HAL open science

LeGOO: An Expertized Knowledge Database For The Model Legume Medicago truncatula

Sébastien Carrere, Marion Verdenaud, Clare Gough, Jerome Gouzy, Pascal Gamas

To cite this version:

Sébastien Carrere, Marion Verdenaud, Clare Gough, Jerome Gouzy, Pascal Gamas. LeGOO: An Expertized Knowledge Database For The Model Legume Medicago truncatula. Plant and Cell Physiology, 2020, 61 (1), pp.203-211. 10.1093/pcp/pcz177 . hal-02290862

HAL Id: hal-02290862

https://hal.science/hal-02290862v1

Submitted on 18 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Title: LeGOO: An Expertized Knowledge Database For The Model Legume *Medicago truncatula*

Short title: The LeGOO *Medicago truncatula* Knowledge Base

Corresponding author: S. Carrère, LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France , sebastien.carrere@inra.fr

Subject areas: regulation of gene expression; genomics, systems biology and evolution

Number of colour figures: 3

Number of tables: 3

LeGOO: An Expertized Knowledge Database For The Model Legume *Medicago truncatula*

Short title: the LeGOO *Medicago truncatula* knowledge base

Sébastien Carrère[1]*, Marion Verdenaud[2], Clare Gough[1], Jérôme Gouzy[1 #], Pascal Gamas[1#]

[1] LIPM, Université de Toulouse, INRA, CNRS, Castanet-Tolosan, France

[2] Laboratoire Reproduction et Développement des Plantes, Univ Lyon, ENS de Lyon, UCB Lyon 1, CNRS, INRA, F-69364, Lyon, France.

*Corresponding author: E-mail Sebastien.Carrere@inra.fr

co-last authors

Abstract

Medicago truncatula was proposed, about three decades ago, as a model legume to study the Rhizobium-legume symbiosis. It has now been adopted to study a wide range of biological questions, including various developmental processes (in particular root, symbiotic nodule and seed development), symbiotic (nitrogen-fixing and arbuscular mycorrhizal endosymbioses) and pathogenic interactions, as well as responses to abiotic stress. With a number of tools and resources set up in *M. truncatula* for omics, genetics, and reverse genetics approaches, massive amounts of data have been produced, as well as four genome sequence releases. Many of these data were generated with heterogeneous tools, notably for transcriptomics studies, and are consequently difficult to integrate. This issue is addressed by the LeGOO knowledge base (https://www.legoo.org), which finds the correspondence between the multiple identifiers of a same gene. Furthermore, an important goal of LeGOO is to collect and represent biological information from peer-reviewed publications, whatever the technical approaches used to obtain this information. The information is modelled in a graph-oriented database, which enables flexible representation, with currently over 200,000 relations retrieved from 298 publications. LeGOO also provides the user with mining tools, including links to the Mt5.0 genome browser and associated information (on gene functional annotation, expression, methylome, natural diversity and available insertion mutants), as well as tools to navigate through different model species. LeGOO is therefore an innovative database that will be useful to the Medicago and legume community to better exploit the wealth of data produced on this model species.

Keywords: Graph-based representation. Knowledge base. *Medicago truncatula.*

Abbreviations: RNAseq, RNA sequencing;

Introduction

The spectacular development of omics approaches in the past 20 years has generated a considerable amount of data, particularly on model species. This has been accompanied by rapid technical advances, notably for transcriptomics analyses. However, this progress has gone together with obvious difficulties for the integration of data obtained with different tools, such as different generations of microarrays and now massive parallel sequencing. An additional layer of complexity and confusion is often encountered when several successive genome sequence versions are released for a same organism, generally with different gene model annotations. As a consequence, much of the transcriptomics data generated with a given tool (generally found in supplementary tables) tend not to be taken into consideration and are usually not compared with those obtained with a different tool. This represents an obvious waste, considering the time, energy and money involved in the production of these data.

Furthermore, besides transcriptomics data management, there are other challenges to try and combine various types of data and to set up tools to easily access and visualize multi-source knowledge, with the corresponding sources of information precisely indicated.

Numerous databases have been developed for plant species in the past years, a number of which gather transcriptomics data and offer tools to analyze them in different ways. For example, TENOR (Kawahara *et al.* 2016) and Plant/OryzaExpress (Kudo *et al.* 2017b) are dedicated to rice RNAseq and microarray-based data respectively, with tools to analyze gene expression networks. Similarly, RNAseq data and mining tools are found for tomato in TOMATOMICS (Kudo *et al.* 2017a) and TomExpress (Zouine *et al.* 2017). For *Arabidopsis thaliana,* Araport (Krishnakumar *et al.* 2015a) and HRGRN (Dai *et al.* 2016) integrate a lot of data, many of them collected from various remote sites. Thus, HRGRN uses graph-search empowered tools to analyze *Arabidopsis* signal transduction, metabolism and gene regulatory networks.

