

## TECHNICAL ADVANCE

# A gene expression atlas of the model legume *Medicago truncatula*

Vagner A. Benedito<sup>1</sup>, Ivone Torres-Jerez<sup>1</sup>, Jeremy D. Murray<sup>1</sup>, Andry Andriankaja<sup>1</sup>, Stacy Allen<sup>1</sup>, Klementina Kakar<sup>2</sup>, Maren Wandrey<sup>2</sup>, Jérôme Verdier<sup>3</sup>, Hélène Zuber<sup>3</sup>, Thomas Ott<sup>4</sup>, Sandra Moreau<sup>4</sup>, Andreas Niebel<sup>4</sup>, Tancred Frickey<sup>5</sup>, Georg Weiller<sup>5</sup>, Ji He<sup>1</sup>, Xinbin Dai<sup>1</sup>, Patrick X. Zhao<sup>1</sup>, Yuhong Tang<sup>1</sup> and Michael K. Udvardi<sup>1,\*</sup>

<sup>1</sup>Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA,

<sup>2</sup>Max-Planck Institute of Molecular Plant Physiology, Am Muhlenberg 1, 14476 Golm, Germany,

<sup>3</sup>INRA-URLEG, Unite de Recherche sur les Legumineuses, BP 86510, F-21065 Dijon Cedex, France,

<sup>4</sup>INRA-CNRS, 31326 Castanet-Tolosan, France, and

<sup>5</sup>Research School of Biological Sciences, The Australian National University, GPO Box 475, Canberra, ACT 2601, Australia

Received 12 February 2008; accepted 13 March 2008.

\*For correspondence (fax +1 580 224 6692; e-mail mudvardi@noble.org).

## Summary

Legumes played central roles in the development of agriculture and civilization, and today account for approximately one-third of the world's primary crop production. Unfortunately, most cultivated legumes are poor model systems for genomic research. Therefore, *Medicago truncatula*, which has a relatively small diploid genome, has been adopted as a model species for legume genomics. To enhance its value as a model, we have generated a gene expression atlas that provides a global view of gene expression in all major organ systems of this species, with special emphasis on nodule and seed development. The atlas reveals massive differences in gene expression between organs that are accompanied by changes in the expression of key regulatory genes, such as transcription factor genes, which presumably orchestrate genetic reprogramming during development and differentiation. Interestingly, many legume-specific genes are preferentially expressed in nitrogen-fixing nodules, indicating that evolution endowed them with special roles in this unique and important organ. Comparative transcriptome analysis of *Medicago* versus *Arabidopsis* revealed significant divergence in developmental expression profiles of orthologous genes, which indicates that phylogenetic analysis alone is insufficient to predict the function of orthologs in different species. The data presented here represent an unparalleled resource for legume functional genomics, which will accelerate discoveries in legume biology.

**Keywords:** *Medicago truncatula*, transcriptome, development, nodule, seed.

## Introduction

Legumes (family Fabaceae) are second only to grasses (Gramineae) in importance to humans as a source of food, feed for livestock, and raw materials for industry (Graham and Vance, 2003). Legumes account for one-third of the world's primary crop production and are key to sustainable agriculture because they can 'fix' nitrogen (reduce N<sub>2</sub> to NH<sub>3</sub>) in a symbiotic association with bacteria called rhizobia, providing crops with a free and renewable source of nitrogen. It is estimated that 40–60 million tonnes of N are fixed annually by cultivated legumes (Smil, 1999), saving about

US\$10 billion on nitrogen fertilizer (Graham and Vance, 2003).

Symbiotic nitrogen fixation (SNF) in legumes takes place in specialized organs called nodules which develop from root cortical cells that start dividing following signal exchanges between the plant roots and rhizobia in the soil (Brewin, 1991; Long, 2001; Oldroyd and Downie, 2004). Rhizobia gain entry into cortical cells of developing nodules via an infection thread, which traverses the epidermal root hair cell that makes first contact with the bacteria, and

subsequently ramifies throughout the cortical tissue. Concomitantly, dividing cortical cells form a nodule primordium inside which a new meristem is initiated that drives nodule organogenesis. Bacteria are then released from infection threads into the cytoplasm of cortical cells via endocytosis, which leaves the bacteria surrounded by a host membrane called the symbiosome membrane (Udvardi and Day, 1997). Within the resulting organelle, called the symbiosome, the bacteria multiply and ultimately differentiate into their nitrogen-fixing 'bacteroid' state. Infected cortical cells typically contain thousands of symbiosomes, each containing one or a few bacteroids. Transcriptome analyses have identified hundreds of plant and bacterial genes that are differentially expressed during nodule development and differentiation (Becker *et al.*, 2004; Colebatch *et al.*, 2002, 2004; El-Yahyaoui *et al.*, 2004; Mitra *et al.*, 2004; Uchiyama *et al.*, 2004), although genome-wide studies of plant gene expression during nodulation are yet to be reported. Likewise, although transcriptome studies have been carried out on different organs in multiple legume species under a variety of experimental conditions (Colebatch *et al.*, 2004; El-Yahyaoui *et al.*, 2004; Zabala *et al.*, 2006), no single study has yet brought together genome-wide data for all major plant organs of a single species.

## Results and discussion

To create a comprehensive gene expression atlas for *Medicago truncatula*, we utilized the new Affymetrix GeneChip Medicago Genome Array, which contains 50 900 probe sets representing the majority of genes in this species. Gene expression values were obtained from three independent biological replicates of each of the major organ systems: roots, nodules, stems, petioles, leaf blades, vegetative buds, flowers, and seed pods. In addition, multiple stages of nodule and seed development were profiled to obtain greater insight into the transcriptional programs that underlie the development of these two organs, which are the foci of most legume research. The fraction of genes for which transcripts were detected (called 'present' by dCHIP; Li and Wong, 2001) was similar in all samples, and ranged from 55.2% in seeds to 63.3% in roots (Figure S1). Similar results were obtained recently for *Arabidopsis* (Schmid *et al.*, 2005). However, *Arabidopsis* cannot nodulate and provides no information about symbiotic nitrogen fixation.

