Universität Bielefeld

Technische Fakultät
Center for Biotechnology (CeBiTec)

Development of a software infrastructure to mine GeneChip expression data and to combine datasets from different Medicago truncatula expression profiling platforms.

Zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften an der Technischen Fakultät der Universität Bielefeld vorgelegte Dissertation

von

Kolja Henckel

19. Januar 2010

Kolja Henckel
Rehhagenhof 57
33619 Bielefeld
khenckel@cebitec.uni-bielefeld.de

Supervisors: Prof. Dr. Ralf Hofestädt
Prof. Dr. Helge Küster

Printed on acid-free paper.

Summary

Medicago truncatula is a model plant for studying legume biology. The ability to interact with beneficial microbial organisms leading to the formation of nitrogen fixing root nodules and to phosphate-acquiring arbuscular mycorriza (AM) is one of the main distinctive features of this family of plants. The two different symbioses of *Medicago truncatula* are investigated by various international research projects.

Oligonucleotide microarrays are a robust technique to examine the expression of thousands of genes in parallel. Affymetrix GeneChips®, more recently designed gene-specifc chips, make it easier for the researcher to compare and evaluate gene expression and thus will most certainly lead to more accurate results. Not surprisingly, Medicago GeneChips® are moving into the focus of gene expression analysis research in this model plant. Software applications for the analysis of GeneChips® are mostly commercial, or implemented as command-line tools without a user interface. Furthermore, a comparison to the analyses of previously perfomed oligonucleotide microarrays is difficult, as analysis pipelines and methods differ in each application. In the scope of this thesis EMMA2, an application for the analysis of oligonucleotide microarrays, was extended to load, store and analyze Affymetrix GeneChips® as compareable as possible to oligonucleotide datasets.

Databases for either sequence, annotation, or microarray experiment datasets are extremely beneficial to the research community, as they centrally gather information from experiments performed by different scientists. However, datasets 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 was no such single data warehouse that integrated all freely available gene sequences, the corresponding gene expression data, and annotation information. The TRUNCATULIX data warehouse is created in the scope of this thesis to store *Medicago truncatula* sequence, annotation, and expression datasets and offer these to the legume community. Different filtersteps allow a precise query for genes and expression values in a database of over 200.000 gene sequences and over 200 hybridizations. For the first time users can now quickly search for specific genes and gene expression datasets in a huge database based on high-quality annotations. The results can be exported as Excel, HTML, or as csv files for further usage.

A multitude of EST and microarray experiments are conducted for *Medicago truncatula* covering different tissues, cell states, and cell types. Under these circumstances the challenge arises to integrate the results of the different expression analysis methods with the goal to discover novel information from the combined datasets. The application MediPlEx is designed to allow an integrated expression analysis for the *Medicago truncatula* datasets stored in SAMS and in the TRUN-

CATULIX data warehouse. After selecting genes of interest by their expression conditions, expression profiles are combined for a hierarchical clustering. The results are presented in a table, as a cluster dendrogram, and in an interactive 3D application.

The three parts of the thesis have been published by Dondrup *et al.* (2009a), Henckel *et al.* (2009), or are submitted (Henckel *et al.* (2010)).

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Gene expression analysis plays a major role in answering biological questions. Using recent biological analysis methods like microarrays, the expression of thousands of genes can be analyzed in parallel. Sequencing based gene expression analysis methods, like Expressed Sequence Tag (EST) library analysis, or 454 mRNA sequencing also offer good analysis results. The task of combining the results of different analysis methods is a challenge for computational biology, providing new insights from already created datasets.

Medicago truncatula (see Figure 1.1) is a model plant for studying legume biology. In addition to the phosphate-acquiring arbuscular mycorrhiza (AM) symbiosis, legumes such as *Medicago truncatula* are characterized by their ability to form a nitrogen-fixing root nodule to interact with beneficial microbial organisms. The two different symbioses of *Medicago truncatula* are investigated by various international research projects. The AM interactions between the host root and the fungal partner are an interesting field of research, since more than 80% of land plants depend on an efficient AM for the uptake of nutrients, primarily phosphate. By recruiting a basic genetic program allowing microbial infection, legumes such as *Medicago truncatula* have evolved the capacity to enter a nitrogen-fixing symbiosis with the soil bacterium *Sinorhizobium meliloti*. Symbiotic nitrogen fixation allows legumes such as *Medicago truncatula* to grow on nitrogen-depleted soils and to develop protein-rich seeds, properties exploited in sustainable agriculture[Baier *et al.* (2007); Gallardo *et al.* (2007); Hohnjec *et al.* (2005, 2006); Barsch *et al.* (2006)].

Figure 1.1: A picture of *Medicago truncatula*. The model legume is capable of forming nitrogen-fixing root nodules in a symbiotic interaction with fungi and to enter a phosphate-acquiring arbuscular mycorrhiza (AM) symbiosis. Picture adopted from http://www.noble.org.

1.1 Motivation

There are many freely available tools for the analysis of cDNA microarrays. Unfortunately, most applications are not able to load and analyze more than one kind of microarray (e.g. conventional oligonucleotide microarrays, Affymetrix GeneChip® microarrays[1], or Agilent chips). As GeneChips® start playing a major role in microarray analysis, this feature is mostly requested by biologists recently. Thus, a free to use application to analyze Affymetrix GeneChips® and compare them to the results obtained with oligonucleotide microarrays is of essential interest.

Sequencing projects often offer their new results (sequences and annotations) as downloadable files, or sometimes in an open access database. As time passes, more and more databases hosting sequences of one organism arise and researchers can get distracted in searching for results of interest. Microarray gene expression experiments are often stored in public access repositories, allowing the download of the datasets, without providing specific analyses or expression queries.
Data warehouses are designed to integrate datasets from different databases, combining information about one specific item (e.g. a gene) from many repositories. As a benefit, cross-resource analyses are possible, allowing to combine queries for attributes of different source databases. In the field of *Medicago truncatula* research, no such data warehouse is available, even though many different sequencing

[1] http://www.affymetrix.com/

projects and various microarray expression experiment datasets are available.

As many EST analysis projects and microarray gene expression analyses were conducted in the recent past, the results of these two transcriptome analysis methods could be combined and used for an integrative analysis. Only one available application, Simcluster created by Vencio *et al.* (2007), tries to combine datasets of different transcriptome analysis methods. Unfortunately, the datasets need to be converted to the simplex space (a theoretical mathematic space), which is used in the analysis. Furthermore, the application is unstable, does neither feature a database connection, nor a user interface, which, altogether, makes it almost unusable.

1.2 Goals

As pointed out in the previous section, the currently available tools for transcriptome analysis concentrate on the analysis of only one kind of expression analysis, which means either EST library analysis, oligonucleotide microarray analysis, or Affymetrix GeneChip® analysis.
The *Medicago truncatula* research community cannot access and query all *Medicago truncatula* sequence and expression datasets at once, as there is no single data warehouse offering these datasets and services. To search all information about a single gene, the researcher has to search manually in different data repositories to find all available information.
As another point, the combination of the results of the different transcriptome experiments for further analyses is not practical at the moment.

The three goals of this thesis can directly be derived from these limitations.

Adaption of EMMA2 for the analysis of Affymetrix GeneChip® expression datasets.
As the Affymetrix GeneChip® microarrays and oligonucleotide microarrays should be analyzed as comparable as possible, EMMA2 is to be enhanced to read, store and analyze Affymetrix GeneChip® microarray datasets. Therefore, the processing of the raw datasets, the analysis of the preprocessed datasets, as well as the expression analyses are to be implemented in a way that they are comparable to the classical oligonucleotide microarrays.

Creation of a data warehouse for *Medicago truncatula* datasets.
For a fast retrieval of sequence and microarray expression datasets in the field of *Medicago truncatula* research, a data warehouse is to be created to store freely available sequences, annotations, and microarray expression datasets. The data warehouse should be useable as stand-alone tool, as well as a service to offer the sequence and expression datasets for other applications via an Application

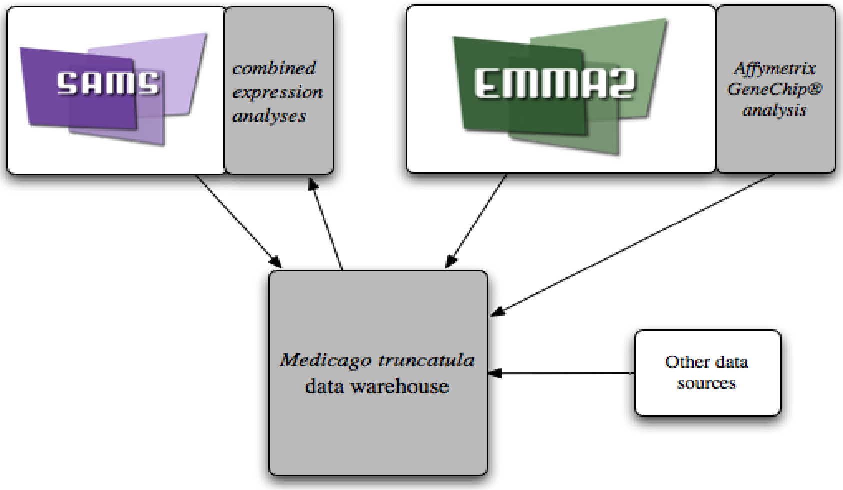

SAMS
combined expression analyses
EMMA2
Affymetrix GeneChip® analysis
Medicago truncatula data warehouse
Other data sources

Figure 1.2: Scheme of the interaction of the proposed applications. EMMA2 is to be extended to analyze Affymetrix GeneChips® and export datasets from these experiments and classical oligonucleotide microarray experiments to the *Medicago truncatula* data warehouse. Sequence and annotation datasets from SAMS are imported into the data warehouse, additionally datasets of other freely available *Medicago truncatula* datasets. The combined expression analysis, as a part of SAMS, can access the data warehouse for a fast data retrieval. The grey boxes indicate the tools to be implemented.

Programming Interface (API).

Integration and analysis of gene expression datasets from different transcriptome experiments in the scope of *Medicago truncatula*.
The main goal of this thesis is to combine EST and microarray expression datasets and analyze them together. For achieving this, an application is to be created on the basis of SAMS that allows to select datasets of these different gene expression analysis methods and to analyze and evaluate them together. The results of this analysis should be presented in a structured way. The resulting datasets should also be available for download.

A scheme of the proposed extension and interaction of the applications is shown in Figure 1.2.

1.3 Structure of the thesis

Following this introduction, Chapter 2 introduces the biological and computational background used within this thesis. In this chapter, the methods used in cDNA library analysis are documented. Afterwards, the techniques of EST expression analysis, as well as the computational EST analysis methods are presented. Subsequently, two different kinds of microarrays are illustrated. Referring to this, computational methods for the analysis of the results of the microarray experiments are pointed out.

Chapter 3 describes the existing systems that are available for the analysis of the different biological data. SAMS is explained in detail for the analysis of EST datasets, different applications for the analysis of microarrays are compared. Simcluster, the only application allowing to combine different expression experiment results is presented.

The fourth chapter deals with the design of an infrastructure to fulfill the previously developed goals. This includes the extension of EMMA2 to store and analyze GeneChip® datasets, the TRUNCATULIX data warehouse, as well as the MediPlEx expression analysis tool.

The next chapter provides the implementation of the previously designed applications and extensions.

Results of the different implementations are presented in Chapter 6. The benefit of each of the implemented tools is demonstrated in the context of *Medicago truncatula*. Additionally, some reslts of *Arabidopsis thaliana* GeneChip® analyses are shown.

Chapter 7 reflects the thesis and provides a summary and a discussion. Finally, an outlook to possible future improvements is given.

CHAPTER 2

Background

This Chapter gives detailed biological and computer science background information used in this thesis. In the first part, the techniques of cDNA analysis are described, among these are cDNA library creation, sequencing, EST expression analysis and computer aided analysis of these datasets. 454 ultra-fast sequencing as a new sequencing method is presented, as it can be used to sequence mRNA faster than using ESTs. Afterwards, gene expression analysis using microarrays is introduced, covering the topics oligonucleotide microarrays, Affymetrix GeneChips®, and gene expression analysis. As a last topic, the techniques of data warehouses storing different types of datasets are presented.

2.1 cDNA analysis

Deoxyribonucleic acid (DNA) stores the information coding for all genes of an organism. During transcription, DNA is transcribed into messenger ribonucleic acid (mRNA), which then is further translated into proteins. Thus, mRNA is the primary indicator of gene expression and therefore used in Expressed Sequence Tag (EST) analysis and for microarray expression analysis [Knippers (2006)]. ESTs are mostly used to gain a first insight into the transcriptome of a species of interest. It has recently become possible to analyze mRNA using ultra-fast sequencing methods, which is much faster and cheaper than EST analysis.

2.1.1 cDNA library creation & EST generation

Complementary DNA (cDNA) libraries provide the biological background that is used for EST analysis. These datasets can be used for *in silico* expression analyses.

To create an EST library for a special tissue, mRNA is extracted from a sample and further processed synthesizing cDNA. The cDNA is used to create the EST library (Figure 2.1): An oligonucleotide made of deoxythymidin-nucleotides (oligo dT) binds at the complementary polyA-tail of the 3'end of the mRNA. This oligo dT operates as a primer for the reverse transcriptase, which synthesizes the first cDNA strand on the mRNA. When this step is finished, RNAseH (a special ribonuclease) is added, hydrolyzing the mRNA. The reaction is stopped before the complete RNA strand is denatured, so that some short pieces of RNA remain at the DNA strand. These pieces act as primers for the now added DNA polymerase. The 3'-ends are used as starting points for this synthesis, while in the same time the remaining RNA is removed by 5'-3'-exonuclease. For further processing in a vector it is necessary to chop the overlaying single-strand parts with the use of 3'-5'exonuclease. The next step is to prepare the ends of the double-stranded cDNA to fit into a cloning-vector. Therefore, adaptors are added to the ends of the cDNA. The adaptors are carefully selected to fit the cleavage site of the target vector-DNA. The double-stranded cDNA is cloned into a plasmid vector. A cDNA library is created by inserting the plasmid vector into a target bacteria by transformation. Afterwards the clones are cultured. Finally the plasmid DNA is extracted from the clones and the cDNA is sequenced: This step is done with the chain terminator sequencing method using dye terminator marking. In this linear PCR-based (polymerase chain reaction) sequencing technology (Sanger sequencing), extension is initiated at a specific site on the template DNA by using a short oligonucleotide primer complementary to the vector. The oligonucleotide primer is extended using a DNA polymerase. Included with the primer and DNA polymerase are the four deoxynucleotide bases, along with a low concentration of a chain terminating nucleotides marked with different fluorescent dyes. Limited incorporation of the chain terminating nucleotide by the DNA polymerase results in a series of related DNA fragments that are terminated only at positions where that particular nucleotide is used. A gel electrophoresis is applied to these DNA fragments (Figure 2.2). The fragments pass a laser (beginning with the shortest fragment), the fluorescence-marked nucleotides emit different wavelengths of light, which are observed and stored as raw chromatogram files.

2.1.2 Expression analysis using pyrosequencing

In the last years, pyrosequencing technologies evolved and revolutionized sequencing all over the world. The probably most widespread pyrosequencing technology is the 454 sequencing developed by 454 Life Sciences (Roche). Due to the experimental setup the sequencing steps for different samples (genomic DNA, PCR products, bacterial artificial chromosomes (BACs), and cDNA) are nearly the same and differ only in preprocessing. Short reads like cDNA are used as they are, longer reads, like genomic DNA and BACs are fractionated into fragments of 300 to 800 basepairs length. Short PCR products are amplified using Genome Sequencer fusion primers. mRNA is transcribed into cDNA, which can subsequently be

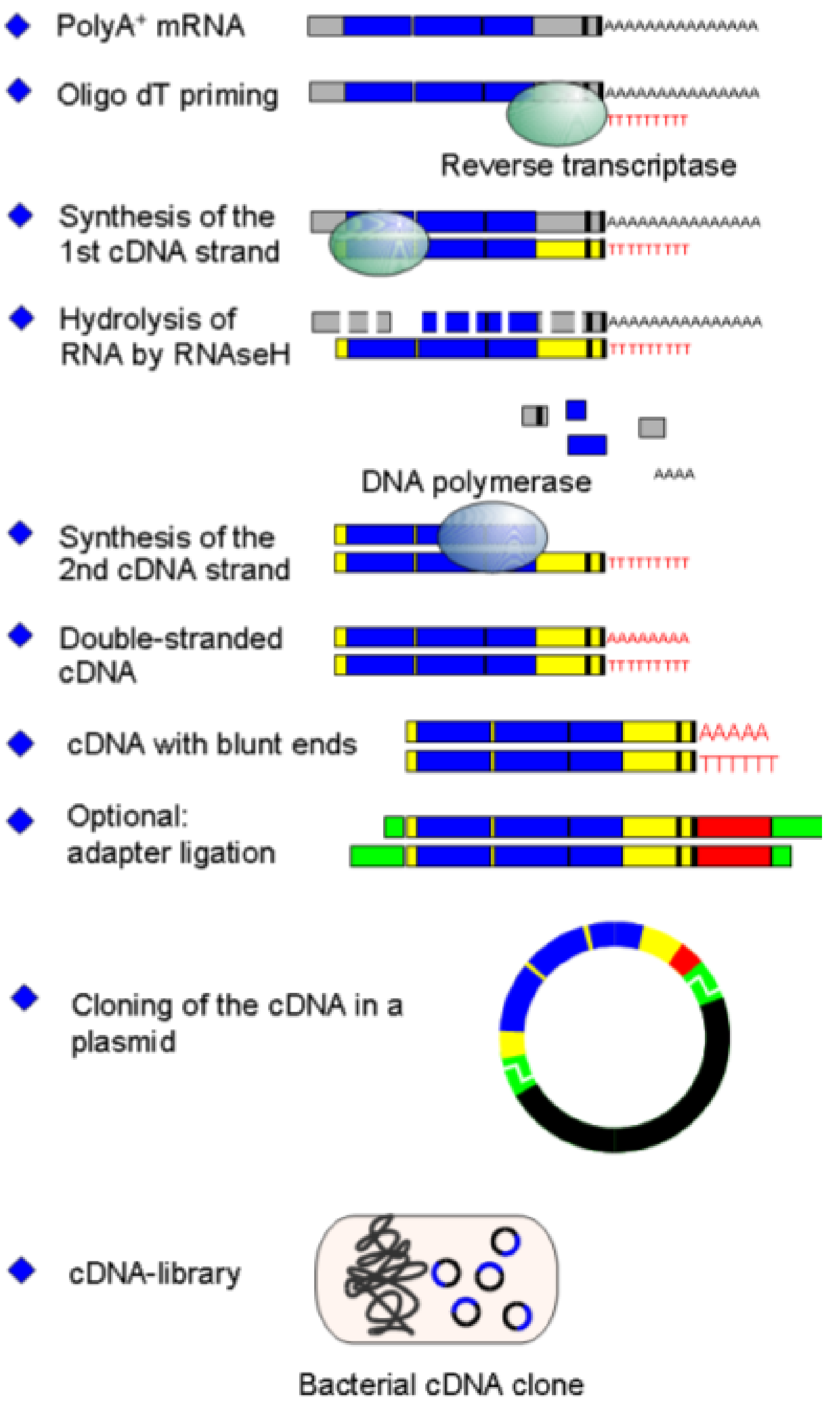


Figure 2.1: Scheme of cDNA library creation from mRNA. The first cDNA strand is synthesized to the mRNA single-strand by reverse transcriptase. Afterwards the mRNA is hydrolyzed and the second cDNA strand is synthesized by DNA polymerase. The cDNA is cloned into a plasmid vector which is then transformed into bacteria. Figure adopted from A.M. Perlick

