Translational Genomics for Alfalfa Varietal Improvement

G. D. May

The XXIV International Grassland Congress / XI International Rangeland Congress (Sustainable Use of Grassland and Rangeland Resources for Improved Livelihoods) takes place virtually from October 25 through October 29, 2021.

Proceedings edited by the National Organizing Committee of 2021 IGC/IRC Congress

Published by the Kenya Agricultural and Livestock Research Organization

Translational genomics for alfalfa varietal improvement

G.D. May
Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, Oklahoma 73401, USA, Email: gdmay@noble.org

Key points

1. *Medicago truncatula* is a model legume with available mapping, genome, and RNA, protein and metabolite profiling databases and genetically diverse populations.
2. Genomics resources developed for *M. truncatula* have application in the study and improvement of alfalfa making it an excellent model for this forage legume.

Keywords: *Medicago*, functional genomics, genome, alfalfa

Introduction

With more than 650 genera and 19,000 species, legumes are one of the two most important crop families in the world. Among cultivated plants, legumes are unique in their ability to fix atmospheric nitrogen through a novel symbiotic relationship with bacteria known as Rhizobia. Since they are not limited for nitrogen, legumes have remarkably high levels of protein, a property that is both biologically and agriculturally significant. Nearly 33% of all human nutritional requirement for nitrogen is derived from legumes, and in many developing countries, legumes serve as the single most important source of protein. Legumes synthesize an impressive array of secondary metabolites, including isoflavonoids and triterpene saponins, shown to possess anti-cancer and other health promoting effects. Not surprisingly, legumes play a central role in nearly all crop rotation systems and are universally viewed as essential for secure and sustainable food production.

All major crop legumes are found in the monophyletic subfamily Papilionoideae. Within this subfamily, the tropical legumes include the economically important soybean (*Glycine max*), common bean (Phaseolus spp.), cowpea (*Vigna unguiculata*), and mung bean (*V. radiata*), while temperate legumes include species such as pea (*Pisum sativum*), alfalfa (*Medicago sativa*), lentil (*Lens culinaris*), and chick pea (*Vicia arietinum*). Papilionoid legumes first appeared around 65 million years ago based on fossil records (reviewed in Doyle, 2002), the same time as other important crop families. Because they form a compact monophyletic evolutionary group, comparative genomics among Papilionoid species has huge potential to increase our understanding of this vitally important group of plants. Indeed, a growing body of evidence demonstrating micro- and macrosynteny among Papilionoids suggests that discoveries made in one species can often be extended to other members of the subfamily.

The uniqueness of a plant family is the product of all of its many traits. Some, notably the diagnostic characters that define the family taxonomically, may be truly unique, but most are found, in different combinations, in unrelated groups of plants. At the level of morphology, anatomy, and chemistry, however, characters shared with other families may be analogous, rather than homologous--functionally similar, but derived independently from different ancestors. The molecular basis for analogous characters may involve different genes, either truly non-homologous genes that have arisen independently, or paralogous members of gene families that were recruited independently in different evolutionary lineages to perform similar roles. The underlying theme of much of genomics research, and the basis for the

highly successful model system approach, is that there are many features shared among genomes even of very distantly related organisms, permitting generalization.

M. truncatula, also referred to as barrel medic because of the shape of its pods is a forage legume commonly grown in Australia and throughout the Mediterranean. It is closely related to the world's major forage legume, alfalfa, but unlike alfalfa, which is a tetraploid, obligate outcrossing species, *M. truncatula* can be self-pollinated and has a simple diploid genome (with eight pairs of homologous chromosomes). *M. truncatula* has been chosen as a model species for genomic studies in view of its small genome, fast generation time (from seed-to-seed), and high transformation efficiency (Cook, 1999, May & Dixon, 2004). Genes from *M. truncatula* share high sequence identity to their orthologs from alfalfa so it serves as an excellent genetically tractable model for alfalfa. Studies on syntenic relationships (comparisons of genome content and organization between organisms) are establishing links between *M. truncatula*, alfalfa, and pea, as well as *Arabidopsis*.