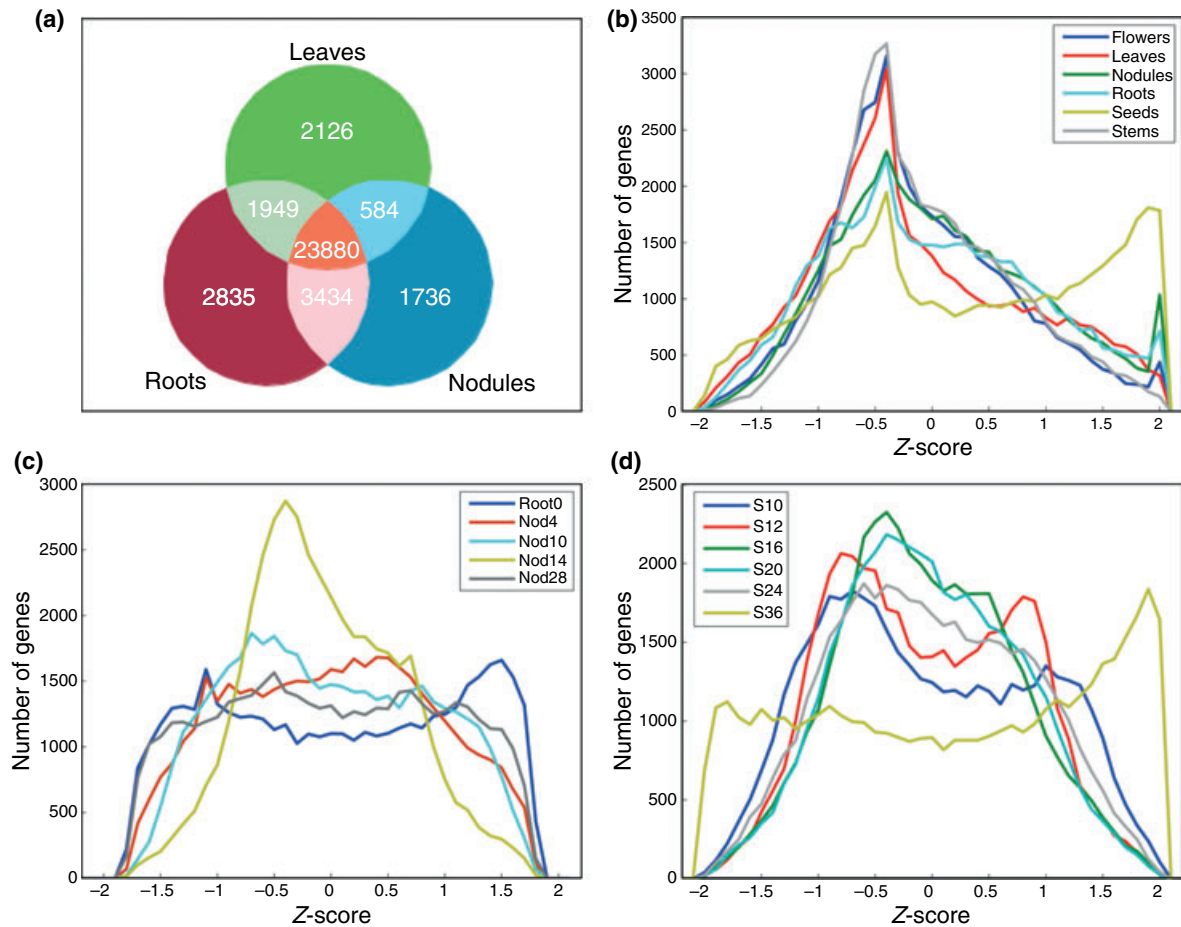
No transcripts were detected for 13.9% of putative genes in any of the organs tested. The majority of these correspond to genes annotated from genomic sequences by the International Medicago Genome Annotation Group (IMGAG; Town, 2006). No evidence for expression of 23.3% of IMGAG-annotated genes was found, corresponding to 4370 out of 18731 probe sets. It is possible that some of these have been incorrectly annotated as genes, although more thorough sampling of the transcriptome, by encom-

passing a wider range of developmental stages, growth, and stress conditions, together with more sensitive measurement devices such as quantitative (q)RT-PCR, will be required to conclude that such DNA is not transcribed under any conditions. In contrast to the results obtained from probe sets designed from IMGAG-annotated genes, only 8.4% of probe sets derived from expressed sequence tag (EST)/cDNA sequences did not detect gene transcripts in any of the organs tested, which presumably reflects the inherent bias in EST data towards more highly expressed genes.

Transcripts were detected in at least one organ for 86.1% of all genes represented by the 50 900 probe sets on the Medicago Genome Chip. There was a high degree of overlap in the sets of genes expressed in different organs: Transcripts of 42% of all expressed genes (36% of all probe sets; Figure S1) were detected in all organs, and this percentage increased in pair-wise comparisons between organs. For example, 79% of genes expressed in roots or nodules of 28-day-old plants were expressed in both organs (Figure 1a). Despite the qualitative similarities in gene expression amongst organs, the dynamics of gene expression differed markedly between organs and within organs over developmental time (Figure 1b–d). The majority of genes were subject to transcriptional and/or post-transcriptional regulation that altered transcript levels during plant development. In fact, at least 73% of all genes exhibited a >100% change in transcript level from the organ with lowest expression to the organ with highest expression. The mean coefficient of variance (CV = standard deviation/mean) of transcript levels for all expressed genes across all organs was 60.6%, ranging from 2.3% to 428.6%, while the mean CV for the three biological replicates of each organ was only 13.3%. In other words, the biological variation in gene transcript levels within an organ, including technical errors associated with measurement, was far less than the biological variation between organs.

Similarity between the transcriptomes of different organs was estimated using Pearson correlation, taking into account all genes expressed in at least one organ. The resulting heatmap of correlations revealed three main groups of organs with similar transcriptomes (Figure 2). The first group consisted of the underground organs, roots and nodules, the second group included seeds at different developmental stages, and the third group contained aerial organs, including leaves, stems, petioles, shoot apices, pods, and flowers. The aerial organs were more closely related to seeds than to underground organs.

Consistent with the large degree of overlap between expressed genes in the various organs, relatively few genes were found to be expressed in an organ-specific manner (Figure 3). The number of organ-specific genes identified in samples from 4-week-old plants ranged from 10 for stems to 322 for nodules. These numbers increased when data were obtained from multiple stages during development, as



**Figure 1.** Gene expression dynamics during *Medicago* development.

(a) Comparison of gene expression (detected transcripts) in roots, nodules, and leaves of 28-day-old plants.

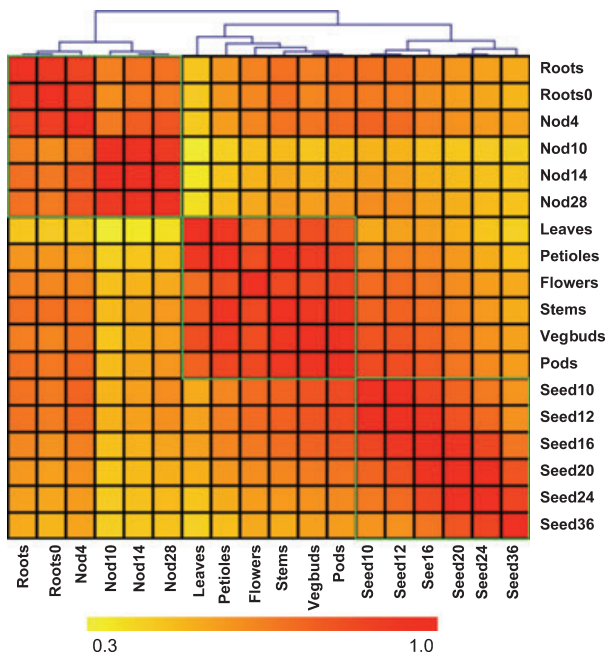
(b–d) Relative gene expression levels (Z scores) in different organs at one developmental stage (b), and at multiple stages for nodules (c) and seeds (d). Transcript levels were  $\log_2$ -transformed before calculation of  $Z = (X - X_{av})/SD$ ; where  $X$  is the mean transcript level for a given gene in the specified organ,  $X_{av}$  is the average transcript level for that gene in all organs, and  $SD$  is the standard deviation of transcript level for that gene across all organs. The number of genes was determined for each  $\Delta Z = 0.1$ . Nod4–Nod28 represent nodules harvested 4–28 days after inoculation with *Sinorhizobium meliloti*, while Root0 represents root tissue immediately prior to inoculation. S10–S36 represent seed harvested from 10 to 36 days after pollination.

exemplified by nodules and seeds in which 473 and 584 genes, respectively, were identified as organ-specific from the corresponding developmental series (Table S1). In some cases, organ-specific genes were expressed only transiently during development, suggesting to us that they play roles in development *per se*, rather than in the maintenance of specialized biochemical or physiological functions of each organ, at least under ideal growth conditions.

