are supported by an automated system for downloading and maintaining up-to-date consistent target information (such as BLAST, HMM, and other targets on the right side of figure), and associated metadata. Core database resources, including BAC and EST sequences, BLAST hits, gene predictions, and other features (left and center of figure) are generally maintained in a relational format, and are accessible to all aspects of the projects. Software resources are generally deployed across compute clusters, and are managed by a scheduling system that directs loading and submission of work to cluster queues.

files for target creation. The use of common source information sets is important for CCGB bioinformatics since targets are created for different platforms (Linux clusters, single machine use, TimeLogic Decypher server) and these different targets must remain coherent and consistent. The processing pipeline is managed with a MySQL RDBMS. The system consists of two schemas: pipeline processing management and post-processing annotation. The collection schema contains sequence, repeats, markers, predicted genes, and transcripts, as well as exons, alignments, protein translations, and annotated BACs. A meta-scheduler, called the grid-manager, has been created for Medicago and other species maintained in the database. The scheduler sits between pipeline jobs described in the database and resources available to complete those jobs. The grid-manager directs loading and submission of work to three distinct cluster queues. Jobs move through the pipeline and their progress is retained within the pipeline database. As results are generated, they populate the collection database and are immediately ready for display via the Web. More details about the pipeline are available at http://decifr.ccgb.umn.edu/Medicago_truncatula/project_status.

DATA MANAGEMENT OPPORTUNITIES AND CHALLENGES

A BAC-by-BAC project like the one under way in Medicago depends on detailed genetic and physical maps, management of the complex relationships among data from these disparate types of maps, and integration into the growing body of genome sequence data. In the case of Medicago, marker-based genetic maps were developed over the course of a decade at multiple labs using different parental crosses. An early task in the sequencing project, therefore, was collecting all marker data, including genotypic scoring information, marker names, marker aliases, marker type, parents, map position, linkage group, and association with BAC clones. Where mapping crosses were made using different parents, approximate map positions had to be extrapolated. In some cases, markers mapped to multiple locations. For example, there is a translocation between chromosomes 4 and 8 in one parent of the primary mapping population (Choi et al., 2004). This means that markers in this region consistently score poorly or map to both 4 and 8. Separately, some markers are associated with multiple BACs, either because the marker is not single-copy or because multiple BACs span the same genomic region. The job of collecting, organizing, and databasing marker and BAC-marker associations is managed in mirrored databases at http://mtgenome.ucdavis.edu and http://medicago.org/genome.

The original Medicago physical map was developed at UCD with restriction digests of nearly 45,000 BACs and construction of a FPC map (Marra et al., 1997; Soderlund et al., 2000). The FPC build produced 1,370 contigs and 2,441 singletons, a solid foundation for genome sequencing with relatively few gaps. Information about BAC membership and contig position is housed and displayed graphically at http://mtgenome.ucdavis.edu and is used internally in databases at UMN and the sequencing centers.

As with genetic marker data, physical map data also presents significant data challenges. Inherent in FPC, a BAC contig may be chimeric, with two or more biologically legitimate contigs joined by BACs that only coincidentally share similar fingerprint patterns. A BAC also can be a member of multiple FPC contigs. Still other contigs overlap, but do not have sufficient shared fingerprint patterns for a successful FPC join. All of these relationships are accommodated in the project relational database at the UMN, but interpreting the relationships is nontrivial.

The Medicago sequencing project therefore employs several approaches to deal with inconsistencies in the FPC contig data. First, the FPC map is treated as suggestive rather than definitive. Once a seed BAC has been sequenced from a contig, BAC-end sequence is used where possible to confirm location and overlap among adjacent BACs. In most cases, BACs are refingerprinted to confirm clone identity. New SSR markers also are designed from BAC or BAC-end sequences on unmapped contigs and for contigs that have only one marker and appear to be weakly joined in the center. Finally, contigs are linked where possible by direct BAC sequence overlap or by multiple, paired BAC-end matches to sequenced BACs from two contigs.

Another important type of relationship to be managed in a clone-by-clone project is that of overlapping BAC sequences, and ultimately, the construction of chromosome arm-scale sequence assemblies. While BAC sequence overlap may seem a straightforward relationship, in practice there are several complications. Sequencing errors and allelic differences result in overlapping regions that usually contain nucleotide substitutions or indels. Large segmental duplications, caused either by transposable elements or events such as slipped-strand mispairing, can generate apparent regions of near-identical matches, either between nonoverlapping clones or in other regions of legitimately overlapping clones. For example, GAG-POL-like retrotransposon sequences occur more than 250 times in Medicago BACs sequenced so far, and are responsible for at least 72 nearly perfect matches (>99% identity) up to nearly 11 kbp in length between nonoverlapping BACs. While RepeatMasker (Smit et al., 2004) helps to avoid spurious associations, additional steps are needed to verify overlaps.