In 1999, a Center for Medicago Genomics Research was established at the Samuel Roberts Noble Foundation. Scientists at Noble have taken a global approach in studying the genetic and biochemical events associated with the growth, development, and biotic and abiotic interactions of the model legume *M. truncatula*. Approaches taken to dissect the genome of *M. truncatula* and its function include; large-scale EST and genome sequencing, gene expression profiling, the generation of *M. truncatula* transposon-tagged and fast-neutron mutagenized populations and high-throughput protein and metabolite profiling. The resulting multidisciplinary data sets developed in our program are being interfaced to provide scientists with an integrated set of tools to address fundamental questions pertaining to legume biology. Our goal has been to establish a research program that will make significant contributions to the areas of legume molecular biology, biochemistry, and genetics research.

Discussion

There is a strong, cohesive and well-organized *Medicago* research community in the US, Europe, Australia and elsewhere and *Medicago* genomics has advanced rapidly in the past five years. This is due in large part to research at the Samuel Roberts Noble Foundation, and also to projects funded under the NSF Plant Genome Program and in the European Union. These efforts have collectively produced a large number of ESTs, the initiation of a Noble Foundation-funded *Medicago* genome sequencing project, a robust physical map that is well-anchored to the genetic map, two generations of expression microarrays using first spotted cDNAs and then oligonucleotide arrays, programs in protein and metabolite profiling, and the generation of EMS, fast-neutron, and transposon-tagged mutant populations. Following on from this, a group of *Medicago* researchers on both sides of the Atlantic formed the *Medicago* Genome Sequencing Consortium/Initiative and was successful in obtaining funding from the NSF, the EU, BBSRC and INRA/Genoscope to complete the sequencing of the euchromatic portion of the *M. truncatula* genome on a chromosome by chromosome basis using a BAC-based strategy by the end of 2006. The underpinning for these successful proposals was the funding provided by the Noble Foundation to Bruce Roe at the University of Oklahoma. This gave the project a jump-start with approximately 63 Mb of BAC sequence in GenBank at the time that the NSF and international projects were reviewed for funding.

M. truncatula as a reference legume: *Medicago* EST and genome sequencing

The Noble Foundation recently released to NCBI more than 27,000 additional *Medicago* ESTs from our databases. As of January 2005, these sequences push the *M. truncatula* EST total to 216,645 - number eight among all plant species on an EST basis. The Foundation's contribution to this total is 114,913 high-quality EST clones -- approximately 53% of the world's efforts. As we continue EST projects for other species, our *Medicago* EST sequencing efforts will now focus upon characterization of full-length ESTs. The first objective of this proposed activity (a collaboration with C. Town, TIGR) is to generate full-length cDNA sequences for in excess of 20,000 genes (i.e. approximately half of the expected transcriptome) from *M. truncatula*. A total of ~ 10,000 candidate *Medicago* FL-cDNAs have been identified. 2,000 of which have been sequenced to completion. Of the ~ 8,000 remaining candidate cDNAs, 6,000 can be found in the EST collection at Noble. The remainder will be obtained from the construction and sequencing of normalized libraries with a high proportion of full-length sequences. A second objective will be to produce full-length sequence-validated cDNA ORF clones for at least 10,000 of these genes in a Gateway recombination vector system for functional analyses.

The international Medicago genome sequencing project

A whole-genome *M. truncatula* sequence program initially began at the University of Oklahoma. The initial goal of the project was to generate an approximately one-fold whole genome shotgun sequence data of the 500 megabase genome from a plasmid-based genomic library and obtain target shotgun clones for additional primer walking-based sequencing. However, preliminary results from the shotgun approach suggest that the *M. trucatula* genome is highly repetitive. As previously predicted, estimates are that approximately 80% of the genome is highly repetitive and that approximately 80% of the gene-rich regions represent only 20% of the total genome. To reduce the amount of redundant sequence, the sequencing strategy was modified to sequence bacterial artificial chromosome (BAC) clones from *M. truncatula* BAC libraries. More than 800 BACs were identified based on DNA markers or gene content and were sequenced to working draft coverage (four- to five-fold) utilizing a BAC-based shotgun sequencing approach, in the first phase of this project.