A more complex picture emerged for genes that were expressed in specific organs at a level at least twice that of any other organ (Figure S2). The number of such genes ranged from 40 in petioles to 908 in roots of 4-week-old plants. Even larger numbers of organ-induced genes were uncovered in the developmental time-course for nodules (1354 genes) and seeds (3228 genes; Table S2). Interestingly, transcript levels of many of the genes induced during nodule or seed development were maintained at relatively high levels in the mature organ, suggesting that they may

play roles in differentiation or the maintenance of specialized organ functions.

To identify all genes that are subject to transcriptional or post-transcriptional regulation during the development of *Medicago*, we chose roots as an arbitrary reference organ and tested the null hypothesis that expression in other organs was not significantly different from that in roots. Using a Bonferroni-corrected (Abdi, 2007)  $P$ -value threshold of  $1.14 \times 10^{-6}$ , the percentage of genes expressed at a different level from roots ranged from 46.7% in nodules to 55.9% in leaves. In total, 81.5% of genes were differentially expressed between roots and one or more of the other seven organs (Table S3). The false discovery rate of these genes was estimated by determining  $Q$ -values for each, using EDGE software (Leek *et al.*, 2006; Storey and Tibshirani, 2003). Clearly, organ development, differentiation, and maintenance in plants are underpinned by massive quantitative changes in gene expression.

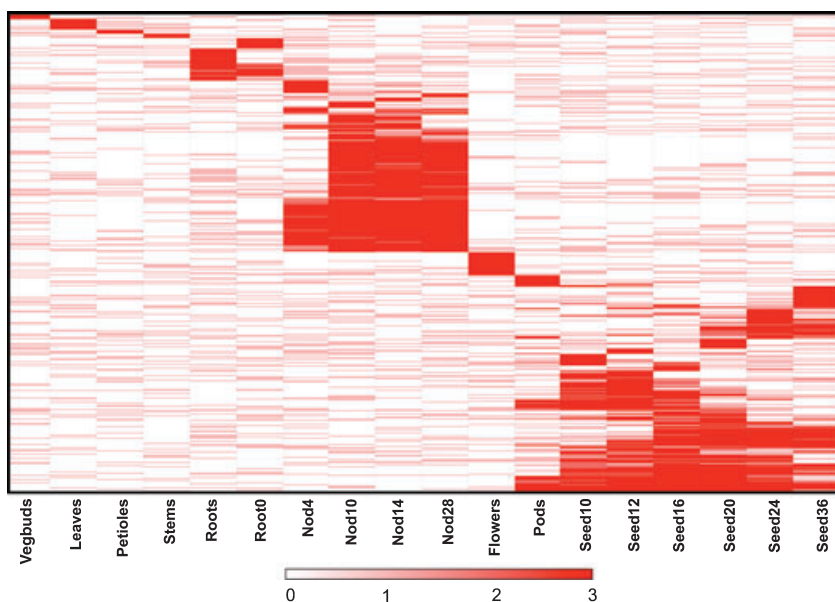


**Figure 2.** Comparison of transcriptomes of various organs. Pair-wise Pearson correlation coefficients were used to generate the heat map. The color scale indicates the degree of correlation. Samples were clustered with Euclidean distance using the MultiExperiment Viewer (MeV, <http://www.tm4.org/mev.html>).

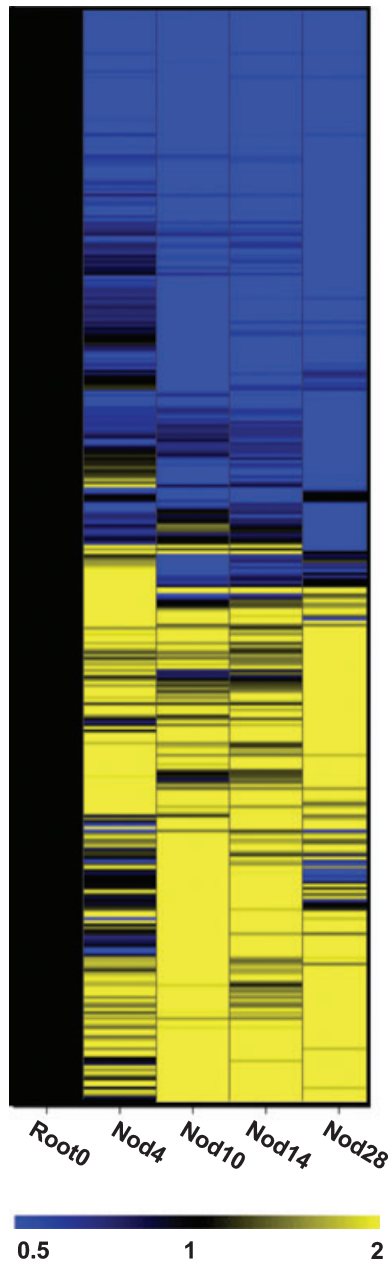
Genes with a constant expression level throughout development and in the face of environmental challenges, which often fulfill housekeeping roles in cells, are useful reference points for comparative gene expression analysis (Czechowski *et al.*, 2005). We identified 102 genes with <16% coefficient of variance for transcript levels amongst all the

organs analyzed (Table S4). Transcript levels of these genes ranged widely, from as high as 14 500 to as low as 100 units, which was used as the minimum threshold level. Amongst the stably expressed genes were several that are used traditionally as reference genes in plants, including glyceraldehyde-3-P-dehydrogenase and ubiquitin. The most stably expressed gene, corresponding to TC97716 with unknown function, had a CV for transcript level amongst all organs of 9.4%. This set of reference genes will be of great value to legume research for normalizing gene expression data derived from qRT-PCR or probe-hybridization approaches.

Symbiotic nitrogen fixation in plants is a process confined largely to the legume (Fabaceae) family. Therefore, well-established, non-legume model species such as *Arabidopsis thaliana* and *Oryza sativa* (rice) cannot be used to learn more about SNF. The data presented here represent the most comprehensive data set to date for gene expression during nodule development in a legume. More than 26 000 genes are expressed during nodule development and 30.2% of these are differentially expressed (transcript levels increase or decrease more than twofold compared with roots with Bonferroni-corrected  $P < 1.14 \times 10^{-6}$ ) at some stage during this development (Figure 4 and Table S5). Visualization of the nodule development data, using MAPMAN to overlay changes in gene expression onto metabolic maps (Goffard and Weiller, 2006; Thimm *et al.*, 2004), confirmed and extended previous, smaller-scale transcriptomics studies (El-Yahyaoui *et al.*, 2004) that showed induction during nodule development of genes involved in glycolysis, carbon fixation, and nitrogen metabolism (Figure S3a–c). Many genes involved in secondary metabolism, such as the terpenoid and flavonoid pathways, were repressed during nodule development (Figure S3d). These results are



**Figure 3.** Heat map of organ-specific genes. The color scale indicates the number of times transcripts for a given gene were detected in the three biological replicates of each organ. Only those genes are shown for which transcripts were detected in all three biological replicates of one organ and no more than once in another organ.



