Published online 17 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D1189–D1194***
doi: 10.1093/nar/gkv1237

LegumeIP 2.0—a platform for the study of gene function and genome evolution in legumes

Jun Li, Xinbin Dai, Zhaohong Zhuang and Patrick X. Zhao*

Bioinformatics Lab, Plant Biology Division, Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway, Ardmore, OK 73401, USA

Received September 15, 2015; Revised November 01, 2015; Accepted November 02, 2015

ABSTRACT

The LegumeIP 2.0 database hosts large-scale genomics and transcriptomics data and provides integrative bioinformatics tools for the study of gene function and evolution in legumes. Our recent updates in LegumeIP 2.0 include gene and protein sequences, gene models and annotations, syntenic regions, protein families and phylogenetic trees for six legume species: *Medicago truncatula*, *Glycine max* (soybean), *Lotus japonicus*, *Phaseolus vulgaris* (common bean), *Cicer arietinum* (chickpea) and *Cajanus cajan* (pigeon pea) and two outgroup reference species: *Arabidopsis thaliana* and *Poplar trichocarpa*. Moreover, the LegumeIP 2.0 features the following new data resources and bioinformatics tools: (i) an integrative gene expression atlas for four model legumes that include 550 array hybridizations from *M. truncatula*, 962 gene expression profiles of *G. max,* 276 array hybridizations from *L. japonicas* and 56 RNA-Seq-based gene expression profiles for *C. arietinum*. These datasets were manually curated and hierarchically organized based on Experimental Ontology and Plant Ontology so that users can browse, search, and retrieve data for their selected experiments. (ii) New functions/analytical tools to query, mine and visualize large-scale gene sequences, annotations and transcriptome profiles. Users may select a subset of expression experiments and visualize and compare expression profiles for multiple genes. The LegumeIP 2.0 database is freely available to the public at http://plantgrn.noble.org/LegumeIP/.

INTRODUCTION

Legumes are the second most important class of crops as a source of food for humans, feed for livestock and raw materials for industry (1). Legumes possess complex secondary metabolisms and have been serving as a lynch pin of sustainable agriculture because they can acquire nitrogen efficiently through nitrogen-fixing endosymbiosis with soil bacteria called rhizobia. It is estimated that about 40–60 million tons of nitrogen are fixed annually by cultivated legumes (2), which is equivalent to about US $40 billion of fertilizer (1). Thus, understanding mechanisms that are fundamental to the legume species, especially the process of nitrogen-fixing endosymbiosis, will be of great value to healthy, low input sustainable agriculture by decreasing the use of fertilizers and improving crop yields.

We have been continuously developing the LegumeIP (3), an integrative database for comparative genomics and transcriptomics of model legumes, for the study of gene function and genome evolution in this important plant family. The original release of LegumeIP, referred to as LegumeIP 1.0 herein, hosted a total of 222,217 protein-coding genes and integrated protein family information, syntenic and phylogenetic context, tissue-specific transcriptomic profiles for three model legumes, *Medicago truncatula*, *Glycine max* and *Lotus japonicus* plus two reference plant species, *Arabidopsis thaliana* and *Populus trichocarpa*. The LegumeIP 1.0 database contained detailed gene annotations based on the UniProt (4), InterProScan (5), Gene Ontology (GO) (6) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) databases (7). It provided interfaces for retrieving gene annotation, systematic synteny analysis across *M. truncatula*, *G. max*, *L. japonicas* and *A. thaliana*, as well as construction and phylogenetic analyses of gene families across the five species. The LegumeIP 1.0 database also contained a small collection of microarray gene expression data, which included 156 *M. truncatula*, 14 *G. max* and 104 *L. japonicus* array hybridization results or RNA-Seq profiles.

Emerging high-throughput technologies, such as Illumina sequencing, have generated large-scale biological data that make genomic data easily available for even more species. In the legume family, more genomes, including *M. truncatula*, *Phaseolus vulgaris* (common bean), *Cicer arietinum* (chickpea) and *Cajanus cajan* (pigeon pea) (8–12), have been completely sequenced since the publication of LegumeIP 1.0. In addition, the sequencing of the *Medicago sativa* (alfalfa), *Arachis duranensis* and *Arachis ipaensis* (the

*To whom correspondence should be addressed. Tel: +1 580 224 6725; Fax: +1 580 224 4743; Email: pzhao@noble.org
Present address: Jun Li, Department of Genomics Medicine, University of Texas MD Anderson Cancer Center, 1901 East Road, Houston, TX 77054, USA.

ancestors of cultivated peanut), *Vigna radiata* (mung bean) (13), *Vigna* angularis (adzuki bean) (14) and *Lupinus angustifolius* (lupin) (15) genomes are almost complete.

Meanwhile, RNA-Seq technology makes large-scale gene expression profiling possible for non-model species, where dedicated commercial microarray chips are not available. Thus, transcriptomic data are now available from various technical platforms for species such as *L. japonicas* (16), *G. max* (17), *M. truncatula* (18), *Medicago sativa* (alfalfa) (19) and *P. vulgaris* (20). These large-scale genomic and transcriptomic datasets enable and facilitate the study of fundamental mechanisms, gene functions and gene and genome evolution using comparative genomic and molecular biology approaches.