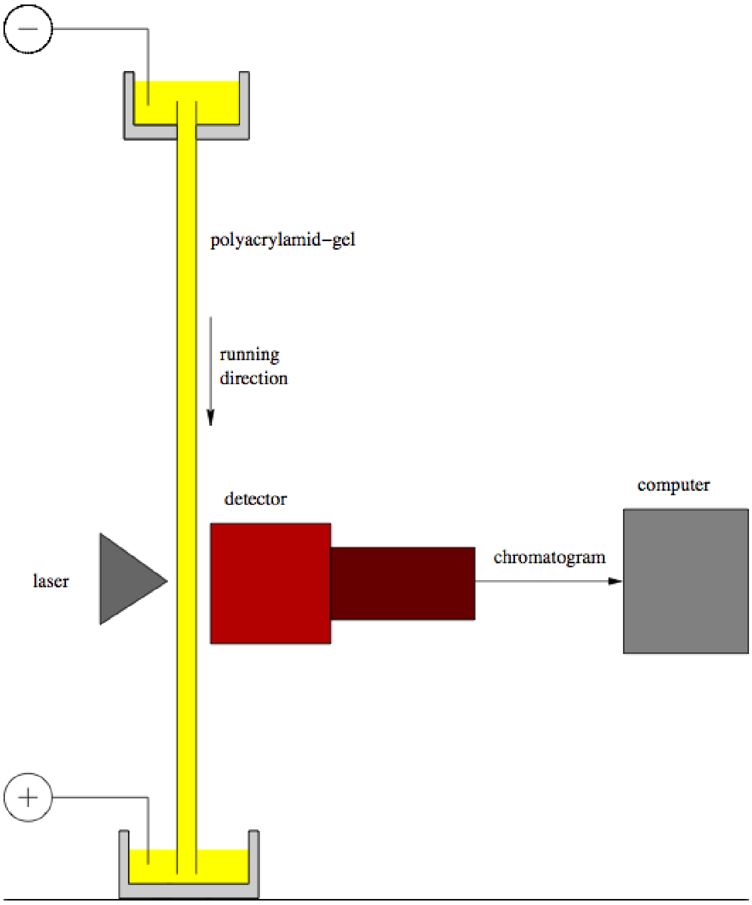


Figure 2.2: Scheme of the fluorescence gel electrophoresis. The fragments created by chain terminated PCR are of different size and mass. They run from the cathode to the anode at different speeds according to their size and pass the laser. The detector absorbs the light emitted by the fluorescent dye and generates a chromatogram file.

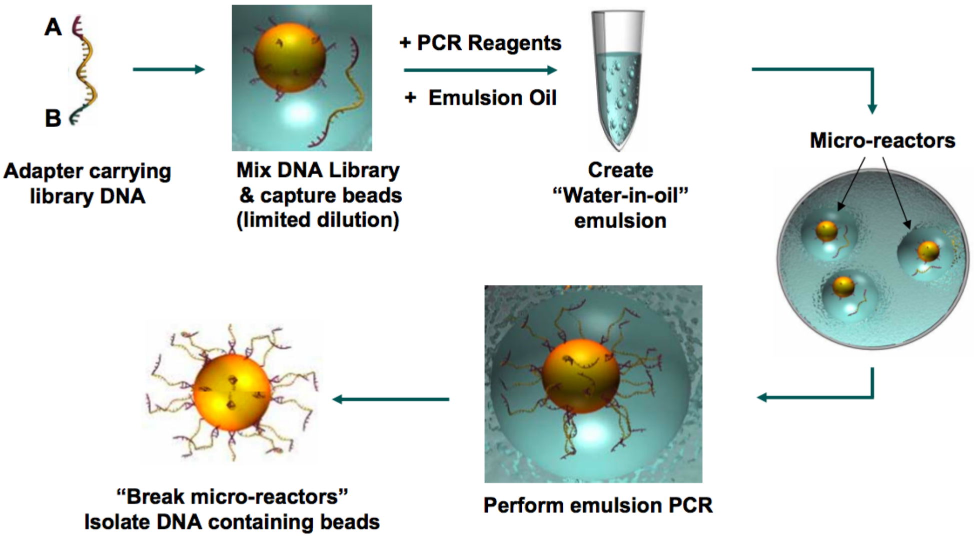

A
B
Adapter carrying library DNA
Mix DNA Library & capture beads (limited dilution)
+ PCR Reagents
+ Emulsion Oil
Create "Water-in-oil" emulsion
Micro-reactors
Perform emulsion PCR
"Break micro-reactors" Isolate DNA containing beads

Figure 2.3: Scheme of the workflow for 454 pyrosequencing - sequencing by synthesis. Two adaptors are added to the cDNA fragments (A and B). The fragments bind on special designed DNA capture beads and are immobilized. By adding amplification reagents in a water-in-oil mixture, the DNA beads are separated, each in one single microreactor. Amplification of the fragments is done in each microreactor separately, all microreactors are processed in parallel. The amplified fragments are loaded onto a PicoTiterPlate for sequencing. Special labeled nucleotides are added to the wells, each carrying exactly one DNA bead. The sequencer detects the emitted light to reconstruct the sequences of millions of fragments at a time. Figure adopted from http://www.454.com

sequenced. The sequencing steps for a 454 sequencing run are described in the following text and visualized in Figure 2.3.

Preparation

Two different adapters (A and B, specified for the 3' and 5' fragment ends) are added to each cDNA fragment. The adapters are used for the purification, amplification and sequencing steps. The single-stranded fragments carrying A and B adapters compose the sample library used afterwards.

Specifically designed DNA Capture Beads® are added, immobilizing the single-stranded DNA fragments. Each bead carries a unique single-stranded fragment. With adding amplification reagents in a water-in-oil mixture, the beads are emulsified and separated resulting in microreactors, each containing exactly one bead with exactly one unique DNA fragment.

Emulsion PCR Amplification

Amplification of the fragments is done for each fragment in its own microreactor,

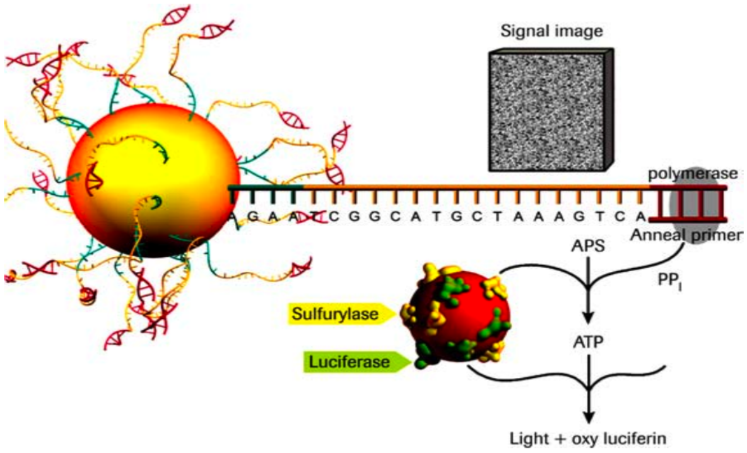


Figure 2.4: Sequencing reaction of the Genome Sequencer System. Millions of copies of a single clonal fragment are contained on each DNA capture bead. During the sequencing progress, nucleotides are flown over the wells in a fixed order. A CCD camera takes an image of each nucleotide adding flow. Figure adopted from http://www.454.com

keeping out contaminating or competing sequences. The entire fragment collection is amplified in parallel, resulting in a copy number of several million per bead. The emulsion PCR is stopped while the amplified fragments are still bound to their specific beads.

Sequencing

The amplified fragments are loaded onto a PicoTiterPlate for sequencing. The wells of the PicoTiterPlate allow only one bead per well due to the well diameter of 44 μm. The Genome Sequencer flows individual nucleotides in a fixed order across all wells on the PicoTiterPlate, resulting in a chemiluminescent signal. The addition of a nucleotide complement to the template strand can be detected by the CCD camera of the Genome Sequencer Instrument. These pictures are stored for further analysis (see Figure 2.4).

2.1.3 Computer aided analysis of sequence datasets (ESTs)

As mentioned in the previous Sections, information about the sequencing runs is stored as raw chromatogram files (EST-sequencing) or as raw picture files (454-sequencing).

In case of a chromatogram file the computer aided analysis starts by obtaining the base sequence for each template from the chromatogram files[Ewing *et al.* (1998)]. The four necessary steps are described in the following.

In the *lane tracking* step the gel lane boundaries are identified and assigned to

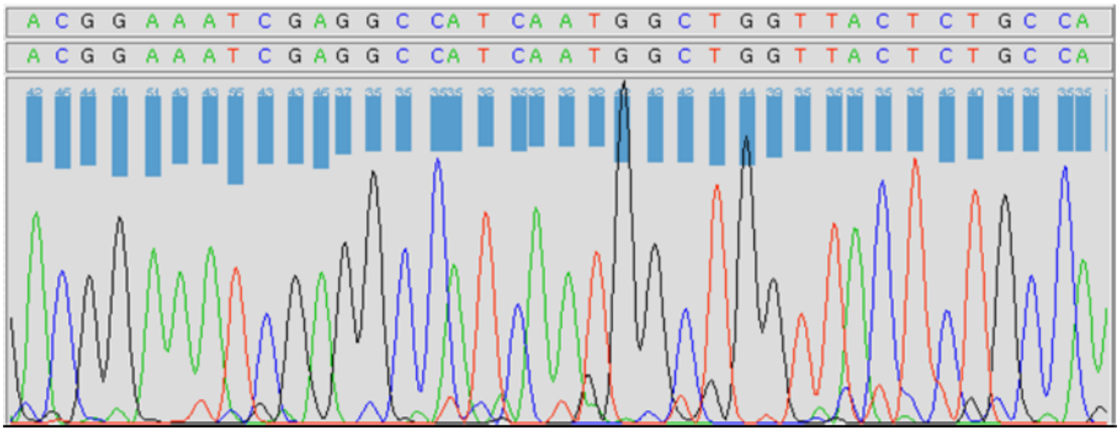


Figure 2.5: This picture shows a trace file of a chromatogram resulting from EST sequencing. The different colors indicate different bases, peaks express the intensity. Additionally the Phred quality of the sequence is displayed as blue bars above the peaks. The base sequence is displayed above.

the right probes. After that, the intensities of the four signals are summed up across the lane width. During this *lane profiling*, a profile (or trace) is created, consisting of four arrays indicating signal intensities during the gel run. Each list consists of the signal intensities of the considered fluorescent dye. In the next step (*trace processing*) signal processing methods are used to deconvolve and smooth the signal estimates. This step also reduces noise and corrects dye effects on fragment mobility. *Base-calling* is the last processing step. Hereby the processed trace is translated into a sequence of bases. Figure 2.5 shows a trace file. The resulting EST sequences are stored in fasta files established by Lipman and Pearson (1985).

In case of 454 sequencing and raw image files, the analysis is performed using the software provided by Roche. The position specific signal intensities allow the software to reconstruct the sequences of each well such that over 1 million reads can be processed in parallel: The raw data from the CCD camera is processed and the intensity for each well is extracted, quantized, and normalized. The series of reads generates a flowgram for each well, similar to the chromatogram files from EST sequencing. The proportional growing signal intensity indicates the number of identical bases incorporated. Thus, the sequence can be generated for each well. The sequences can be assembled afterwards using different bioinformatic applications, concerning the individual purpose (see Figure 2.6).

To reduce redundancy, the sequences are grouped (*clustered*) on sequence level using a clustering tool (e.g. tgicl by Pertea *et al.* (2003)). Afterwards the clusters are assembled to Tentative Consensus sequences (TCs) (*assembly*), or in case of only one remaining read, this read is stored as singlet. This is commonly done using CAP3 by Huang and Madan (1999) or the Genome Sequencer *De*

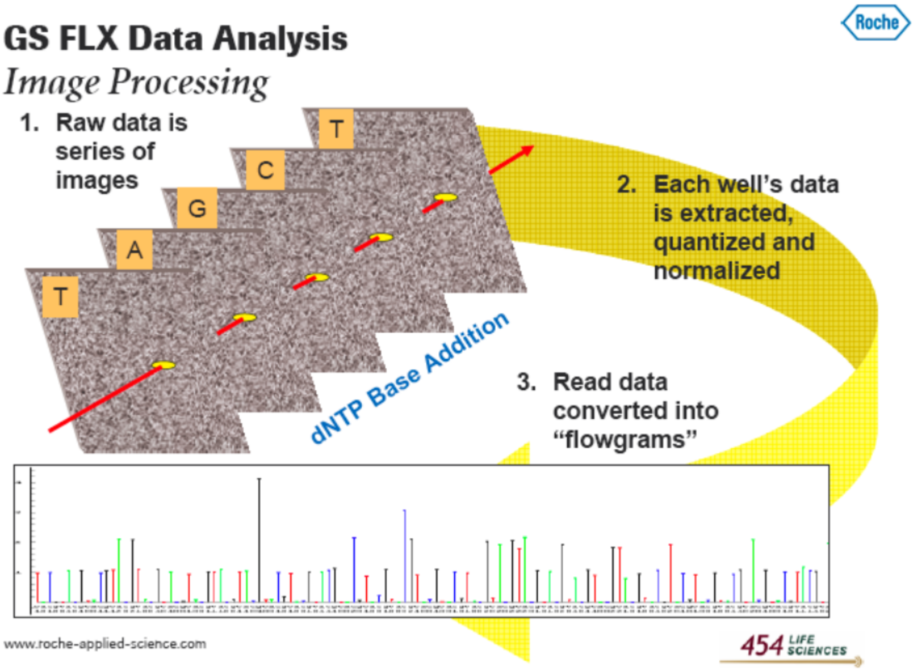

Roche
GS FLX Data Analysis
Image Processing
1. Raw data is series of images
T
A
G
C
T
2. Each well's data is extracted, quantized and normalized
dNTP Base Addition
3. Read data converted into "flowgrams"
www.roche-applied-science.com
454 LIFE SCIENCES

Figure 2.6: Obtaining the base sequence for a fragment from the raw images files. For each of the microreactors all images are analyzed, the intensity values are extracted, quantified and normalized. This data is then stored as a flowgram from which the sequence is obtained. Figure adopted from http://www.454.com

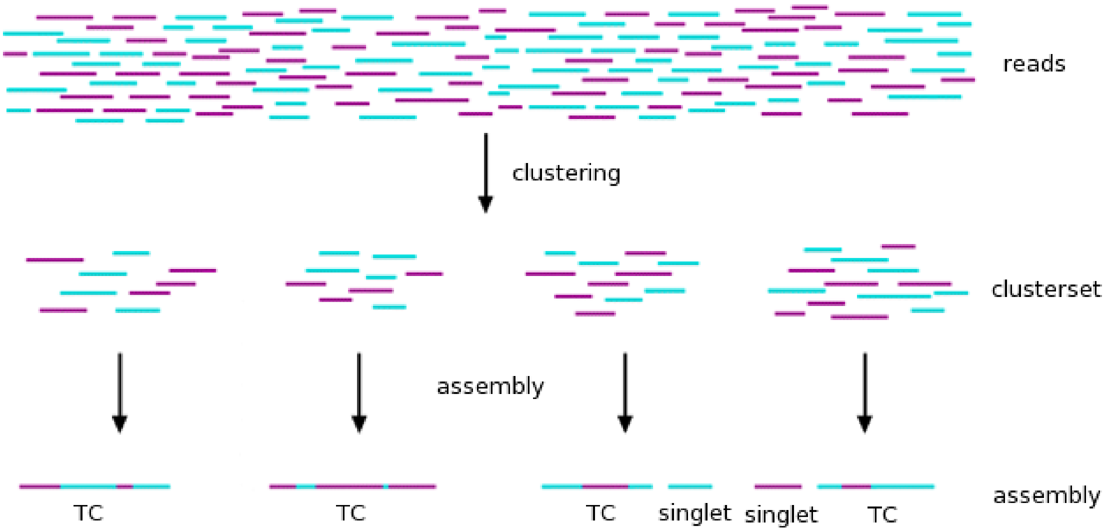


Figure 2.7: This figure shows the processing steps to generate TCs from sequencing reads (ESTs or 454 reads). The reads are clustered according to their base sequence. The clusters are assembled gaining TCs and singlets. The different colors indicate the different libraries of the reads.

Novo Assembler Software (Roche Applied Science, Mannheim, Germany). It is possible to cluster and assemble reads from more than one EST/454 library together, so that sequences occurring in both libraries are assembled to one TC. Figure 2.7 shows the clustering and assembly of reads to TCs. The resulting TCs and singlets can be analyzed functionally using different bioinformatic applications.

2.1.4 cDNA expression analysis

In order to compare gene expression of different samples *in-silico*, it is fundamental to define a formula which calculates comparable values for the expression rate of genes. For all types of gene expression analysis in cDNA libraries the assembly information of each TC has to be known (which reads from which library were assembled). There are different approaches in defining this formula. One approach by Audic and Claverie (1997) is to compare the expression in two different cDNA libraries, or two sets of cDNA libraries.

A second approach calculates an expression value for TCs according to the number of libraries clustered, the size of the libraries, the size and composition of the TC. For this so-called logarithmic likelihood ratio, only one set or subset of libraries is used[Stekel *et al.* (2000)].

Enhancing this formula, Journet *et al.* (2002) developed the likelihood ratio & frequency ratio, which compares the expression of a gene in two sets of libraries according to the logarithmic likelihood ratio.

In contrast to the proposed formulas from Audic *et al.* and Journet *et al.*, the formula of Stekel *et al.* is not limited to two libraries or sets of libraries, but can contain numerous libraries that are used for the expression analysis. Because of this feature, the logarithmic likelihood ratio is described here in detail.

The logarithmic likelihood ratio (R-value) expresses the contribution of the TC from reads of different libraries. The formula for the R-value is denoted as follows: Let $x_{i,j}$ be the number of reads for gene j in the i-th library and N_i the total number of reads in the i-th library. The equation

$$R_j = \sum_{i=1}^{m} x_{i,j} log \left(\frac{x_{i,j}}{N_i f_j} \right) \tag{2.1}$$

is calculated for the number of cDNA libraries, m, and for the frequency of gene product, f_j, defined by

$$f_j = \frac{\sum\limits_{i=1}^{m} x_{i,j}}{\sum\limits_{i=1}^{m} N_i}. \tag{2.2}$$

Unfortunately there is no universal scale for the R-value, as there are many factors in this formula which differ for experiment and library sizes. However, the expression values within one analysis are comparable to each other. The larger the logarithmic likelihood is, the more significant is the expression of the gene.