**Figure 4.** Hierarchical clustering of genes that were differentially expressed during nodule development. Clustering of genes was based on Pearson correlation. The heat map portrays transcript levels in nodules 4, 10, 14, and 28 days after inoculation (Nod4–Nod28) relative to that in uninoculated roots (Root0).

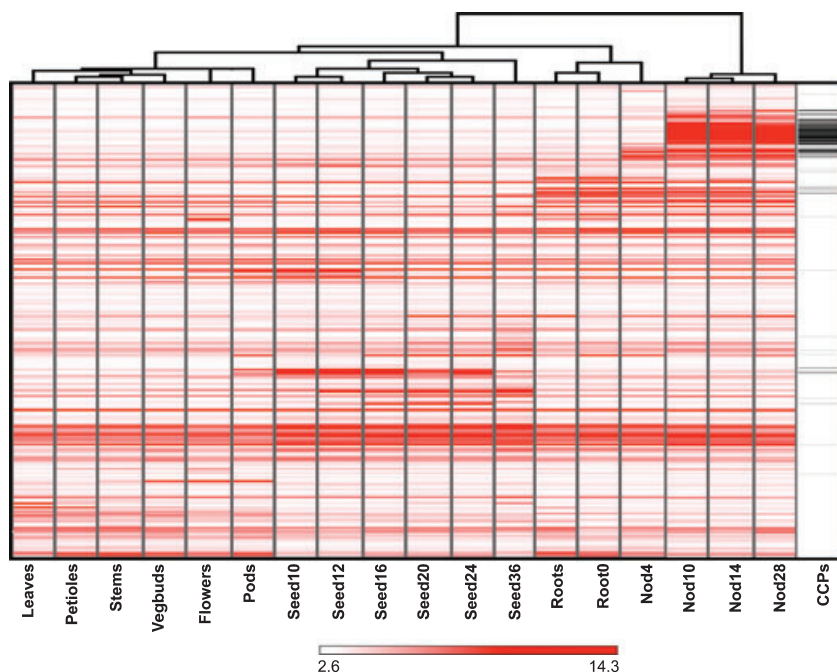
interesting in light of the fact that many secondary compounds play roles in plant defense (Dixon, 2001), a process that is presumably suppressed in nodules in order to maintain a quasi-stable symbiosis with the nitrogen-fixing rhizobia. We provide the nodule development data described here in a form that can be imported into MAPMAN for exploration by the reader of other pathways and processes (Table S6).

The massive changes in transcript abundance that occur during the development of nodules and other organs indicate tremendous regulatory activity at the transcriptional and/or post-transcriptional levels. Transcription factors (TFs) are DNA-binding proteins that interact with specific *cis*-elements of genes to regulate transcription, either positively or negatively. Plants such as *Arabidopsis* may possess as many as 2000 TF genes, representing more than 6% of all their genes (Riechmann and Ratcliffe, 2000; Riechmann *et al.*, 2000). To identify TF genes that control development and differentiation in *Medicago*, we first created a list of putative TF genes by screening predicted protein sequences for the presence of known or suspected DNA-binding domains, using InterPro (<http://www.ebi.ac.uk/interpro/>) and Pfam (<http://pfam.sanger.ac.uk/>) domain identification, additional hidden Markov model (HMM) predictions (Guo *et al.*, 2005; Sonnhammer *et al.*, 1997), and a BLASTX search of the NCBI NR database to support annotations. In this way, we previously identified 1298 putative TF genes represented by probe sets on the Affymetrix *Medicago* GeneChip (Udvardi *et al.*, 2007). Most of these (1169) fall into the 45 known families of plant TF genes or other transcriptional regulators, while 129 may define novel TF families in plants (Table S7). Five hundred and thirty-two of the putative TF genes are differentially expressed (more than twofold change;  $P < 1.14 \times 10^{-6}$ ) during nodule development and may therefore play important roles in SNF (Table S5). The vast majority (>1100) of TFs are differentially expressed in other organs (Table S3). These data are a rich source of information and a sound platform for future experimental work aimed at unraveling genetic regulatory networks that govern organ development in *Medicago*.

The ability to form a nitrogen-fixing symbiosis appears to have evolved relatively recently in land plants, approximately 65 million years ago (Doyle, 1998), and as a result SNF is restricted to legumes and a few non-legume species. Interestingly, some of the genes required to establish SNF in legumes appear to have been recruited from a more ancient set of genes that are required for mycorrhizal symbiosis (Kistner and Parniske, 2002). Mycorrhizal symbioses are believed to have evolved when plants first colonized land 450 million years ago (Redecker *et al.*, 2000; Remy *et al.*, 1994), and it has been suggested that these fungal symbionts served as an extension of the primitive plant root system. Indeed, mycorrhizal fungi extend the reach of plant roots and aid in plant nutrition, especially phosphorus uptake (Harrison, 1999). The ancient origin and importance of mycorrhizal symbioses to land plants is reflected by the fact that approximately 90% of all land plant species are able to form such symbioses. A number of genes, mostly encoding signaling proteins, have been discovered in legumes that are required for both mycorrhizal symbiosis and nodulation/SNF (Kistner and Parniske, 2002; Parniske, 2004). Additional genes are

essential for SNF but not for mycorrhizal symbiosis (Kalo *et al.*, 2005; Radutoiu *et al.*, 2003; Schauser *et al.*, 1999; Smit *et al.*, 2005). Each of these additional genes has so far been found to have one or more homologs in non-legume, non-nitrogen-fixing plant species, such as Arabidopsis. This raises an important question: Were all genes required for nodule development and SNF simply recruited from a pre-existing stock of plant genes, or did novel genes evolve as a result of natural selection for SNF? Legume-specific genes (LSGs) that appear to be absent from the genomes of non-legumes have been identified in a number of species. Several classes of LSGs have been identified in Medicago that encode short proteins, including over 300 cysteine cluster proteins (CCPs; Alunni *et al.*, 2007; Fedorova *et al.*, 2002; Graham *et al.*, 2004; Mergaert *et al.*, 2003), 63 proline-rich proteins (PRPs; Graham *et al.*, 2004; Sherrier *et al.*, 2005) and 21 glycine-rich proteins (GRPs; Alunni *et al.*, 2007; Kevei *et al.*, 2002; Silverstein *et al.*, 2006). Five thousand eight hundred and forty-two probe sets representing LSGs were identified on the Medicago GeneChip. This included sequences for 355 CCPs, 5 PRPs, and 7 GRPs (Table S8). Analysis of LSG expression revealed that a subset of 322 CCPs (called NCRs for nodule-specific cysteine rich) and all seven GRPs were expressed in a nodule-specific manner consistent with roles for these genes in nodule development and/or function (Figure 5 and Table S8). Some PRPs were expressed in nodules but were also detected in other tissues such as seeds or flowers (Figure 5). These results confirm and extend earlier work on LSGs (Alunni *et al.*, 2007; Fedorova *et al.*, 2002; Graham *et al.*, 2004; Mergaert *et al.*, 2003).