In this paper, we present the '*LegumeIP 2.0 - A Platform for the Study of Gene Function and Genome Evolution in Legumes*' with a focus on the two most important new features of the database. First, we integrated genomic sequences for six legume species. These include *M. truncatula*, *G. max* and *L. japonicus* that were in the LegumeIP 1.0, three newly sequenced crop legumes, *P. vulgaris*, *C. arietinum* and *C. cajan*, and two reference plant species, *A. thaliana* and *P. trichocarpa*. The updated LegumeIP 2.0 covers 391,107 protein-coding gene sequences in total. All sequences were annotated based on sequence similarities and domain features by referring to the UniProt TrEMBL, InterProScan, GO and KEGG databases. Second, we built a brand new integrative gene expression atlas, which consists of large-scale gene expression profiles from multiple technologies for four model legumes, including 550 array hybridizations from *M. truncatula*, 962 gene expression profiles from *G. max*, 276 array hybridizations from *L. japonicas*, and 56 RNA-Seq transcriptome profiles from *C. arietinum*. All expression profiles were manually curated, annotated using free text descriptions, keywords and further hierarchically organized following the Experimental Ontology (EO) (http://bioportal.bioontology.org/ontologies/PECO?p=classes) (21,22) and Plant Ontology (PO) (http://www.plantontology.org/) (23), allowing users to browse, search and retrieve gene expression profiles for the experiments of their choice using both the EO tree view browser and comprehensive keyword searches. We believe this integrative gene atlas is essential and extremely valuable for the study of gene function and provide valuable large-scale gene expression data to biologists through user-friendly web interfaces.

DATABASE PRODUCTION: DATA EXPANSION AND FEATURE IMPROVEMENTS

Compilation and processing of genomic data

In LegumeIP 2.0, we have compiled and integrated gene sequences and comprehensive gene annotations for six model or crop legume genomes, including *M. truncatula*, *G. max*, *L. japonicus*, *P. vulgaris*, *C. arietinum*, *C. cajan* and two reference plant species, *A. thaliana* and *P. trichocarpa*. LegumeIP 2.0 contains 391,107 protein-coding gene sequences in total from eight plant species. Table 1 lists the included plant species and the sources of gene sequence and annotation data.

To annotate these genes, the sequences were searched against the reference sequences from the UniProt, TrEMBL, GO and KEGG databases using BLASTP with an expect value <1e-06. Furthermore, these sequences were searched for conserved domains for inclusion of functional annotation using the InterProScan program. All transcription factors (TFs) were predicted using our previously published method (24) for improved TF annotation.

We performed systematic syntenic analysis and reconstructed cross-species gene family and phylogenic trees for the six legume genomes and the two reference plant genomes using the same methods that were used in the development of LegumeIP 1.0 (3). Supplementary Figure S1 (Supplemental Material) shows an example of three complex micro synteny views for six legume species. As we described in the LegumeIP 1.0, multiple leucine-rich repeat receptor kinases are reportedly involved in the signaling pathway that mediates early root responses to bacterial and fungi infections in epidermal tissues of root nodules (25). In LegumeIP 2.0, multiple genes (e.g., *Ca_11537*, *C.cajan_12295* and *Phvul.002G143400.1*) with the same functions could also be identified with high confidence in *C. arietinumx*, *C. cajan* and *P. vulgaris* based on micro synteny analysis across the six legume species.