BAC overlaps in the project are represented as regions of perfect or high quality overlap (HQOR) and flanking regions of overhang. For example, two BACs might have an HQOR of 9,000 nucleotides (nt), extending from the first through the 9,000th nt of one of the overlapping BACs, followed by a short indel, followed by an overhang of 1,000 nt. Because of the

likelihood of transpositions and local duplications, corroborating support for an overlap generally needs to be provided by information such as paired end sequences from other BACs or by FPC contig proximity. Overlap data is stored in two types of files: a local file representing overlaps, orientations, and evidence for triplets of overlapping BACs; and a global file representing the start and end coordinates of each BAC in a chromosomal pseudo-assembly. Overlap relationships are calculated and maintained within each center and are also calculated and displayed genome-wide at http://medicago.org/genome, both as text files (at the ''Overlap tables'' page) and in the java BAC overlap and map information viewer (Fig. 1).

PERSPECTIVES

A genome sequencing project depends on comprehensive data integration. Certainly, the long-range biological goals for both the sequencing and transcriptome projects are integrative: to help identify all genes, regulatory environment, and evolutionary history of an organism, and, eventually, to extend this information to understand the biology of whole systems far beyond any sequenced model genome. The information that needs to be integrated includes physical map (BAC fingerprints), genetic map (markers, genotypic records, marker-BAC associations), BAC-end sequences, BAC-BAC overlaps and BAC-BAC-end relationships, ESTs, EST contigs and metadata, and complex predicted and external data relationships required for annotation. In a multi-institution sequencing project (of a genome, transcriptome, or any other -ome), an additional layer of integration is essentially sociological. Project databases and associated architectures and Web technologies need to accommodate change and to foster communication and data exchange. As in any large, public, multi-year project, the databases and interfaces must provide biologists access to intermediate data products.

The present Medicago sequencing and EST databases have both drawn from and may provide lessons for other sequencing projects. The BAC-by-BAC sequencing approach depended on particular conditions and extensive cumulative resources: strong genetic and physical maps, a relatively small and gene-rich euchromatic genespace, deep EST and BAC-end sequencing, and the experiences of genome sequencing initiatives that came before. Common, project-wide databases with eb access have been important, including the mtgenome.ucdavis.edu.edu physical map and marker database/interface and the project BAC Registry and portal at medicago.org/genome. Even as the project has depended on central databases, it has been important for each sequencing and informatics center to be able to develop specialized working databases and annotation viewers. To avoid fragmentation, a central portal attempts to provide transparent access to public resources and viewers.

Beyond the goal of producing high quality raw genome sequence, the highest value in the sequence information will be in genes and transcript information and in the record of conservation and change with respect to other plant genomes. Some comparative genomic tools are now part of the Medicago Ensembl implementation. These will be expanded as the genome sequence nears completion, and initial comparisons with other plant genomes are already possible at least over portions of the Medicago and Lotus genomes (Young et al., 2005). Gene identification and transcript information depends, in large part, on EST data. Much of the utility in EST data comes through comparing transcript frequencies in different tissues and in related species. The MtDB2 and Nimbus database/interface provide query options that link information on homology, tissue expression, and species origin. Gene prediction and annotation depends both on ab initio predictions and use of external information such as homologies to other genes and matches to transcripts. Several annotation pipelines and viewers are active in the sequencing project. This provides both local control of analysis as well as several independent analyses and data displays. As with the project databases, to avoid fragmentation, a consensus consortium annotation will be available, broadcast using DAS and displayed on each browser. The annotation pipelines are complex and computationally intensive, and in some senses, are the most integrative stage in the project, drawing on diverse data and analyses. The consortium IMGAG annotation pipeline likewise draws on processing steps at several institutions, producing a consensus proteome for the entire project.

ACKNOWLEDGMENTS

We are grateful to many participants in the international Medicago community. A comprehensive list of participants is found at http://www.medicago.org/genome/people.php. Particular thanks to the Samuel Roberts Noble Foundation for seed money to begin Medicago genome sequencing. Collaborators meriting special mention include: Philippe Bardou, Christine Burton, Frederic Debelle, Rene Geurts, Jerome Gouzy, Gyorgy Kiss, Klaus Mayer, Giles Oldroyd, Frances Quetier, Jane Rogers, Pierre Rouze, Thomas Schiex, Heiko Schoof, and Manuel Spannagl. Apologies and thanks to the many others whose work has been important in support of the development, sequencing, and analysis of this genomic data.

Received December 31, 2004; returned for revision March 4, 2005; accepted March 21, 2005.