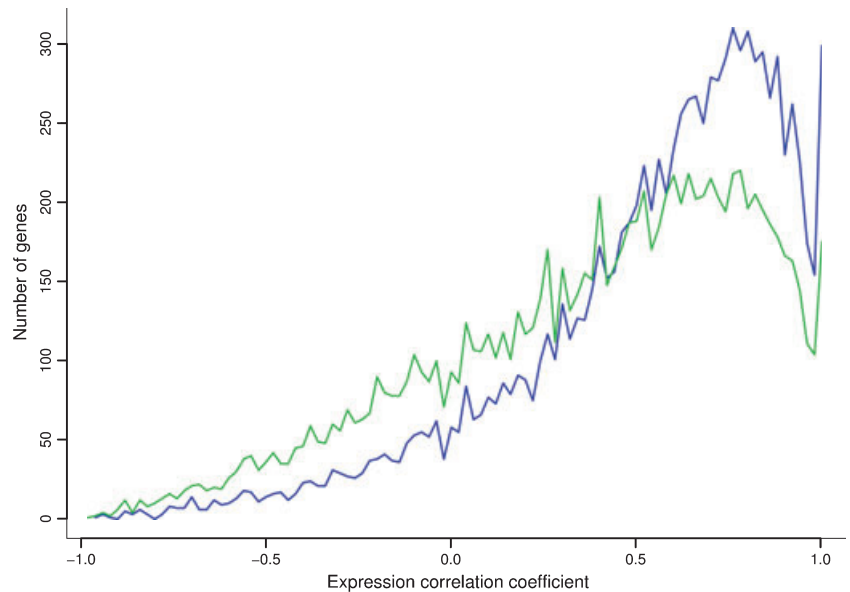
Large data sets of the type presented here for Medicago and elsewhere for different plant species, such as Arabidopsis (Schmid *et al.*, 2005), enable us to address other important questions about gene evolution in plants. One of these questions relates to the conservation of gene function in different plant lineages. To gain insight into the possible role(s) of a gene in a crop species, for instance, plant biologists often turn to a related model species, such as Arabidopsis, rice, or Medicago, and ask: what is the function of the ortholog(s) of the crop gene(s)/protein(s) in the model species? The implicit assumption is that orthologs in different species perform similar, if not identical, physiological functions despite millions of years of evolution. For this to be true, orthologous genes must have similar expression profiles in the two organisms. To test whether this is generally the case, we identified homologs of Medicago genes in Arabidopsis and made pair-wise comparisons of gene expression between the two species, using matching data from roots, stems, leaves, flowers, seeds, petioles, and vegetative buds. Pearson correlation analysis was used to rank Arabidopsis homologs based on the similarity of gene expression profiles in the various organs of the two species. The top-ranking homolog determined from correlation analysis of gene expression matched the putative ortholog, based on phylogenetic analysis in only 62% of cases (Figure 6 and Table S9). Thus, transcriptional regulation of putative orthologs has diverged substantially between Medicago and Arabidopsis, indicating that many orthologs may not perform the same range of functions in these two species. It should be noted that the plant growth conditions and developmental states of Medicago and Arabidopsis



**Figure 5.** Expression of legume-specific genes in Medicago organs.

Genes were clustered based on Pearson correlation. Cysteine cluster proteins (CCPs) are indicated in the far-right column. The color scale shows  $\log_2$ -transformed transcript levels for each gene.

**Figure 6.** Correlation between expression profiles of homologous and orthologous genes in *Arabidopsis thaliana* and *Medicago truncatula*. Normalized transcript levels of roots, stems, leaves, flowers, seeds, petioles, and vegetative buds were compared between the two species. Pearson (linear) correlation coefficients were determined for all pairs of sequences regarded as either homologous or orthologous between *Medicago* and *Arabidopsis*. Histograms of the number of sequences (Y-axis) over the correlation coefficient (X-axis) of *Medicago* sequences and the corresponding *Arabidopsis* sequences are shown. The blue line represents the best-correlating *Arabidopsis* sequence identifiable within the set of sequence homologs for each *Medicago* sequence. The green line represents the best-correlating *Arabidopsis* sequence identifiable within the set of putative sequence orthologs for each *Medicago* sequence.



organs compared here were not identical, which would tend to decrease the correlation between gene expression patterns in the two species. Nonetheless, given that the expression patterns of paralogs often exhibited higher correlation than predicted orthologs (Figure 6), phylogenetic analysis alone may yield inaccurate predictions for the physiological functions of many genes, at least for comparisons between families as divergent as legumes and crucifers. This underscores the importance of the gene expression data collected here as a tool for *Medicago* and legume functional genomics.

In summary, we have produced a comprehensive gene expression atlas for the model legume *M. truncatula*, which encompasses all organs of this species, including detailed time-courses through nodule and seed development. In addition to being a rich source of information for legume biologists, this data set enables large-scale comparisons between the transcriptomes of different plant species. Analysis of the data presented here shows that differences between plant organs result mainly from quantitative rather than qualitative changes in global gene expression. Relatively few genes are organ-specific in this species. Amongst these are subsets of legume-specific genes that appear to be expressed exclusively or preferentially in nodules. This implies that evolution of symbiotic nitrogen fixation was accompanied, and possibly facilitated, by the evolution of novel genes in legumes. These genes are clear targets for future studies aimed at identifying the core set of genes required for SNF in legumes. Finally, comparative functional genomics studies of the type presented here for *Medicago* and *Arabidopsis* will add a new dimension to studies of gene evolution in plants and other organisms.

We plan to expand the *Medicago* Gene Expression Atlas to encompass transcript data from wild-type and mutant plants exposed to various biotic and abiotic challenges, and to increase the spatial resolution of expression data by analyzing specific tissues and cell types. We invite the scientific community to collaborate with us in this venture by submitting raw *Medicago* Affymetrix data from complementary experiments together with metadata describing the experimental material, either directly to us (contact [mudvardi@noble.org](mailto:mudvardi@noble.org)) or to ArrayExpress (<http://www.ebi.ac.uk/miamexpress/>).