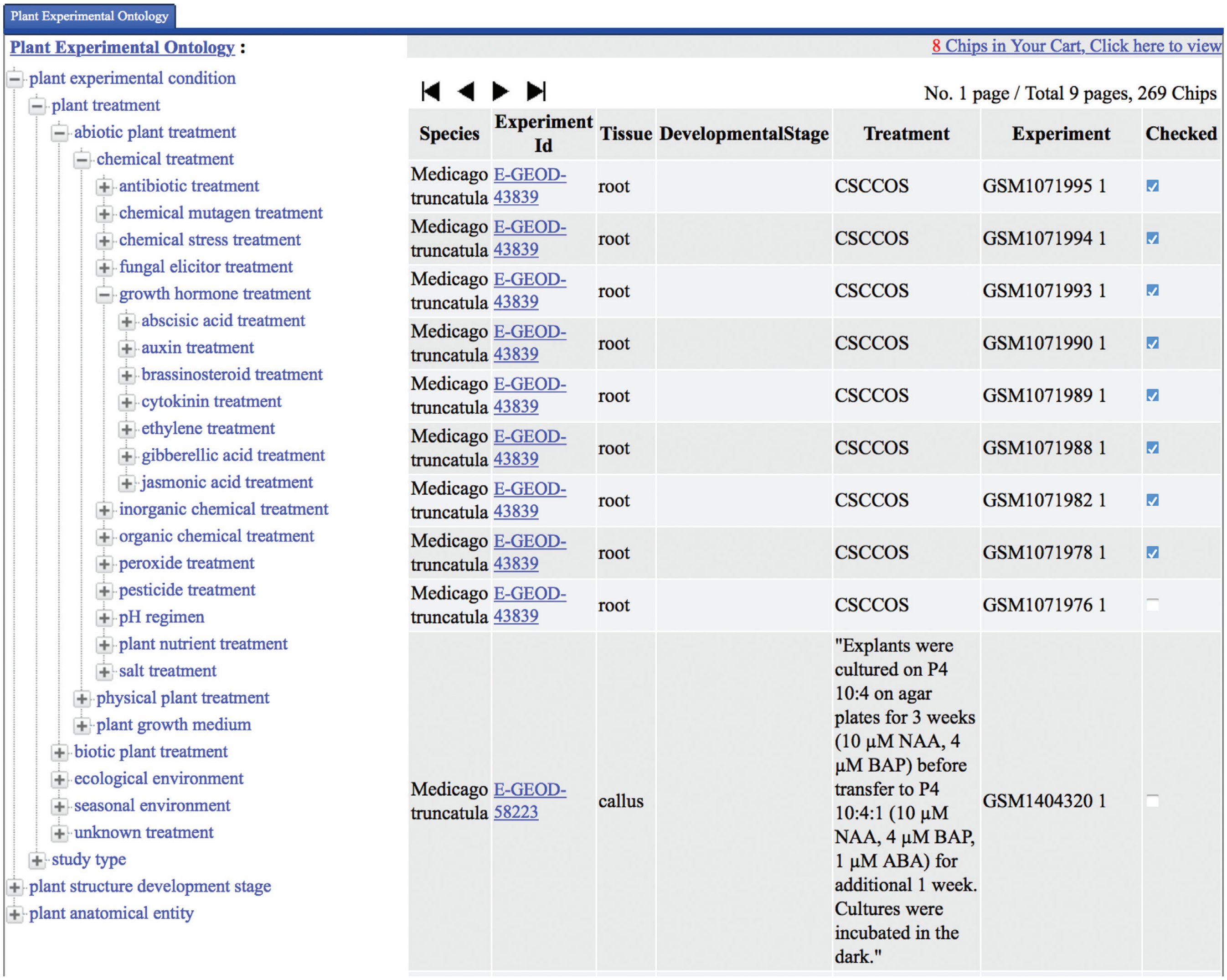
Compilation and processing of transcriptomic data

We have developed a brand new integrative gene expression atlas in LegumeIP 2.0. Because there is abundant microarray data for *M. truncatula*, *G. max* (soybean) and *L. japonicus* in public repositories, those datasets were collected, compiled and further curated for inclusion in LegumeIP 2.0. All of the expression profiles were downloaded from ArrayExpress (26), the NCBI GEO data repository (27) and other legume-specific gene expression atlases (16–18,20). The raw data were normalized by the Robust Multichip Average method (28) using the 'affy' module of the R (https://www.r-project.org/) software package. In addition, we included RNA-Seq-based gene expression profile data for chickpea from the NCBI SRA (29) database because there is no microarray chip designed for this organism. We adopted Bowtie (30) and RSEM (31) to map reads on the genomes, then estimated gene expressions in Fragments Per Kilobase of transcript per Million reads (FPKM) and further used edgeR (32) to normalize the final expression values. Currently, the updated gene expression atlas consists of 550 hybridizations using the Affymetrix A-AFFY-71 GeneChip for *M. truncatula*, 962 hybridizations using the Affymetrix GeneChip Soybean Genome Array A-AFFY-59, 276 hybridizations using the Affymetrix GeneChip Lotus Gene Array A-AFFY-59 and 56 Illumina RNA-Seq profiles for chickpea.

In addition to the integration of large-scale gene expression data for four legumes, one of the significant improvements in LegumeIP 2.0 is that we curated all transcriptomic datasets according to the descriptions of the original experimental designs, and further re-annotated the microarray assays (or Illumina runs) using the newly developed EO terms and PO terms. The former is a popular ontology that defines experimental conditions based on the published lit-

Table 1. A list of plant species and their gene sequence and annotation data sources in LegumeIP 2.0

Plant Species	Release Version	URL for Downloading
Medicago truncatula	Mt4.0	http://www.jcvi.org/medicago/display.php?pageName=General&section=Download
Glycine max	1.1	http://genome.jgi.doe.gov/PhytozomeV9/download/
Lotus japonicus	2.5	ftp://ftp.kazusa.or.jp/pub/lotus/lotus_r2.5/
Cicer arietinum	1.0	http://cicar.comparative-legumes.org
Phaseolus vulgaris	1.0	http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV9
Cicer arietinum	V5.0	http://cajca.comparative-legumes.org
Arabidopsis thaliana	V10	https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FSequences
Populus trichocarpa	V3.0	http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=PhytozomeV9