The whole-genome shotgun approach resulted in the sequencing of the *M. truncatula* chloroplast genome, since the total genomic DNA preparation not only contained the nuclear genome, but also a significant level of the chloroplast DNA. The DNA sequence of the *M. truncatula* chloroplast genome has now been completed and consists of one contiguous 124,039 base pair circle. Artificially linearizing the sequence at the histidine tRNA prior to the psbA gene allows the *Medicago* chloroplast genomic sequence to be co-linear with the *Arabidopsis*, tobacco, and most other chloroplast genomes. The semi-automated annotation of the *M. truncatula* chloroplast genome using Web-Artemis has been completed, and can be viewed at: http://www.genome.ou.edu/medicago_chloroplast/med_chloro_art.html.

Currently the *Medicago* genome sequencing program involves researchers at the University of Minnesota, The Institute for Genomic Research and the University of Oklahoma in the U.S. and Sanger and Genoscope in the U.K. and Europe, respectively in a chromosome by chromosome approach. As of early 2005, more than 800 BACs are finished and almost 1,400 total BACs are at sequencing phase 1, 2, or 3 with the phase 2 and 3 BAC sequences comprising more than 135 Mbp of non-redundant *M. truncatula* genome sequence. All

sequences are available through GenBank and EMBL databases. The anticipated completion date for the *Medicago* genome sequencing project is December 2006.

The other *Medicago* -omics: profiling transcripts, proteins and metabolites

Expression analyses

Two generations of DNA microarray technologies have been established for expression profiling in *Medicago* species. *M. truncatula* genome-wide microarrays are being generated using the *Medicago* Array-Ready Oligonucleotide Set (GS-1700-02) Version 1.0 (Operon). Approximately 16,000, amino-linked, 70-mer oligonucleotides are being printed onto aminosilane-coated "Superamine" slides (Telechem), using Telechem type SMP3 printing pins in Dr. David Galbraith's laboratory at the University of Arizona. Preliminary results in our groups suggest that *M. truncatula* oligonucleotide arrays hybridize well with targets synthesized using *M. sativa* mRNA as a template. These arrays should provide a valuable tool to study complex traits in alfalfa.

The design of an Affymetrix *Medicago* GeneChip array, the composition of which was arrived at after consultations between Affymetrix and the international *Medicago* community has been completed. The array contains probe sets to profile approximately 60,000 gene sequences that were derived by combining all *M. truncatula* EST and annotated genomic sequence data. *M. truncatula* sequences included on the array are; 1) International *Medicago* Genome Annotation Group (IMGAG) high-quality gene prediction from *Medicago* BAC sequences, 2) FGENESH gene predictions from all Phase II and Phase III M. truncatula BACs sequences, and 4) chloroplast ORFs. For tentative consensus sequences (TCs) and singletons that could not be orientated both strands were tiled. In addition, the array includes *M. sativa* sequences that do not have corresponding orthologs in the *M. truncatula* data sets. Also included on the array are *Sinorhizobium meliloti* predicted ORFs from genome and plasmid sequences. It is anticipated that the *Medicago* GeneChip array will be released early summer 2005.

To supplement the expression analyses data generated by microarrays, we have added an "open system" serial analysis of gene expression (SAGE) to our set of transcript profiling tools. Such open system approaches allow for the identification and analysis of genes not previously characterized. With "closed systems" such as microarrays, analysis is limited to only those species previously identified and assigned to an array. SAGE analyses have already been used to study gene expression in plant systems (Matsumura *et al*., 1999). Among the high-throughput, comprehensive technological methods used to analyze transcript expression levels, array-based hybridization and SAGE are currently the most common approaches.