2.2 Microarray gene expression analysis

This Section focuses on explaining the main principles of microarray gene expression experiments and analyses.

The first experiments attaching cDNA to a glass surface were made by Schena and Davis (1992) and further more by Schena *et al.* (1998). Since then, a variety of different microarray types evolved, the two most interesting ones are cDNA microarrays and oligonucleotide microarrays.
These two cover more than 90% of the hybridized microarrays (65% cDNA & 26% oligonucleotide microarrays [Schena (2002)]). Other microarray types to be mentioned here are protein microarrays and tissue microarrays. The length of the spotted reporter sequences for microarrays may vary from 15 nucleotides (shortest oligonucleotide fragment) to 2500 nucleotides (longest cDNA fragment), common lengths range between 150 to 300 nucleotides.

The main principles of DNA microarrays can be summarized as short reporters complementary to the genes to be analyzed are spotted on a surface; extracted

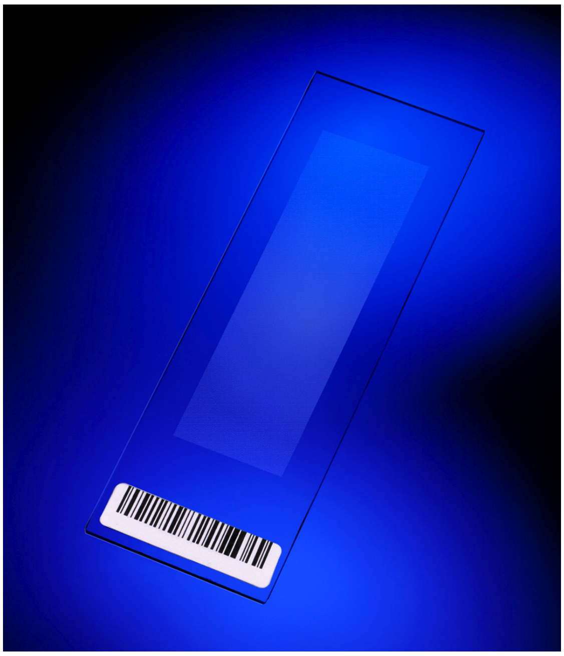


Figure 2.8: This picture shows an oligonucleotide microarray using a glass slide.

mRNA from cells of interest is washed, marked with dye and hybridized on the microarray.
Pictures taken from the hybridized microarray indicate genes expression levels for the spotted reporters.

In contrast to EST/454 analyses, microarrays are no sequencing based technology and the base sequence of the genes to be analyzed have to be known before an analysis can be performed. Mircoarrays can be regarded as a quantitative analysis, whereas EST/454 analysis datasets normally are normally not used for quantitative analyses.

2.2.1 Oligonucleotide microarrays

For the analysis and profiling of gene expression, oligonucleotide microarrays are frequently used. Allowing thousands of hybridizations in parallel, microarrays can be used to detect genes to be expression under different cell conditions. The oligonucleotides are synthesized using PCR and afterwards spotted on the glass surface using a robotic spotter with print-tips or ink-jet like printing.
A picture of an oligonucleotide microarray is shown in Figure 2.8.

Longer oligonucleotide probes are more specific to individual target genes, whereas shorter probes may be spotted in higher density across the array and are cheaper to manufacture.

Oligonucleotide microarrays normally use a two color system, meaning that two different sample mRNAs are marked with Cy3 (light emission at 570nm = green)

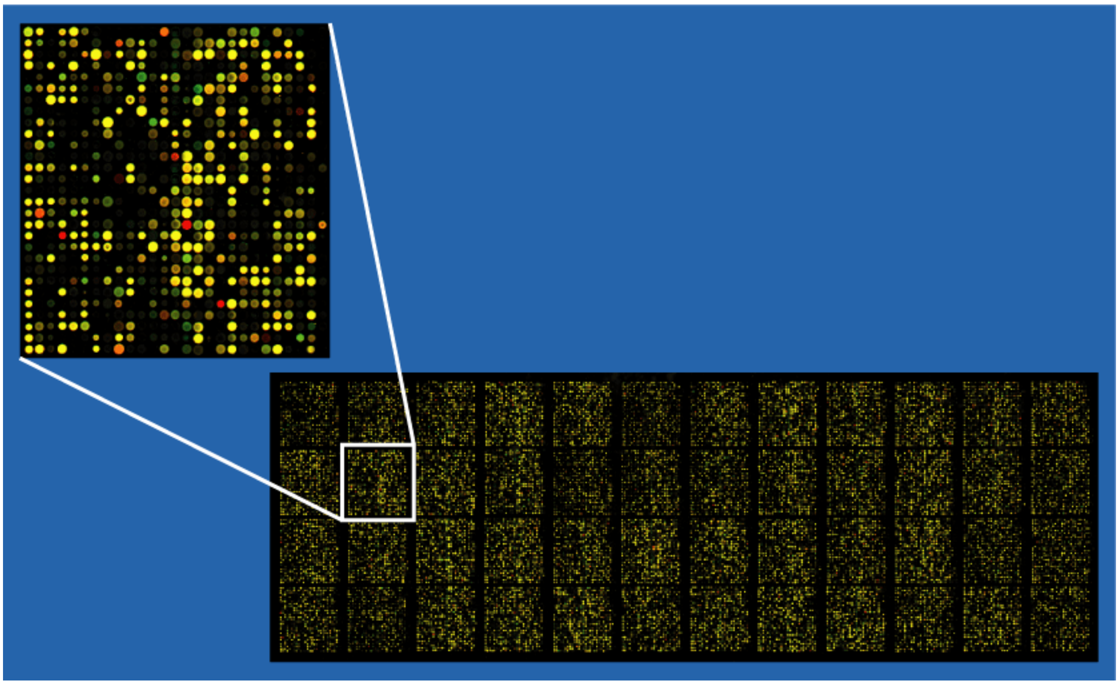


Figure 2.9: This picture shows an image taken from the expression of an oligonucleotide microarray. The red dots indicate genes expressed in one tissue, the green dots represent genes expressed in the other tissue. Yellow spots mark genes expressed in both samples. Picture adopted from *http://www.wikipedia.com*

and Cy5 (light emission at 670nm = red) respectively. The Cy-labeled cDNA targets are used to detect the probes on the microarrays. Both marked cDNA samples are washed over the chip and hybridized. After the hybridization step, the microarray can be excited with a laser beam and the emitted fluorescence can be captured by a CCD camera (see Figure 2.9 for an example of a resulting image).

The expression of the different genes can be read as green, red, and yellow (red and green in combination) colors which are normalized using special spotted RNA spike-ins and added control probes. This two-color technique allows to compare the expression in one single organism under two different conditions, e.g. healthy vs. diseased, growing vs. fully-grown, or two different organism types against each other, e.g. wildtype vs. mutant. The results are relative values, as the expression intensities (emitted light) are unique to the actual hybridized microarray.
The intensities can be used to identify up-regulated and down-regulated genes in the two probes.

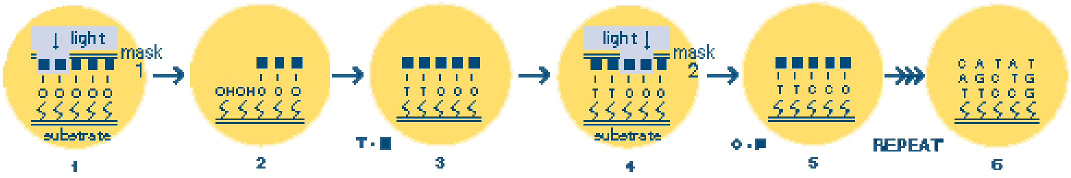


Figure 2.10: This schema demonstrates the photolithographic spotting. Reporters are protected such that no base can bind to them. The protection is removed using UV light where the new base can be added. The new base again carries a protection. Schema adopted from http://surf-chuck.com/research/page11/page11.html

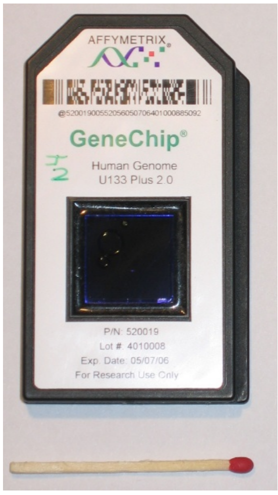


Figure 2.11: This picture shows an Affymetrix GeneChip® and a match as size comparison. Picture adopted from *http://themedicalbiochemistrypage.org*

2.2.2 Affymetrix GeneChip® microarrays

The Affymetrix GeneChip® microarray is a commercially preproduced oligonucleotide microarray. The reporters are synthesized directly in the surface of the slide using UV-masks and photoactivated chemistry (see Figure 2.10): At first, all reporters sites are protected so that no base can bind to it. Reporter sites and reporters that should be extended are lightened by UV light, the others are masked. The UV light removes the protection so that one base (A, C, T, or G) can be added, carrying a new protection at the end. This procedure continues until all reporters are completely spotted. This fast and accurate method allows to spot reporters in parallel on the whole array.

Each GeneChip® is embedded in a special cartridge, preventing it from contaminations and allowing easy handling and transport (see Figure 2.11). There are

Figure 2.12: An image of a hybridized GeneChip® taken by a CCD camera.

currently GeneChips® for 75 species available. In most cases one array is sufficient to carry all reporters for all genes of one species, sometimes related organisms share one GeneChip®.

The length of the reporters is fixed to 25 basepairs, one gene is represented by 22 to 40 spotted reporters. As a control, one half of the reporters are complemented at the 13th base, named mismatch probes (vs. perfect match probes). In contrast to the commonly used oligonucleotide microarrays, the Affymetrix GeneChip® seizes an enormous number of reporters (up to 1.000.000 reporters representing over 60.000 genes).

GeneChips® are designed to hybridize only one single mRNA probeset. This techniques requires to hybridize at least two chips to compare the expression from one chip to the other. This offers the advantage to compare the gene expression from newly hybridized GeneChips® to experiments performed before, or to GeneChip® experiments performed in different research labs.

An image taken from a hybridized GeneChip® is shown in Figure 2.12.

2.2.3 Methods of microarray gene expression analysis

The main principles of microarray gene expression analysis are explained in this section:

Starting with raw image files, the analysis of the expression values begins with background-correction, log-ratio computing, and normalization:

Background correction is based on the assumption that the measured signal consist of the sum of the foreground signal and an unspecified signal of the microarray surface. Different suggestions on how to deal with the background fluorescence were made in the past[Chen *et al.* (1997); Yang *et al.* (2001); Quackenbush (2002); Attoor *et al.* (2004); Yin *et al.* (2005)].

Ratio computing is used to compute the ratio between the two spotted conditions, in one oligonucleotide microarray (two-color microarrays), or in two different arrays (e.g. GeneChips®).

$$T_i = R_i/G_i \tag{2.3}$$

with ration T_i for the i-th gene and comparing measurement of a treatment R_i against the measurement of a control condition G_i. Using this formula, one has to keep in mind that the amount of mRNA used for the hybridization can lead to different results.

The result needs to be logarithmized to reduce noise (the noise error is multiplicative - the higher the expression is higher the noise error gets) and to make the up- and downregulation comparable (0.5 is half the expression and 2.0 is double the expression)[Chen *et al.* (1997); Li and Wong (2001); Sásik *et al.* (2002); Quackenbush (2002)].

To make different microarrays experiments comparable to each other, normalization is used to remove systematic bias from the datasets [Quackenbush (2002); Smyth and Speed (2003)]. This bias may originate from differences in RNA-concentrations between samples, differences in scanner settings, and differences in labeling, bleaching, and detection behavior of the fluorophores.

Many normalization algorithms have been established in the last years, specializing on two-color or on single-color microarrays (in this case mostly normalizing all arrays of an experiment together). The most commonly used normalizations are the lowess normalization by Cleveland and Devlin (1988) for two-color arrays, which has been optimized by Dudoit *et al.* (2002) and Yang *et al.* (2002). The algorithm has been adopted for the use with single-color arrays by Bolstad *et al.* (2003), using a pairwise comparison of the intensities of all microarrays in one experiment(cyclic-loess).

Affymetrix GeneChips are mostly normalized using one of the normalizations MAS5, RMA, MBEI, or GCRMA [Bolstad *et al.* (2003); Gautier *et al.* (2004)].

The next step in microarray data analysis is mostly the identification of significant expressed genes. Using a fixed cut-off for ratios or log-ratios is understandably a bad practice [Quackenbush (2002)]. Statistical tests can bring insight into significant gene expression variations, testing if the expression change occurred by chance, or may be caused by actual expression change. A variety of statistical tests can be used for the analysis (Student's T-Test, Wilcoxon's Rank-Sum Test by Siegel (1956), CyberT[Baldi and Long (2001)], LIMMA[Smyth (2004, 2005)], SAM[Tusher *et al.* (2001)]), where the Student's T-Test is the mostly used statistical test for microarray gene expression analysis. Dondrup *et al.* (2009b) compared these and more statistical tests on the data of specially hybridized microarrays.

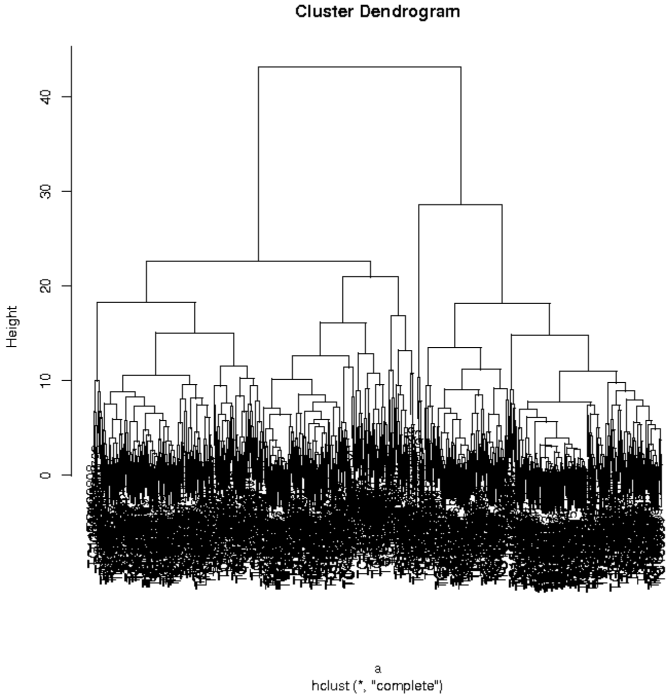


Figure 2.13: A dendrogram of a hierarchical clustering (hclust, complete linkage clustering). On the x-axis the genes are listed, on the y-axis (Height) the similarity of the expression profiles of the genes is shown. The clustering is illustrated by the tree structure from top to bottom.

The study revealed a good usability for the T-test, which does not need many assumptions for an analysis. Another recommendation is the SAM method, delivering a very good false-positive rate. This is related to the special design of the SAM method, as it is a special microarray evaluation method.

Often a subset of genes is connected to some biological pathway, activated or deactivated by some treatment of the cells. A clustering can be performed to find genes with corresponding expression profiles. Typical clustering methods are Ward's clustering, complete and single linkage clustering, McQuitty clustering, median clustering, and average clustering.

These analyses can be visualized as cluster dendrograms (see Figure 2.13), as M/A plots(see Figure 2.14), or as cluster heatmaps (see Figure 2.15).

2.2.4 Standards for microarray expression datasets

Due to the complexity and amount of data gathered in a microarray experiment, standardized data storage and data handling is an optimal goal. The MGED

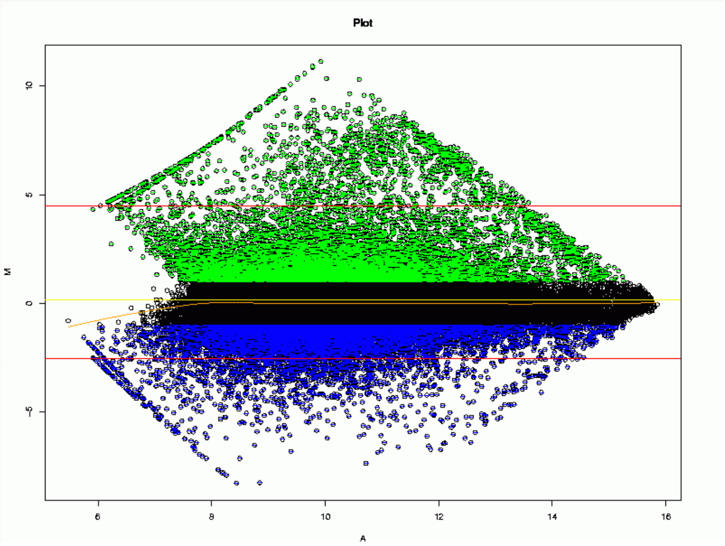


Figure 2.14: An M/A plot. Each dot represents a spotted reporter, where M (x-axis) is the intensity ratio and A (y-axis) is the average intensity of the spot in the plot.

Society (Microarray Gene Expression Database Society[1]) is an international organization of biologists, computer scientists, and data analysts that aims to facilitate biological and biomedical discovery through data integration. Within the MGED, different groups are set up to solve the problems of standardization and deliver rules for storage and modelling of microarray datasets. The Minimal Information About a Microarray Experiment (MIAME) describes the information "needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment." [Brazma *et al.* (2001)]. These information can be projected using the MAGE-OM (Microarray Gene Expression - Object Model) [Whetzel *et al.* (2006)] and can be exchanged using the MAGE-ML (Microarray Gene Expression - Markup Language) data exchange format described by Spellman *et al.* (2002).
The MAGE-OM schema covers 17 packages, containing 132 classes with 123 attributes. The classes are connected via 223 relations. MAGE-OM has been modelled using the Unified Modelling Language (UML), MAGE-ML has been implemented using XML (eXtensible Markup Language[2]).