Plant Experimental Ontology
8 Chips in Your Cart, Click here to view
⏮ ◀ ▶ ⏭

Plant Experimental Ontology :

- ⊟ plant experimental condition
  - ⊟ plant treatment
    - ⊟ abiotic plant treatment
      - ⊟ chemical treatment
        - ⊞ antibiotic treatment
        - ⊞ chemical mutagen treatment
        - ⊞ chemical stress treatment
        - ⊞ fungal elicitor treatment
        - ⊟ growth hormone treatment
          - ⊞ abscisic acid treatment
          - ⊞ auxin treatment
          - ⊞ brassinosteroid treatment
          - ⊞ cytokinin treatment
          - ⊞ ethylene treatment
          - ⊞ gibberellic acid treatment
          - ⊞ jasmonic acid treatment
        - ⊞ inorganic chemical treatment
        - ⊞ organic chemical treatment
        - ⊞ peroxide treatment
        - ⊞ pesticide treatment
        - ⊞ pH regimen
        - ⊞ plant nutrient treatment
        - ⊞ salt treatment
      - ⊞ physical plant treatment
      - ⊞ plant growth medium
    - ⊞ biotic plant treatment
    - ⊞ ecological environment
    - ⊞ seasonal environment
    - ⊞ unknown treatment
  - ⊞ study type
- ⊞ plant structure development stage
- ⊞ plant anatomical entity

No. 1 page / Total 9 pages, 269 Chips

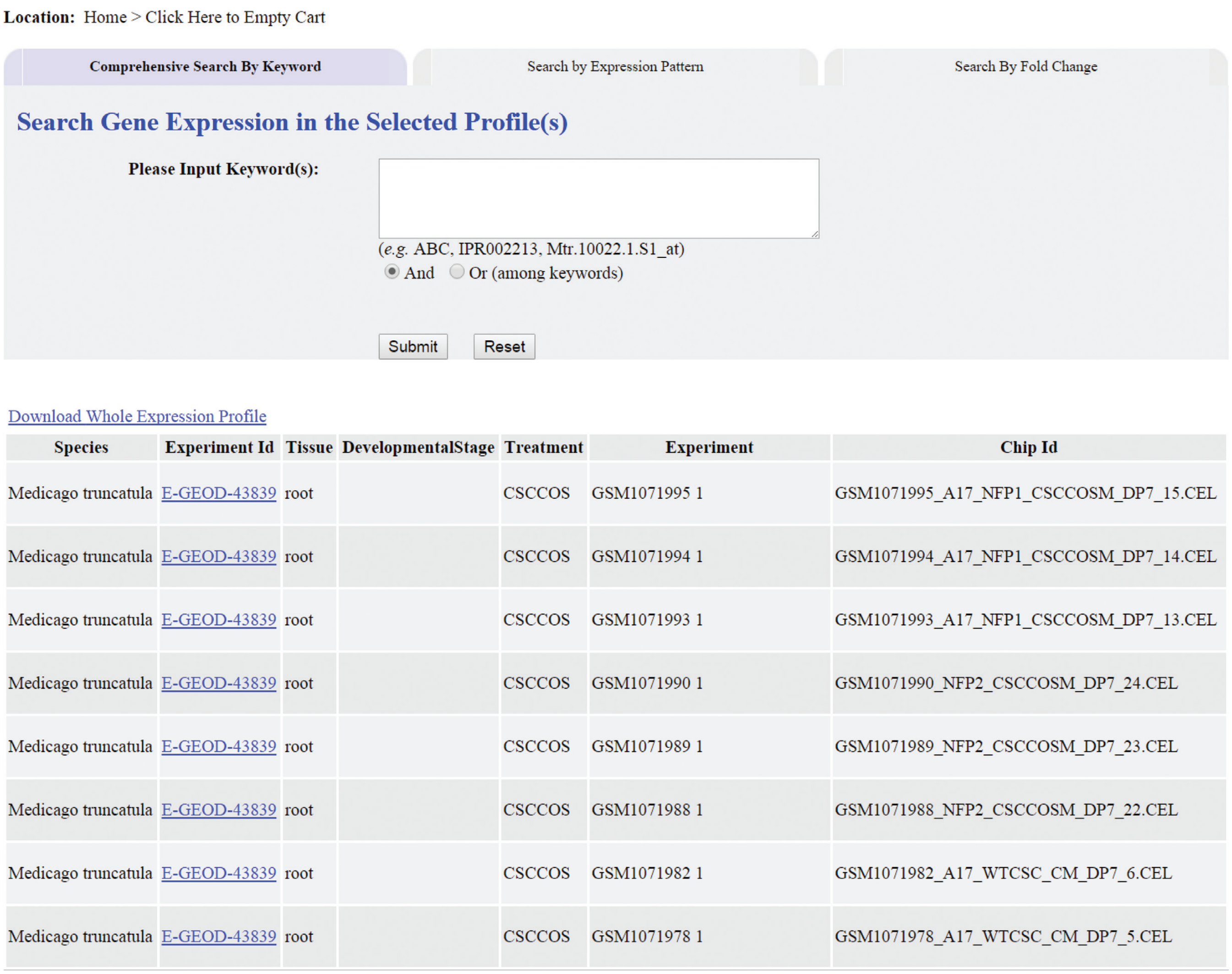
Species	Experiment Id	Tissue	DevelopmentalStage	Treatment	Experiment	Checked
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071995 1	☑
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071994 1	☑
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071993 1	☑
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071990 1	☑
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071989 1	☑
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071988 1	☑
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071982 1	☑
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071978 1	☑
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071976 1	□
Medicago truncatula	E-GEOD-58223	callus		"Explants were cultured on P4 10:4 on agar plates for 3 weeks (10 μM NAA, 4 μM BAP) before transfer to P4 10:4:1 (10 μM NAA, 4 μM BAP, 1 μM ABA) for additional 1 week. Cultures were incubated in the dark."	GSM1404320 1	□

