BMC Plant Biology

BioMed Central

Database

TRUNCATULIX – a data warehouse for the legume community

Kolja Henckel*[1,2,3,4,5], Kai J Runte[1,4,5], Thomas Bekel[1,4,5], Michael Dondrup[1,4,5], Tobias Jakobi[1,3,5], Helge Küster[2,6,7,5] and Alexander Goesmann[1,4,5]

Address: [1]Bioinformatics Resource Facility, Center for Biotechnology, Bielefeld University, Bielefeld, Germany, [2]International Graduate School in Bioinformatics and Genome Research, Bielefeld University, Bielefeld, Germany, [3]Technical Faculty, Bielefeld University, Bielefeld, Germany, [4]Computational Genomics, Center for Biotechnology, Bielefeld University, Bielefeld, Germany, [5]Faculty for Biology and Genetics, Bielefeld University, Bielefeld, Germany, [6]Genomics of Legume Plants, Institute for Genome Research and Systems Biology, Center for Biotechnology, Bielefeld University, Bielefeld, Germany and [7]Unit IV – Plant Genomics, Institute for Plant Genetics, Leibniz Universität Hannover, Germany

Email: Kolja Henckel* - khenckel@cebitec.uni-bielefeld.de; Kai J Runte - krunte@cebitec.uni-bielefeld.de; Thomas Bekel - tbekel@cebitec.uni-bielefeld.de; Michael Dondrup - mdondrup@cebitec.uni-bielefeld.de; Tobias Jakobi - tjakobi@cebitec.uni-bielefeld.de; Helge Küster - helge.kuester@genetik.uni-bielefeld.de; Alexander Goesmann - agoesman@cebitec.uni-bielefeld.de

* Corresponding author

Published: 11 February 2009

BMC Plant Biology 2009, **9**:19 doi:10.1186/1471-2229-9-19

Received: 27 October 2008
Accepted: 11 February 2009

This article is available from: http://www.biomedcentral.com/1471-2229/9/19

Abstract

Background: Databases for either sequence, annotation, or microarray experiments data are extremely beneficial to the research community, as they centrally gather information from experiments performed by different scientists. However, data from different sources develop their full capacities only when combined. The idea of a data warehouse directly adresses this problem and solves it by integrating all required data into one single database – hence there are already many data warehouses available to genetics. For the model legume *Medicago truncatula*, there is currently no such single data warehouse that integrates all freely available gene sequences, the corresponding gene expression data, and annotation information. Thus, we created the data warehouse TRUNCATULIX, an integrative database of *Medicago truncatula* sequence and expression data.

Results: The TRUNCATULIX data warehouse integrates five public databases for gene sequences, and gene annotations, as well as a database for microarray expression data covering raw data, normalized datasets, and complete expression profiling experiments. It can be accessed via an AJAX-based web interface using a standard web browser. For the first time, users can now quickly search for specific genes and gene expression data in a huge database based on high-quality annotations. The results can be exported as Excel, HTML, or as csv files for further usage.

Conclusion: The integration of sequence, annotation, and gene expression data from several *Medicago truncatula* databases in TRUNCATULIX provides the legume community with access to data and data mining capability not previously available. TRUNCATULIX is freely available at http://www.cebitec.uni-bielefeld.de/truncatulix/.

Background

Medicago truncatula is a model plant for studying legume biology. Legumes are mainly characterized by their ability to interact with beneficial microbial organisms, leading to the formation of nitrogen-fixing root nodules and to phosphate-acquiring arbuscular mycorriza. Various inter-

national research projects are investigating these different symbioses of *Medicago truncatula*. The arbuscular mycorrhiza (AM) interaction between the host root and the fungal partner is an interesting field of research because more than 80% of land plants depend on an efficient AM for the uptake of nutrients, primarily phosphate. Apart from AM, *Medicago truncatula* is capable of entering a nitrogen-fixing symbiosis with the soil bacterium *Sinorhizobium meliloti*. The capacity for symbiotic nitrogen fixation allows legumes such as *Medicago truncatula* to grow on nitrogen-depleted soils and to develop protein-rich seeds, which are properties exploited in sustainable agriculture [1-5].

In recent years, more and more databases for the storage of microarray expression data (Arrayexpress [6], PEPR [7], The Stanford MicroArray Database [8], PlexDB [9]), and data from different sequencing projects (EST sequencing(dbEST [10]), BAC sequencing (GenMapDB [11]), ultrafast sequencing (Short Read Archive [12])), have been developed to store the exponentially growing amount of data. However, scientists have to search for the specific information in each and every database separately.

Directly adressing this issue, data warehouses, specially designed databases, offer an approach to store different aspects of a certain data object in an optimized data schema. This provides fast data access and enables return of query results in minimal time [13,14].

To overcome the problem of distributed data sources in the field of *Medicago truncatula* research, we created TRUNCATULIX, a data warehouse storing sequence data, annotations, and expression experiments of the model legume *Medicago truncatula*.

Construction and content

The construction of the TRUNCATULIX data warehouse is divided in three main parts: 1) the database schema, 2) the data integration, 3) and the source data. The following sections outline these three aspects in detail.

Database schema

The intention of the TRUNCATULIX data warehouse is to store information on gene sequences, functional annotations, and expression data. Thus, we created a relational data schema (see Figure 1), containing all these aspects. The different tables that store the information are linked via unique keys.

The TRUNCATULIX data warehouse is based on the IGetDB data warehouse engine [15] using Java [16]. It uses MySQL [17] as the database management system and is based on the Biomart API [18], while providing additional functionality such as full-text searches and direct access to SQL-functions (e.g. SUM, AVG, ROUND). The primary gene information such as sequence data, start codon, stop codon, length of the encoded open reading frame, name, or gene_id are stored in the main table, whereas data such as gene expression values from different experiments, GO numbers, or KOG categories are stored in extra tables refering to the main table.

Data integration

High throughput technologies yield vast amounts of data. However, in order to investigate, for example, regulatory pathways, further standard genetics approaches are yet most effective. High throughput data provide an excellent means to reduce the number of possible mutation targets. Such data is found, in general, in different sources and the screening of (most likely) thousands of candidates is, when performed manually, a rather tedious task. On the other hand, if the data was integrated and preprocessed for querying, such a task can be performed in a matter of minutes. As has been shown recently, such data integration strategies saves both time and money [19].

For the integration and import of data we use the extract, transform and load (ETL) approach commonly used in data warehousing [20]. Sequence and annotation data are extracted from SAMS (cf. next section), and are subsequently transformed and loaded into the TRUNCATULIX database. Expression data is exported from EMMA (cf. Section "Expression data") via an export script that can be initiated by a human curator within the EMMA web interface. During the transformation step, the expression data are linked to the sequence data. This enables the user of TRUNCATULIX to conveniently search across annotation and expression data.