Molecular mechanisms underlying the initiation and maintenance of embryonic pathways in plants are largely unknown. To gain better insight into these processes, serial analysis of gene expression (SAGE) was used to profile transcript accumulation levels and to identify differentially expressed genes in early stage somatic embryos of *M. truncatula*. A total of more than 131,000 SAGE tags were sequenced and 30,329 unique tags were identified in non-embryogenic, pro-embryo and globular embryo cell cultures. These studies illustrate the power of SAGE technology as a tool for both transcript profiling and gene discovery and its use in examining global changes in plant gene expression patterns. As additional plant genomes such as *M. truncatula* are sequenced and plant-specific SAGE databases become

publicly available, the use of SAGE in understanding fundamental changes in gene expression should gain broad appeal in the plant research community.

Protein and metabolite profiling

The protein complement of the genome, the proteome, serves as a biological counterpart to the *Medicago* EST and gene expression analyses. Given that many biological phenomena lack the requirement for *de novo* gene transcription, proteomics studies provide a mechanism to study proteins and their modifications under developmental changes and in response to environmental stimuli.

Two-dimensional polyacrylamide gel electrophoresis (2-D PAGE) has been established as the dominant technique for analysis of complex protein mixtures since its introduction in 1975 (O'Farrell, 1975; Blackstock & Weir, 1999). The technique utilizes isoelectric focusing and polyacrylamide gel electrophoresis for first and second dimension separation, respectively. Currently, 2-D PAGE technology is capable of resolving some 10,000 proteins, with 2,000 proteins being typical experimental results (Klose & Kobalz, 1995). A recent review describes the role of 2-D PAGE in proteomic and genetic studies of plant systems, including its use as a tool to investigate genetic diversity, phylogenetic relationships, mutant characterization, and drought tolerance (Thiellement *et al*., 1999).

Although 2-D PAGE analysis has been used for the last 20 years in protein profiling, it provides limited information on protein identification. Recent advances in mass spectrometry and the establishment of protein databases have substantially increased the ease and speed with which proteins can be identified (Yates, 1998). The union of these technologies is the foundation for modern proteomic studies. The typical experiment begins with comparative digital imaging of the 2-D gels to detect variations in protein concentration or elution profile. These protein spots are excised, extracted, and identified by using a variety of mass spectrometry techniques.

Basically, two mass spectrometry (MS) techniques are used for protein identification. The first is peptide mass-mapping of proteolytic digested fragments (Wolf *et al*., 1998; Yates, 1998). The observed mass fragments can be searched against a theoretical list of proteolytic peptide maps predicted by a given database. Increased peptide mass accuracy has increased the success and selectivity of such searches (Jensen *et al*., 1996). If the database query is unsuccessful, the protein can be sequenced by using tandem mass spectrometry (MS/MS) (Yates, 1998). During the MS/MS experiment, only the peptide mass of interest is isolated or transmitted, thus discriminating against all other components of the mixture with different mass values. After isolation, the peptide is further fragmented by using a unimolecular or bimolecular (collision gas) strategy. Fragments observed in the isolated peptide MS/MS spectrum can then be rationalized to a sequence.

Initial proteome profiling at the Noble Foundation has been performed to generate representative 2-D PAGE protein profiles for stems, leaves, seedpods, roots, flowers, tissues, and suspension cell cultures. Proteins were systematically identified and cataloged by using peptide mass mapping and database searching. An interactive database of the results of these analyses can be found at the following web address: http://www.noble.org/2dpage/search.asp. Analytical and biological variances associated with the 2-D PAGE proteomics approach for *M. truncatula* have been determined and will function as baseline measurements for comparative protein profiling in elicitor-induced *M. truncatula* cell cultures.

Metabolic profiling is the key to understanding how changes at the transcriptional and translational levels affect cellular function. Unlike proteomics, a single analytical technique does not exist that is capable of profiling all the low molecular weight metabolites of the cell. Our approach is to profile metabolites of control and treatment tissues by using an assortment of analytical techniques including: high-performance liquid chromatography (HPLC), capillary electrophoresis (CE), gas chromatography (GC), mass spectrometry (MS), and various combinations of the above techniques such as GC/MS, LC/MS, and CE/MS.