Figure 1. The Experimental Ontology and Plant Ontology treeview browser for selecting a subset of microarray or RNA-Seq experiments.

erature and the latter defines plant anatomical structures, which is analogous to the GO (6). Incorporating information about the experimental conditions and biological samples significantly simplifies searches of transcriptome profiling data. Furthermore, we constructed genome-scale networks for these species using our published tools (33,34).

Database production and user-friendly web interfaces for data access

The LegumeIP 2.0 system runs on a Linux-based Resin Java web server using MySQL as its backend data management system. Compared with the previous version, we have significantly improved the web site of LegumeIP 2.0



Location: Home > Click Here to Empty Cart

Comprehensive Search By Keyword
Search by Expression Pattern
Search By Fold Change

Search Gene Expression in the Selected Profile(s)

Please Input Keyword(s):

(*e.g.* ABC, IPR002213, Mtr.10022.1.S1_at)

● And ○ Or (among keywords)

Submit
Reset
Download Whole Expression Profile

Species	Experiment Id	Tissue	DevelopmentalStage	Treatment	Experiment	Chip Id
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071995 1	GSM1071995_A17_NFP1_CSCCOSM_DP7_15.CEL
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071994 1	GSM1071994_A17_NFP1_CSCCOSM_DP7_14.CEL
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071993 1	GSM1071993_A17_NFP1_CSCCOSM_DP7_13.CEL
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071990 1	GSM1071990_NFP2_CSCCOSM_DP7_24.CEL
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071989 1	GSM1071989_NFP2_CSCCOSM_DP7_23.CEL
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071988 1	GSM1071988_NFP2_CSCCOSM_DP7_22.CEL
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071982 1	GSM1071982_A17_WTCSC_CM_DP7_6.CEL
Medicago truncatula	E-GEOD-43839	root		CSCCOS	GSM1071978 1	GSM1071978_A17_WTCSC_CM_DP7_5.CEL

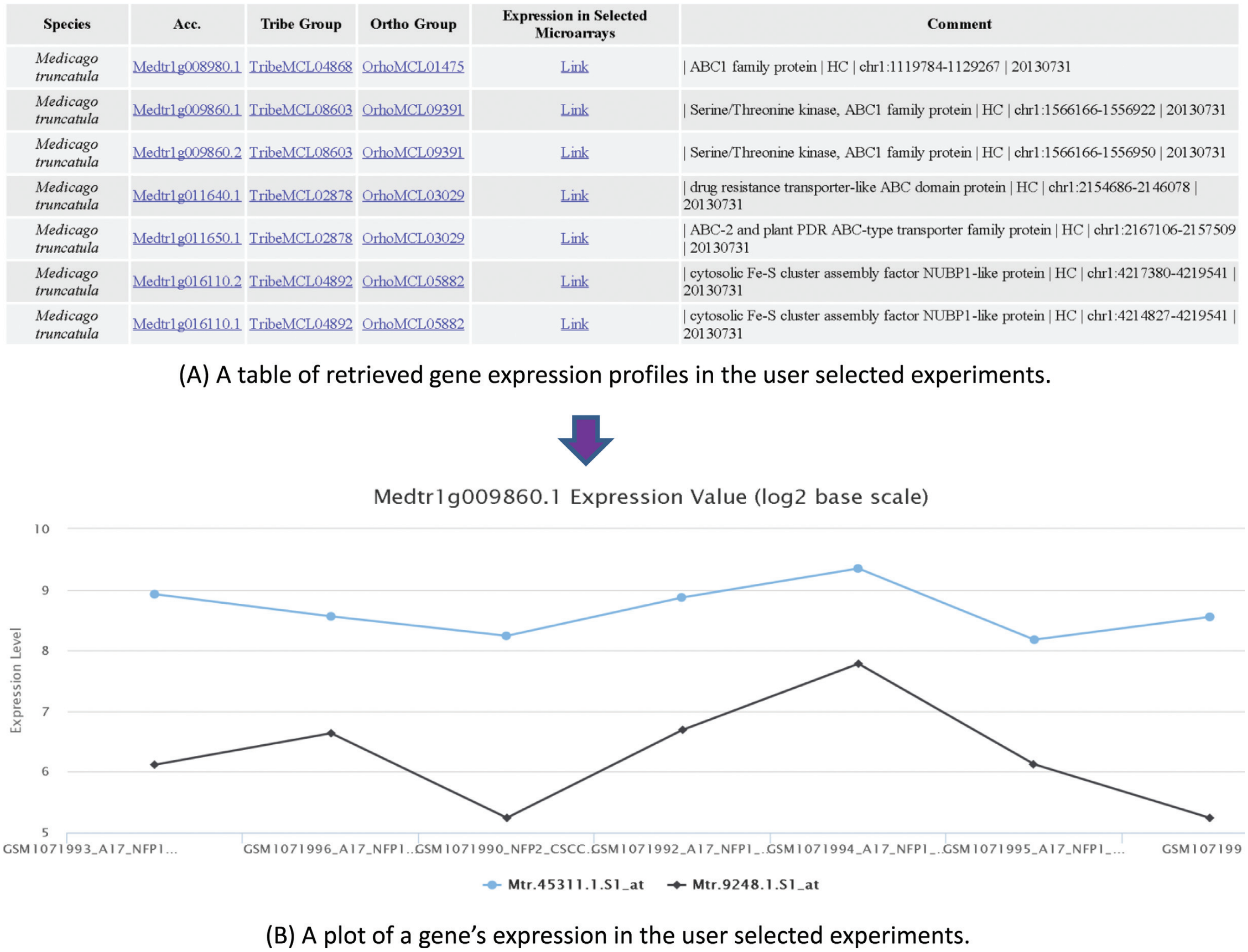
Figure 2. The comprehensive keyword search interface for retrieving genes' expression in the user-selected experiments.