For *Medicago truncatula,* a model legume species (Kang *et al.* 2016), a widely used gene atlas (MtGEA; https://mtgea.noble.org/v3/) gathers transcriptomic analyses performed with Affymetrix chips from a wide range of biological conditions (Benedito *et al.* 2008; He *et al.* 2009), while LegumeGRN generates gene network predictions, based on the Affymetrix data (Wang *et al.* 2013). Another database, SYMbiMICS (https://iant.toulouse.inra.fr/symbimics/), presents RNAseq data obtained so far from roots and symbiotic nodules, notably laser-microdissected samples (Roux *et al.* 2014; Jardinaud *et al.* 2016). Yet, other transcriptomics tools have been used, such as Mt6k (Kuster *et al.* 2004), Mt16k (Hohnjec *et al.* 2005) and Nimblegen microarrays (Verdier *et al.* 2013). Four *M. truncatula* genome sequences and corresponding gene annotations have been released, namely Mt3.5 (Young *et al.* 2011)**,** JCVI Mt4.0 (Tang *et al.* 2014), Mt20120830 (Roux *et al.* 2014) and Mt5.0 (Pecrix *et al.* 2018). Corresponding databases and genome browsers are the Medicago genome database [Mt3.5 and Mt4.0; http://www.medicagogenome.org/; (Krishnakumar et al. 2015b)], the Legume Information System [https://legumeinfo.org/ ; (Dash *et al.* 2016)], the *M. truncatula* A17 r5.0 genome portal [https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/; (Pecrix *et al.* 2018)], as well as a portal to explore the natural diversity of *M. truncatula* genomes [hapmap project;

http://www.medicagohapmap2.org/ (Kang *et al.* 2015)]. The Mt5.0 genome browser integrates data from numerous sources for thorough and extended analyses: gene, transposon and long non-coding RNA annotations; protein annotation (blastp and blast2GO results, interproscan analyses); gene models from all previous genome sequence releases; positions of Affymetrix and Nimblegen probe sets, as well as *Tnt1* insertions in the mutant collection generated at the Noble Foundation (Cheng *et al.* 2014); natural diversity data [hapmap project; (Kang *et al.* 2015)]; mRNAseq, sRNAseq and methylome data.

The LeGOO knowledge base presented here is based on different principles. It does not collect "raw" data but rather information manually extracted from publications. Thus, for transcriptomics data, only statistically relevant differentially regulated genes are used to feed LeGOO. This allows reliable knowledge generated with a variety of acquisition tools to be considered and integrated. This also implies that correspondences between all gene IDs and gene probes (used for microarrays) must be defined, which was performed using the last release (Mt5.0) of the *M. truncatula* genome sequence as the pivotal nomenclature. Furthermore, LeGOO is not limited to transcriptomics and incorporates a variety of data and information, e.g. obtained by genetics or molecular biology approaches. This information is modelled in a graph-oriented database (Miller 2013), similarly to HRGRN (Dai *et al.* 2016), with pairwise relations (edges) between objects (nodes).

So far, data from about 300 publications, from different fields (symbiotic and pathogenic interactions, response to abiotic stress, plant developmental pathways) have been manually retrieved, leading to more than 200,000 relations available in LeGOO. The LeGOO user can thus take advantage of mining tools to discover useful information on any gene of interest. The knowledge that is integrated and made accessible via LeGOO is a valuable addition for the Medicago community, as well as for other plant spp. by inference based on orthology groups.

Results

Tracking *Medicago truncatula* genes amongst different generations of datasets

The exploration of the *M. truncatula* transcriptome was initiated years before the first release of the *M. truncatula* genome sequence, in contrast to *Arabidopsis thaliana*, for which a high quality genome sequence was available prior to the development of array technologies, thereby enabling the establishment and use of a stable nomenclature. Moreover, over almost two decades, several generations of macro and microarrays were produced along with several *M. truncatula* genome versions, with consequently several nomenclatures being used in articles reporting on mid to large-scale gene expression analyses. This led us to develop an "ID Converter" service to quickly find the different identifiers (IDs) representing a same gene. The ID Converter tool is an essential part of LeGOO, and it can be used independently to find the various names and codes given to the same gene.

The ID Converter service manages most datasets that have been used for global *M. truncatula* gene expression analyses. It includes datasets built from Sanger Expressed Sequence Tags (EST) [TIGR/Dana Faber *M. truncatula* Gene Indices (MtGI) releases 5 to 11, EST MENS releases MtCDJan2003 (Journet *et al.* 2002) and

MtSCDJun2006 (Godiard *et al.* 2007)] and several sets of microarray probes, namely the Mt6k (Kuster *et al.* 2004), Mt16k (Hohnjec *et al.* 2005), Affymetrix (Benedito *et al.* 2008) and Nimblegen microarrays (Verdier *et al.* 2013). The database also includes all successive *M. truncatula* genome sequence and annotation releases, namely Mt3.5 (Young *et al.* 2011)**,** JCVI Mt4.0 (Tang *et al.* 2014), Mt20120830 (Roux *et al.* 2014) and Mt5.0 (Pecrix *et al.* 2018). The current reference transcriptome from pea (Alves-Carvalho *et al.* 2015), *Pisum sativum,* a legume species closely related to *M. truncatula*, is also included to facilitate knowledge transfer between a model and a crop legume species. This particular dataset was processed like *M. truncatula* reference datasets with specific overlap parameters (see Methods).

The method for establishing correspondences between all the IDs for a given gene (Table 1) was to map all corresponding sequences on the last reference genome sequence, Mt5.0 (see Methods). All "synonymy" links (Fig. 1) were then loaded into the LeGOO database to enable knowledge inference across the different generations of sequence datasets. Each pairwise comparison of ID sets is available as a spreadsheet table. It is also possible to search for all IDs of any single gene through the main search box or a dedicated search form (termed “ID converter”).