Data sources

Sequence data

• *Medicago truncatula* Gene Index 8.0

The Institute for Genomic Research (TIGR – J. Craig Venter Institute since October 2006) clustered and assembled 226,923 high-quality ESTs from over 60 different *Medicago truncatula* EST-libraries sequenced in laboratories all over the world. Using the clustering software *tgicl* [21], the *Medicago truncatula* Gene Index (MtGI, hosted at the Dana-Farber Cancer Institute – DFCI), was built. The MtGI 8.0 contains 18,612 Tentative Consensus sequences (TCs) and 18,238 singletons (Jan. 2005) [22]. The sequences were imported into the Sequence Analysis and Management System (SAMS) [23], an annotation software created at the Center for Biotechnology (CeBiTec) in Bielefeld. The SAMS system contains an automatic annotation pipeline (Metanor), which runs several bioinformatics tools for gene annotation (Blast, Interpro, TMHMM) [24-26]. A high quality consensus annotation

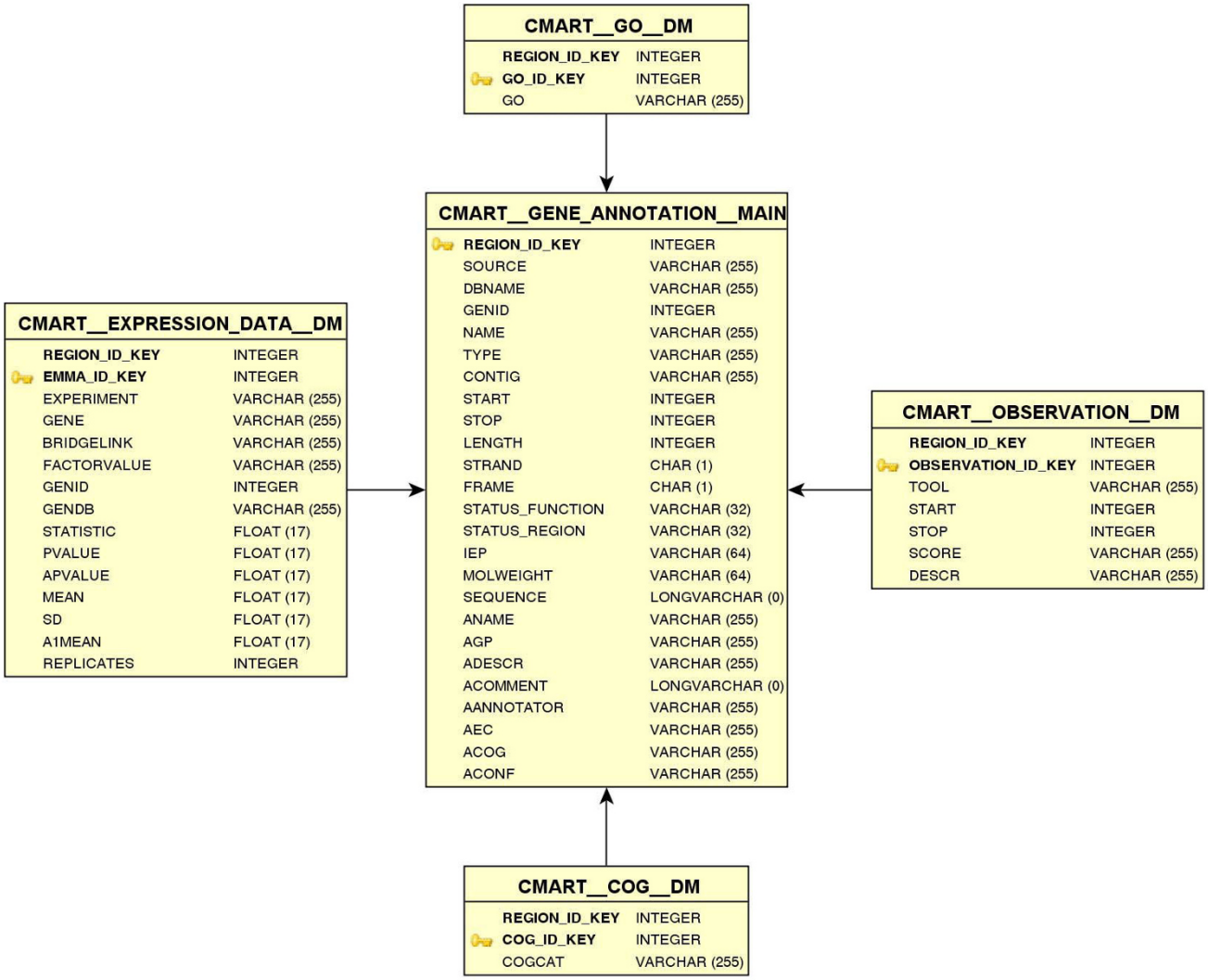

Figure 1
The TRUNCATULIX database schema. The TRUNCATULIX database schema. The main table stores the sequence data. All other information is stored in different tables refering to the main table.

is created, covering EC numbers [27], KEGG functions [27], GO numbers [28], KOG numbers [29], putative gene functions, and gene names.

• *Medicago truncatula* Gene Index 9.0

Recently, the J. Craig Venter Institute released a new version of the *Medicago truncatula* Gene Index, now covering over 70 EST-libraries. The assembly of the 259,642 ESTs led to 29,273 TCs, while 26,696 ESTs remained as singletons. In addition to the previous Gene Index 8.0, TIGR used 25,600 mature transcripts (ETs) from the qcGene Database [30] for the EST assembly, whereof 11,494 ETs remained as singletons. The new sequences were downloaded from the DFCI websites and imported into SAMS, where a complete automatic annotation was performed.

• *Medicago* genome project

The *Medicago* Genome Sequence Consortium (MGSC) sequenced the *Medicago truncatula* genome using a classical BAC sequencing approach [31,32]. Starting in 2005, they released a assembly of the sequences in October 2007 (release 2.0). This release contains 38,759 coding sequences (CDS) and the same number of translated protein sequences. The CDS's were downloaded from the project website and afterwards imported into SAMS. Using SAMS, a complete automatic annotation was performed.

Table 1: Sequence data integrated into TRUNCATULIX.

Project	Sequences	EC numbers	KOG numbers	GO numbers
MtGI 8.0 TCs & Singletons	36,878	6,174 (16.74%)	12,746 (34.56%)	10,268 (27.84%)
MtGI 9.0 TCs & Singletons	67,463	11,253 (16.68%)	20,570 (30.49%)	19,008 (28.18%)
Mt Genome 2.0	38,759	5,938 (15.32%)	3,434 (8.85%)	10,444 (26.95%)
Affymetrix Medicago GeneChip® probes	61,103	12,044 (19.71%)	18,731 (30.65%)	19,775 (32.36%)
Medicago 454 sequencing project	3,619	911 (25.17%)	1,798 (49.68%)	519 (14.34%)
Total	**207,822**	**36,320 (17.48%)**	**57,278 (27.56%)**	**60,014 (28.88%)**

The table shows the number of sequences integrated in TRUNCATULIX via the SAMS software. The annotation information was calculated using various bioinformatic tools.