with more intuitive and user-friendly interfaces for searching and exploring genes, gene families and syntenic regions. For example, we developed treeview interfaces for browsing the GO-based gene annotations (Supplementary Figure S2), KEGG-based enzyme annotations, transcription factors and transporters. Furthermore, we developed user-friendly interfaces, which can be accessed under the menu '*Expression Atlas*', for exploring and searching the gene expression atlas of LegumeIP 2.0, in which users can select a subset of microarray or RNA-Seq experiments and further compare expression profiles for multiple genes through EO and PO treeview browsers (Figure 1), comprehensive keyword search (Figure 2), differential expression gene analysis (Supplementary Figure S3) and advanced gene expression pattern search (Supplementary Figure S4) interfaces. Figure 3 shows a table of retrieved gene expression profiles (A) and a plot of a gene's expression (B) in the user selected experiments. We have adopted HTML 5 technologies for visualizing data in rich formatted text, tables and figures, which makes LegumeIP 2.0 compatible to all of the modern internet browsers. All of the search results can be downloaded to end-users local computers for further analysis.

CONCLUSION AND FUTURE DEVELOPMENT

The research in legume biology has progressed significantly. Using Affymetrix GeneChip and next-generation sequencing technologies have allowed researchers to generate large-scale genomic and transcriptomic data for both model and economically important crop legumes in recent years, which necessitated an important update to our LegumeIP database. We significantly improved the database by integrating more legume species based on the latest legume genome sequencing efforts. Such improvements make LegumeIP 2.0 a more useful tool for comparative genomic study utilizing its hosted syntenic regions, cross-species gene families and phylogenetic models. In this regard, we will continuingly enhance the LegumeIP database as more legume genomes become available.

In LegumeIP 2.0, we mapped microarray and RNA-Seq experiments to the graph structures of both EO and PO. In computer science, ontology is the formal naming and defi-



Species	Acc.	Tribe Group	Ortho Group	Expression in Selected Microarrays	Comment
Medicago truncatula	Medtr1g008980.1	TribeMCL04868	OrhoMCL01475	Link	| ABC1 family protein | HC | chr1:1119784-1129267 | 20130731
Medicago truncatula	Medtr1g009860.1	TribeMCL08603	OrhoMCL09391	Link	| Serine/Threonine kinase, ABC1 family protein | HC | chr1:1566166-1556922 | 20130731
Medicago truncatula	Medtr1g009860.2	TribeMCL08603	OrhoMCL09391	Link	| Serine/Threonine kinase, ABC1 family protein | HC | chr1:1566166-1556950 | 20130731
Medicago truncatula	Medtr1g011640.1	TribeMCL02878	OrhoMCL03029	Link	| drug resistance transporter-like ABC domain protein | HC | chr1:2154686-2146078 | 20130731
Medicago truncatula	Medtr1g011650.1	TribeMCL02878	OrhoMCL03029	Link	| ABC-2 and plant PDR ABC-type transporter family protein | HC | chr1:2167106-2157509 | 20130731
Medicago truncatula	Medtr1g016110.2	TribeMCL04892	OrhoMCL05882	Link	| cytosolic Fe-S cluster assembly factor NUBP1-like protein | HC | chr1:4217380-4219541 | 20130731
Medicago truncatula	Medtr1g016110.1	TribeMCL04892	OrhoMCL05882	Link	| cytosolic Fe-S cluster assembly factor NUBP1-like protein | HC | chr1:4214827-4219541 | 20130731

(A) A table of retrieved gene expression profiles in the user selected experiments.