As the program has progressed, the development of methods to extend the profiling range to include metabolite classes such as phenylpropanoids, lignins, terpenoids saponins, soluble sugars, sugar phosphates, complex carbohydrates, amino acids, and lipids has continued. Method development also includes procedures for sequential extraction and parallel analysis. Sequential extraction segregates the metabolome into more manageable classes of chemical compounds with similar physical/chemical properties thereby facilitating the use of parallel analytical profiling techniques. Profiling of elicitor-induced cell cultures and *M. truncatula* natural variants for flavonoids, lignins, other phenylpropanoids and triterpenoids, especially saponins, has been performed as a component of an NSF-funded *Medicago* functional genomics project.

Forward and reverse genetics approaches in *M. truncatula*

Forward and reverse genetic systems for *M. trucatula* are being developed at the Noble Foundation and elsewhere in the *Medicago* research community. Forward genetic systems facilitate efficient identification of genes underlying phenotypic traits of interest, while reverse genetics systems enable the isolation of mutations in genes of known sequence. Kiran Mysore's laboratory at the Noble Foundation, in collaboration with Dr. Pascal Ratet, CNRS, Gif sur Yvette, France, are developing a large-scale, transposon-tagged mutant library of *M. truncatula* using the tobacco retrotransposon Tnt1 (Tadege *et al*., 2005). Approximately 20,000 tagged *M. truncatula* lines will be generated during the next five years. Transposon-plant genome junctions will be isolated and characterized through DNA sequence analyses. A database of these junction sequences is being created, and these sequences are being mapped to the *M. truncatula* genome for a reverse genetics approach to determine gene function.

Fast-neutron irradiation induces DNA damage and chromosomal deletions. Deletions that occur in known genes can be detected by a shift in the size of PCR amplification products of genes of interest. Dr. Rujin Chen's group at the Noble Foundation is developing a fast-neutron mutagenized population of *M. truncatula*. Of the approximately 10,000 fast-neutron irradiated M1 *M. truncatula* plants generated thus far, two percent display a visible mutant phenotype. It is anticipated that 100,000 M1 *M. truncatula* plants will be screened within the next three years.

Databases and genomics resources

With approximately 200,000 genomics-related visits in 2004, the Foundation's web site continues to benefit the *M. truncatula* and legume research communities by providing access to data and resources developed at Noble (Table 1.). More that 1,200 EST clones from the Foundation's *Medicago* EST collection have been distributed free of charge to researchers in 18 countries. The identity of all requested ESTs are confirmed by 5'-end sequencing. The Foundation is also providing legume researchers access to oligonucleotide microarrays on a cost recovery basis.

A large number of public world wide web-based *Medicago* databases are available (Table 1.) These databases provide the research community with access to the tools and data developed in the *Medicago* DNA sequencing and functional genomics programs.

Table 1 Web addresses for *Medicago* genomics resources

Site	URL
The Center for *Medicago* Genomics Research	http://www.noble.org/medicago/index.html
The Consensus Legume Database	http://www.legumes.org
The Legume Information System	http://www.comparative-legumes.org
The *M. truncatula* Gene Index	http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=Medicago
TIGR *M. truncatula* Genome Resources	http://www.tigr.org/tdb/e2k1/mta1/
M. truncatula Sequencing Resources	http://www.medicago.org/genome/
Medicago Bioinformatics at the University of California – Davis	http://medicago.plantpath.ucdavis.edu/
European Research Programmes on the Model Legume M. truncatula	http://medicago.toulouse.inra.fr/
Medicago Genome Sequencing University of Oklahoma	http://www.genome.ou.edu/medicago.html
Medicago truncatula Functional Genomics and Bioinformatics	http://medicago.vbi.vt.edu/data.html

Conclusion

M. truncatula is a highly developed model legume, with a large research community, that serves as an excellent model for developing new forage varieties. What is still necessary are laser-capture microdissection (LCM) techniques to enable the isolation of specific cells (i.e. specific zones within a nodule or root cap) from complex tissues for subsequent molecular analyses. Tissue preparation and microextraction protocols are being established to allow LCM microsamples to undergo quantitative transcript, protein and metabolite profiling. The application of LCM techniques to established functional genomics technologies has the potential to enhance our understanding of diverse plant cell type-specific biological processes.