• Affymetrix *Medicago* GeneChip® probes

Affymetrix [33] offers a GeneChip® microarray holding probes primarily for genes of *Medicago truncatula*, but also for the related legume *Medicago sativa* and their symbiontic *Sinorhizobium meliloti*. The sequences used by Affymetrix to construct the *Medicago* Genome GeneChip® were downloaded from the Affymetrix website and imported into SAMS. That way, 61,103 sequences containing the Affymetrix annotations were integrated into SAMS and were automatically re-annotated using the Metanor pipeline.

• *Medicago truncatula* 454 sequencing project

In 2006, Cheung *et al*. used the pyrosequencing approach to generate 292,465 cDNA reads of *Medicago truncatula* using a GS20 sequencer [34]. The reads were assembled forming 3,619 sequences. These sequences were downloaded from the project website and imported into SAMS. Using SAMS, a complete automatic annotation was performed.

Using Blast homology search, the sequences of all five projects were compared against the sequences of the other projects (one-by-one). This way, the sequences corresponding to each other in the different datasets could be found (using an e-value cutoff of e^{-5}). The complete sequence and annotation data from all five projects was integrated into the TRUNCATULIX data warehouse. Table 1 presents the number of sequences imported into TRUNCATULIX from the five projects, giving additional information about the automatic annotation.

Table 2: Microarray expression data imported from EMMA into TRUNCATULIX.

Experiment	Number of microarrays	Number of transformed datasets
Nitrogen-fixing root nodules in *Medicago truncatula**	4	10
Nod-Factor response in *Medicago truncatula* roots [43]	9	13
Root endosymbiosis in *Medicago truncatula* [5]	10	23
Uromyces pathogenesis in *Medicago truncatula***	3	4
AHL treatment of *Medicago truncatula roots****	11	17
LMW EPS I treatment of *Medicago truncatula* roots I*****	6	8
LMW EPS I treatment of *Medicago truncatula* roots II*****	6	8
LMW EPS I treatment of *Medicago truncatula* roots III*****	24	32
Nod-factor treatment of *Medicago truncatula* roots I*****	6	8
Nod-factor treatment of *Medicago truncatula* roots II*****	18	24
Seed development in *Medicago truncatula* [44]	22	51
Early Salt Stress in *Medicago truncatula* [45]	4	5
Cold stress in *Medicago truncatula******	8	11
Medicago truncatula wild type roots vs. TN1_11 mutant roots after 1 h of salt stress******	16	20
Response to phosphate in *Medicago truncatula* roots [5]	3	4
Total	**150**	**248**

The table shows the amount of microarray expression data integrated into TRUNCATULIX via EMMA. Currently there are 15 different experiments with a total of 150 microarrays (248 transformed datasets) integrated. Some of the integrated data is unpublised: H. Küster(*), M. Hahn, N. Hohnjec, H. Küster(**), D. Hinse, A. Becker, H. Küster(***), C. Hogekamp, H. Küster(****) and F. Frugier(*****). Abbreviations: AHL: acetylhomoserine lactone, LMW: low molecular weight, EPS: exopolysaccharide, Nod-factor: Nodulation factor

Table 3: GeneChip data integrated into TRUNCATULIX.

Experiment	Number of GeneChip arrays
Mature organs series	**24**
Leaf: 4-week old trifolia were harvest without their petioles (but with their petiolule) [46]	3
Petiole: Petioles from 4-week old plant [46]	3
Stem: Stems of 4-week old plants (without vegetative buds) [46]	3
Vegetative Bud: Vegetative buds of 4-week old plants [46]	3
Root: 4-week old non-inoculated roots [46]	3
Nodule: Nodules from 4-week old plants [46]	3
Flower: Fully open flowers were harvest at the day of anthesis [46]	3
Pod: Mix of small, medium and physiologically mature pods [46]	3
Nodulation development series	**12**
Root-0d: Roots at 0 dpi (control for nodule developmental series) [46]	3
Nod4d: Nodules at 4 dpi (root lumps with residual roots) [46]	3
Nod10d: Developing nodules at 10 dpi [46]	3
Nod14d: Mature nodules at 14 dpi [46]	3
Seed development series	**18**
Seed10d: Developing seeds at early embryogenesis – 10 dap [46]	3
Seed12d: Developing seeds at 12 dap (transition between embryogenesis and seed filling) [46]	3
Seed16d: Developing seeds at 16 dap (accumulation of storage proteins) [46]	3
Seed20d: Developing seeds at 20 dap (seed filling) [46]	3
Seed24d: Developing seeds at 24 dap (maturation phase) [46]	3
Seed36d: Developing seeds at 36 dap (physiologically mature seeds, desiccation) [46]	3
Total	**54**

The table shows the experiments and number of GeneChip arrays® directly imported into TRUNCATULIX. The three experiments address major topics: Mature organs covering the whole plant, nodulation development, and seed development.

Expression data

• Oligo-microarray expression data

In recent years, almost 1,000 oligo-microarrays studying *Medicago truncatula* gene expression in different conditions were hybridized in the framework of various international projects [35]. These microarrays used two chip layouts designated Mt16kOLI1 [5] and Mt16kOLI1Plus [36] (Arrayexpress ID: A-MEXP-85/A-MEXP-138). These arrays are associated to more than 50 different expression profiling experiments that were analyzed via the EMMA [37,38] software. EMMA is a tool that analyses microarray expression data and stores it in a MIAME compliant way [39]. It is also MAGE compliant [40] and the analysis pipelines used are well documented and evaluated. EMMA supports various image analyses software, such as ImaGene [41] and GenePix [42]. Within EMMA it is possible to group different mircoarrays into experiments. The user can decide to filter the differentially expressed genes using different significance tests. Some of the results of these analyses can be found in [1-3,35]. Table 2 lists the currently available *Medicago truncatula* oligo-microarray experiments, as well as the number of microarrays and the number of transformed datasets integrated into TRUNCATULIX via EMMA. Some of the integrated data is already published in [43-45], some more is unpublished data by H. Küster(*), M. Hahn, N. Hohnjec, H. Küster(**), D. Hinse, A. Becker, H. Küster(***), C. Hogekamp, H. Küster(****) and F. Frugier(*****). More microarray experiments will be integrated as soon as the researchers in charge have approved their integration into the data warehouse.

• GeneChips® expression data

In 2008, Benedito *et al*. hybridized more than 50 Affymetrix *Medicago truncatula* GeneChip arays® [46], addressing three major topics: mature organs covering the whole plant, nodule development, and seed development. For each of these topics, four to eight experiments were perfomed in three replicates each (see Table 3). The expression data of the GeneChip arrays was downloaded and directly migrated into the TRUNCATULIX data warehouse.