LITERATURE CITED

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389–3402

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, et al (2004) UniProt: the Universal Protein knowledgebase. Nucleic Acids Res **32:** D115–D119

Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Rep **9:** 208–218

Bell CJ, Dixon RA, Farmer AD, Flores R, Inman J, Gonzales RA, Harrison

MJ, Paiva NL, Scott AD, Weller JW, et al (2001) The Medicago Genome Initiative: a model legume database. Nucleic Acids Res **29:** 114–117

Birney E, Andrews D, Bevan P, Caccamo M, Cameron G, Chen Y, Clarke L, Coates G, Cox T, Cuff J, et al (2004) An overview of Ensembl. Nucleic Acids Res **32:** D468–D470

Blondon F, Marie D, Brown S, Kondorosi A (1994) Genome size and base composition in Medicago sativa and M. truncatula species. Genome **37:** 264–275

Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res **33:** D212–D215

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

Choi HK, Kim D, Uhm T, Limpens E, Lim H, Mun JH, Kalo P, Penmetsa RV, Seres A, Kulikova O, et al (2004) A sequence-based genetic map of Medicago truncatula and comparison of marker colinearity with M. sativa. Genetics **166:** 1463–1502

Cook DR (2004) Unraveling the mystery of Nod factor signaling by a genomic approach in Medicago trunactula. Proc Natl Acad Sci USA **101:** 4339–4340

Degroeve S, Saeys Y, De Baets B, Rouze P, Van de Peer Y (2005) SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics **21:** 1332–1338

Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L (2001) The distributed annotation system. BMC Bioinformatics **2:** 7

Foissac S, Bardou P, Moisan A, Cros MJ, Schiex T (2003) EUGENE'HOM: A generic similarity-based gene finder using multiple homologous sequences. Nucleic Acids Res **31:** 3742–3745

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31:** 5654–5666

Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S (1996) Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. Nucleic Acids Res **24:** 3439–3452

Journet EP, van Tuinen D, Gouzy J, Crespeau H, Carreau V, Farmer MJ, Niebel A, Schiex T, Jaillon O, Chatagnier O, et al (2002) Exploring root symbiotic programs in the model legume Medicago truncatula using EST analysis. Nucleic Acids Res **30:** 5579–5592

Kato T, Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S (2003) Structural analysis of a Lotus japonicus genome. V. Sequence features and mapping of sixty-four TAC clones which cover the 6.4 mb regions of the genome. DNA Res **10:** 277–285

Kulikova O, Geurts R, Lamine M, Kim DJ, Cook DR, Leunissen J, de Jong H, Roe BA, Bisseling T (2004) Satellite repeats in the functional centromere and pericentromeric heterochromatin of Medicago truncatula. Chromosoma **113:** 276–283

Kulikova O, Gualtieri G, Geurts R, Kim DJ, Cook D, Huguet T, de Jong JH, Fransz PF, Bisseling T (2001) Integration of the FISH pachytene and genetic maps of Medicago truncatula. Plant J **27:** 49–58

Lamblin AF, Crow JA, Johnson JE, Silverstein KA, Kunau TM, Kilian A, Benz D, Stromvik M, Endre G, VandenBosch KA, et al (2003) MtDB: a database for personalized data mining of the model legume Medicago truncatula transcriptome. Nucleic Acids Res **31:** 196–201

Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, Antonescu V, Chan A, Cheung F, Quackenbush J (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res **33:** D71–D74

Limpens E, Bisseling T (2003) Signaling in symbiosis. Curr Opin Plant Biol **6:** 343–350

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. Genome Res **7:** 1072–1084

May GD, Dixon RA (2004) Medicago truncatula. Curr Biol **14:** R180–R181

Okagaki RJ, Phillips RL (2004) Maize DNA-sequencing strategies and genome organization. Genome Biol **5:** 223

Palmer LE, Rabinowicz PD, O'Shaughnessy AL, Balija VS, Nascimento LU, Dike S, de la Bastide M, Martienssen RA, McCombie WR (2003) Maize genome sequencing by methylation filtration. Science **302:** 2115–2117

Pedrosa A, Sandal N, Stougaard J, Schweizer D, Bachmair A (2002) Chromosomal map of the model legume Lotus japonicus. Genetics **161:** 1661–1672

Rhee SY, Beavis W, Berardini TZ, Chen G, Dixon D, Doyle A, Garcia-Hernandez M, Huala E, Lander G, Montoya M, 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

Salamov AA, Solovyev VV (2000) Ab initio gene finding in Drosophila genomic DNA. Genome Res **10:** 516–522

Smit AFA, Hubley R, Green P (2004) RepeatMasker Open-3.0. http://repeatmasker.org

Soderlund C, Humphray S, Dunham A, French L (2000) Contigs built with fingerprints, markers, and FPC V4.7. Genome Res **10:** 1772–1787

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al (2002) The generic genome browser: a building block for a model organism system database. Genome Res **12:** 1599–1610

Whitelaw CA, Barbazuk WB, Pertea G, Chan AP, Cheung F, Lee Y, Zheng L, van Heeringen S, Karamycheva S, Bennetzen JL, et al (2003) Enrichment of gene-coding sequences in maize by genome filtration. Science **302:** 2118–2120

Young ND, Roe BA, Town CD, Cannon SB, Sato S, Tabta S (2005) Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. Plant Physiol **137:** 1174–1181