(B) A plot of a gene's expression in the user selected experiments.

Figure 3. A table of retrieved gene expression profiles (**A**) and a plot of a gene's expression (**B**) in the user selected experiments.

nition of the types, properties and inter-relationships of the entities. Both EO and PO define terms and organize information about experiment design and plant samples in graph structures. The introduction of EO and PO can standardize the search of a transcriptomic dataset based on experimental conditions and plant samples and treatments used in the experiments, which are the main challenges in the development of large gene atlas. This development will also simplify our efforts toward continuously updating and expanding transcriptomic data in the LegumeIP system.

The LegumeIP 2.0 database features an integrative gene-atlas of four legume species and comparative genomic data, such as synteny and protein family information for all hosted legume and reference genomes. We collected raw data from the genome sequencing projects of legumes (Table 1) and transcriptomic data repositories such as the ArrayExpress (26) and the NCBI SRA (29). We performed extensive computational analysis and curation to generate these unique information which is valuable addition to other public legume databases, such as the Cool Season Food Legume Genome Database (https://www.coolseasonfoodlegume.org/), Legume Information System (35), PhaseolusGenes (http://phaseolusgenes.bioinformatics.ucdavis.edu/), SoyBase (36) and Soybean Knowledge Base (37). In future, we plan to develop application program interfaces to provide the Representational State Transfer (REST) web services to facilitate the information sharing among the legume databases.

AVAILABILITY

LegumeIP 2.0 is free to the public and available at http://plantgrn.noble.org/LegumeIP/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation (Grant DBI: 0960897 and DBI: 1458597 to P.X.Z.) and the Samuel Roberts Noble Foundation. Funding for the open access charge: National Science Foundation; Samuel Roberts Noble Foundation.