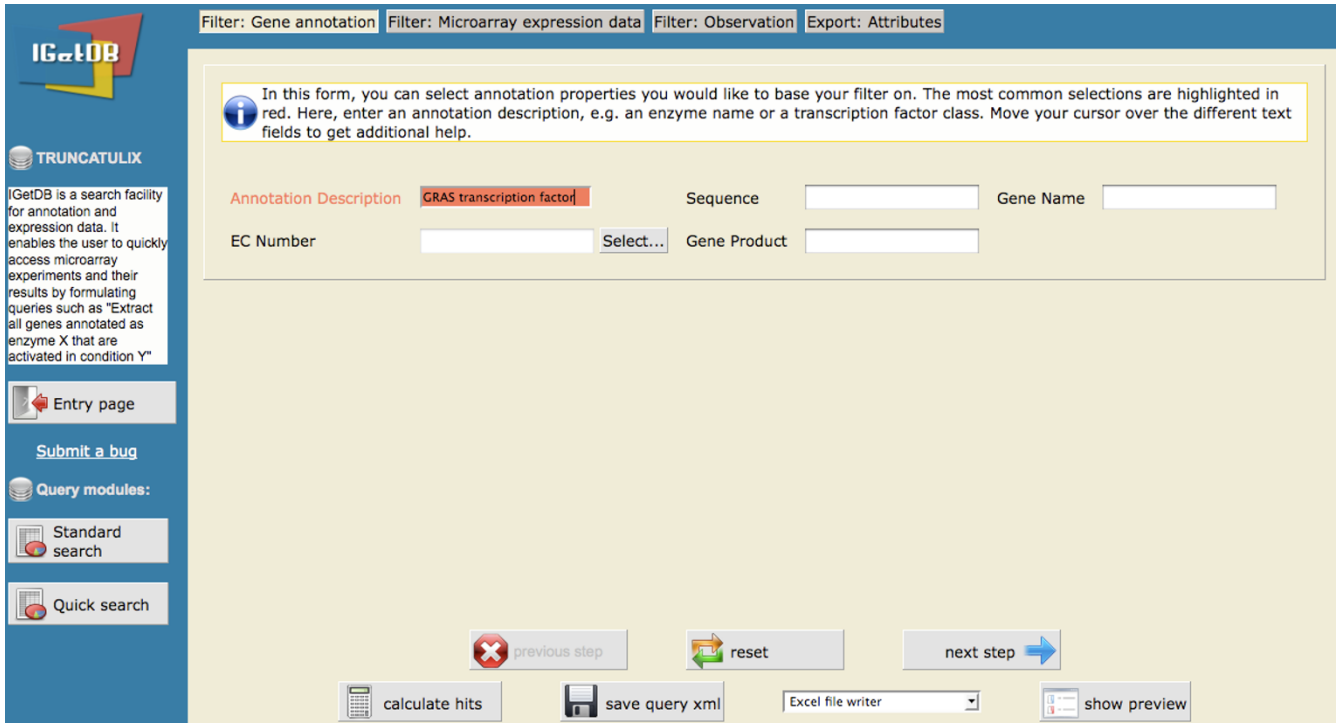

Figure 2
The first filterstep – gene annotation. The screenshot shows the filter page for the gene annotation data. The annotation descriptions can be queried, as well as the gene names, sequences and EC numbers.

The integrated data (sequence and expression data) is curated before it is imported into TRUNCATULIX. No user is allowed to import any datasets, this is reserved to the curators, but can be done upon request. In case of an update of one of the source databases the integrated data are updated by the curators.

Utility and Discussion

TRUNCATULIX is a relational, integrated database of sequence data, annotation information and microarray expression data, which is specially created to store data of the model legume *Medicago truncatula*.

Web interface

The interface for the TRUNCATULIX data warehouse can be accessed via a web browser. It provides a user-friendly interface built with Echo2 [47].

Case study

Consider a query for genes concerning GRAS transcription factors, suggesting that these genes are activated during nodulation [48,49]. As an example, we search for genes annotated as GRAS transcription factors in microarray experiments covering nodulation with three or more replicates.

To calculate this query, a user first specifies the standard search dialog on the left (see Figure 2). The user is guided through the different filter steps with an indicator bar on the top of the web pages. Help texts give examples and describe the options given by the different filters. Because some filters are more commonly used than others, the most commonly selected filters are highlighted in red.

For our example, the Annotation Description field is set to "GRAS transcription factor", due to our interest in GRAS genes. This way only genes that are annotated as GRAS transcription factor genes remain for the next filterstep.

The next page directly informs the user that only 222 entries passed the first filterstep and asks for a filtering concerning microarray expression data. The user now selects all nodulation related experiments ("Nitrogen-fixing root nodules in *Medicago truncatula*", "Nod-factor response in *Medicago truncatula* roots", "Root endosymbiosis in *Medicago truncatula*", "AHL treatment of *Medicago truncatula* roots", "LMW EPS I treatment of *Medicago truncatula* roots I", "LMW EPS I treatment of *Medicago truncatula roots* II", "LMW EPS I treatment of *Medicago truncatula* roots III", "Nod-factor treatment of *Medicago truncatula* roots I", "Nod-factor treatment of *Medicago truncatula* roots II", "Response to phosphate in *Medicago truncatula* roots", "Nodulation development series, Mt oligo-dT primed") as Experiments and sets the No. of Replicates (>) to "2" (Figure 3).