## Experimental procedures

### *Plant material, RNA isolation, probe preparation and array hybridization*

*Medicago truncatula* cv. Jemalong, line A17 seeds were scarified with concentrated sulfuric acid, rinsed, sterilized with 2% sodium hypochlorite, and vernalized at 4°C for 3 days on moist, sterile filter paper. Germinated seedlings were transplanted to pots containing Surface MVP calcined (illite) clay (Profile Products, <http://www.profileproducts.com/>) and placed in a growth chamber set to the following conditions: 16-h/8-h light/dark regime, 200  $\mu\text{E m}^{-2} \text{sec}^{-1}$  light irradiance, 24°C and 40% relative humidity. Plants were fertilized daily with half-strength B&D solution (Broughton and Dilworth, 1971) containing 2 mM  $\text{KNO}_3$  and 2 mM  $\text{NH}_4\text{NO}_3$ . A subset of plants were inoculated with *Sinorhizobium meliloti* strain 1021 at 1 and 7 days after sowing and fertilized with half-strength B&D solution containing 0.5 mM  $\text{KNO}_3$ . Vegetative organs (roots, stems, petioles, leaves, and shoot buds from uninoculated plants and nodules from inoculated plants) were harvested from multiple plants 28 days after planting and pooled for individual biological replicates in a completely randomized design. All experiments were performed with three biological replicates planted on separate days. For flowers and pods, plants were vernalized for 2 weeks to decrease

time to flowering. Flowers were harvested on the first day that they opened fully, while pods were collected at various stages of development (length ranged from 2.5 to 9.0 mm) within 21 days after the appearance of the floral bud. Harvesting of all organs occurred at the same time each morning, approximately 3 h after 'dawn', to avoid as far as possible diurnal changes in gene expression. All harvested material was frozen immediately in liquid nitrogen and stored at  $-80^{\circ}\text{C}$  prior to RNA isolation. Total RNA was extracted using TRIzol reagent (Invitrogen, <http://www.invitrogen.com/>; Chomczunski and Mackey, 1995), treated with DNaseI (Ambion, <http://www.ambion.com/>), and column purified with a RNeasy MinElute CleanUp Kit (Qiagen, <http://www.qiagen.com/>).

Material for the nodule developmental series was harvested from plants grown aeroponically at  $22^{\circ}\text{C}$ , 75% hygrometry, a light irradiance of  $200\ \mu\text{E m}^{-2}\ \text{sec}^{-1}$ , and a 16-h/8-h light/dark regime. Plants were grown initially for 11 days using a nitrogen-rich medium (Journet *et al.*, 2001) then deprived of nitrogen for 4 days before being inoculated with *S. meliloti* strain 2011. Roots were harvested 0, 4, 10, and 14 days post-inoculation and nodules were dissected from roots prior to freezing in liquid nitrogen, storage at  $-80^{\circ}\text{C}$ , and RNA isolation using a Nucleospin RNA kit (Macherey-Nagel, <http://www.macherey-nagel.com/>).

Material for the seed developmental series was harvested from plants grown in pots containing attapulgate (50%) and clay beads (50%) at  $22^{\circ}\text{C}/19^{\circ}\text{C}$  day/night, 16-h photoperiod at  $220\ \mu\text{E m}^{-2}\ \text{sec}^{-1}$  light irradiance, and 60–70% relative humidity. Plants were fertilized with nutrient solution three times a week and watered on intervening days. Seeds were excised from pods 10, 12, 16, 20, 24, and 36 days after pollination, frozen in liquid nitrogen, and stored at  $-80^{\circ}\text{C}$  prior to RNA isolation (Chang *et al.*, 1993).

Ribonucleic acid was quantified using a Nanodrop Spectrophotometer ND-100 (NanoDrop Technologies, <http://www.nanodrop.com/>) and evaluated for purity with a Bioanalyzer 2100 (Agilent, <http://www.home.agilent.com/>). The Affymetrix GeneChip® Medicago Genome Array (Affymetrix, <http://www.affymetrix.com/>) was used for expression analysis. The RNA from three independent biological replicates was analyzed for each organ/developmental stage. Probe labeling using  $10\ \mu\text{g}$  RNA, array hybridization and scanning were performed according to the manufacturer's instructions (Affymetrix) for eukaryotic RNA, using a one-cycle protocol for cDNA synthesis.

#### Data extraction and normalization

For each Affymetrix array hybridized, the resulting .cel file was exported from GeneChip Operating Software Version 1.4 (Affymetrix) and imported into Robust Multiarray Average (RMA; Irizarry *et al.*, 2003) for global normalization. Presence/absence call for each probe set was obtained using dCHIP (Li and Wong, 2001). Gene selections based on an associative *t*-test (Dozmorov and Centola, 2003) were made using Matlab (MathWorks, <http://www.mathworks.com/>). Using this method, the background noise presented between replicates and the technical noise generated during hybridization were measured by the residual presented among a group of genes whose residuals are homoscedastic within the control group. Only genes whose residuals between compared sample pairs are significantly higher than the measured noise level will be considered to be differentially expressed. Since the residual was obtained from thousands of genes on the chip, the *P*-values obtained by this method are corrected towards a large sampling size, thus enabling the use of Bonferroni corrections without being overly stringent. The advantage of this methodology is that it takes into consideration technical noise and internal variations between replicates within a sample group and provides a baseline for

selecting biologically significant genes. A selection threshold of two for transcript ratios (where applicable) and a Bonferroni-corrected *P*-value threshold of  $1.14 \times 10^{-6}$  were used. Bonferroni-corrected  $P = 0.05/N$ , where *N* is the number of genes in the comparison, which was 43 836 in the experiments reported here. To monitor the false discovery rate of differentially expressed genes, the *Q*-value of each gene was obtained by EDGE software (Leek *et al.*, 2006; Storey and Tibshirani, 2003).

#### Z scores

The Z score was calculated as follows:  $Z = (X - X_{\text{av}})/\text{SD}$ ; where *X* is the  $\log_2$ -transformed mean transcript level for a given gene in a specific organ,  $X_{\text{av}}$  is the  $\log_2$ -transformed mean transcript level for that gene in all organs, and SD is the standard deviation of transcript level for that gene across all organs. Mean transcript levels were determined from three biological replicates of each organ.

#### Hierarchical clustering analysis (HCA)

Hierarchical clustering analysis was conducted with Spotfire DecisionSite 8.1 (Spotfire Inc., <http://spotfire.tibco.com/>). For clustering analysis of data from different organs, data were transformed into  $\log_2$  and clustered using Pearson correlation analysis (Zar, 1999). For the nodule developmental series, transcript levels were expressed relative to the level in roots just prior to inoculation with rhizobia (Root 0) before constructing clusters using the Pearson correlation coefficient.

#### Legume-specific genes

Legume-specific genes (LSGs) represented on the Affymetrix Medicago GeneChip were identified by a series of in-house BLAST searches that were used to eliminate probe sets representing sequences with homology to any non-legume plant sequence in GenBank. Briefly, starting with all 50 900 sequences upon which the Medicago GeneChip was based, consecutive BLAST searches were used to filter out homologs from *O. sativa*, *A. thaliana*, *Populus trichocarpa*, and *Chlamydomonas reinhardtii*. Finally, using the reduced list, a final search against all remaining non-legume sequences in GenBank NR and EST databases was made using the BLAST Network Client 'BLASTcl3'. (<ftp://ftp.ncbi.nlm.nih.gov/blast/executables/>). For all these searches both TBLASTX and BLASTN (*E*-value cutoff of  $\leq 1 \times 10^{-4}$ ) were used. Subsets of LSGs (CCPs and GRPs) represented on the Affymetrix Medicago GeneChip were identified based on homology to known family members using a TBLASTX search (match criteria: *E*-value  $\leq 1 \times 10^{-5}$ ). Probe sets corresponding to known PRPs were identified based on perfect or near perfect matches from a BLASTN search employing a complexity filter.