The mapping criteria were defined to take into account the variable accuracy of the datasets (see Methods). Each *M. truncatula* dataset could be mapped over 88% with a global mapping rate of 95%. The most recent datasets (Mt4.0, Mt20120830 and small secreted peptide genes [SSPs; (de Bang *et al.* 2017)] have the highest mapping rates, confirming at the same time their high quality and the completeness of the Mt5.0 reference genome. We can also observe a good specificity of the mapping (more than 96% of mapped sequences are mapped at only one location) for all the datasets except for the Nimblegen array probes, probably because they also included a set of transposable elements. Overall 93% of the *M. truncatula* mapped sequences have at least one additional corresponding ID in another dataset, while the average number of synonymous IDs is 9.1 per Mt4.0 locus.

Elaborating a knowledge base associated with *M. truncatula* genes

The aim of the LeGOO base is to retrieve knowledge, i.e. published information focused on scientific questions and validated through the publication process (with *ad hoc* statistical analyses and cross-validations when necessary). Non-curated raw data (e.g. raw transcriptomics, proteomics or interactome data) are therefore not taken into consideration. Importantly, results of a variety of targeted approaches, such as forward and reverse genetics or molecular biology (e.g. quantitative RT-PCR analyses, protein-protein or protein-DNA interactions…) are used to feed the LeGOO knowledge base, in addition to omics data.

The way LeGOO is conceived therefore implies managing heterogeneous objects / data sets and defining appropriate representation tools. We chose a graph-based representation, structured as graph nodes and edges (Fig. 2), with Cytoscape.js library (Franz *et al.* 2016) to show relations (“encodes”, “induces”, “represses”, “is required for”, “modifies”, “binds”…) between different types of biological objects and processes (oligonucleotide, EST, gene, RNA, protein, metabolite, biological process…).

Content of the LeGOO knowledge base

As of April 2019, data from 298 publications from peer-reviewed journals concerning *M. truncatula* were manually retrieved, the list of which is provided on the LeGOO site (KnowledgeBase / PubMed List). In total more than 200,000 relations have been entered in the database, using controlled vocabulary for objects, relations and processes (based on Gene Ontology terms whenever possible). Additional related information, especially details of experimental conditions, the magnitude of observed effects (e.g. fold change for induction or repression) and plant or interacting microbe species and genotypes, have also been indicated by the curators.

Furthermore, continuing efforts have been made to compile gene names used in publications, whether defined from phenotypes, genetic screens or systematic surveys of gene families and to establish their correspondence with genomic loci (> 2,500 genes are currently listed with corresponding publication identifiers as well as Mt5.0, Mt4.0, Mt3.5, Mt20120830, Affymetrix and Nimblegen probe IDs, a list which is regularly updated). This resource (termed Gene Acronyms vs. IDs, in the ID converter tool) is very valuable when analyzing any transcriptome or proteome data set, to quickly find relevant information on regulated genes of interest, and also to avoid renaming entities already reported in the literature.

A large fraction (140) of the publications used to feed LeGOO deal with transcriptomics analyses, which generated 98% of the relations and involved 88,250 different biological entities. Symbiosis-related publications (N-fixing and arbuscular mycorrhizal symbioses) are the most represented in the knowledge base with 125 publications and 71,626 relations, due to the scientific interests of the curators, but other processes are also documented (Table 2).

Retrieving gene-associated knowledge

LeGOO is queried using a full-text search (gene name or genomic locus). A small object list is obtained in return, corresponding to the subtype(s) (gene, transcript, protein) documented in LeGOO, based on the number of curated relations. Once one or several object(s) has/have been selected, the entire set of synonymous IDs is collected by LeGOO, and used to produce a knowledge graph, integrating the information collected for all synonymous identifiers.

To visually simplify the graphs, by default LeGOO collapses all equivalent targets (i.e. involving a same relation), and all synonymous entities into a single meta-node, which can be unraveled. Similarly, when a hub (i.e. an object with a high number of relations of the same kind) is retrieved, its targets are not collected but their number is displayed in red in an information panel next to the graph (Fig. 2). This information panel, posted when clicking on any node or relation, makes a series of metadata directly available, including a link to the original dataset and to the publication used to document the relation.

Organizing the layout of the knowledge graph

The output graph can be easily modified by the user to make it simpler or clearer (e.g. by moving or deleting any element). While the system offers by default a condensed view of synonymous objects and multiple targets

(display of the most informative object in terms of knowledge), several layouts are available to automatically organize the graph, and the user still has the ability to move, delete and cancel changes on the fly.

A contextual menu on all graph nodes makes several options available, namely building a new graph centered on this node and linking an object to the Mt5.0 genome browser (to access numerous additional types of information). It is also possible to expand a knowledge graph by simply "double-clicking" on a node, leading to the addition of all the relations corresponding to this object.

Finally, a search box makes it possible to highlight nodes or edges matching a word.

Discovering new indirect links between genes

LeGOO uses the shortest path algorithm (implemented in the Neo4J database engine) to provide an original service. This functionality aims at finding a shortest path between two biological entities without any *a priori* about their connection. To run it, the user selects two objects and the maximum number of paths that should be retrieved. The result is a knowledge graph representing all the paths and intermediate nodes that connect the two objects (Fig. 3), as well as a table indicating all paths when several are found. As with all LeGOO graphical outputs, additional information on any objects (e.g. list of best blastp hits in several legume and non-legume plant genomes, Interpro domains, existence of Tnt1 insertion mutants, RNAseq and methylome data…) can be obtained via a link to the Mt5.0 genome browser (using a contextual menu; Fig.3). In spite of obvious limitations (e.g. the existence of hubs that connect hundreds of objects, or the fact that relation orientation is not taken into account), this tool may allow unexpected relations to be discovered or co- or anti-regulated genes to be identified. Such information is generally difficult to find directly from publications, due to nomenclature issues and the mass of information to be integrated.