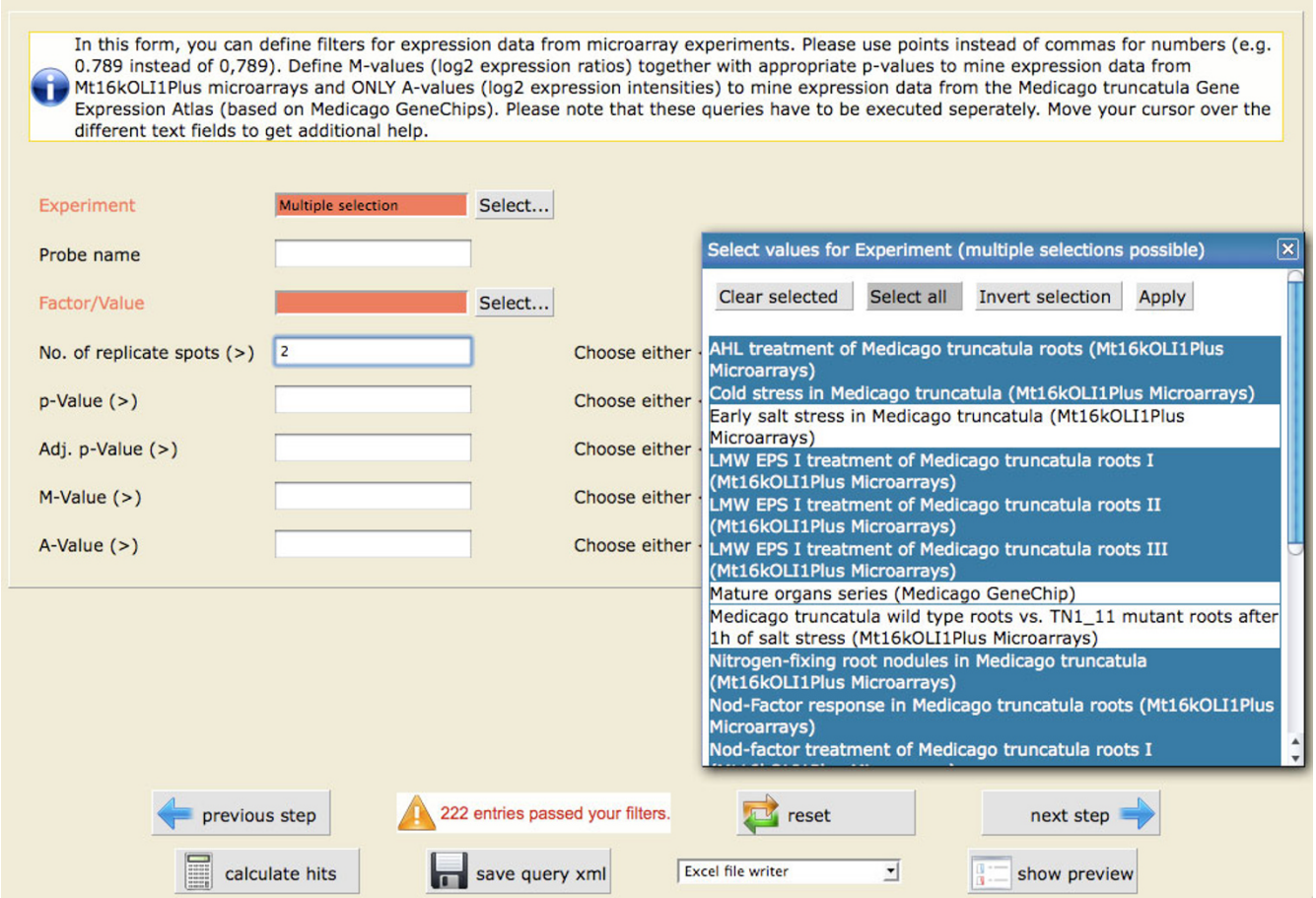

Figure 3
The second filterstep – gene expression. This screenshot shows the filter page for the expression data. The different integrated experiments can be selected, as well as different expression values and the number of replicates.

In the next step, the user can specify results for the different functional annotation tools, as well as for KOG Categories and Gene Ontology numbers. They remain unused in our example, but more complex queries could use these to filter for specific KOG categories or GO numbers.

The last page shows the possible export options that can be selected for the remaining 35 entries. The previously selected filter criteria are preselected for the export.

In this case, the export covers Database, Gene Name, Region Type, EC Number, and Annotation Description for the gene annotation; the Experiment name, Factor/Value, Probe name, p-Value, Adj. p-Value, M-Value, and no. of Replicates for the gene expression data from EMMA; the GO Number and the KOG Category, as well as the annotation Observations made by different Tools (Figure 4). Hitting the "calculate hits" button shows the number of datasets to be exported. This number may differ from the number of hits calculated previously based on sequence information because the new value is based on both sequence and expression information. To sneak a peek at the data, the user can use the preview option and browse through the first 100 results, or export the data as Excel, HTML, or csv file. A quick search – a more simple search interface found on the left side of the navigation panel – offers options to search for gene names, gene annotations, gene products, and repoter names using a google-like search interface (Figure 5).

After sumbitting the query, the details of the export can be selected as shown before using the complete query dialog (see Figure 4).

With the TRUNCATULIX data warehouse, it is now possible to have a look at the specific gene expressions for the different experiments and to find possibly interesting candidates for further experiments. Our example covers GRAS transcription factors. Some of them having already been reported to be key components of symbiotic signal transduction during nodulation [48,50,51].

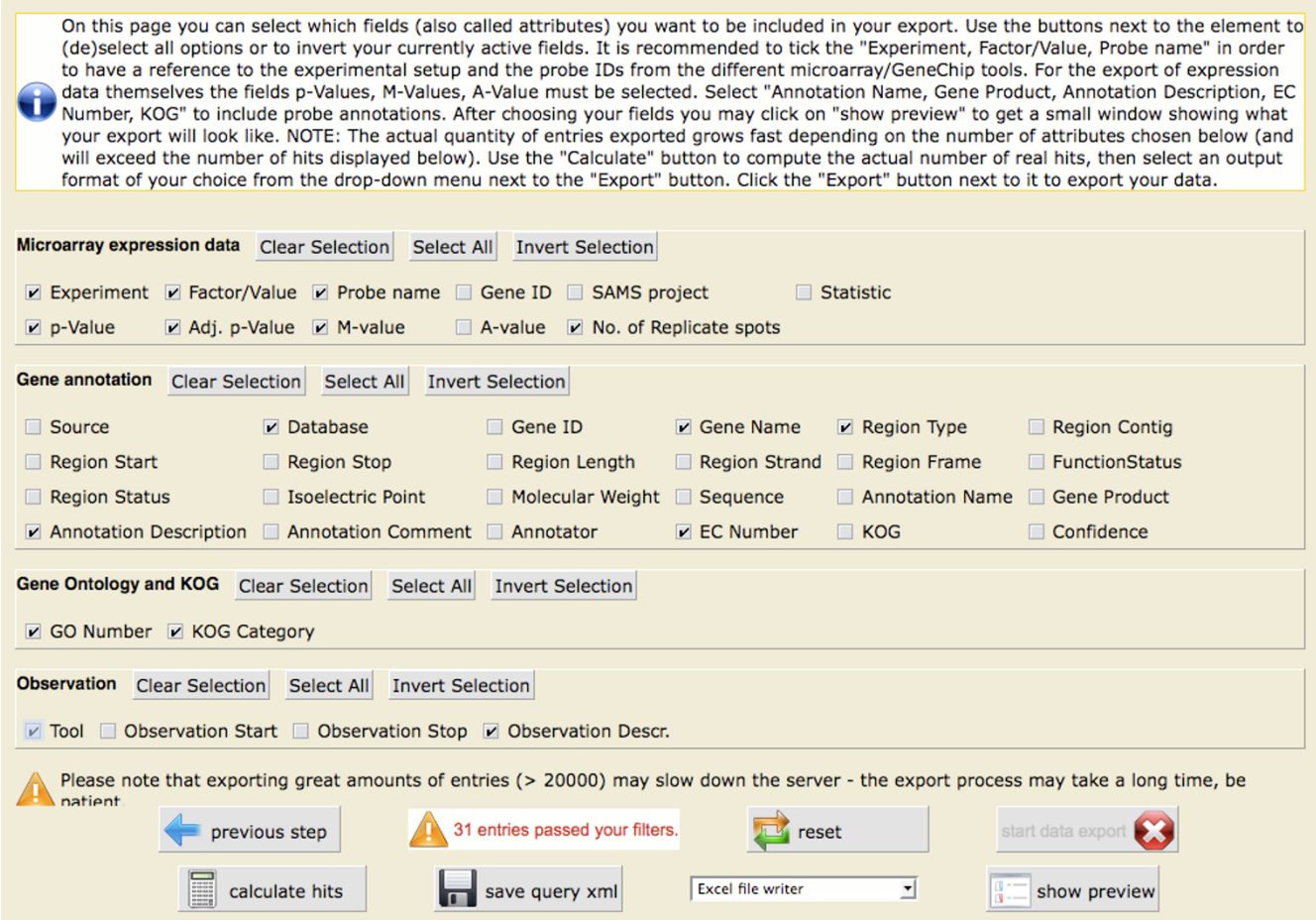

Figure 4
Export options of TRUNCATULIX. The screenshot shows the last page of the query dialog. The user can select which data and details should be exported, receiving an Excel, HTML, or csv file as result.