Software applications used for the analysis of microarray data should necessarily be compliant to the MIAME standard, and be able to import and export MAGE-ML data files. The best case is a software architecture using the MAGE-OM model

[1] http://www.mged.org
[2] http://www.w3.org/TR/xml/

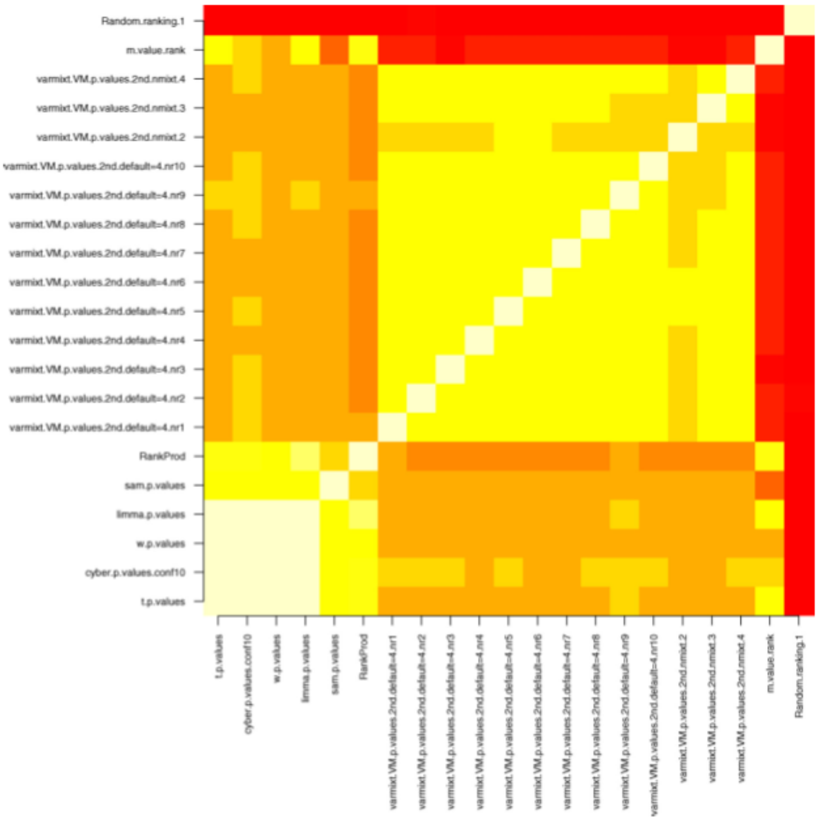


Figure 2.15: A heatmap of a clustering. The x-axis and the y-axis list the clustered genes, the matrix in the middle indicates the expression correlation in a white (similar) to red (not similar) scale.

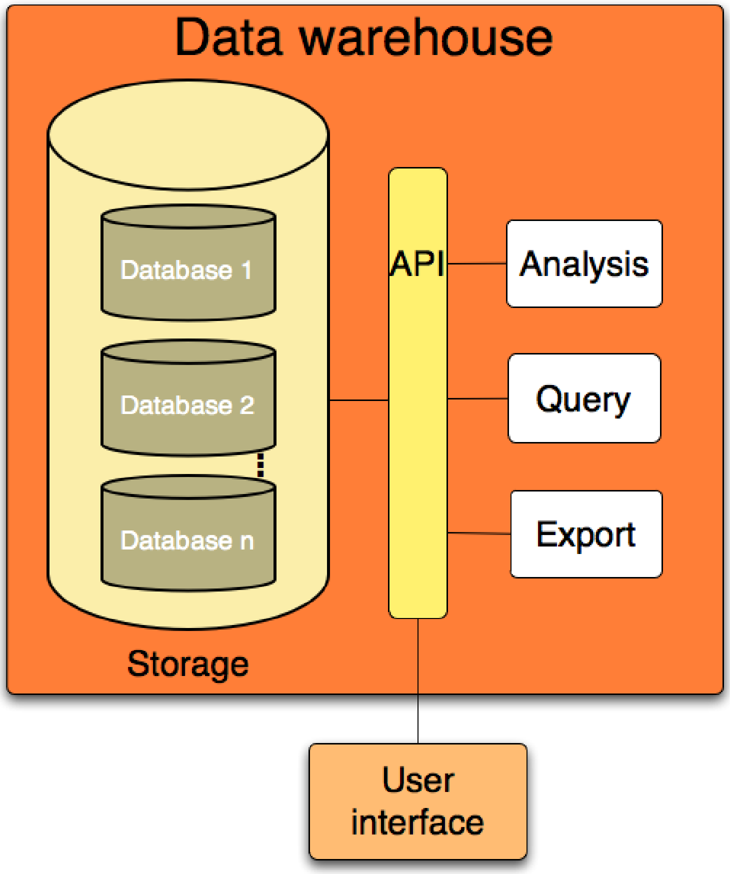


Figure 2.16: Data warehouse structure. A data warehouse integrates datasets from other databases to combine the knowledge. Additionally, a data warehouse offers analysis tools, queries, and export options. The user interface uses the API to connect to the database of the warehouse to use analysis, query, and export options.

to be completely MAGE-compliant.

2.3 Data warehousing

In computational biology datasets are often stored in special databases dedicated to a certain species, or to certain biological units (proteins, genes, etc.). Collecting all information about one special gene often demands for manual work, because the databases storing the desired information have to be queried manually.

The main goal of a data warehouse is the combination of datasets from different data sources and a fast data access to this data repository. Users should be able to find datasets they are searching for and be able to extract all information they

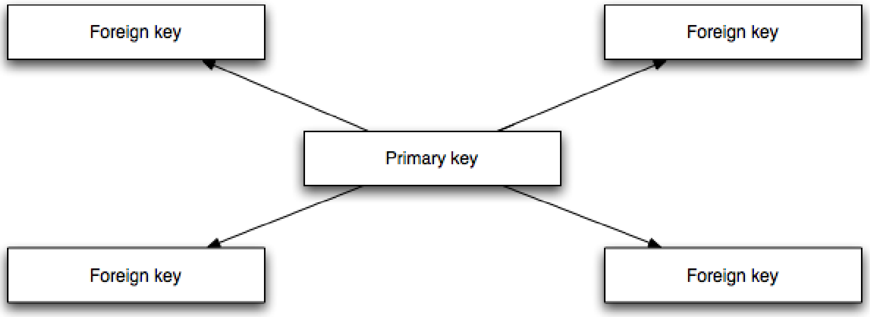


Figure 2.17: Star database schema. Each dataset stored in the main table keeps references to the foreign keys of the datasets stored in the secondary tables.

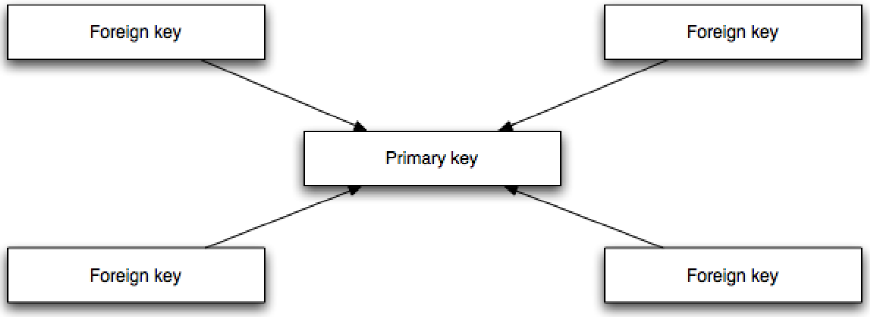


Figure 2.18: Reversed-star database schema. The main table stores the primary key, which is referenced by the entries of the secondary tables.

need. Data warehouses may offer analyses for the datasets, like summing up values of a query, clustering of the datasets, or combining values of different experiments (see Figure 2.16) [Kimball and Margy (2002); Kimball and Caserta (2004)].

Most data warehouses use a star- or reversed-star data schema design (see Figure 2.17 and Figure 2.18). The star data schema defines keys in the main table referring to the data in the dimension tables. In contrast to the star data schema, the reversed-star data schema uses one primary key in the main table, all foreign keys in the dimension tables are referring to this primary key. The benefit of a reversed-star data schema is the ability to easily add and delete referenced datasets and associate the datasets to the already existing ones, as they are always connected to the primary class via the stored primary key. Using a star schema the primary entry always has to be edited because the associated data changes.

The design of the data import into a data warehouse is characterized in the three steps export, transform, and load (ETL).

In the export step, the relevant data is exported from different source databases. In the transformation step, the data from the different sources is transformed such that a consistent data structure is created (e.g. one database uses abbreviation whereas another database does not). The datasets from the different databases are connected, such that datasets for one object can be stored as one object in the data warehouse, or as one object with references on the detail information. The load step inserts the complete data structure into the database of the data warehouse.

A special created user interface allows to query the database for specific datasets and offers analysis and export options of the results.

CHAPTER 3

Existing systems

This Chapter focuses on the existing systems relevant for the thesis. At first, applications for the analysis of EST datasets are presented, focusing on the SAMS system. Afterwards, microarray gene expression analysis applications are introduced, with the main focus on EMMA2. Different data warehouse solutions are presented in Section 3.3. In the end, the only so far existing system for the combination of different gene expression analysis methods is outlined.

3.1 Computer applications for the analysis of EST datasets

A set of different tools is required to obtain Tentative Consensus sequences (TCs) from raw chromatogram files or sequence files.

Different applications that combine these tools are available, here only to mention **EST2uni** developed by Forment *et al.* (2008) at the Polythechnical University of Valencia, Spain, **ESTExplorer** developed by Nagaraj *et al.* (2007) at Macquarie University, Sydney, Australia, and **SAMS** (Sequence Analysis and Management System) developed by Bekel *et al.* (2009) at Bielefeld University, Germany. All these applications nearly use the same subset of tools and the same pipeline driven approach to analyze the datasets. A comparison of the three applications can be found in Table 3.1.

EST2Uni is a local inastallabel application without user authentication and group management. Providing import of raw datasets as well as Fasta files, it allows clustering, assembly and automatic annotations. Unfortunately, no manual annotation editing functionality is available. GO categories and annotations are implemented,

Feature	EST2Uni	ESTExplorer	SAMS
Installation	local	web	web
User authentication			✓
User groups			✓
Data storage	permanent	1 week	permanent
Import formats	Fasta or raw	Fasta or raw	Fasta or raw
Import pipelines	✓	✓	✓
Clustering	✓	✓	✓
Assembly	✓	✓	✓
Automatic annotation	✓	✓	✓
Manual annotation			✓
GO	✓	✓	✓
Blast sequences against database	✓		✓
KEGG pathways			✓
Expression analysis			✓
Export of sequences / annotations	✓/ ✓	✓/	✓/ ✓

Table 3.1: Comparison of the three EST analysis applications EST2Uni, ESTExplorer and SAMS.

just as a possibility to blast new sequences against the imported genes. It is not possible to project the genes to KEGG pathways, or to perform a gene expression analysis. Export functions for the sequence and annotation datasets are available.
ETSExplorer is a web based EST analysis application that does not feature a user authentication or user groups. Datasets are available using shortcuts like ”John_123” and are stored for one week after analysis. Featuring a raw and fasta import, as well as clustering and assembly functionality and an automatic annotation. The absence of a manual annotation, KEGG pathways, no possibility to blast against the sequence database, and no expression analysis features make the application less attractive to use.
As the only application with a user authentication and group management, **SAMS** features a permanent data storage. The imported raw or fasta files are processed in a clustering and assembly pipeline, fillowed by an automatic annotation and the possibility to manual edit and add annotations. A KEGG pathway integration allows a visualization of the genes in the respective pathways and an expression analysis offers library specific queries. All sequences and annotations can be exported.

To illustrate an EST analysis, focuses on SAMS, developed at Bielefeld University. SAMS is designed to handle not only cDNA datasets, but also whole-genome-shotgun reads, metagenome datasets, and other already preprocessed sequence

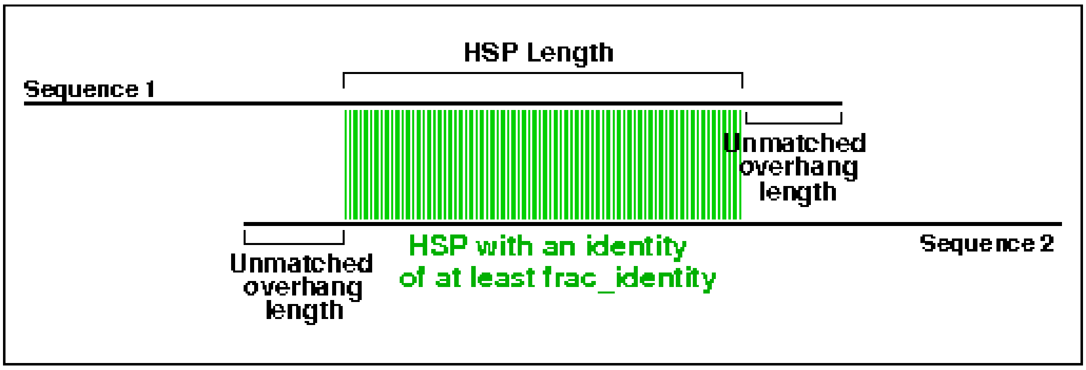

HSP Length
Sequence 1
Unmatched overhang length
HSP with an identity of at least frac_identity
Sequence 2
Unmatched overhang length

Figure 3.1: This scheme depicts the clustering parameters. Two reads have to have at least 95 percent identity for at least 40 base pairs. The unmatched overhang must not exceed 20 basepairs. Scheme adopted from T.Bekel.

datasets (gene and protein sequences).

Mostly used for the analysis of EST experiments, SAMS is designed to import raw chromatogram files as well as already preprocessed (quality clipped and vector clipped) sequence files in FASTA format. EST datasets are processed as described previously (cf. Section 2.1.1) using phred as a quality clipping tool [Ewing *et al.* (1998); Ewing and Green (1998)]. For vector clipping, the sequences are blasted against a database consisting of the EMBL standard vector database EMVec[1], the NCBI vector database UniVec[2] and some in-house vector and adaptor sequences. The sequences are then trimmed off the vectors for further analysis.

For the clustering and assembling process, SAMS uses a pipeline based approach. The pipeline by default uses a set of standard parameters for the clustering, defined by the J. Craig Venter Institute (JCVI - previously called The Institute for Genome Research - TIGR). Using these parameters, reads are clustered into one group if the following similarity conditions are fulfilled: First, two reads must show an alignment of not less than 40 base pairs with at least 95 percent identity in a pairwise comparison. Second, flanking unmatched overhangs next to the alignment must not exceed a length of 20 bp (Figure 3.1).

These parameters can be changed by the user if necessary. After calculating the clusters, they are assembled using the application CAP3 by Huang and Madan (1999). This application calculates the TCs and leaves some non-matching reads as singlets. The TCs and singlets together form a nearly non-redundant representation of the sequenced data.

On the basis of this data an automatic annotation pipeline is started to find a putative annotation for each TC and singlet. The automatic annotation pipeline

[1] ftp://ftp.ebi.ac.uk/pub/databases/emvec/

[2] http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html

consists of several bioinformatic tool, namely BLAST[Altschul *et al.* (1990)] homology searches against standard sequence databases (NT, NR, Swiss-Prot[Boeckmann *et al.* (2003)], KEGG[Kanehisa and Goto (2000)], KOG[Tatusov *et al.* (2003)]), as well as Interproscan[Mulder and Apweiler (2007)] and HMMer[Eddy (1998)]. A manual annotation can be performed on the basis of the observations of the different tools, afterwards.

3.2 Microarray analysis software

There are different freely available applications for the storage and analysis of microarray datasets, here mentioning **Arrayexpress** developed by Parkinson *et al.* (2005, 2007, 2009), **MayDay**[Dietzsch *et al.* (2006)], and **EMMA2**[Dondrup *et al.* (2009a)]. The main features of these widely used tools are compared in Table 3.2.
Arrayexpress is a web based application to mainly store microarray expression datasets. It allows to import MAGE-ML datasets using a user authentication. Normalization and analysis of the datasets is available using the tool "Expression Profiler". Arrayexpress uses the MAGE-OM schema to model the datasets in a MySQL or Orcale database with the addition of NetCDF file storage. The datasets are manually curated in the import step. Export options allow to export all uploaded datasets as csv or raw files.
MayDay is a Java Webstart based application that can be run local or with the public webserver as backend server. Due to this, no user authenitication or group management is needed. The import of raw datasets and MAGE-ML files is supported, even if no MAGE-OM schema is used. Datasets can be normalized and a gene expression analysis can be performed. The datasets are stored in a relational MySQL database and can be exported as MAGE-ML or csv files. There is no KEGG pathway integration, but due to a plugin-system it could possibly be added in the future.
EMMA2 can be locally nistalled, or run via the web interface hosted at Bielefeld University. A user and group management allow to analyze datasets in a group of scientists providing different rights and roles for the data access. The datasets are stored in a relational MySQL database and HDF5 files. EMMA2 uses a LIMS system for raw microarray file storage (ArrayLIMS[3]). The complete MAGE-OM is used to provide a MAGE-ML compatibility. Various customizable normalization and gene expression analysis pipelines are implemented in EMMA2. A KEGG integration allows to map the gene expression to the KEGG pathway maps and visualize the expression experiments. MAGE-ML, MAGE-TAB[Rayner *et al.* (2006)] and csv export options are provided by this open source system. None of the three mentioned systems supported one-color microarrays (GeneChips®) at the start of the project. As EMMA2 is developed at Bielefeld University and offers the most interesting criteria in the comparison, this project will extends EMMA2

[3] https://www.cebitec.uni-bielefeld.de/groups/brf/software/arraylims/

Microarray analysis applications	Arrayexpress	MayDay	EMMA2
Installation	web	local & web	local & web
Interface	web	Java WebStart	web
Import	MAGE-ML	raw data	MAGE-ML & raw data
User authentication	✓		✓
User groups			✓
Data normalization	✓	✓	✓
Expression analysis	✓	✓	✓
two-color microarrays	✓	✓	✓
one-color microarrays			
KEGG pathways			✓
MAGE-ML	✓	✓	✓
MAGE-OM	✓		✓
Database backend	NetCDF & Oracle or MySQL	MySQL	MySQL & ADF5 files
Curation of datasets	✓		
Export	MAGE-ML & csv	MAGE-ML & csv	MAGE-ML & csv & MAGE-TAB
Access control	rudimental		✓
Open source		✓	✓

Table 3.2: Comparison of three different microarray analysis applications: ArrayExpress, MayDay and EMMA2.

to load, store, and analyze Affymetrix GeneChip® datasets.

3.3 Data warehouses

Currently there are many different data warehouses and data warehouse systems available. Main features of a data warehouses are integrating datasets of different types and from different resources, rapid and flexible data access, support for easy integration with third-party programs, and an intuitive user interface. Analyzing and querying the stored datasets, data warehouses offer their combined knowledge to the researcher (cf. Section 2.3).

A widely used data warehouse system is the **BioMart system**, developed by Smedley *et al.* (2009) at the Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI). The BioMart system offers a data warehouse design tool for the design of the database classes and for the creation of the MySQL tables. Moreover, a Perl and a Java API are available for an integration into already existing software applications. A web interface called MartView offers an easy access to the integrated datasets and allows to process simple analysis like counting results and exporting of queries datasets.

The **HapMap data warehouse** is one of the largest instances of the BioMart data warehouse system. It stores and administers datasets to identify and catalog genetic similarities and differences in human beings (Haplotype Map of the Human Genome) [International HapMap Consortium (2003, 2004, 2005, 2007)]. The HapMap database contains over 26 million entries and is uses by researchers from all over the world, as the project is a collaboration among scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria, and the United States.

Other widely used data warehouses built upon the BioMart data warehouse system are **WormBase**, storing datasets of the organism *Caenorhabditis elegans* and related nematodes [O'Connell (2005); Harris and Stein (2006); Harris *et al.* (2009); Schwarz *et al.* (2006); Bieri *et al.* (2007); Girard *et al.* (2007)], **dictyBase**, storing datasets of the amoeba *Dictyostelium discoideum* [Kreppel *et al.* (2004); Chisholm *et al.* (2006); Fey *et al.* (2006, 2009)], and the **rat genome database**, storing genetic datasets of diverse rat sequencing and expression analysis projects [Twigger *et al.* (2006); Dwinell *et al.* (2009)].