Mining tools and interoperability with Medicago community resources

Users can enter LeGOO using any keyword, gene name, domain annotation, Gene Ontology term to retrieve biological entities. Pubmed IDs can also be used to get the list of relations extracted from a specific publication.

Thanks to an automatic detection of patterns corresponding to known IDs, LeGOO proposes links to reference resources such as the Medicago Gene Atlas (Benedito *et al.* 2008) or MedicMine (Krishnakumar *et al.* 2015b). These links are displayed on the main page of search results in the form of an identity card or in additional links on the information panel of the knowledge graph.

A tool for mining open access publications, complementary to other tools such as PubMed, is also provided in LeGOO for gene queries. The system collects the set of corresponding synonyms and uses the API of Europe PMC to retrieve up to last 100 publications in which at least one of the synonyms appears. As an example of the usefulness of this tool, when using MtEFD as a query, three publications were found when searching in PubMed (as of May 29th 2019) vs. 10 using LeGOO. Publications already curated and available in the LeGOO database are indicated with the number of relations that were extracted per publication by the curators.

Navigating through different plant model species

In addition to the current reference transcriptome from pea (http://bios.dijon.inra.fr/FATAL/cgi/PsUniLowCopy.cgi), LeGOO offers a service to search for potential orthologs or recent paralogs to *M. truncatula* proteins (from the latest annotation release (Pecrix *et al.* 2018)) of identifiers from reference organisms (Table 3) such as *A. thaliana*. This functionality provides a way to benefit from the knowledge acquired on *M. truncatula* for the purpose of translational research, or reciprocally to help curators and users to identify relations missing in the LeGOO system. For any other organism that is not available among the reference list, users can enter the system via a blast search against all *M. truncatula* datasets used for the ID Converter service.

Discussion

Here, we describe LeGOO, a database centered on information retrieved from publications, with a graph-based structuration of knowledge acquired by the *M. truncatula* community. LeGOO provides: a comprehensive ID converter that considerably facilitates the mining and comparison of data produced over about two decades with different tools; a list of gene acronyms used in publications, along with the corresponding gene identifiers in successive genome sequence releases and microarrays; a text mining tool to find publications where a gene of interest is cited (with a search of all synonymous IDs corresponding to the gene).

The interests of a knowledgebase such as LeGOO are that: (i) information and data obtained by a variety of approaches can be integrated and represented; (ii) information is already filtered and validated by the publication process, which ensures quality standards and reliability, and decreases the level of "noise" caused by spurious data that may be found in raw data. Thanks to LeGOO, the user may find out valuable information outside his/her field of expertise or data of direct interest that is buried in supplementary tables of publications. For example, while *MtEFD* is known to be important for the regulation of nodulation (Vernié *et al.* 2008) and pathogenesis induced by the bacteria *Ralstonia solanacearum* (Moreau *et al.* 2014), data mining via LeGOO revealed that *MtEFD* expression is also activated by the oomycete *Aphanomyces euteiches*, inhibited during flower and fruit development, and positively or negatively impacted by phytohormone treatments (abscisic acid and jasmonic acid, respectively) (Fig. 2).

LeGOO is currently updated and maintained by manual curation by biologists, which is quite time consuming. In the future it would be very interesting to develop semi-automated text mining methods to facilitate the regular integration of published results.

Materials and Methods

Computing of synonymy links

The gene models (release 5.1.6) of the high quality reference genome (Pecrix *et al.* 2018) were used as pivotal sequences and nomenclature. All datasets were mapped on the Mt5.0 genome using GMAP (Wu and Watanabe 2005) (gmap.version = 2017-09-05, gmap.parameters = --gff3-add-separators=0 --mapboth --npaths=10 --suboptimal-score=1.0 -L 100000 --min-intronlength=35 -K 25000 --trim-end-exons=25). Then,

corresponding biological objects were identified with the bedtools intersect command combined with custom Perl scripts to define the following spanning thresholds: minimal query coverage of 30% for EST and gene queries and 80% for oligonucleotide, Affymetrix and Nimblegen probe queries; minimal subject coverage of 30% for EST and genes (same orientation as queries). For the *P. sativum* transcriptome data set (assemblies from short RNAseq reads), the threshold for subject coverage (i.e. the *P. sativum* sequences) was limited to 150% to avoid artefactual gene fusions between co-localized paralog genes.

Knowledge extraction from publication and formatting

The "knowledge" extracted from each publication was modelled as a relation between one source object (e.g. a gene, a protein or a biological process) and a list of targets (biological objects or processes). Each object is defined with a type (controlled vocabulary), a name, an organism and a genotype. Additional optional fields can be attached to the object, such as a subtype (e.g. mRNA or ncRNA for the RNA type), a description or a link to a database. The relation is labelled using a controlled vocabulary describing the type of the relation (e.g. induces). Additional metadata are attached to the relation, such as experimental conditions, the presence of other organisms (e.g. pathogen/symbiont) or a molecule of biological interest, as well as metadata related to the score (p-value, fold change), the source of information within the publication or the curator's name. To facilitate the entry of new relations, Excel©-formatted templates are used, containing pre-filled options with frequently used values (organism names, genotypes) and controlled vocabulary for object and relation types. Controlled vocabulary relies as much as possible on ontologies (Sequence Ontology (Eilbeck et al. 2005) for object types) and established terms from Gene Ontology (Harris et al. 2004) to describe biological processes whenever possible.