The TRUNCATULIX data warehouse allows users easy access to sequencing, annotation, and expression research done at many laboratories worldwide.

Discussion

Combining different data sources for fast searching and filtering is a widely used and common approach to overcome the immense manual work of searching for related data in every available database [52]. Related examples for this technique are **GeWare** [13] and the *Medicago truncatula* **Gene Expression Atlas** [46]. GeWare is a data warehouse that stores Affymetrix GeneChip® microarray data combined with manually added annotations and data from public databases like GO, Ensembl, LocusLink and NetAffx [28,53-55]. The GeWare data warehouse focuses on the processing and analysis of microarrays, mostly containing clinical data. Various filter and export options are implemented.

The *Medicago truncatula* Gene Expression Atlas stores previously processed and analyzed gene expression and annotation data from Affymetrix GeneChip® experiments. The data published is also available on ArrayExpress. The annotation data from Affymetrix is integrated into the data warehouse, but it can only be used for queries, it cannot be exported or viewed. In the same way, the GO numbers, KEGG functions, and the annotations of the genes can be queries. A homology search using Blast offers the opportunity to find genes according to their similarity to the Affymetrix GeneChip® reporter or consensus sequences. The results of a query can be downloaded, but only the names of the reporters and the expression values are listed, the annotations are not shown and cannot be extracted.

In contrast to the other data warehouses, TRUNCATLIX not only stores the sequence and annotation data of the Affymetrix GeneChip arrays, but also sequence data from other genetic projects and institutes. The annotations that are provided are calculated using a well evaluated pipeline, offering KEGG, KOG, and GO numbers, if found in the annotations of homologues. TRUNCATULIX offers

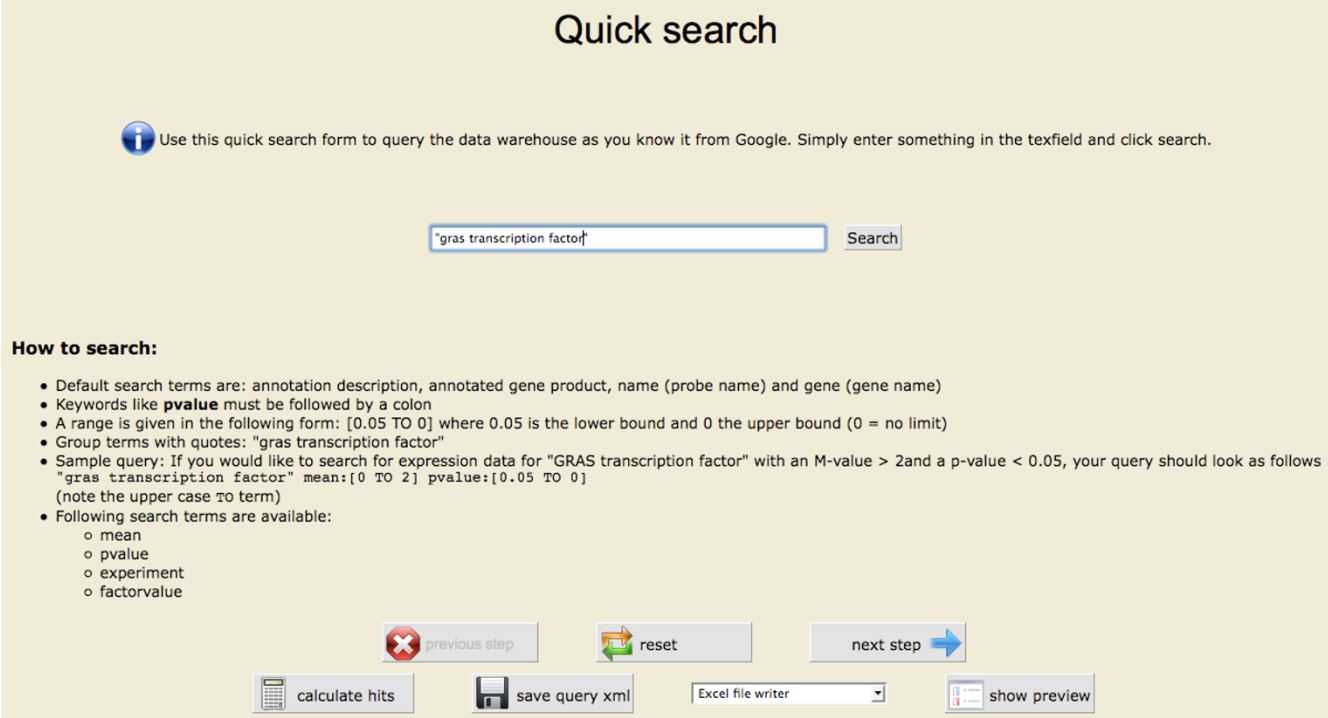

Figure 5
The simple search dialog. The screenshot shows the simple search dialog. The user can search for text fragments in the annotation of the genes, the gene names, the gene products, and the reporter names.

the option to export the results of a query, but in contrast to the the export of the *Medicago truncatula* Gene Expresison Atlas, not only the gene names and the expression values are exported, but (on request) also any other information that is integrated into the data warehouse. Table 4 compares the main features of the three data warehouses. The immense amount of intergrated sequence, annotation, and expression information makes the TRUNCATULIX data warehouse a very convenient resource in the field of *Medicago truncatula* research. TRUNCATULIX is freely available to the research community and additional expression data can be integrated upon request.

Future development

As more *Medicago trunculata* data becomes available from oligo-microarray and Affymetrix GeneChip® experiments, it will be integrated into TRUNCATULIX. Additionally, a homology search using Blast will be implemetend.

Table 4: A comparison of TRUNCATULIX to other data warehouses

Data Warehouse feature	GeWare	*Medicago truncatula* Gene Expression Atlas	TRUNCATULIX
Target organism	*homo sapiens*	*Medicago truncatula*	*Medicago truncatula*
static/dynamic data	dynamic	Static	static
number of microarray experiments	unknown	18	18
number of microarrays mircoarrays	unknown	54	204
automatic annotation information	yes	searchable, but not visible	yes
KEGG-mapping	no	yes	yes
GO-numbers	no	yes	yes
search	yes	yes	yes
blast homology search	no	yes	no, planned
export options	yes	yes	yes
free use	yes	yes	yes
free access	no	yes	yes

The table shows a comparison of the three data warhouses GeWare, the *Medicago truncatula* Gene Expression Atlas, and TRUNCATULIX.