The development of bioinformatics tools for the processing, visualization and integration of transcript, protein and metabolite profiles and datasets with the evolving genome sequence is still required. These tools will lead to a correlated view of gene expression and cellular response. The long-term impact will be the integration of transcript, protein, and metabolite data for plant mutants and natural variants, that will advance all aspects of fundamental and applied legume research. This information will be used to develop agronomically important legume species, such as alfalfa that (i) are more resistant to cold, drought and fungal and viral diseases, (ii) will provide higher crop yields while reducing needs for chemical inputs, and (iii) will produce natural chemicals that promote human and animal health and nutrition. Higher yields and lower production costs will enhance the economy of rural agriculture, especially in developing nations, while a reduction in chemical usage will benefit the environment. Value-added traits such as increased levels of nutraceuticals will provide

farmers with new crop alternatives and allow them to participate in the high value niche markets.

Acknowledgements

Richard A. Dixon, Maria J. Harrison, Lloyd W. Sumner, Kiran Mysore, Rujin Chen and members of their laboratory teams are to be acknowledged for their efforts. Bruce Roe, Nevin Young and Chris Town are acknowledged for their contributions to U.S. component of the *M. truncatula* genome project. Pedro Mendes for his extensive contributions to the bioinformatics portions of the NSF-funded program "An Integrated Approach to Functional Genomics and Bioinformatics in a Model Legume" (DBI-0109732). This program is supported by the National Science Foundation (DBI-0109732 and DBI-0110206), Forage Genetics International, and the Samuel Roberts Noble Foundation.

References

Blackstock, W.P. & M.P. Weir (1999). Proteomics: quantitative and physical mapping of cellular proteins. *Trends in Biotechnology*, 17, 121-127.

Cook, D.R. (1999). *Medicago truncatula* - a model in the making! *Current Opinion in Plant Biology*, 2, 301-304.

Doyle, J.J., J.L. Doyle, A.H.D. Brown & R.G. Palmer. (2002). Genomes, multiple origins, and lineage recombination in the *Glycine tomentella* (Leguminosae) polyploid complex: histone H3-D gene sequences. *Evolution*, 56, 1388-1402.

May, G.D. & R.A. Dixon (2004). Medicago truncatula. *Current Biology*, 14, 180-181.

Jensen, O.N., A. Podtelejnikov, M. Matthias-Mann (1996). Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rapid Communications in Mass Spectrometry*, 10, 1371-1378.

Klose, J. & U. Kobalz, (1995). Two-dimensional electrophoresis of proteins: An updated protocol and implications for a functional analysis of the genome. *Electrophoresis*, 16, 1034-1059.

Matsumura, H., S. Nirasawa & R. Terauchi, R. (1999). Transcript profiling in rice (Oryza sativa L.) seedlings using serial analysis of gene expression (SAGE). *Plant Journal*, 20, 719-726.

O'Farrell, P.H. (1975). High resolution two-dimensional electrophoresis. *Journal of Biological Chemistry*, 250, 4007-4021.

Tadege, M., Ratet, P. & K.S. Mysore (2005). Insertional mutagenesis: a Swiss Army knife for functional genomics of *Medicago truncatula*. *Trends in Plant Science*, in press.

Thiellement, H., N. Bahrman, C. Damerval, C. Plomion, M. Rossignol, V. Santoni, D. Devienne, & M. Zivy (1999). Proteomics for genetic and physiological studies in plants. *Electrophoresis*, 20, 2013-2026.

Wolf, B.P., L.W. Sumner, S.J. Shields, K. Nielsen, K.A. Gray & D.H. Russell, D.H. (1998). Characterization of proteins utilized in the desulfurization of petroleum products by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. *Analytical Biochemistry*, 260, 117-127.

Yates, J.R. (1998). Mass spectrometry and the age of the proteome. *Journal of Mass Spectrometery*, 33, 1-19.