Data wrangling and database setup

The construction of a new release of the database is left to the LeGOO administrator. All files uploaded by curators, as well as lookup tables, reference files, and annotation files are collected. Automatic corrections are made to fix case errors or use common identifiers (using mapping files). Biological objects are identified through the calculation of a unique signature composed of the type, name, organism and genotype in which the object is observed. Synonymy links are automatically inserted to link analog objects in different genotypes, or to model the *Gene -> Transcript -> Protein* relations. The set of relations is loaded into a Neo4J graph database (https://neo4j.com/). Nodes and edges are annotated and indexed using ElasticSearch (https://www.elastic.co/) allowing full-text searches.

Web service implementation

The web service is developed in Perl and accesses the database using the Neo4J Rest API and the Elasticsearch CPAN module. The data is exposed as JSON documents via a Rest-like API. Hypermedia links provide pointers to related pages such as the external resources. Documents are formatted in Javascript using the JQuery library for interactivity and Cytoscape.js components for the graph layout.

Computation of orthology links

Orthology links between the proteins of the *M. truncatula* genome (proteome release 5.1.6) and a set of 14 selected plant proteomes were computed with Orthofinder software (v 2.2.0) (Emms and Kelly 2015). The files of the putative orthologs of *M. truncatula* (https://github.com/davidemms/OrthoFinder#results-files-orthologues) were analyzed with a Perl script (https://framagit.org/LIPM-BIOINFO/KGB/blob/master/bin/int/kgb_convert_orthofinder-files.pl) and loaded into a SQLite database allowing queries with non *Medicago* protein accessions

Funding

This work was supported by the Agence nationale de la recherche [project LeGOO: ANR-06-GPLA-0005] and conducted in the "Laboratoire d'Excellence (LABEX)" TULIP [ANR-10-LABX-41].

Acknowledgments

We thank Stefano Collela and Marc Lepetit (LMGM, Montpellier), Sandra Bensmihen, Françoise Jardinaud, Andreas Niebel and Nicolas Pauly (LIPM, Castanet Tolosan) for valuable discussions on the LeGOO template and their contribution to the curation work. We also thank Ludovic Legrand and Ludovic Cottret (LIPM) for technical advice.

Disclosures

The authors have no conflict of interest to declare.

References

Alves-Carvalho, S., Aubert, G., Carrère, S., Cruaud, C., Brochot, A.L., Jacquin, F., et al. (2015) 'Full-length de novo assembly of RNA-seq data in pea (Pisum sativum L.) provides a gene expression atlas and gives insights into root nodulation in this species', *Plant J*, 84(1), 1-19.

Benedito, V.A., Torres-Jerez, I., Murray, J.D., Andriankaja, A., Allen, S., Kakar, K., et al. (2008) 'A gene expression atlas of the model legume Medicago truncatula', *Plant J*, 55(3), 504-13.

Cheng, X., Wang, M., Lee, H.K., Tadege, M., Ratet, P., Udvardi, M., Mysore, K.S. and Wen, J. (2014) 'An efficient reverse genetics platform in the model legume Medicago truncatula', *New Phytol*, 201(3), 1065-76.

Cheng, X., Wen, J., Tadege, M., Ratet, P. and Mysore, K.S. (2011) 'Reverse genetics in medicago truncatula using Tnt1 insertion mutants', *Methods Mol Biol*, 678, 179-90.

Dai, X., Li, J., Liu, T. and Zhao, P.X. (2016) 'HRGRN: A Graph Search-Empowered Integrative Database of Arabidopsis Signaling Transduction, Metabolism and Gene Regulation Networks', *Plant Cell Physiol*, 57(1), e12.

Dash, S., Campbell, J.D., Cannon, E.K., Cleary, A.M., Huang, W., Kalberer, S.R., et al. (2016) 'Legume information system (LegumeInfo.org): a key component of a set of federated data resources for the legume family', *Nucleic Acids Res*, 44(D1), D1181-8.

de Bang, T., Lundquist, P.K., Dai, X., Boschiero, C., Zhuang, Z., Pant, P., et al. (2017) 'Genome-wide Identification of Medicago Peptides involved in Macronutrient Responses and Nodulation', *Plant Physiol*, 175(4):1669-1689.

Eilbeck, K., Lewis, S.E., Mungall, C.J., Yandell, M., Stein, L., Durbin, R., et al. (2005) 'The Sequence Ontology: a tool for the unification of genome annotations', *Genome Biol*, 6(5), R44.

Emms, D.M. and Kelly, S. (2015) 'OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy', *Genome Biol*, 16, 157.

Franz, M., Lopes, C.T., Huck, G., Dong, Y., Sumer, O. and Bader, G.D. (2016) 'Cytoscape.js: a graph theory library for visualisation and analysis', *Bioinformatics*, 32(2), 309-11.

Godiard, L., Niebel, A., Micheli, F., Gouzy, J., Ott, T. and Gamas, P. (2007) 'Identification of new potential regulators of the Medicago truncatula-Sinorhizobium meliloti symbiosis using a large-scale suppression subtractive hybridization approach', *Mol Plant Microbe Interact*, 20(3), 321-32.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004) 'The Gene Ontology (GO) database and informatics resource', *Nucleic Acids Res*, 32(Database issue), D258-6.