Another exemplary data warehouse is the **Genevestigator** data warehouse introduced by Zimmermann *et al.* (2004), storing genes and gene expression datasets of the model organism *Arabidopsis thaliana*. Nowadays, the focus lies on the evaluation of the imported gene expression (over 30.000 hybridized microarray datasets) datasets of ten different model organisms [Zimmermann *et al.* (2005, 2008); Laule *et al.* (2006); Grennan (2006)]. Different analysis tools are implemented to analyze gene expression in the stored microarray hybridizations. The tools cover an expression analysis (*Meta-profile Analysis*), a *Biomarker Search*, allowing to find genes expressed under specific condition, a *Clustering Analysis*, identifying groups of genes that have similar expression profiles, and a tool called *Pathway Projector*, which projects found genes on the metabolic and regulatory pathways of *Arabidopsis thaliana*.

In the scope of *Medicago truncatula* research, there is no single data warehouse storing genes, annotations, and expression datasets, which leads to the idea of creating a comprehensive data warehouse.

3.4 Combination of different gene expression analysis methods

Currently, the only application that is able to combine the results of different gene expression analysis methods with each other is **Simcluster**, developed by Vencio *et al.* (2007) at the Institute for Systems Biology, Seattle, USA.

Simcluster may receive different expression experiment datasets, which include SAGE[Velculescu *et al.* (1995)], MPSS[Brenner *et al.* (2000)], and Digital Northern powered by traditional[Okubo *et al.* (1992)] or, recently developed, EST sequencing-by-synthesis (SBS) technologies[Bainbridge *et al.* (2006)], and analyzes them using the simplex space[Aitchison (1988, 2001)].

The expression datasets have to be transferred into the simplex space before they are combined for the analysis. This transfer should make the data from different data sources and methods more comparable, as the simplex space does not use absolute values and scales, but relative ones (relative values to the overall expression for single experiments). With the combined datasets a hierarchical clustering is performed and the results are presented.

The application neither provides a database connection, nor does it allow to use expression values "as they are", the values have to be transferred to the simplex space before they can be loaded and analyzed. Due to these two issues in usability, Simcluster is not useable for the research community. Picking up the idea of combining gene expression methods, this thesis will create an application useable for *Medicago truncatula* expression analyses.

CHAPTER 4

System Design

This Chapter describes the design of the applications stated as goals in Chapter 1.2.

For this purpose, this chapter firstly expounds the extension of the microarray expression analysis software EMMA2 (cf. Section 2.2.4) to store and analyze Affymetrix GeneChip® expression data in the same way as conventional oligonucleotide microarrays.
Secondly, the design of a data warehouse named TRUNCATULIX for *Medicago truncatula* datasets is presented, focusing on data types and on data storage.
The last part of this chapter describes the design of the tool MediPlEx (MEdicago truncatula multiPLe EXpression tool), which combines datasets of different gene expression analysis methods and analyzes these datasets together.

4.1 Extension of EMMA2 to store and analyze Affymetrix GeneChip® expression datasets

One of the features of EMMA2 is the MIAME and MAGE compliancy (cf. Section 2.2.4). This implies that there is no limitation in storing and processing any MAGE dataset describing any kind of microarray experiment. Anyhow, the Affymetrix GeneChip® layout differs from the classical oligonucleotide layout (see Section 2.2.2). Thus, a new importer for the GeneChip® array layout has to be designed according to these specialties.
Fortunately, there is no change needed in the EMMA2 database schema to store the new layout.

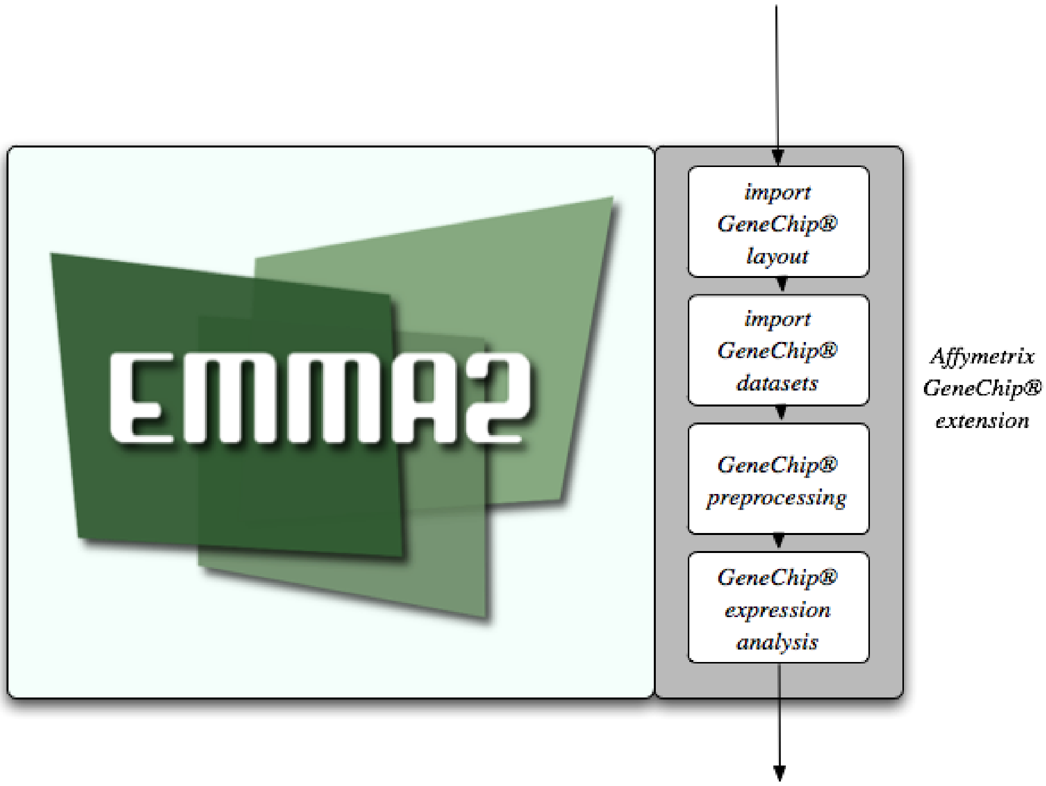

EMMA2
import GeneChip® layout
import GeneChip® datasets
GeneChip® preprocessing
GeneChip® expression analysis
Affymetrix GeneChip® extension

Figure 4.1: This scheme demonstrates the extension of the EMMA2 software. Affymetrix GeneChip® datasets should be analyzed, wherefore the layout of the GeneChip® array has to be imported and the analysis pipelines have to be adopted.

Another issue is the creation of experiments in EMMA2, by combining the datasets of different microarrays and replicates. For the use of GeneChips®, this setup has to be extended, allowing to combine two (or more) sets of GeneChips® to form one experiment. Each of these sets contains the slides for the hybridization of one sample and its replicates. Additionally, the interface of EMMA2 has to be adjusted for this experimental design. A scheme of this extension is shown in Figure 4.1.

Microarray layout

The MAGE-OM schema containing the attributes and relations for an `ArrayLayout` is shown in Figure 4.2 and the schema for the `DesignElement` (to model reporters) is shown in Figure 4.3. The layout of Affymetrix GeneChips® is different from classical oligonucleotide microarrays (see Chapter 2.2.1 and Chapter 2.2.2): Each gene to be analyzed is represented by 22 - 40 spotted reporters, of which the first half are perfect match probes (PM) and the second half are mismatch probes (MM). Missmatch probes have the same sequence as PM probes, with the exception that

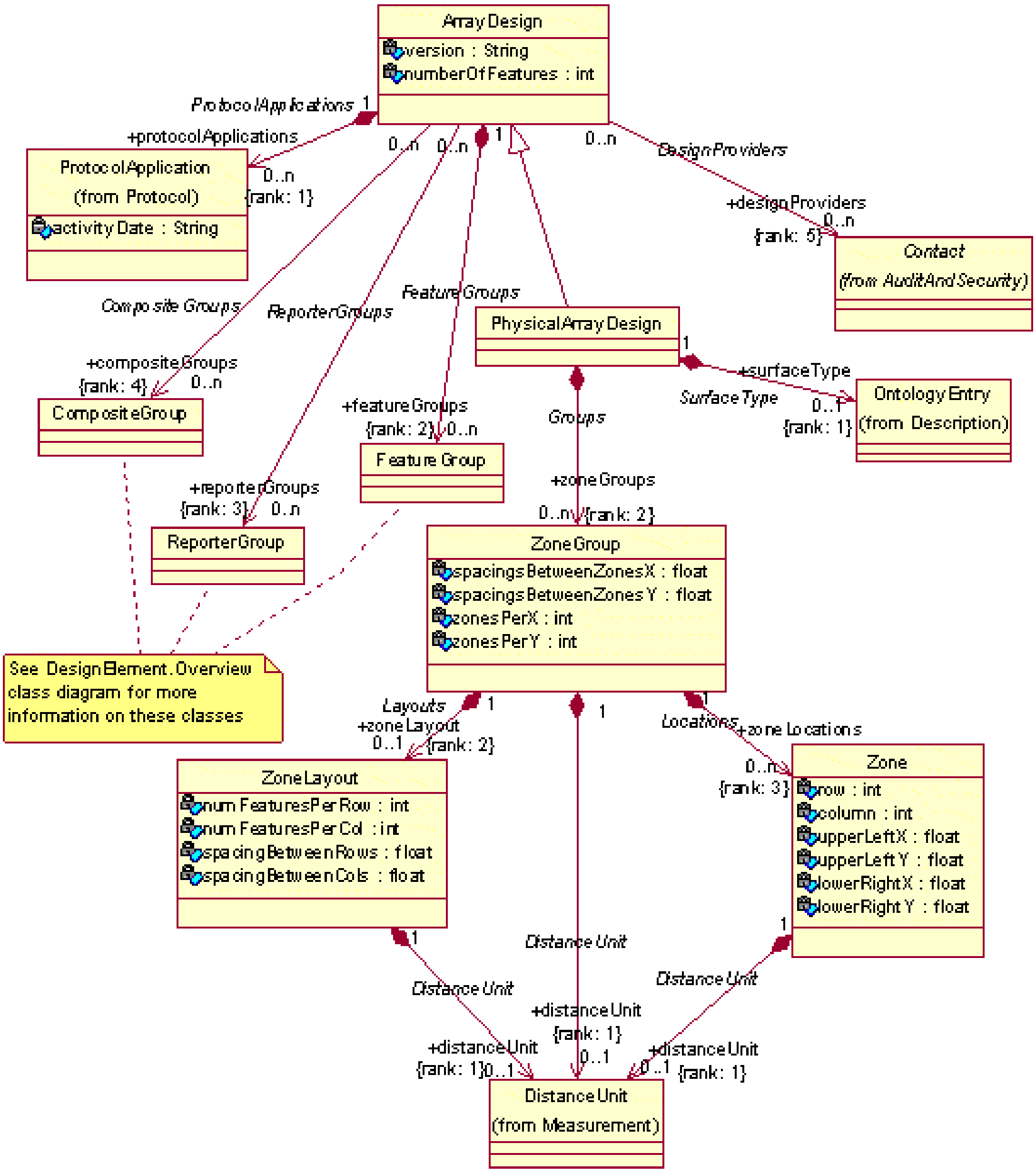


Figure 4.2: This diagram shows the MAGE-OM scheme for an ArrayDesign class. The main classes and relations of the scheme are shown. Scheme adopted from *http://www.ebi.ac.uk/*.

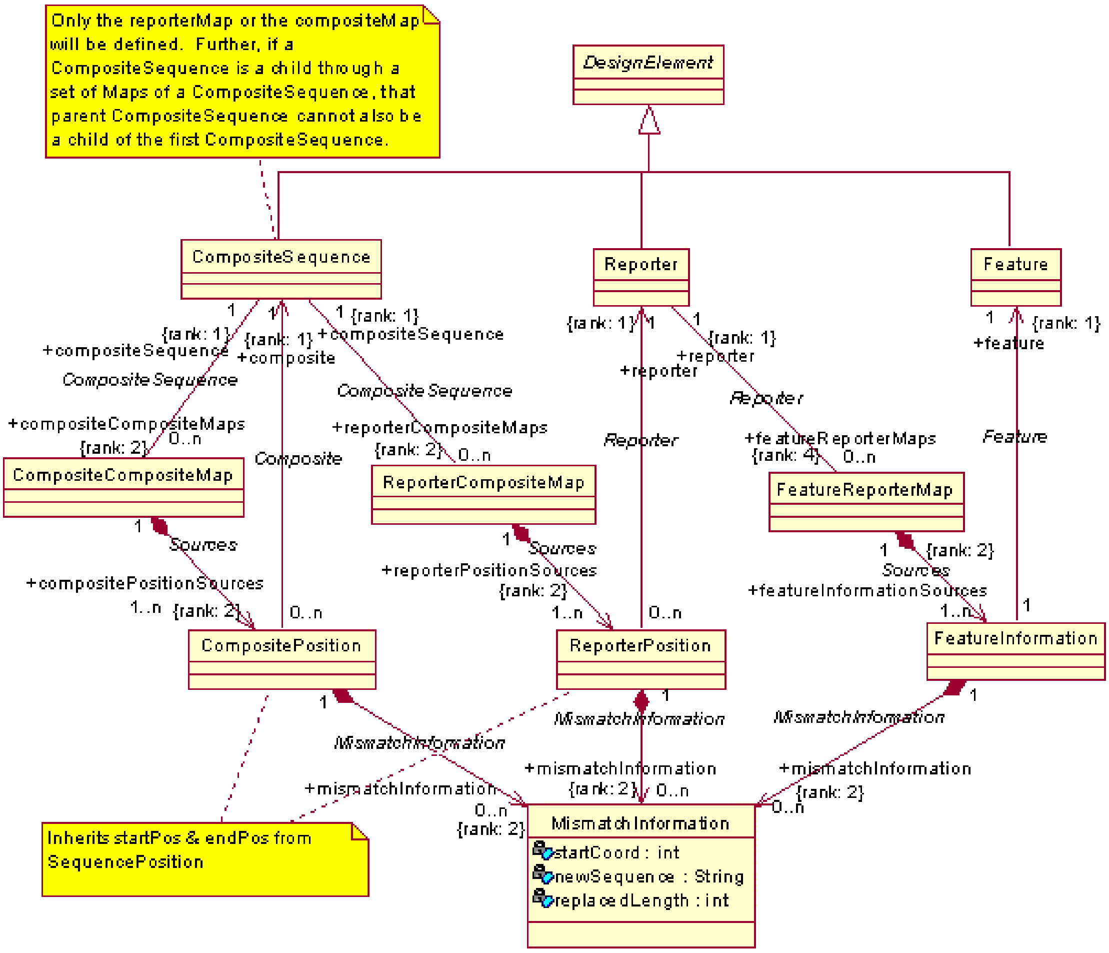


Figure 4.3: This diagram shows the MAGE-OM scheme for the DesignElement class. Reporters, CompositeSequences and Features are stored and combined in DesignElement objects. The attributes and the relations of theobjects can be checked in the scheme. Scheme adopted from *http://www.ebi.ac.uk/*.

the 13th of the 25 bases is complemented. This information, combined with the positional information (x and y coordinates) and the sequence information is stored in the layout files provided by Affymetrix (CDF, SIF, _probe_tab). The CDF file stores the main layout information, containing the reporter positions, the information which reporter is a PM or a MM probe, and the reporter names. The SIF file stores the names and the corresponding sequences of the genes in FASTA format. The _probe_tab file contains the probe names, the x and y coordinates and the sequences of the spotted reporters (25 bases).

A new layout importer should handle this new data and create the required objects in the EMMA2 database.

The import of the microarray layout should be divided in two steps, because of the complex data structure and the memory management.

Import of GeneChip® datasets

In EMMA2, the datasets of hybridized microarrays are stored according to the referred layout. This allows to easily use the previous imported GeneChip® layout to store all expression values from the CEL file of a GeneChip® hybridization. The design of the GeneChip® data import is kept simple:
Load all expression values from the CEL file (which is stored in ArrayLIMS) and store the raw intensity values into the EMMA2 database as MBAD objects (Measured BioAssay Dataset).

Preprocessing of GeneChip® datasets

The preprocessing of the GeneChip® expression datasets should be handled in a similar way to the preprocessing of the oligonucleotide microarrays in EMMA2, to make a comparison of the results easier. This means that the expression datasets in one experiment are preprocessed together in one step.
The preprocessing should be designed as pipeline job, equal to the preprocessing of the oligonucleotide microarray datasets. There are different algorithms available for preprocessing GeneChip® raw expression datasets, the ones typically used should be integrated (MAS5, RMA, MBEI, and GCRMA (see Section 2.2.2)). The raw datasets (MBAD - Measured BioAssay Dataset) should be read from the database, normalized using the integrated functions and stored in the database as DBAD objects (Derived BioAssay Dataset).

Expression analysis of GeneChip® datasets

As the datasets are normalized and stored in the database like the oligonucleotide microarray datastes (as DBAD objects), the expression analyses should be usable as for oligonucleotide microarrays before. As has become clear in Section 2.2.3, many significance tests are available for the analysis of gene expression in microarrays. For Affymetrix GeneChips®, the two-sample t-statistic, as well as an Affymetrix optimized version thereof, as well as the LIMMA test should be implemented as pipeline tools to calculate the significant gene expression in the experiment.

For clustering expression datasets, the same pipeline tools should be used as for conventional oligonucleotide microarray datasets(Hclust pipeline tool).

4.2 TRUNCATULIX - a data warehouse for the legume community

In Chapter 1.2, the need for a data warehouse in the field of *Medicago truncatula* research is pointed out. This section focuses on the design of this data warehouse, called TRUNCATULIX.

TRUNCATULIX should be designed as stand-alone tool for the legume research community, hosting sequence and expression datasets of the model plant *Medicago truncatula*. It should also be useable as a data repository offering the complete backend query functionality via API to be used from other applications.

For the TRUNCATULIX data warehouse, the Sophia data warehouse backend developed by Runte (2010) and the IgetDB data warehouse frontend[1] should be used. The Sophia backend is BioMart[Durinck *et al.* (2005)] compatible and uses a reversed-star schema (see Section 2.3), which makes it easy to add additional datasets to the data warehouse, afterwards. The database schema has to be created such that information about gene sequences, annotations and expression datasets can be stored and queried fast and easily. The IgetDB web frontend is modular and should be adjusted to the TRUNCATULIX data warehouse needs. Therefore, interfaces for filtering, presentation, and export of the sequence and expression datasets should be created. As the TRUNCATULIX data warehouse is created for data integration and fast access, data analysis functions should be not be integrated in the initial version.

A scheme of the data warehouse and the source databases is shown in Figure 4.4.