#### Correlation analysis of expression profiles for homologous genes in Medicago and Arabidopsis

All *M. truncatula* consensus sequences (<http://www.affymetrix.com/support/technical/byproduct.affx?product=medicago>), the sequences from which the probe sets for the corresponding Affymetrix chip were derived, were translated into protein in all six frames. The longest open reading frame (ORF) for each sequence was compared, using BLASTP, against six-frame translations of the consensus sequences of other Affymetrix chips as well as the NCBI non-redundant sequence database 'nr'. Consensus sequences for the *M. truncatula* and Arabidopsis chips were accepted as sequence

homologs if BLASTP hits connected the sequences with *E*-values better than  $1 \times 10^{-30}$  and bit-scores greater than 150. Putative phylogenetic orthologs between *M. truncatula* and *A. thaliana* were identified using AffyTrees (Frickey *et al.*, 2008). AffyTrees is based on PhyloGenie (Frickey and Lupas, 2004) and provides a repository of neighbor-joining (Saitou and Nei, 1987) trees for Affymetric consensus sequences in plants.

To compare the gene expression of homologous sequences, *A. thaliana* microarray data for stems (ATGE\_27), petioles (ATGE\_19), leaves (ATGE\_14), vegetative buds (ATGE\_8), flowers (ATGE\_39), roots (ATGE\_9), and seeds (ATGE\_79; Schmid *et al.*, 2005) were compared against the corresponding organs provided by this atlas (stem, petiole, leaf, vegetative bud, flower, root, and seed20d). All expression data were normalized using GCRMA (Wu *et al.*, 2004). The Pearson (linear) correlation coefficient of the expression values was calculated for all pairs of sequence homologs and used to quantify the similarity of expression of homologous sequences for the two species. As the number of *Medicago* consensus sequences for which BLAST and AffyTrees could determine homologs or orthologs differed, we restricted the analysis to the 10 243 sequences for which both methods were able to produce results. Sequence pairs between *M. truncatula* and *A. thaliana* with the highest correlation coefficient within the set of sequence homologs, as determined by BLASTP, and the set of putative orthologs, as determined by AffyTrees, were compared.

Our project web site is <http://bioinfo.noble.org/gene-atlas/>. All gene expression data have been deposited in the ArrayExpress Database (<http://www.ebi.ac.uk/miameexpress/>) under accession number E-MEXP-1097.

### Acknowledgements

We thank the USDA CSREES-NRI, the Samuel Roberts Noble Foundation, the Max Planck Society, the European Union FP6 Program, and the Australian Research Council Centre for Integrative Legume Research for support of this work.

### Supplementary Material

The following supplementary material is available for this article online:

**Figure S1.** Fraction of genes expressed in different organs.

**Figure S2.** Heat map of organ-induced genes. Transcripts levels for each of these genes were at least twice as high in one organ as in any other organ.

**Figure S3.** Transcriptional dynamics during nodule development of genes encoding enzymes involved in: (a) glycolysis, (b) carbon fixation, (c) nitrogen metabolism; and (d) flavonoid biosynthesis.

**Table S1.** List of organ-specific genes and their presence calls in all organs.

**Table S2.** List of organ-induced genes and their transcript levels in all organs.

**Table S3.** Differentially expressed genes in *Medicago*.

**Table S4.** Stably expressed 'reference' genes for transcript normalization in *Medicago*.

**Table S5.** Genes differentially expressed during nodule development.

**Table S6.** Nodule developmental data for visualization with MAPMAN software.

**Table S7.** Differential expression of transcription factor genes in various organs of *Medicago*.

**Table S8.** Legume-specific genes and their expression level in different organs.

**Table S9.** Information about the various BLAST hits (better than  $1 \times 10^{-30}$  and bit-scores of 150) as well the sequences present in the AffyTrees phylogenies and the corresponding Pearson (linear) correlation of the expression values.

This material is available as part of the online article from <http://www.blackwell-synergy.com>.

Please note: Blackwell publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

### References

- Abdi, H. (2007) Bonferroni and Sidak corrections for multiple comparisons. In *Encyclopedia of Measurement and Statistics* (Salkind, N.J., ed.). Thousand Oaks: Sage, pp. 103–107.
- Alunni, B., Kevei, Z., Redondo-Nieto, M., Kondorosi, A., Mergaert, P. and Kondorosi, E. (2007) genomic organization and evolutionary insights on GRP and NCR genes, two large nodule-specific gene families in *Medicago truncatula*. *Mol. Plant Microbe Interact.* **20**, 1138–1148.
- Becker, A., Berges, H., Krol, E. *et al.* (2004) Global changes in gene expression in *Sinorhizobium meliloti* 1021 under microoxic and symbiotic conditions. *Mol. Plant Microbe Interact.* **17**, 292–303.
- Brewin, N.J. (1991) Development of the legume root nodule. *Annu. Rev. Cell Biol.* **7**, 191–226.
- Broughton, W.J. and Dilworth, M.J. (1971) Control of leghaemoglobin synthesis in Snake Beans. *Biochem. J.* **125**, 1075–1080.
- Chang, S., Puryear, J. and Cairney, J. (1993) A simple and efficient method for isolating RNA from pine trees. *Plant Mol. Biol. Rep.* **11**, 113–116.
- Chomczunski, P. and Mackey, K. (1995) Modification of the TRI TM Reagent procedure for isolation of RNA from polysaccharide- and proteoglycan-rich sources. *Biotechniques*, **19**, 942–945.
- Colebatch, G., Trevaskis, B. and Udvardi, M. (2002) Symbiotic nitrogen fixation research in the postgenomics era. *New Phytol.* **153**, 37–42.
- Colebatch, G., Desbrosses, G., Ott, T., Krusell, L., Montanari, O., Kloska, S., Kopka, J. and Udvardi, M.K. (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in *Lotus japonicus*. *Plant J.* **39**, 487–512.
- Czechowski, T., Stitt, M., Altmann, T., Udvardi, M.K. and Scheible, W.R. (2005) Genome-wide identification and testing of superior reference genes for transcript normalization in *Arabidopsis*. *Plant Physiol.* **139**, 5–17.
- Dixon, R.A. (2001) Natural products and plant disease resistance. *Nature*, **411**, 843–847.
- Doyle, J.J. (1998) Phylogenetic perspectives on nodulation: an evolving view of plants and symbiotic bacteria. *Trends Plant Sci.* **3**, 473–478.
- Dozmorov, I. and Centola, M. (2003) An associative analysis of gene expression array data. *Bioinformatics*, **19**, 204–211.
- El-Yahyaoui, F., Küster, H., Amor, B.B. *et al.* (2004) Expression profiling in *Medicago truncatula* identifies more than 750 genes differentially expressed during nodulation, including many potential regulators of the symbiotic program. *Plant Physiol.* **136**, 3159–3176.
- Fedorova, M., van de Mortel, J., Matsumoto, P.A., Cho, J., Town, C.D., VandenBosch, K.A., Gantt, J.S. and Vance, C.P. (2002) Genome-wide identification of nodule-specific transcripts in the model legume *Medicago truncatula*. *Plant Physiol.* **130**, 519–537.
- Frickey, T. and Lupas, A.N. (2004) PhyloGenie: automated phylogeny generation and analysis. *Nucleic Acids Res.* **32**, 5231–5238.