He, J., Benedito, V.A., Wang, M., Murray, J.D., Zhao, P.X., Tang, Y., et al. (2009) 'The Medicago truncatula gene expression atlas web server', *BMC Bioinformatics*, 10, 4411.

Hohnjec, N., Vieweg, M.F., Puhler, A., Becker, A. and Kuster, H. (2005) 'Overlaps in the transcriptional profiles of Medicago truncatula roots inoculated with two different Glomus fungi provide insights into the genetic program activated during arbuscular mycorrhiza', *Plant Physiol*, 137(4), 1283-301.

Jardinaud, M.F., Boivin, S., Rodde, N., Catrice, O., Kisiala, A., Lepage, A., et al. (2016) 'A Laser Dissection-RNAseq Analysis Highlights the Activation of Cytokinin Pathways by Nod Factors in the *Medicago truncatula* Root Epidermis', *Plant Physiol*, 171(3), 2256-76.

Journet, E.P., van Tuinen, D., Gouzy, J., Crespeau, H., Carreau, V., Farmer, M.J., et al. (2002) 'Exploring root symbiotic programs in the model legume Medicago truncatula using EST analysis', *Nucleic Acids Research*, 30(24), 5579-5592.

Kang, Y., Li, M., Sinharoy, S. and Verdier, J. (2016) 'A Snapshot of Functional Genetic Studies in Medicago truncatula', *Front Plant Sci*, 7, 1175.

Kang, Y., Sakiroglu, M., Krom, N., Stanton-Geddes, J., Wang, M., Lee, Y.C., et al. (2015) 'Genome-wide association of drought-related and biomass traits with HapMap SNPs in Medicago truncatula', *Plant Cell Environ*, 38(10), 1997-2011.

Kawahara, Y., Oono, Y., Wakimoto, H., Ogata, J., Kanamori, H., Sasaki, H., et al. (2016) 'TENOR: Database for Comprehensive mRNA-Seq Experiments in Rice', *Plant Cell Physiol*, 57(1), e7.

Krishnakumar, V., Hanlon, M.R., Contrino, S., Ferlanti, E.S., Karamycheva, S., Kim, M., et al. (2015a) 'Araport: the Arabidopsis information portal', *Nucleic Acids Res*, 43(Database issue), D1003-9, available: http://dx.doi.org/10.1093/nar/gku1200.

Krishnakumar, V., Kim, M., Rosen, B.D., Karamycheva, S., Bidwell, S.L., Tang, H., et al. (2015b) 'MTGD: The Medicago truncatula genome database', *Plant Cell Physiol*, 56(1), e1.

Kudo, T., Kobayashi, M., Terashima, S., Katayama, M., Ozaki, S., Kanno, M., et al. (2017a) 'TOMATOMICS: A Web Database for Integrated Omics Information in Tomato', *Plant Cell Physiol*, 58(1), e8.

Kudo, T., Terashima, S., Takaki, Y., Tomita, K., Saito, M., Kanno, M., et al. (2017b) 'PlantExpress: A Database Integrating OryzaExpress and ArthaExpress for Single-species and Cross-species Gene Expression Network Analyses with Microarray-Based Transcriptome Data', *Plant Cell Physiol*, 58(1), e1.

Kuster, H., Hohnjec, N., Krajinski, F., El Yahyaoui, F., Manthey, K., Gouzy, J., et al. (2004) 'Construction and validation of cDNA-based Mt6k-RIT macro- and microarrays to explore root endosymbioses in the model legume Medicago truncatula', *Journal of Biotechnology*, 108(2), 95-113.

Miller, J.J. (2013) 'Graph Database Applications and Concepts with Neo4J', *Proceedings of the Southern Association for Information Systems Conference, Atlanta, GA, USA March 23rd-24th*, 141-147.

Moreau, S., Fromentin, J., Vailleau, F., Vernié, T., Huguet, S., Balzergue, S., et al. (2014) 'The symbiotic transcription factor MtEFD and cytokinins are positively acting in the Medicago truncatula and Ralstonia solanacearum pathogenic interaction', *New Phytol*, 201(4), 1343-57.

Pecrix, Y., Staton, S.E., Sallet, E., Lelandais-Briere, C., Moreau, S., Carrere, S., et al. (2018) 'Whole-genome landscape of Medicago truncatula symbiotic genes', *Nat Plants*, 4(12), 1017-1025.

Roux, B., Rodde, N., Jardinaud, M.F., Timmers, T., Sauviac, L., Cottret, L., et al. (2014) 'An integrated analysis of plant and bacterial gene expression in symbiotic root nodules using laser-capture microdissection coupled to RNA sequencing', *Plant J*, 77(6), 817-37.

Tadege, M., Wen, J., He, J., Tu, H., Kwak, Y., Eschstruth, A., et al. (2008) 'Large-scale insertional mutagenesis using the Tnt1 retrotransposon in the model legume Medicago truncatula', *Plant J*, 54(2), 335-47.

Tang, H., Krishnakumar, V., Bidwell, S., Rosen, B., Chan, A., Zhou, S., et al. (2014) 'An improved genome release (version Mt4.0) for the model legume Medicago truncatula', *BMC Genomics*, 15, 312.