Data sources

Sequence datasets

- ***Medicago truncatula*** **GeneIndex 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 by Pertea *et al.* (2003), the *Medicago truncatula* GeneIndex (MtGI, hosted at the Dana-Farber Cancer Institute - DFCI) was built. The MtGI 8.0 contains 18,612

[1] http://www.cebitec.uni-bielefeld.de/groups/brf/software/igetdb_info/

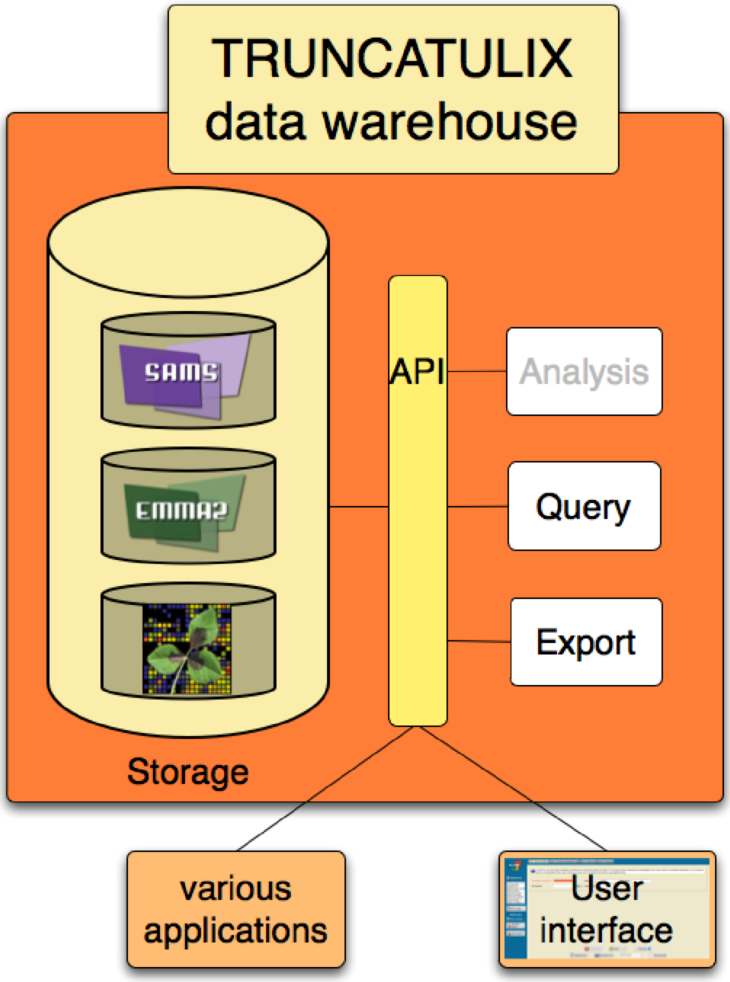


Figure 4.4: This scheme denotes the data sources to be integrated into the TRUNCATULIX data warehouse. Queries allow to search for datasets of interest and an exporter allows to save the datasets externally. Sequence and annotation datasets are integrated from various SAMS projects, expression datasets are imported from EMMA2 and the Medicago gene expression atlas. The API is used by the user interface for the interaction with the database, it can also be used by other applications to retrieve datasets from the warehouse.

Tentative Consensus sequences (TCs) and 18,238 singletons (Jan. 2005 [Quackenbush *et al.* (2001)]). The sequences were imported into the Sequence Analysis and Management System (SAMS) (see Section 3.1). The SAMS system contains an automatic annotation pipeline (Metanor), which runs several bioinformatics tools for gene annotation (BLAST[Altschul *et al.* (1990)], Interproscan[Mulder and Apweiler (2007)], TMHMM[Sonnhammer *et al.* (1998)]). A high quality consensus annotation is created, covering EC numbers[Kanehisa and Goto (2000)], KEGG functions[Kanehisa and Goto (2000)], GO numbers[Ashburner *et al.* (2000)], KOG numbers[Tatusov *et al.* (2003)], putative gene functions, and gene names.

- ***Medicago truncatula*** **GeneIndex 9.0**
  Recently, the J. Craig Venter Institute released a new version of the *Medicago truncatula* GeneIndex, 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 (http://compbio.dfci.harvard.edu/tgi/qcGene.html) for the EST assembly, whereof 11,494 ETs remained as singletons. The new sequences were downloaded from the DFCI web pages and imported into SAMS, a complete automatic annotation was performed.

- ***Medicago truncatula*** **genome project**
  The Medicago Genome Sequence Consortium (MGSC[2]) sequenced the *Medicago truncatula* genome using a classical BAC sequencing approach[Cannon *et al.* (2006); Young *et al.* (2005)]. The project started in 2005, in October 2007 the second sequence assembly was released (version 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 web page and afterwards imported into SAMS. Using SAMS, a complete automatic annotation was performed.

- **Affymetrix Medicago GeneChip® probes**
  Affymetrix 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 web page and imported into SAMS. That way, 61,103 sequences containing the Affymetrix annotations were integrated into SAMS and automatically re-annotated using the Metanor pipeline.

- ***Medicago truncatula*** **454 sequencing project**
  Cheung *et al.* (2006) used the pyrosequencing approach to generate 292,465

[2] http://www.medicago.org/genome/about.php

cDNA reads of *Medicago truncatula* using a GS20 sequencer. The reads were assembled into 3,619 sequences. These sequences were downloaded from the project web page and imported into SAMS. Using SAMS, a complete automatic annotation was performed.

Expression datasets

- **Oligonucleotide microarray expression datasets**
  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[Küster *et al.* (2007)]. These microarrays used two chip layouts designated Mt16kOli1[Hohnjec *et al.* (2005)] and Mt16kOli1Plus[Thompson *et al.* (2005)] (Arrayexpress ID: A-MEXP-85/A-MEXP-138). These arrays are associated to more than 50 different expression profiling experiments that were analyzed with the EMMA 2 (see Section 3.2) software. Results of these analyses are for example published by Baier *et al.* (2007), Gallardo *et al.* (2007), Hohnjec *et al.* (2006), and Küster *et al.* (2007).

- **Affymetrix GeneChip® expression data**
  Benedito *et al.* (2008) hybridized more than 50 Affymetrix Medicago GeneChips®, 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 performed in three replicates each. The expression datasets of the GeneChips® should be downloaded and integrated into the TRUNCATULIX data warehouse.

  As the EMMA2 software should be extended to analyze Affymetrix GeneChips®, the results of these hybridizations should be integrated into the data warehouse.

Database schema

To store information about genes, annotations, GO Categories (GeneOntology), COG groups (Clusters of Orthologous Groups of proteins), and expression datasets, five classes representing the different aspects are designed, pointed out in the following:

The main class in the reversed-star schema of the data warehouse is the class `GENE_ANNOTATION_MAIN` (see Figure 4.5). An object of the class `GENE_ANNOTATION_MAIN` stores the `REGION_ID_KEY`, which is the primary key for the reversed-star schema. Additionally, the `SOURCE` of the data (e.g. SAMS) and the name of the database (`DBNAME`) are stored. The other attributes of an object of the class `GENE_ANNOTATION_MAIN` are the `GENEID`, the `NAME` of the gene, the `TYPE` of the gene, the `LENGTH` of the gene, the functional annotation status (`STATUS_FUNCTION`), and the regional annotation status (`STATUS_REGION`).

GENE_ANNOTATION_MAIN	
REGION_ID_KEY:	INTEGER
SOURCE:	VARCHAR (255)
DBNAME:	VARCHAR (255)
GENEID:	INTEGER
NAME:	VARCHAR (255)
TYPE:	VARCHAR (255)
SEQUENCE:	TEXT
LENGTH:	INTEGER
STATUS_FUNCTION:	VARCHAR (32)
STATUS_REGION:	VARCHAR (32)
ANNOTATION_NAME:	VARCHAR (255)
ANNOTATION_GENEPRODUCT:	VARCHAR (255)
ANNOTATION_DESCRIPTION:	VARCHAR (255)
ANNOTATION_COMMENT:	VARCHAR (255)
ANNOTATION_ANNOTATOR:	VARCHAR (255)
ANNOTATION_EC:	VARCHAR (255)
ANNOTATION_COG:	VARCHAR (255)
ANNOTATION_CONFIDENCE:	VARCHAR (255)

Figure 4.5: The class GENE_ANNOTATION_MAIN. The unique key of an object of the class GENE_ANNOTATION_MAIN is the attribute REGION_ID_KEY. Each gene stored in the warehouse is represented by an object of the class GENE_ANNOTATION_MAIN, which stores all information about the gene that is imported from SAMS, including the annotation (attributes starting with ANNOTATION_).

If the stored gene has been annotated (automatically or manually), this information should also be stored in the object. For this purpose, the attributes ANNOTATION_NAME, ANNOTATION_GENEPRODUCT, ANNOTATION_DESCRIPTION, ANNOTATION_COMMENT, ANNOTATION_ANNOTATOR, ANNOTATION_EC, ANNOTATION_COG, and ANNOTATION_CONFIDENCE store the entitled values.

The class EXPRESSION_DATA handles information about microarray gene expression experiments (see Figure 4.6). An object of this class refers to exactly one GENE_ANNOTATION_MAIN object by storing the REGION_ID_KEY of that object. This way, the results of many different expression experiments can be referenced to one GENE_ANNOTATION_MAIN object. Each EXPRESSION_DATA object stores a unique EXPRESSION_ID_KEY, the name of the respective EXPERIMENT, the name of the AUTHOR who performed the experiment, the name of the represented GENE, an internal BRIDGELINK to a linked GenDB or SAMS gene if available, the FACTORVALUE of the experiment, the GENEID, the name of the applied STATISTICal analysis, the calculated expression values (PVALUE, APVALUE, MEAN, SD, A1MEAN) and the number

EXPRESSION_DATA	
REGION_ID_KEY:	INTEGER
EXPRESSION_ID_KEY:	INTEGER
EXPERIMENT:	VARCHAR (255)
AUTHOR:	VARCHAR (255)
GENE:	VARCHAR (255)
BRIDGELINK:	VARCHAR (255)
FACTORVALUE:	VARCHAR (255)
GENEID:	INTEGER
STATISTIC:	FLOAT (17)
PVALUE:	FLOAT (17)
APVALUE:	FLOAT (17)
MEAN:	FLOAT (17)
SD:	FLOAT (17)
A1MEAN:	FLOAT (17)
REPLICATES:	INTEGER

Figure 4.6: The class EXPESSSION_DATA. The attributes of an object of the class EXPESSSION_DATA are the REGION_ID_KEY (which connects each object of the class with one object from the main table), the EMMA_ID_KEY, the information about the performed experiment, and the resulting expression values.

of REPLICATES used in the experiment.

An object of the class OBSERVATION (see Figure 4.7) stores information about the prediction of functional tools for a single gene. The observation refers to a gene via a stored REGION_ID_KEY from the class GENE_ANNOTATION_MAIN. This way one or more observations are connected to one GENE_ANNOTATION_MAIN object. An *OBSERVATION* stores the following information: The attribute OBSERVATION_ID_KEY holds a unique key for each OBSERVATION. The other attributes are the TOOL that created the observation, the START and the STOP of the observation, the SCORE the tool rated the observation, and the DESCRIPTION of the result.

An object of the class GO (see Figure 4.8) stores information about a GeneOntology number. The GeneOntology number is associated to a gene via the REGION_ID_KEY. The attributes of a GO object are defined as a unique GO_ID_KEY and the GO number.

An object of the class COG stores a REGION_ID_KEY to the associated gene, a unique COG_ID_KEY, and the COG category itself (COGCAT) using the COG category identifier (Figure 4.9).

The class schema of the TRUNCATULIX data warehouse is shown in Figure 4.10.

OBSERVATION	
REGION_ID_KEY:	INTEGER
OBSERVATION_ID_KEY:	INTEGER
TOOL:	VARCHAR (255)
START:	INTEGER
STOP :	INTEGER
SCORE:	VARCHAR (255)
DESCRIPION:	VARCHAR (255)

Figure 4.7: The class `OBSERVATION`. Attributes of an object of the class `OBSERVATION` are the `REGION_ID_KEY`, the `OBSERVATION_ID_KEY`, the name of the `TOOL`, the `START` and `STOP` of the observation, the `SCORE` of the tool, and the `DESCRIPTION` of the tool results.

GO	
REGION_ID_KEY:	INTEGER
GO_ID_KEY:	INTEGER
GO	VARCHAR (255)

Figure 4.8: The class `GO`. Attributes of an object of the class `GO` are the `REGION_ID_KEY`, the `GO_ID_KEY`, and the `GO` number.

COG	
REGION_ID_KEY:	INTEGER
COG_ID_KEY:	INTEGER
COGCAT:	VARCHAR (255)

Figure 4.9: The class `COG`. Attributes of an object of the class `COG` are the `REGION_ID_KEY`, the `COG_ID_KEY`, and the COG category identifier (`COGCAT`).

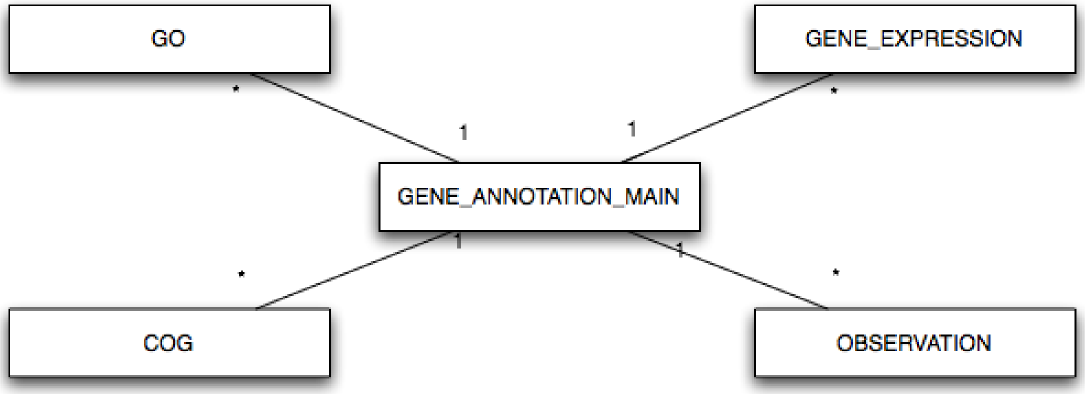


Figure 4.10: The class scheme of the designed classes of the TRUNCATULIX data warehouse. The `REGION_ID_KEY` is the connection for each of the other tables to the `GENE_ANNOTATION_MAIN` table.

Data import design:
Typically, data warehouses use the ETL approach for the import of datasets (extract - transform - load, see Section 2.3). As TRUNCATULIX uses a reversed-star schema, it is possible to split the three steps. This has the positive effect that the datasets from different data sources can be connected to each other when already imported into the data warehouse, and to import additional data later on without the need to reimport all datasets.

ETL:
For each source database an export script has to be created. Due to the previously described possibility of the reversed-star schema to link the datasets after the import, it is possible to create combined export and import scripts for each source database and to link the imported datasets afterwards. SAMS stores the sequence, annotation, and observation datasets of five different *Medicago truncatula* projects. The SAMS database can be accessed via the O2DBI2 Perl API and the TRUNCATULIX database can be connected to via the Perl DBI module (Perl Database Interface Module) or the BioMart Perl API. This provides the opportunity to export the sequence and annotation information from SAMS and directly import them into the TRUNCAULTIX database within one script.

Most of the microarray expression datasets that should be integrated into the data warehouse are stored in the EMMA2 database, which can also be accessed via the O2DBI2 Perl API. The modular pipeline system of EMMA2 allows to create an export-import script that can be started within EMMA2. The script can be configured within the web interface. Once started, it gathers the selected microarray expression datasets and directly imports them into the TRUNCATULIX database. Additional microarray expression datasets are downloaded from the *Medicago*

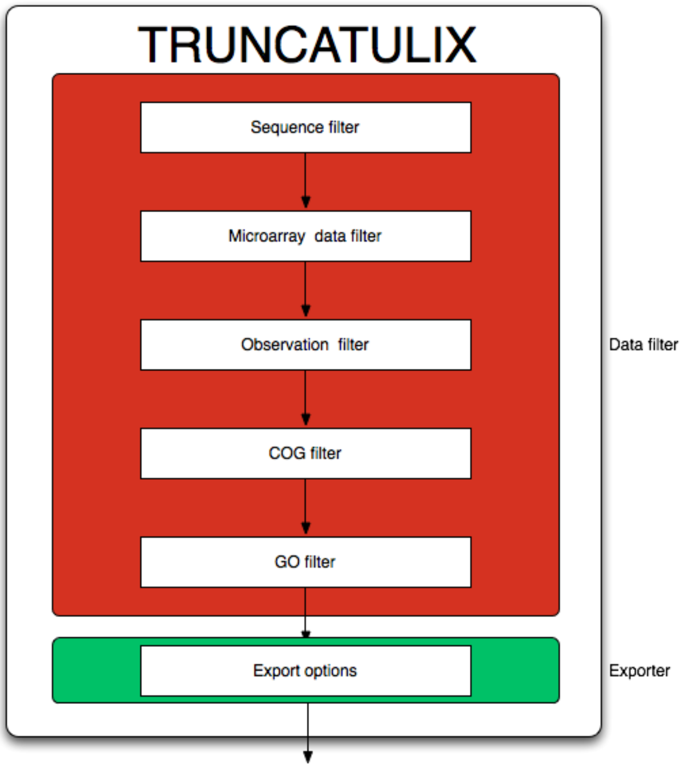


Figure 4.11: Workflow for a standard TRUNCATLIX query. At first, all datasets are filtered according to sequence and annotation, expression experiments, observations, COG, and GO numbers. Afterwards the results can be exported according to the export options.

truncatula gene expression atlas and are stored as csv files. A script should be created to import these expression datasets.

Link datasets:

After the import of all sequence and expression datasets, the linking of the datasets should be completed with an extra script. This script should be given a file containing the information which gene to link to which expression dataset.

Frontend:

The TRUNCATULIX data warehouse should be accessible via a frontend for users all over the world. This suggests to design a web-based frontend for easy access of the warehouse, requiring nothing more than a conventional webbrowser. As the user does not want to see all stored datasets, filters are used to separate interesting datasets from the complete data repository. More filters result in a smaller and more precise output. These filters should be arranged in a clear manner, so that the user is not glutted by the filter options. A pipeline for the workflow of a standard query of TRUNCATULIX is displayed in Figure 4.11.

After all filtering steps are completed, an export page should allow to select which attributes of the found datasets should be exported and what kind of file format is to be created for the export. A preview should demonstrate how the datasets look like (export attributes and data values).

4.3 MediPlEx - a tool to combine in silico & experimental gene expression values of the model legume Medicago truncatula

The idea to combine expression datasets from different gene expression analysis methods, such as microarray gene expression datasets and EST expression information is a central goal of this thesis. This should be implemented for the plant *Medicago truncatula*, as it is a model organism for legume biology and many datasets were created in the past. The design for the tool MediPlEx (MEDIcago truncatula multiPLe EXpression tool) is described in this section.