FUNDING

National Science Foundation [Grant DBI: 0960897 and DBI: 1458597 to P.X.Z.]; Samuel Roberts Noble Foundation. Funding for the open access charge: National Science Foundation; Samuel Roberts Noble Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Graham,P.H. and Vance,C.P. (2003) Legumes: importance and constraints to greater use. *Plant Physiol.*, **131**, 872–877.
2. Smil,V. (1999) Nitrogen in crop production: An account of global flows. *Global Biogeochem. Cycles*, **13**, 647–662.
3. Li,J., Dai,X., Liu,T. and Zhao,P.X. (2012) LegumeIP: an integrative database for comparative genomics and transcriptomics of model legumes. *Nucleic Acids Res.*, **40**, D1221–D1229.
4. The UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
5. Mitchell,A., Chang,H.Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.
6. Gene Ontology Consortium. (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056.
7. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.
8. Young,N.D., Debelle,F., Oldroyd,G.E., Geurts,R., Cannon,S.B., Udvardi,M.K., Benedito,V.A., Mayer,K.F., Gouzy,J., Schoof,H. *et al.* (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.
9. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
10. Sato,S., Nakamura,Y., Kaneko,T., Asamizu,E., Kato,T., Nakao,M., Sasamoto,S., Watanabe,A., Ono,A., Kawashima,K. *et al.* (2008) Genome structure of the legume, *Lotus japonicus*. *DNA Res.*, **15**, 227–239.
11. Schmutz,J., McClean,P.E., Mamidi,S., Wu,G.A., Cannon,S.B., Grimwood,J., Jenkins,J., Shu,S., Song,Q., Chavarro,C. *et al.* (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.*, **46**, 707–713.
12. Varshney,R.K., Song,C., Saxena,R.K., Azam,S., Yu,S., Sharpe,A.G., Cannon,S., Baek,J., Rosen,B.D., Tar'an,B. *et al.* (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nat. Biotechnol.*, **31**, 240–246.
13. Kang,Y.J., Kim,S.K., Kim,M.Y., Lestari,P., Kim,K.H., Ha,B.K., Jun,T.H., Hwang,W.J., Lee,T., Lee,J. *et al.* (2014) Genome sequence of mungbean and insights into evolution within Vigna species. *Nat. Commun.*, **5**, 5443.
14. Kang,Y.J., Satyawan,D., Shim,S., Lee,T., Lee,J., Hwang,W.J., Kim,S.K., Lestari,P., Laosatit,K., Kim,K.H. *et al.* (2015) Draft genome sequence of adzuki bean, *Vigna angularis*. *Sci. Rep.*, **5**, 8069.
15. Yang,H., Tao,Y., Zheng,Z., Zhang,Q., Zhou,G., Sweetingham,M.W., Howieson,J.G. and Li,C. (2013) Draft genome sequence, and a sequence-defined genetic linkage map of the legume crop species *Lupinus angustifolius L. PloS One*, **8**, e64799.
16. Verdier,J., Torres-Jerez,I., Wang,M., Andriankaja,A., Allen,S.N., He,J., Tang,Y., Murray,J.D. and Udvardi,M.K. (2013) Establishment of the *Lotus japonicus* Gene Expression Atlas (LjGEA) and its use to explore legume seed maturation. *Plant J.*, **74**, 351–362.
17. Libault,M., Farmer,A., Joshi,T., Takahashi,K., Langley,R.J., Franklin,L.D., He,J., Xu,D., May,G. and Stacey,G. (2010) An integrated transcriptome atlas of the crop model *Glycine max*, and its use in comparative analyses in plants. *Plant J.*, **63**, 86–99.
18. Benedito,V.A., Torres-Jerez,I., Murray,J.D., Andriankaja,A., Allen,S., Kakar,K., Wandrey,M., Verdier,J., Zuber,H., Ott,T. *et al.* (2008) A gene expression atlas of the model legume *Medicago truncatula*. *Plant J.*, **55**, 504–513.
19. O'Rourke,J.A., Fu,F., Bucciarelli,B., Yang,S.S., Samac,D.A., Lamb,J.F., Monteros,M.J., Graham,M.A., Gronwald,J.W., Krom,N. *et al.* (2015) The *Medicago sativa* gene index 1.2: a web-accessible gene expression atlas for investigating expression differences between *Medicago sativa* subspecies. *BMC Genomics*, **16**, 502.
20. O'Rourke,J.A., Iniguez,L.P., Fu,F., Bucciarelli,B., Miller,S.S., Jackson,S.A., McClean,P.E., Li,J., Dai,X., Zhao,P.X. *et al.* (2014) An RNA-Seq based gene expression atlas of the common bean. *BMC Genomics*, **15**, 866.
21. Shimayoshi,T., Amano,A. and Matsuda,T. (2007) A generic representation format of physiological experimental protocols for computer simulation using ontology. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, 382–385.
22. Malone,J., Holloway,E., Adamusiak,T., Kapushesky,M., Zheng,J., Kolesnikov,N., Zhukova,A., Brazma,A. and Parkinson,H. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.
23. Avraham,S., Tung,C.W., Ilic,K., Jaiswal,P., Kellogg,E.A., McCouch,S., Pujar,A., Reiser,L., Rhee,S.Y., Sachs,M.M. *et al.* (2008) The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res.*, **36**, D449–D454.
24. Dai,X., Sinharoy,S., Udvardi,M. and Zhao,P.X. (2013) PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool. *BMC Bioinform.*, **14**, 321.
25. Markmann,K., Giczey,G. and Parniske,M. (2008) Functional adaptation of a plant receptor-kinase paved the way for the evolution of intracellular root symbioses with bacteria. *PLoS Biol.*, **6**, e68.
26. Parkinson,H., Sarkans,U., Kolesnikov,N., Abeygunawardena,N., Burdett,T., Dylag,M., Emam,I., Farne,A., Hastings,E., Holloway,E. *et al.* (2011) ArrayExpress update—an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res.*, **39**, D1002–D1004.
27. Barrett,T., Suzek,T.O., Troup,D.B., Wilhite,S.E., Ngau,W.C., Ledoux,P., Rudnev,D., Lash,A.E., Fujibuchi,W. and Edgar,R. (2005) NCBI GEO: mining millions of expression profiles–database and tools. *Nucleic Acids Res.*, **33**, D562–D566.
28. Irizarry,R.A., Bolstad,B.M., Collin,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.
29. Kodama,Y., Shumway,M. and Leinonen,R. (2012) The sequence read archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**, D54–D56.
30. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
31. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.*, **12**, 323–323.
32. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
33. Li,J., Wei,H., Liu,T. and Zhao,P.X. (2014) GPLEXUS: enabling genome-scale gene association network reconstruction and analysis for very large-scale expression data. *Nucleic Acids Res.*, **42**, e32.
34. Li,J., Wei,H. and Zhao,P.X. (2013) DeGNServer: deciphering genome-scale gene networks through high performance reverse engineering analysis. *BioMed Res. Int.*, 856325.
35. Gonzales,M.D., Archuleta,E., Farmer,A., Gajendran,K., Grant,D., Shoemaker,R., Beavis,W.D. and Waugh,M.E. (2005) The Legume Information System (LIS): an integrated information resource for comparative legume biology. *Nucleic Acids Res.*, **33**, D660–D665.
36. Grant,D., Nelson,R.T., Cannon,S.B. and Shoemaker,R.C. (2010) SoyBase, the USDA-ARS soybean genetics and genomics database. *Nucleic Acids Res.*, **38**, D843–D846.
37. Joshi,T., Fitzpatrick,M.R., Chen,S., Liu,Y., Zhang,H., Endacott,R.Z., Gaudiello,E.C., Stacey,G., Nguyen,H.T. and Xu,D. (2014) Soybean knowledge base (SoyKB): a web resource for integration of soybean translational genomics and molecular breeding. *Nucleic Acids Res.*, **42**, D1245–D1252.