Verdier, J., Lalanne, D., Pelletier, S., Torres-Jerez, I., Righetti, K., Bandyopadhyay, K., et al. (2013) 'A regulatory network-based approach dissects late maturation processes related to the acquisition of desiccation tolerance and longevity of Medicago truncatula seeds', *Plant Physiol*, 163(2), 757-74

Vernié, T., Moreau, S., de Billy, F., Plet, J., Combier, J.P., Rogers, C., et al. (2008) 'EFD Is an ERF Transcription Factor Involved in the Control of Nodule Number and Differentiation in *Medicago truncatula*', *Plant Cell*, 20(10), 2696-2713.

Wang, M., Verdier, J., Benedito, V.A., Tang, Y., Murray, J.D., Ge, Y., et al. (2013) 'LegumeGRN: a gene regulatory network prediction server for functional and comparative studies', *Plos One*, 8(7), e67434.

Wu, T.D. and Watanabe, C.K. (2005) 'GMAP: a genomic mapping and alignment program for mRNA and EST sequences', *Bioinformatics*, 21(9), 1859-75.

Young, N.D., Debelle, F., Oldroyd, G.E., Geurts, R., Cannon, S.B., Udvardi, M.K., *et al.* (2011) 'The Medicago genome provides insight into the evolution of rhizobial symbioses', *Nature*, 480(7378), 520-4.

Zouine, M., Maza, E., Djari, A., Lauvernier, M., Frasse, P., Smouni, A., et al. (2017) 'TomExpress, a unified tomato RNA-Seq platform for visualization of expression data, clustering and correlation networks', *Plant J*, 92(4), 727-735.

Tables

Dataset	Size [1]	Mapped [2]	Uniquely Mapped [3]	ID linked [4]
affx-1 (*S. meliloti* and *M. sativa* probes excluded)	51063	48597	47361	45037
DeBang_PlantPhysiol2017-Currated-SSP	1970	1970	1925	1879
IMGA-Mt3.5.5-gene	77464	76730	74828	71717
JCVI-Mt4.0v2-gene	50444	50119	49301	47393
Mt16kOLIplus-2004	16780	16201	15654	15764
Mt20120830.gene-ncrna-missing	82940	82678	79150	67371
Mt6kRIT-Jan2003	4143	3941	3899	3929
MtCDJan2003	37413	33730	32638	33407
MtGI5	33765	30161	29352	29861
MtGI6	36235	31906	31040	31708
MtGI7	36976	32509	31593	32259
MtGI8	36878	33355	32479	33068
MtGI9	67463	63189	61227	62859
MtGI10	68848	64519	62543	64272
MtGI11	68814	64496	62522	64227
MtSCDJun2006	43398	39260	37769	38264
NCR-2003	311	311	300	311
Nimblegen-GPL16373-IRHS_Medtr_102K_v1	102123	101634	94179	82110
Clusters_PsUniLowCopy	40395	30571	26783	12312

Table 1. Sources of identifiers used in the LeGOO knowledge base; the Pea transcriptome dataset (Clusters_PsUniLowCopy; (Alves-Carvalho *et al.* 2015)) has been compared only to Mt5.0 gene models. [1] number of unique IDs in fasta file; [2] number of sequences mapped onto the genome; [3] number of sequences mapped at only one position with highest score; [4] number of IDs with a correspondence in at least one dataset.

Category	Publications	Relations
N-fixing and arbuscular mycorrhizal symbioses	125	71627
organ development, except nodule	38	13820
response to abiotic factors	31	46112
phytohormone-related	31	10429
response to biotic factors, except N-fixing and arbuscular mycorrhizal symbioses	22	35130
biosynthetic process	4	6
suppression of host defenses	3	3
developmental process	2	678
response to endogenous stimulus	2	3
growth	2	2
metabolic processes, except phytohormone-related	1	1
ANNOTATION	193	2424

Table 2- Number of curated publications and relations extracted per category.

Organism	Genotype	Version
Brassica oleracea var. oleracea	TO1000	2.1.31
Brachypodium distachyon	Bd21	Phytozome.12.314_3.1
Arabidopsis thaliana	Col-0	Araport11
Zea mays	B73	Phytozome.12.284
Vitis vinifera	PN40024	Phytozome.12
Solanum lycopersicum	Heinz 1706	ITAG.3.2
Phaseolus vulgaris	BAT93	EnsemblPlants.38
Nicotiana benthamiana	Nb-1	Solgenomics.1.0.1
Lotus japonicus	Miyakojima MG-20	3.0
Helianthus annuus	XRQ	1.2
Glycine max	Williams 82	Phytozome.12
Brassica rapa	Chiifu-401	EnsemblPlants.38
Triticum aestivum	Chinese Spring	EnsemblPlants.38
Hordeum vulgare	Morex	EnsemblPlants.38

Table 3: List of target proteomes used to infer orthologs or recent paralogs.