The desired workflow of MediPlEx is depicted in Figure 4.12, the different steps are described in the following sections:

4.3.1 Gene selection

The first step in a combined expression analysis should be to select genes of interest. For this purpose, EST libraries should be selected such that genes expressed under these library conditions can be found. The assembly information which is stored in SAMS should be used to find these genes. The logarithmic likelihood ratio (see Section 2.1.4) should be calculated for this set of genes based on the assembly information and the selection of EST libraries. The genes and the logarithmic likelihood ratio are then used for further analysis.

4.3.2 Selection of microarray expression datasets

As a second step, microarray gene expression datasets should be selected and combined with the previously calculated logarithmic likelihood ratio. The user is presented a complete lists of microarray gene expression experiments stored in the data warehouse TRUNCATULIX (Section 4.2), from which he can select the experiments he want to use for the combined analysis.

4.3.3 Clustering of expression datasets

All expression datasets should be clustered hierarchically, supposing that gene clusters show correlating expression profiles for the selected expression experiments.

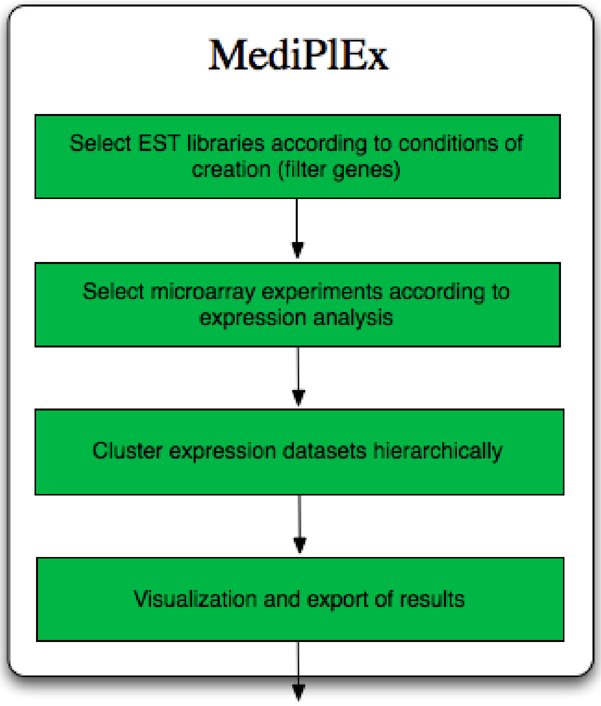

MediPlEx
Select EST libraries according to conditions of creation (filter genes)
Select microarray experiments according to expression analysis
Cluster expression datasets hierarchically
Visualization and export of results

Figure 4.12: Suggested workflow of MediPlEx. In the first step, EST libraries created under certain conditions should be selected so that genes expressed under these conditions can be found. The second step allows to select which microarray gene expression datasets should be used for the expression analysis. Afterwards the datasets are combined and clustered hierarchically. The results can be browsed and downloaded.

Genes of one cluster share a similar expression profile and may belong to the same pathway or are needed in the same reactions.
Ward's clustering algorithm should be used for this purpose, as the algorithm tries to minimize the loss of information while creating the clusters.

4.3.4 Visualization of results

For the presentation of the results, a table should show up all expression values for the set of genes found. Additionally, the cluster dendrogram should be presented. An interactive 3D-visualization should make the clustering more traceable for the user. The results should be downloadable and contain all original expression datasets and annotations.

The interface of MediPlEx should be accessible to users all over the world via a web browser and should be easy to use.

CHAPTER 5

Implementation

This Chapter describes the implementation of the previously designed applications. First, this Chapter illustrates the implementations to store and analyze Affymetrix GeneChip® expression datasets with EMMA2 (cf. Section 2.2.4) the same way as conventional oligonucleotide microarrays. Second, the implementation of the TRUNCATULIX data warehouse is presented, focusing on data handling, data access, and frontend visualization. As a last part of this chapter, the implementation of the tool MediPlEx, combining different gene expression analyses, is outlined in detail.

5.1 Extension of EMMA2 to store and analyze Affymetrix GeneChip® expression datasets.

EMMA2 is implemented in Perl[1], using a relational MySQL database[2], and an objectrelational mapping between Perl objects and relational data storage (O2BDI2 introduced by Clausen (2002)). An Apache2 webserver hosts the html-based frontend (dynamic HTML generated with Perl, combined with some Java applets, JavaScript, and dynamic Ajax elements). Microarray datasets are stored efficiently using HDF5 files (Hierarchical Data Format[3]). The Perl Data Language (PDL) is used as an interface for the HDF5 files. Statistical computations are performed using the R statistic programming language[R Development Core Team (2008)], compute jobs are calculated on a compute cluster.

[1] http://www.perl.org

[2] http://www.mysql.com/

[3] http://www.hdfgroup.org/

Layout import

The import of an array layout of GeneChips$^{®}$ is more complex than importing a layout of classical oligonucleotide microarrays, because the data to be imported is not stored in one single file, but spread over 3 files (for details see Section 4.1).

Due to the size of a GeneChip$^{®}$ dataset (about 1.000.000 reporter), the import of the layout is divided into two successive steps:
In the first step the essential information of the layout is imported. This includes the reporters, genes, and basic layout information. In a second step, all additional information are added to the array layout, like sequences of the genes and reporters, but also the information of x and y coordinates and PM and MM information.

This two-step method offers the possibility to create the layout fastly and to add all additional data afterwards, which is more memory-efficient.

Implementation of the GeneChip$^{®}$ array layout importer:

The GeneChip$^{®}$ array layout importer is implemented in Perl so that it can be integrated in the existing pipeline framework of EMMA2. The main steps of the import process are:

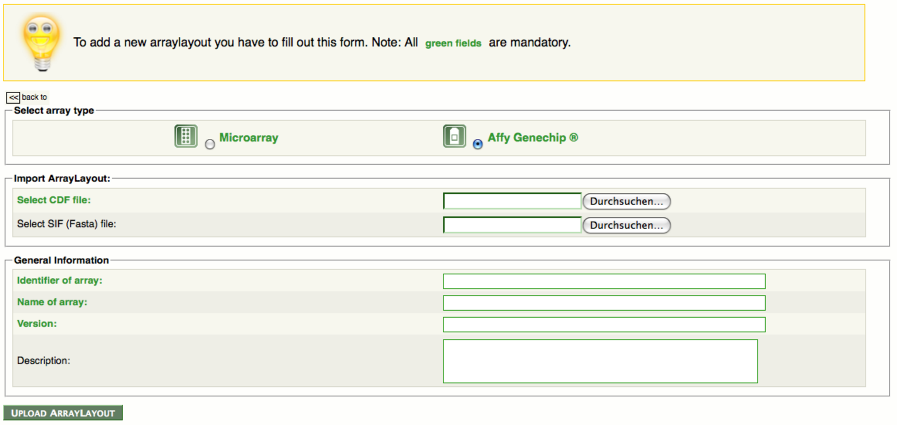
- Read the CDF file, load only basic layout information and reporter names.
- Create an ArrayLayout in the database of the EMMA2 project, containing the basic information loaded before.
- Read the CDF file again, load all stored information (including the reporter sequences and the x and y coordinates of the spots).
- Store these information and link the objects in the ArrayLayout.
- Read the SIF file, containing the gene names and fasta sequences.
- Store these information in the ArrayDesign and link it to the existing objects.

The web interface of EMMA2 is extended to load Affymetrix GeneChip layouts as shown in Figure 5.1.

As an additional option, a script is implemented to import the sequences of the spotted reporters stored in the _probe_tab file. This information is general not of interest and thus there is no option integrated into the web interface of EMMA2. If a user wants to add this information to the imported ArrayLayout, an administrator can start the script to do so.

Import of Affymetrix GeneChip$^{®}$ datasets

The raw GeneChip$^{®}$ microarray expression datasets can be uploaded into the ArrayLims application in the same way as oligonucleotide microarrays. The raw files are stored and administered internally. The EMMA2 software can connect to the ArrayLIMS application and load these raw datasets during the experiment creation step. In this step, the microarray layout has to be chosen. In case of an



To add a new arraylayout you have to fill out this form. Note: All green fields are mandatory.

Select array type

○ Microarray ● Affy Genechip ®

Import ArrayLayout:

Select CDF file:

Select SIF (Fasta) file:

General Information

Identifier of array:

Name of array:

Version:

Description:

Figure 5.1: A screenshot of the EMMA2 web interface focusing the import of an Affymetrix GeneChip$^{®}$ layout.

Affymetrix GeneChip$^{®}$ layout, it is only allowed to import GeneChip$^{®}$ datasets from ArrayLims. The datasets are loaded via the R statistic programming language (using the Bioconductor package, and the affy library). Transferred back to the Perl O2DBI API (using RSPerl) the raw values are stored in the EMMA2 database in `MBAD` (BioAssayData → MeasuredBioAssayData) objects. A screenshot of the web interface for the import of Affymetrix GeneChip$^{®}$ datasets into EMMA2 is shown in Figure 5.2.

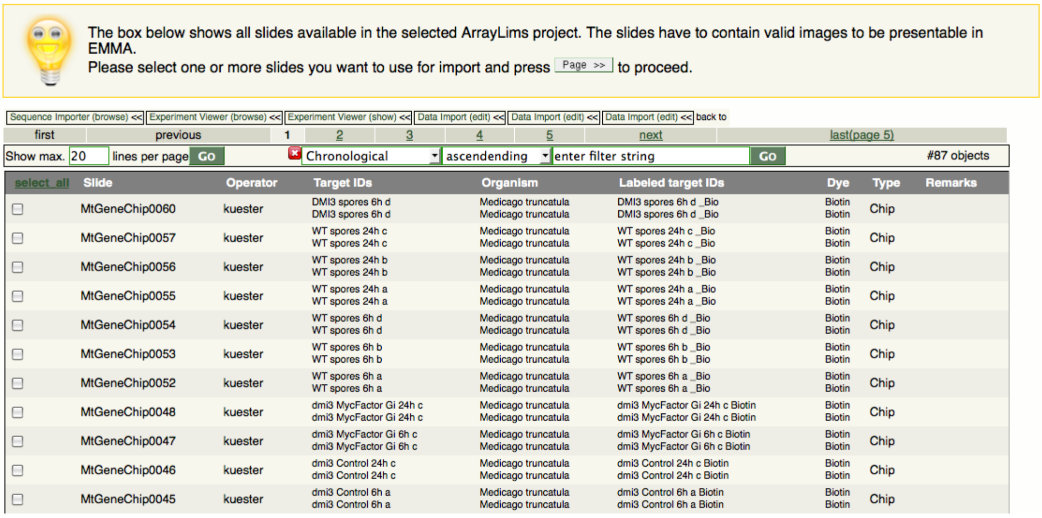
Data pre-processing and processing

For Affymetrix GeneChip$^{®}$ microarrays there exists a set of different pre-processing and normalization methods that have been developed in recent years. For the implementation in the EMMA2 system, the most commonly used are integrated as pipeline tools to be computed for all arrays in one experiment. As for oligonucleotide microarrays, the R statistic programming language is used for the computation, as it is very fast and efficient. The functions adapted for EMMA2 resort on the functions provided by the affy package by Gautier *et al.* (2004) (MAS5, RMA, and MBEI) from Bioconductor and the expresso package by Wu *et al.* (2003)(GCRMA) . The different normalization functions offer different options that can be used to fine-tune the calculations.

These functions are explained here, the options can be adjusted in the web-interface:

MAS5.0:

MAS5.0 normalization is performed on each of the GeneChips$^{®}$ separate using one GeneChip$^{®}$ as a reference. A background correction is performed using perfect-match (PM) and mismatch (MM) probes.



The box below shows all slides available in the selected ArrayLims project. The slides have to contain valid images to be presentable in EMMA.
Please select one or more slides you want to use for import and press Page >> to proceed.

Show max. 20 lines per page

#87 objects

select_all	Slide	Operator	Target IDs	Organism	Labeled target IDs	Dye	Type	Remarks
□	MtGeneChip0060	kuester	DMI3 spores 6h d DMI3 spores 6h d	Medicago truncatula Medicago truncatula	DMI3 spores 6h d _Bio DMI3 spores 6h d _Bio	Biotin Biotin	Chip	
□	MtGeneChip0057	kuester	WT spores 24h c WT spores 24h c	Medicago truncatula Medicago truncatula	WT spores 24h c _Bio WT spores 24h c _Bio	Biotin Biotin	Chip	
□	MtGeneChip0056	kuester	WT spores 24h b WT spores 24h b	Medicago truncatula Medicago truncatula	WT spores 24h b _Bio WT spores 24h b _Bio	Biotin Biotin	Chip	
□	MtGeneChip0055	kuester	WT spores 24h a WT spores 24h a	Medicago truncatula Medicago truncatula	WT spores 24h a _Bio WT spores 24h a _Bio	Biotin Biotin	Chip	
□	MtGeneChip0054	kuester	WT spores 6h d WT spores 6h d	Medicago truncatula Medicago truncatula	WT spores 6h d _Bio WT spores 6h d _Bio	Biotin Biotin	Chip	
□	MtGeneChip0053	kuester	WT spores 6h b WT spores 6h b	Medicago truncatula Medicago truncatula	WT spores 6h b _Bio WT spores 6h b _Bio	Biotin Biotin	Chip	
□	MtGeneChip0052	kuester	WT spores 6h a WT spores 6h a	Medicago truncatula Medicago truncatula	WT spores 6h a _Bio WT spores 6h a _Bio	Biotin Biotin	Chip	
□	MtGeneChip0048	kuester	dmi3 MycFactor Gi 24h c dmi3 MycFactor Gi 24h c	Medicago truncatula Medicago truncatula	dmi3 MycFactor Gi 24h c Biotin dmi3 MycFactor Gi 24h c Biotin	Biotin Biotin	Chip	
□	MtGeneChip0047	kuester	dmi3 MycFactor Gi 6h c dmi3 MycFactor Gi 6h c	Medicago truncatula Medicago truncatula	dmi3 MycFactor Gi 6h c Biotin dmi3 MycFactor Gi 6h c Biotin	Biotin Biotin	Chip	
□	MtGeneChip0046	kuester	dmi3 Control 24h c dmi3 Control 24h c	Medicago truncatula Medicago truncatula	dmi3 Control 24h c Biotin dmi3 Control 24h c Biotin	Biotin Biotin	Chip	
□	MtGeneChip0045	kuester	dmi3 Control 6h a dmi3 Control 6h a	Medicago truncatula Medicago truncatula	dmi3 Control 6h a Biotin dmi3 Control 6h a Biotin	Biotin Biotin	Chip	

Figure 5.2: A screenshot of the EMMA2 dialog for importing Affymetrix GeneChip® arrays into a new experiment in EMMA2.

RMA:

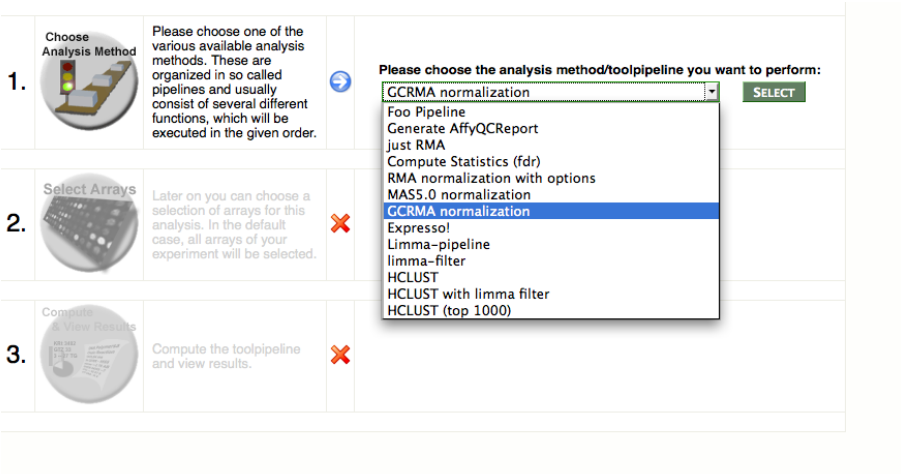
Using the RMA (Robust Multichip Average) normalization defined by Irizarry *et al.* (2003), all GeneChips® in the experiment are normalized together. The algorithm uses a pool of perfect-match (PM) probes to normalize each value. As background correction, the PM distribution is used to get an overall background level. Then a transformation based on a background noise and signal model is applied.

GCRMA:

GCRMA uses the RMA normalization with the help of probe sequence and with GC-content background correction. The perfect-match (PM) values are background-corrected, normalized and finally summarized resulting in a set of expression measures.

Expresso:

The expresso package offers more options that can be adjusted to the datasets. The expresso package implements nearly all available algorithms for background correction, normalization, PM adjustment measures, and expression value transformation. Available background correction methods are rma, rma2, mas, and none. For the normalization, the user can select from the following algorithms: quantiles, scaling(mas5 like), constant, invariant set (aka dChip), paired loess, contrast, quantiles.probeset, qspline, and quantiles.robust. The available PM adjustment methods are pmonly, substactmm(mas4), and mas(mas5). For the calculation of an expression value, the algorithms mas, medianpolish(rma), playerout, liwong(aka. dChip), and avgdiff are available.



1. Choose Analysis Method — Please choose one of the various available analysis methods. These are organized in so called pipelines and usually consist of several different functions, which will be executed in the given order.

2. Select Arrays — Later on you can choose a selection of arrays for this analysis. In the default case, all arrays of your experiment will be selected.

3. Compute & View Results — Compute the toolpipeline and view results.

Please choose the analysis method/toolpipeline you want to perform:

Figure 5.3: The screenshot shows the EMMA2 interface for preprocessing and normalization of Affymetrix GeneChip® microarrays.

The user can also decide to logarithmize the results of the computation.

The setup for the normalization functions for the GeneChip® in one experiment:

- Load all `MBAD` (MeasuredBioAssayData) objects from the database

- Create a job to be computed on the compute cluster using R, starting the selected normalization function with the selected options and the complete datasets of the experiment.

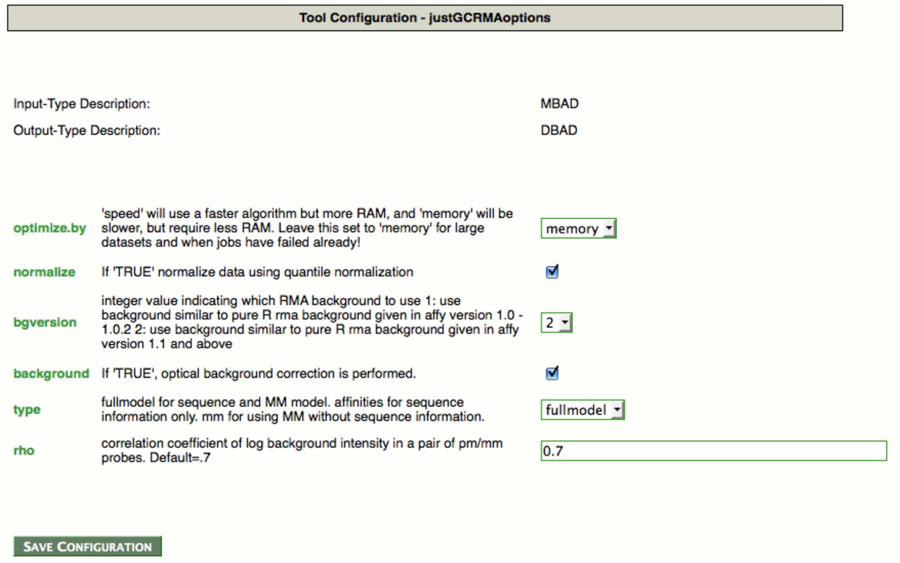
- Store expression datasets in `DBAD` (Derived BioAssayData) objects in the database

A screenshot of the web interface of EMMA2 for selecting the preprocessing and normalization method is presented in Figure 5.3, a screenshot presenting the selectable options is shown in Figure 5.4.