Conclusion

We created TRUNCATULIX, a data warehouse that combines data from microarray experiments with sequence data and high quality annotations in the area of *Medicago truncatula*. TRUNCATULIX is the first data warehouse in the field or *Medicago truncatula* research that offers the opportunity to search in all publicly available *Medicago truncatula* sequence data and expression data for different criteria and as a result to get a complete list of sequences, expression data, and annotations. Thus, a researcher can save much time and work finding interesting genes and results of previously conducted expression experiments. As the application uses an AJAX-based web interface, it can be used via a web browser and is platform independent. The results can be exported as Excel, HTML, or as csv files.

Availability and requirements

The TRUNCATULIX data warehouse is freely available at http://www.cebitec.uni-bielefeld.de/truncatulix/.

Authors' contributions

KH initialized the project, computed the annotations for the sequence data and is the main author of the manuscript. KR was responsible for the data schema, the data warehouse backend, and supervised the design of the web interface. TB helped on annotating the sequence data using SAMS. MD contributed in extracting of the microarray data from EMMA. TJ designed the web interface. HK analyzed some of the microarray data and supervised the export. AG supervised the integration of the different data sources. All authors revised and approved the final manuscript.

Acknowledgements

KH thanks the International NRW Graduate School in Bioinformatics and Genome Research for funding the project. We thank Florian Frugier (ISV, Gif-sur-Yvette) for making available data prior to publication.

References

1. Baier M, Barsch A, Küster H, N H: **Antisense repression of the Medicago truncatula nodule-enhanced sucrose synthase leads to a handicapped nitrogen fixation mirrored by specific alterations in the symbiotic transcriptome and metabolome.** *Plant Physiol* 2007, **145(4):**1600-1618.
2. Gallardo K, Firnhaber C, Zuber H, Héricher D, Belghazi H, Henry C, Küster H, R T: **A combined proteome and transcriptome analysis of developing Medicago truncatula seeds: Evidence for metabolic specialization of maternal and filial tissues.** *Mol Cell Proteomics* 2007, **6(12):**2165-2179.
3. Hohnjec N, Henckel K, Bekel T, Gouzy J, Dondrup M, Goesmann A, Küster H: **Transcriptional snapshots provide insights into the molecular basis of arbuscular mycorrhiza in the model legume Medicago truncatula.** *Functional Plant Biology* 2006, **33(8):**737-748.
4. Barsch A, Tellström V, Patschkowski T, Küster H, K N: **Metabolite profiles of nodulated alfalfa plants indicate that distinct stages of nodule organogenesis are accompanied by global physiological adaptations.** *Mol Plant-Microbe Interact* 2006, **19(9):**998-1013.
5. Hohnjec N, Vieweg M, Pühler A, Becker A, Küster H: **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* 2005, **137:**1283-1301.
6. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A: **ArrayExpress-a public database of microarray experiments and gene expression profiles.** *Nucleic Acids Res* 2007, **35(Database issue):**D747-D750.
7. Chen J, Zhao P, Massaro D, Clerch L, Almon R, DuBois D, Jusko W, Hoffman E: **The PEPR GeneChip data warehouse, and implementation of a dynamic time series query tool (SGQT) with graphical interface.** *Nucleic Acids Res* 2004, **32(Database issue):**578-581.
8. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese J, Dwight S, Kaloper M, Weng S, Jin H, Ball C, Eisen M, Spellman P, Brown P, Botstein D, Cherry J: **The Stanford Microarray Database.** *Nucleic Acids Res* 2001, **29(1):**152-155.
9. Wise RP, Caldo RA, Hong L, Shen L, Cannon E, Dickerson JA: **BarleyBase/PLEXdb.** *Methods Mol Biol* 2007, **406:**347-363.
10. Boguski M, Lowe T, CM T: **dbEST-database for "expressed sequence tags".** *Nat Genet* 1993, **4(4):**332-3.
11. Morley M, Arcaro M, Burdick J, Yonescu R, Reid T, Kirsch IR, Cheung VG: **GenMapDB: a database of mapped human BAC clones.** *Nucleic Acids Res* 2001, **29:**144-147.
12. **Short Read Archive** [http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi]
13. Rahm E, Kirsten T, Lange J: **The GeWare data warehouse platform for the analysis of molecular-biological and clinical data.** *Journal of Integrative Bioinformatics* 2007, **4:**47.
14. Hu H, Brzeski H, Hutchins J, Ramaraj M, Qu L, Xiong R, Kalathil S, Kato R, Tenkillaya S, Carney J, Redd R, Arkalgudvenkata S, Shahzad K, Scott R, Cheng H, Meadow S, McMichael J, Sheu SL, Rosendale D, Kvecher L, Ahern S, Yang S, Zhang Y, Jordan R, Somiari SB, Hooke J, Shriver CD, Somiari RI, Liebman MN: **Biomedical informatics: development of a comprehensive data warehouse for clinical and genomic breast cancer research.** *Pharmacogenomics* 2004, **5(7):**933-941.
15. **IGetDB** [http://www.cebitec.uni-bielefeld.de/groups/brf/software/igetdb_info/]
16. **Java** [http://java.sun.com/]
17. **MySQL** [http://www.mysql.com/]
18. Durinck S, Moreau Y, Kasprzyk A, Davis S, Moor BD, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 2005, **21(16):**3439-40.
19. Zimmermann P, Hennig L, Gruissem W: **Gene expression analysis and network discovery using Genevestigator.** *Trends Plant Sci* 2005, **9(10):**407-409.
20. Kimball R, Caserta J: *Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. 0-7645-6757-8, John Wiley* 2004.
21. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J, Quackenbush J: **TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets.** *Bioinformatics* 2003, **19(5):**651-652.
22. Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R, White J: **The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species.** *Nucleic Acids Res* 2001, **29:**159-164.
23. Bekel T, Henckel K, Dondrup M, Küster H, Meyer F, Mittard-Runte V, Neuweger H, Paarmann D, Rupp O, Zakrzewski M, Pühler A, Stoye J, Goesmann A: **The Sequence Analysis and Management System-SAMS-2.0: Data management and sequence analysis adapted to changing requirements from traditional sanger sequencing to ultrafast sequencing technologies.** *J Biotechnol* in press.
24. Altschul S, Gish W, Miller W, Myers E, Lipman D: **Basic Local Alignment Search Tool.** *J Mol Biol* 1990, **215(3):**402-410.
25. Mulder N, Apweiler R: **InterPro and InterProScan: tools for protein sequence classification and comparison.** *Methods Mol Biol* 2007, **396:**59-70.
26. Sonnhammer E, von Heijne G, Krogh A: **A hidden Markov model for predicting transmembrane helices in protein sequences.** *Proc Int Conf Intell Syst Mol Biol* 1998, **6:**175-82.

27. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.
29. Tatusov R, Fedorova N, Jackson J, Jacobs A, Kiryutin B, Koonin E, Krylov D, Mazumder R, Mekhedov S, Nikolskaya A, Rao B, Smirnov S, Sverdlov A, Vasudevan S, Wolf Y, Yin J, Natale D: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **11(4):**41.
30. **The qcGene Database** [http://compbio.dfci.harvard.edu/tgi/qcGene.html]
31. Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T, Spannagl M, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KFX, Rogers J, Quetier F, Oldroyd GE, Debelle F, Cook DR, Retzel EF, Roe BA, Town CD, Tabata S, Peer Y Van de, Young ND: **Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes.** *Proc Natl Acad Sci USA* 2006, **103(40):**14959-14964.
32. Young ND, Cannon SB, Sato S, Kim D, Cook DR, Town CD, Roe BA, Tabata S: **Sequencing the genespaces of Medicago truncatula and Lotus japonicus.** *Plant Physiol* 2005, **137(4):**1174-1181.
33. **Affymetrix** [http://www.affymetrix.com/]
34. Cheung F, Haas B, Goldberg S, May G, Xiao Y, Town C: **Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology.** *BMC Genomics* 2006, **7:**272.
35. Küster H, Becker A, Firnhaber C, Hohnjec N, Manthey K, Perlick A, Bekel T, Dondrup M, Henckel K, Goesmann A, Meyer F, Wipf D, Requena N, Hildebrandt U, Hampp R, Nehls U, Krajinski F, Franken P, Pühler A: **Development of bioinformatic tools to support EST-sequencing, in silico- and microarray-based transcriptome profiling in mycorrhizal symbioses.** *Phytochemistry* 2007, **68:**19-32.
36. Thompson R, Ratet P, Küster H: **Identification of gene functions by applying TILLING and insertional mutagenesis strategies on microarray-based expression data.** *Grain Legumes* 2005, **41:**20-22.
37. Dondrup M, Goesmann A, Bartels D, Kalinowski J, Krause L, Linke B, Rupp O, Sczyrba A, Pühler A, Meyer F: **EMMA: a platform for consistent storage and efficient analysis of microarray data.** *J Biotechnol* 2003, **106(2–3):**135-46.
38. Dondrup M, Albaum S, Griebel T, Henckel K, Jünemann S, Kahlke T, Kleindt C, Küster H, Linke B, Mertens D, Mittard-Runte V, Neuweger H, Runte K, Tauch A, Tille F, Pühler A, Goesmann A: **EMMA 2 – A MAGE-compliant system for the collaborative analysis and integration of microarray data.** *BMC Bioinformatics* in press.
39. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball H, Causton C, Gaasterland T, Glenisson P, Holstege F, Kim I, Markowitz V, Matese J, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M: **Minimum information about a microarray experiment (MIAME) – toward standards for microarray data.** *Nature Genetics* 2001, **29:**365-371.
40. Spellman P, Miller M, Stewart J, Troup C, Sarkans U, Chervitz S, Bernhart D, Sherlock G, Ball C, Lepage M, Swiatek M, Marks W, Goncalves J, Markel S, Iordan D, Shojatalab M, Pizarro A, White J, Hubley R, Deutsch E, Senger M, Aronow B, Robinson A, Bassett D, Stoeckert C Jr, Brazma A: **Design and implementation of microarray gene expression markup language (MAGE-ML).** *Genome Biology* 2002, **3:**research0046.1-0046.9.
41. **Biodiscovery** [http://www.biodiscovery.com/]
42. **MDS Analytical Technologies** [http://www.moleculardevices.com/]
43. Combier JP, Kuster H, Journet EP, Hohnjec N, Gamas P, Niebel A: **Evidence for the involvement in nodulation of the two small putative regulatory peptide-encoding genes MtRALFL1 and MtDVL1.** *Mol Plant Microbe Interact* 2008, **21(8):**1118-1127.
44. Gallardo K, Firnhaber C, Zuber H, Héricher D, Belghazi M, Henry C, Küster H, Thompson R: **A combined proteome and transcriptome analysis of developing Medicago truncatula seeds: Evidence for metabolic specialization of maternal and filial tissues.** *Mol Cell Proteomics* 2007, **6(12):**2165-2179.
45. Gruber V, Blanchet S, Diet A, Zahaf O, Boualem A, Kakar K, Alunni B, Udvardi M, Frugier F, Crespi M: **Identification of transcription factors involved in root apex responses to salt stress in Medicago truncatula.** *Mol Genet Genomics* 2009, **281:**55-66.
46. Benedito V, Torres-Jerez I, Murray J, Andriankaja A, Allen S, Kakar K, Wandrey M, Verdier J, Zuber H, Ott T, Moreau S, Niebel A, Frickey T, Weiller G, He J, Dai X, Zhao P, Tang Y, Udvardi M: **A gene expression atlas of the model legume Medicago truncatula.** *Plant J* 2008, **55(3):**504-513.
47. **Echo Web Framework** [http://echo.nextapp.com/site/]
48. Kaló P, Gleason C, Edwards A, Marsh J, Mitra R, Hirsch S, Jakab J, Sims S, Long S, Rogers J, Kiss G, Downie J, Oldroyd G: **Nodulation signaling in legumes requires NSP2, a member of the GRAS family of transcriptional regulators.** *Science* 2005, **308:**1786-1789.
49. Smit P, Raedts J, Portyanko V, Debelle F, Gough C, Bisseling T, Geurts R: **NSP1 of the GRAS protein family is essential for rhizobial Nod factor-induced transcription.** *Science* 2005, **308:**1789-1791.
50. Heckmann A, Lombardo F, Miwa H, Perry J, Bunnewell S, Parniske M, Wang T, Downie J: **Lotus japonicus Nodulation Requires Two GRAS Domain Regulators, One of Which Is Functionally Conserved in a Non-Legume.** *Plant Physiol* 2006, **142(4):**1739-1750.
51. Limpens E, Bisseling T: **Signaling in symbiosis.** *Current Opinion in Plant Biology* 2003, **6(4):**343-350.
52. Kimball R, Margy R: *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. ISBN 0471200247* John Wiley & Sons Inc; 2002.
53. Liu G, Loraine A, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose M: **NetAffx: Affymetrix probesets and annotations.** *Nucleic Acids Res* 2003, **31(1):**82-86.
54. Birney E, Andrews T, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T, Down T, Eyras E, Fernandez-Suarez X, Gane P, Gibbins B, Gilbert J, Hammond M, Hotz H, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Lehvaslaiho H, McVicker G, Melsopp C, Meidl P, Mongin E, Pettett R, Potter S, Proctor G, Rae M, Searle S, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Ureta-Vidal A, Woodwark K, Cameron G, Durbin R, Cox A, Hubbard T, Clamp M: **An Overview of Ensembl.** *Genome Res* 2004, **14:**925-928.
55. Pruitt K, Maglott D: **RefSeq and LocusLink: NCBI gene-centered resources.** *Nucleic Acids Research* 2001, **29:**137-140.