Figure legends

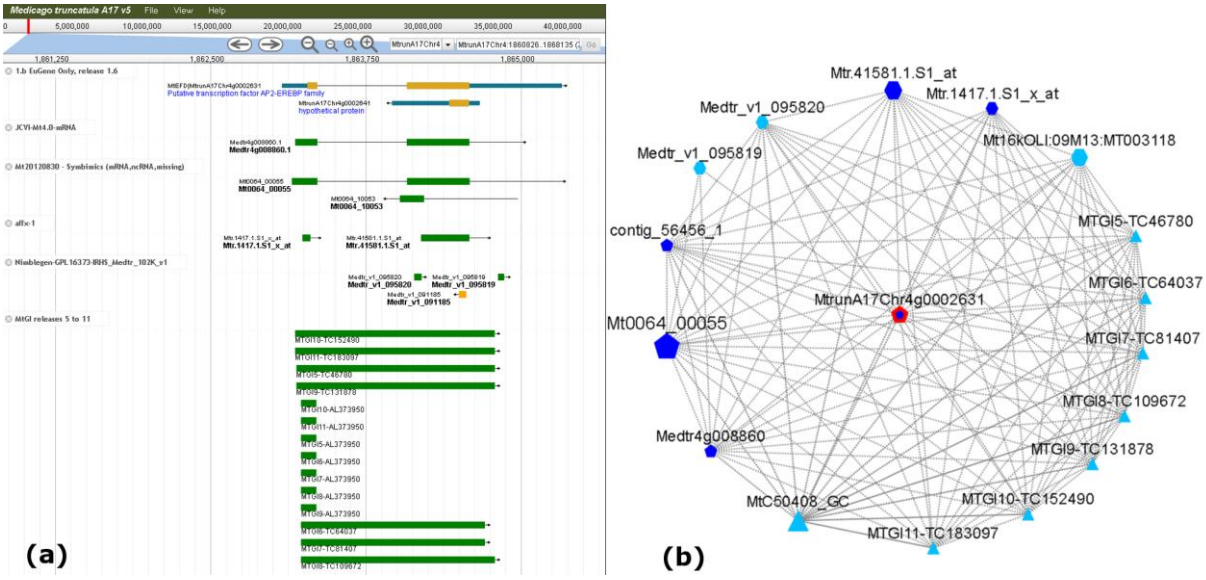

Figure 1. Example of various identifiers (ESTs, microarray probes and locus from different genome sequence releases) corresponding to a same gene, Mt*EFD* [*Ethylene response Factor required for nodule Differentiation*; (Vernié *et al.* 2008)]. Each relation corresponds to an overlap between objects based on their mapping onto the Mt5.0 genome sequence (https://medicago.toulouse.inra.fr/MtrunA17r5.0-ANR/).

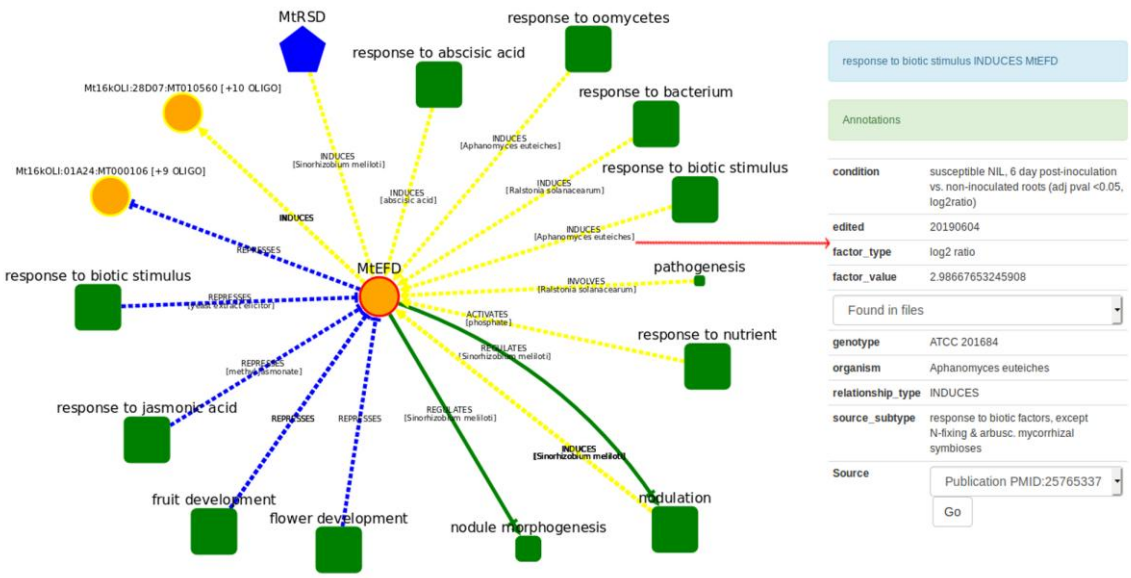

Figure 2. Example of relations retrieved from the LeGOO base, depicted using a graph representation, here using the MtEFD transcription factor as a query. The relation types along with associated information are indicated on the edges and depicted with different colors (e.g. blue for represses and yellow for induces).

Additional information (such as experimental conditions, a link to the corresponding publication…) can be found on a panel on the side of the graph. Dashed lines indicate several relations of a same type; details on those relations can be obtained by expanding the synonymous nodes ("Synonymous Display" menu).

Figure 3. Example of shortest paths (top right panel) found between two proteins (MtEFD and MtDME) (number of paths set to five). A contextual menu (long click on any object; here Mt0127_00021) gives access to the Mt5.0 genome browser and numerous associated data: here Mt5.0 and Mt4.0 gene models, Affymetrix and Nimblegen gene probe location, position of TnT1 insertions in the TnT1 mutant population (Tadege *et al.* 2008; Cheng *et al.* 2011), nodule and root RNAseq data; three bottom right panels: annotation information accessed through a right click on the Mt5.0 gene model, with links to different databases (here ThaleMine, to access information on *Arabidopsis thaliana* BLASTp hits).