- Frickey, T., Benedito, V.A., Udvardi, M. and Weiller, G. (2008) AffyTrees: facilitating comparative analysis of Affymetrix plant microarray chips. *Plant Physiol.* **146**, 377–386.
- Goffard, N. and Weiller, G. (2006) Extending MapMan: application to legume genome arrays. *Bioinformatics*, **22**, 2958–2959.
- Graham, P.H. and Vance, C.P. (2003) Legumes: importance and constraints to greater use. *Plant Physiol.* **131**, 872–877.
- Graham, M.A., Silverstein, K.A.T., Cannon, S.B. and VandenBosch, K.A. (2004) Computational identification and characterization of novel genes from legumes. *Plant Physiol.* **135**, 1179–1197.
- Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L. and Luo, J. (2005) DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, **21**, 2568–2569.
- Harrison, M.J. (1999) Molecular and cellular aspects of the arbuscular mycorrhizal symbiosis. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 361–389.
- Irizarry, R.A., Hobbs, B. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.
- Journet, E.P., El-Gachtouli, N., Vernoud, V., de Billy, F., Pichon, M., Dedieu, A., Arnould, C., Morandi, D., Barker, D.G. and Gianinazzi-Pearson, V. (2001) Medicago truncatula ENOD11: a novel RPRP-encoding early nodulin gene expressed during mycorrhization in arbuscule-containing cells. *Mol. Plant Microbe Interact.* **14**, 737–748.
- Kalo, P., Gleason, C., Edwards, A. et al. (2005) Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators. *Science*, **308**, 1786–1789.
- Kevei, Z., Vinardell, J.M., Kiss, G.B., Kondorosi, A. and Kondorosi, E. (2002) Glycine-rich proteins encoded by a nodule-specific gene family are implicated in different stages of symbiotic nodule development in Medicago spp. *Mol. Plant Microbe Interact.* **15**, 922–931.
- Kistner, C. and Parniske, M. (2002) Evolution of signal transduction in intracellular symbiosis. *Trends Plant Sci.* **7**, 511–518.
- Leek, J.T., Monsen, E., Dabney, A.R. and Storey, J.D. (2006) EDGE: extraction and analysis of differential gene expression. *Bioinformatics*, **22**, 507–508.
- Li, C. and Wong, W. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl Acad. Sci. USA*, **98**, 31–36.
- Long, S.R. (2001) Genes and signals in the Rhizobium-legume symbiosis. *Plant Physiol.* **125**, 69–72.
- Mergaert, P., Nikovics, K., Kelemen, Z., Maunoury, N., Vaubert, D., Kondorosi, A. and Kondorosi, E. (2003) A novel family in *Medicago truncatula* consisting of more than 300 nodule-specific genes coding for small, secreted polypeptides with conserved cysteine motifs. *Plant Physiol.* **132**, 161–173.
- Mitra, R.M., Shaw, S.L. and Long, S.R. (2004) Six nonnodulating plant mutants defective for Nod factor-induced transcriptional changes associated with the legume-rhizobia symbiosis. *Proc. Natl Acad. Sci. USA*, **101**, 10217–10222.
- Oldroyd, G.E.D. and Downie, J.A. (2004) Calcium, kinases, and nodulation signalling in legumes. *Nat. Rev. Mol. Cell Biol.* **5**, 566–576.
- Parniske, M. (2004) Molecular genetics of the arbuscular mycorrhizal symbiosis. *Curr. Opin. Plant Biol.* **7**, 414–421.
- Radutoiu, S., Madsen, L.H., Madsen, E.B. et al. (2003) Plant recognition of symbiotic bacteria requires two LysM receptor-like kinases. *Nature*, **425**, 585–592.
- Redecker, D., Kodner, R. and Graham, L.E. (2000) Glomalean fungi from the Ordovician. *Science*, **289**, 1920–1921.
- Remy, W., Taylor, T.N., Hass, H. and Kerp, H. (1994) Four hundred-million-year-old vesicular arbuscular mycorrhizae. *Proc. Natl Acad. Sci. USA*, **91**, 11841–11843.
- Riechmann, J.L. and Ratcliffe, O.J. (2000) A genomic perspective on plant transcription factors. *Curr. Opin. Plant Biol.* **3**, 423–434.
- Riechmann, J.L., Heard, J., Martin, G. et al. (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
- Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.
- Schauser, L., Roussis, A., Stiller, J. and Stougaard, J. (1999) A plant regulator controlling development of symbiotic root nodules. *Nature*, **402**, 191–195.
- Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D. and Lohmann, J.U. (2005) A gene expression map of Arabidopsis thaliana development. *Nat. Genet.* **37**, 501–506.
- Sherrier, D.J., Taylor, G.S., Silverstein, K.A.T., Gonzales, M.B. and VandenBosch, K.A. (2005) Accumulation of extracellular proteins bearing unique proline-rich motifs in intercellular spaces of the legume nodule parenchyma. *Protoplasma*, **225**, 43–55.
- Silverstein, K.A.T., Graham, M.A. and VandenBosch, K.A. (2006) Novel paralogous gene families with potential function in legume nodules and seeds. *Curr. Opin. Plant Biol.* **9**, 142–146.
- Smil, V. (1999) Nitrogen in crop production: an account of global flows. *Global Biogeochem. Cycles*, **13**, 647–662.
- Smit, P., Raedts, J., Portyanko, V., Debelle, F., Gough, C., Bisseling, T. and Geurts, R. (2005) NSP1 of the GRAS protein family is essential for rhizobial Nod factor-induced transcription. *Science*, **308**, 1789–1791.
- Sonnhammer, E.L.L., Eddy, S.R. and Durbin, R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins*, **28**, 405–420.
- Storey, J.D. and Tibshirani, R. (2003) Statistical significance for genome wide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.
- Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L.A., Rhee, S.Y. and Stitt, M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.* **37**, 914–939.
- Town, C.D. (2006) Annotating the genome of *Medicago truncatula*. *Curr. Opin. Plant Biol.* **9**, 122–127.
- Uchiumi, T., Ohwada, T., Itakura, M. et al. (2004) Expression islands clustered on the symbiosis island of the *Mesorhizobium loti* genome. *J. Bacteriol.* **186**, 2439–2448.
- Udvardi, M.K. and Day, D.A. (1997) Metabolite transport across symbiotic membranes of legume nodules. *Annu. Rev. Plant Physiol. Plant Molec. Biol.* **48**, 493–523.
- Udvardi, M.K., Kakar, K., Wandrey, M. et al. (2007) Legume transcription factors: global regulators of plant development and response to the environment. *Plant Physiol.* **144**, 538–549.
- Wu, Z., Irizarry, R.A., Gentleman, R., Murillo, F.M. and Spencer, F. (2004) *A Model Based Background Adjustment for Oligonucleotide Expression Arrays*. Technical Report. Baltimore, MD: John Hopkins University, Department of Biostatistics Working Papers.
- Zabala, G., Zou, J., Tuteja, J., Gonzalez, D.O., Clough, S.J. and O., V.L. (2006) Transcriptome changes in the phenylpropanoid pathway of *Glycine max* in response to Pseudomonas syringae infection. *BMC Plant Biol.* **6**, 26.
- Zar, J.H. (1999) *Biostatistical Analysis*. Upper Saddle River, NJ: Prentice Hall.