For quality control, the R package AffyQCReport is integrated and can create PDF documents with various statistics and plots.

Significance tests

The significance tests to be used with Affymetrix GeneChip® datasets are nearly the same that were used for the classical oligonucleotide microarrays. One new significance test was added to the EMMA2 system, and Affymetrix optimized two-sample t-test (Affy two sample-test). The pipelines load the normalized datasets (`DBAD` objects) and run the selected significance test in the R environment. The



Tool Configuration - justGCRMAoptions

Input-Type Description:	MBAD
Output-Type Description:	DBAD

optimize.by	'speed' will use a faster algorithm but more RAM, and 'memory' will be slower, but require less RAM. Leave this set to 'memory' for large datasets and when jobs have failed already!	memory
normalize	If 'TRUE' normalize data using quantile normalization	☑
bgversion	integer value indicating which RMA background to use 1: use background similar to pure R rma background given in affy version 1.0 - 1.0.2 2: use background similar to pure R rma background given in affy version 1.1 and above	2
background	If 'TRUE', optical background correction is performed.	☑
type	fullmodel for sequence and MM model. affinities for sequence information only. mm for using MM without sequence information.	fullmodel
rho	correlation coefficient of log background intensity in a pair of pm/mm probes. Default=.7	0.7

Figure 5.4: A screenshot showing the EMMA2 interface so select normalization options for the integrated GCMRA normalization.

calculated values are stored in the database as DBAD objects afterwards. Figure 5.5 show a screenshot of the EMMA2 web databrowser and the loaded GeneChip® expression datasets.

Clustering

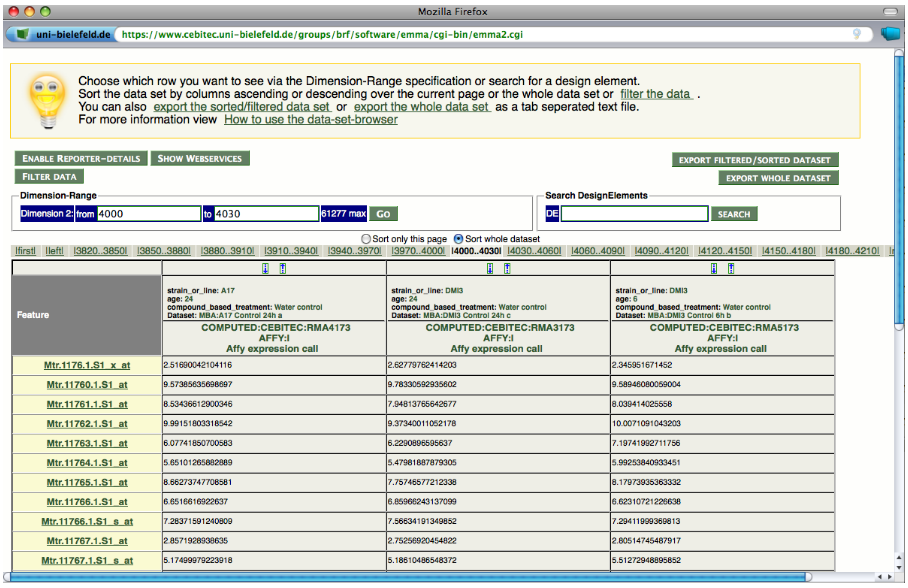
The clustering pipelines used for conventional oligonucleotide microarrays can be used for GeneChip® datasets as well, because the DBAD objects can be handled as classical oligonucleotide datasets. The pipeline ”Hierarchical clustering (Top 1000) ” is the common clustering pipeline, allowing to tweak the clustering according to prefiltering options, significance filters, distance method and clustering method.

5.2 TRUNCATULIX

This Chapter describes the implementation of the data warehouse TRUNCATULIX designed in Section 4.2.

The backend of the TRUNCATULIX data warehouse is based on the Sophia data warehouse backend developed by Runte (2010). MySQL is used as relation database management system.

As Sophia is compatible to BioMart, the database schema is generated using the BioMart designer[Durinck *et al.* (2005)]. An API for the database and query functionality is available in JAVA and Perl (using the BioMart API). Intensive tests showed that the BioMart API is well designed, but for the purpose of querying for specific datasets it is too complex and too slow. A self-implemented rudimental


Mozilla Firefox
uni-bielefeld.de
https://www.cebitec.uni-bielefeld.de/groups/brf/software/emma/cgi-bin/emma2.cgi
ENABLE REPORTER-DETAILS
SHOW WEBSERVICES
FILTER DATA
EXPORT FILTERED/SORTED DATASET
EXPORT WHOLE DATASET
Go
SEARCH
|first| |left| |3820..3850| |3850..3880| |3880..3910| |3910..3940| |3940..3970| |3970..4000| |4000..4030| |4030..4060| |4060..4090| |4090..4120| |4120..4150| |4150..4180| |4180..4210| |r

Choose which row you want to see via the Dimension-Range specification or search for a design element.
Sort the data set by columns ascending or descending over the current page or the whole data set or filter the data .
You can also export the sorted/filtered data set or export the whole data set as a tab seperated text file.
For more information view How to use the data-set-browser

Dimension-Range
Dimension 2: from 4000 to 4030 61277 max

Search DesignElements
DE

○ Sort only this page ● Sort whole dataset

Feature	strain_or_line: A17 age: 24 compound_based_treatment: Water control Dataset: MBA:A17 Control 24h a COMPUTED:CEBITEC:RMA4173 AFFY:I Affy expression call	strain_or_line: DMI3 age: 24 compound_based_treatment: Water control Dataset: MBA:DMI3 Control 24h c COMPUTED:CEBITEC:RMA3173 AFFY:I Affy expression call	strain_or_line: DMI3 age: 6 compound_based_treatment: Water control Dataset: MBA:DMI3 Control 6h b COMPUTED:CEBITEC:RMA5173 AFFY:I Affy expression call
Mtr.1176.1.S1_x_at	2.51690042104116	2.62779762414203	2.345951671452
Mtr.11760.1.S1_at	9.57385635698697	9.78330592935602	9.58946080059004
Mtr.11761.1.S1_at	8.53436612900346	7.94813765642677	8.039414025558
Mtr.11762.1.S1_at	9.99151803318542	9.37340011052178	10.0071091043203
Mtr.11763.1.S1_at	6.07741850700583	6.2290896595637	7.19741992711756
Mtr.11764.1.S1_at	5.65101265882889	5.47981887879305	5.99253840933451
Mtr.11765.1.S1_at	8.66273747708581	7.75746577212338	8.17973935363332
Mtr.11766.1.S1_at	6.6516616922637	6.85966243137099	6.62310721226638
Mtr.11766.1.S1_s_at	7.28371591240809	7.56634191349852	7.29411999369813
Mtr.11767.1.S1_at	2.8571928938635	2.75256920454822	2.80514745487917
Mtr.11767.1.S1_s_at	5.17499979223918	5.18610486548372	5.51272948895852

Figure 5.5: The screenshot shows the processed datasets of an Affymetrix GeneChip® in the EMMA2 dataset browser.

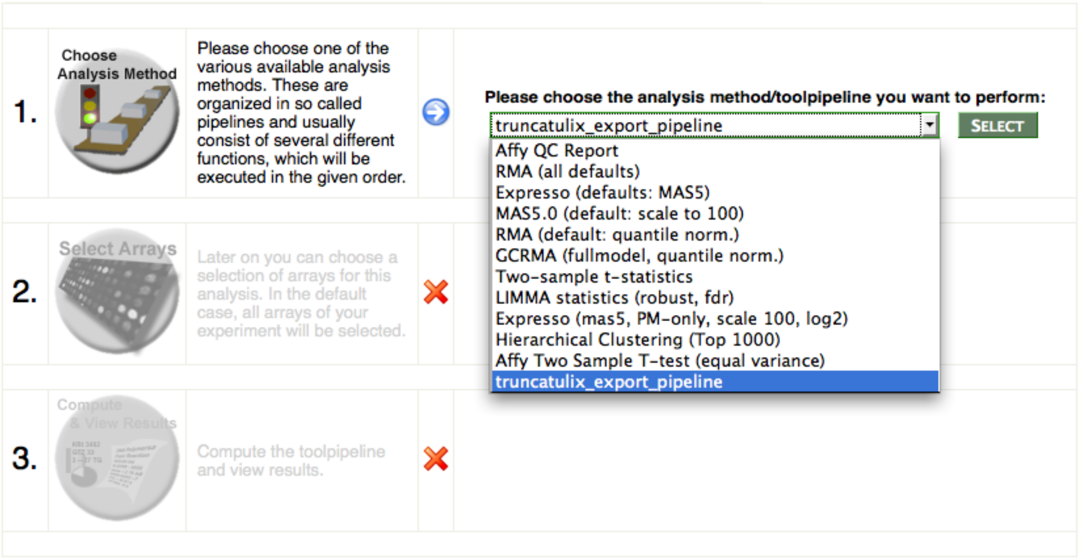
Perl API is created for a fast and effective query of the gene expression datasets (documented in the Appendix A.2).

As designed in Section 4.2, three importer tools need to be implemented to gain to ability to import every dataset connected to *Medicago truncatula*. One importer should import sequence and annotation datasets from SAMS. This importer also fetches the observations, GO numbers, and COG categories to import them into the warehouse. The second importer collects the microarray gene expression datasets from the EMMA2 database and transfers them into the database of TRUNCATULIX. The third importer should read csv-files storing microarray gene expression datasets. This kind of data is available for download in most online repositories.

Only an administrator has the privileges to import the datasets into the TRUNCATULIX data warehouse.

Importer for sequence and annotation datasets from SAMS

The importer to load sequence and annotation datasets from SAMS is implemented as a Perl command-line script. This way, the O2DBI API can be used to access the SAMS database. The project from which the datasets should be exported is selected as a command-line option. Using the O2DBI API, all gene objects of the selected project are fetched from the database. For each gene, all relevant data is fetched from the database and transferred into the database of the TRUNCATULIX data warehouse: The attributes of the (`GENE_ANNOTATION_MAIN`) table are taken from the `$gene` object in SAMS, and from the `$latest_annotation_function` object from this gene.



1. Choose Analysis Method — Please choose one of the various available analysis methods. These are organized in so called pipelines and usually consist of several different functions, which will be executed in the given order.

2. Select Arrays — Later on you can choose a selection of arrays for this analysis. In the default case, all arrays of your experiment will be selected.

3. Compute & View Results — Compute the toolpipeline and view results.

Please choose the analysis method/toolpipeline you want to perform: truncatulix_export_pipeline

SELECT
Affy QC Report
RMA (all defaults)
Expresso (defaults: MAS5)
MAS5.0 (default: scale to 100)
RMA (default: quantile norm.)
GCRMA (fullmodel, quantile norm.)
Two-sample t-statistics
LIMMA statistics (robust, fdr)
Expresso (mas5, PM-only, scale 100, log2)
Hierarchical Clustering (Top 1000)
Affy Two Sample T-test (equal variance)
truncatulix_export_pipeline

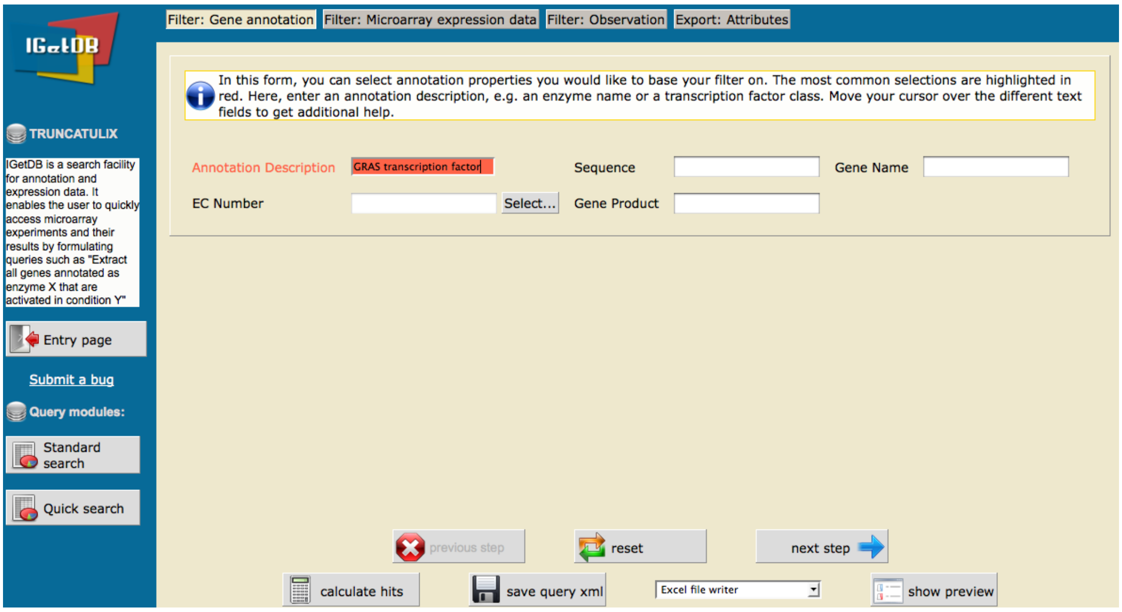
Figure 5.6: A Screenshot showing the EMMA2 web interface and the option to export expression data via pipeline into the database of the TRUNCATULIX data warehouse.

Importer for microarray expression datasets from EMMA2

For the import of microarray expression datasets, an importer is implemented as an EMMA2 pipeline tool to export all expression information of a selected experiment to the TRUNCATULIX data warehouse (see Figure 5.6). When starting the pipeline tool, the user has to select which datasets should be exported. The user can select which of the available values should be exported. All datasets are preselected by default, but if some datasets should not be opened to the public (as the TRUNCATULIX data warehouse is a public data source) they can be deselected. A standard export creates one entry in the `EXPRESSION_DATA` table per reporter and per hybridization. As the expression data has to be linked to the `GENE_ANNOTATION_MAIN` table entries, the `Reporter` object in EMMA2 stores a reference to a `BioSequence` object, which has an attribute called `_GenDBRegion`. This attribute may store a linked SAMS gene name (e.g. TC00012). If this name is stored, the importer looks up this gene in the TRUNCATULIX database and links the gene and the expression dataset via the `REGION_ID_KEY`. If the attribute is not used, the expression data can be linked afterwards in a manual started linking script (see below). The `EMMA_ID_KEY` is a unique and auto-incrementing key for the expression dataset. The `EXPERIMENT` is set by the experiment name from EMMA ($experiment → name()), the name of the gene is taken from the `reporter` object ($reporter → name()), if an internal bridge-link is stored in the attribute $reporter → _bridgelink, it is stored in the `BRIDGELINK` attribute of the table. The attribute `GENEID` stores the $reporter → _id(), the `GENDB` attribute stores the previously used linked GenDBRegion name. The name of statistic used in the data analysis is taken from the DBAD object (name) and stored in the attribute STATISTIC. The calculated statistical values are taken from the derived BioAssayData objects and its attributes and stored in the respective EXPRESSION_DATA attributes (PVALUE, APVALUE(adjusted p-value), MEAN, SD, and A1MEAN). The number of microarray replicates used in the experiment is stored in the REPLICATES attribute.

Importer for microarray gene expression datasets from a csv-file
For the import of microarray gene expression datasets from other sources than EMMA2, a script is implemented in Perl to import datasets from csv-files (comma separated values). The script receives the name of the EXPERIMENT, the FACTORVALUE, the AUTHOR, the STATISTIC used and the filename of the file storing the expression information as arguments. The csv-file has to be stored in the following format:
Each row of the csv file is imported as one dataset in the EXPRESSION_DATA table. The first column contains the name of the reporter and is stored in the GENE attribute. The second column contains the number of replicates, stored in the REPLICATES attribute. The following rows contain the data to be stored in the attributes A1MEAN, and if available, the MEAN, SD, PVALUE and APVALUE. The given reporter name (GENE) is checked to match a NAME in the GENE_ANNOTATION_MAIN table and link the expression dataset to this gene. The csv-file is iterated row-by-row importing the expression information and linking to existing GENE_ANNOTATION_MAIN datasets in one step.

Linking script
A script for the linking of expression datasets to the stored genes is implemented, in case of expression datasets storing outdated gene names (e.g. references to an old version of the *Medicago truncatula* GeneIndex), or references to other gene names and databases. The script should be started manually after the import of the microarray gene expression datasets were linking errors occurred, or if it is known that the stored gene names are not valid to be linked. The script receives the name of the EXPERIMENT, the FACTORVALUE, and the name of the linking file, which contains two columns. The first column contains the name of the reporter of the gene name, the second column contains the reporter name of the microarray expression dataset. The script looks up the database for the entry of the gene and the entries of the expression datasets containing the repotername. If the script finds a microarray expression dataset for the given EXPERIMENT and FACTORVALUE which does not contain a link to a GENE, the expression dataset is linked to the respective gene. If the microarray expression dataset already contains a link to a gene different to the one given in the link-file, the microarray expression dataset is copied, given a new EXPRESSION_ID_KEY and linked to the gene from the link-file. This way, redundant data is stored in the database, as some datasets may be linked to more than one gene (e.g. to the same gene in the MTGI 8.0 and 9.0), but the reversed-star schema does not allow a linking of the secondary tables to the main table. This way the number of datasets increases due



IGetDB is a search facility for annotation and expression data. It enables the user to quickly access microarray experiments and their results by formulating queries such as "Extract all genes annotated as enzyme X that are activated in condition Y"

In this form, you can select annotation properties you would like to base your filter on. The most common selections are highlighted in red. Here, enter an annotation description, e.g. an enzyme name or a transcription factor class. Move your cursor over the different text fields to get additional help.

Annotation Description: GRAS transcription factor
Sequence:
Gene Name:
EC Number: Select...
Gene Product:

Excel file writer

Figure 5.7: A screenshot of the TRUNCATULIX web interface.The filter panel for the gene and annotation information is shown. The user can search the database for the annotation description, the sequence, the gene name and the EC number. The next filter step will then only search within the results of the first filter step.

to the stored redundancy, but as the speed of the database queries is nearly not affected, this tradeoff is accepted.

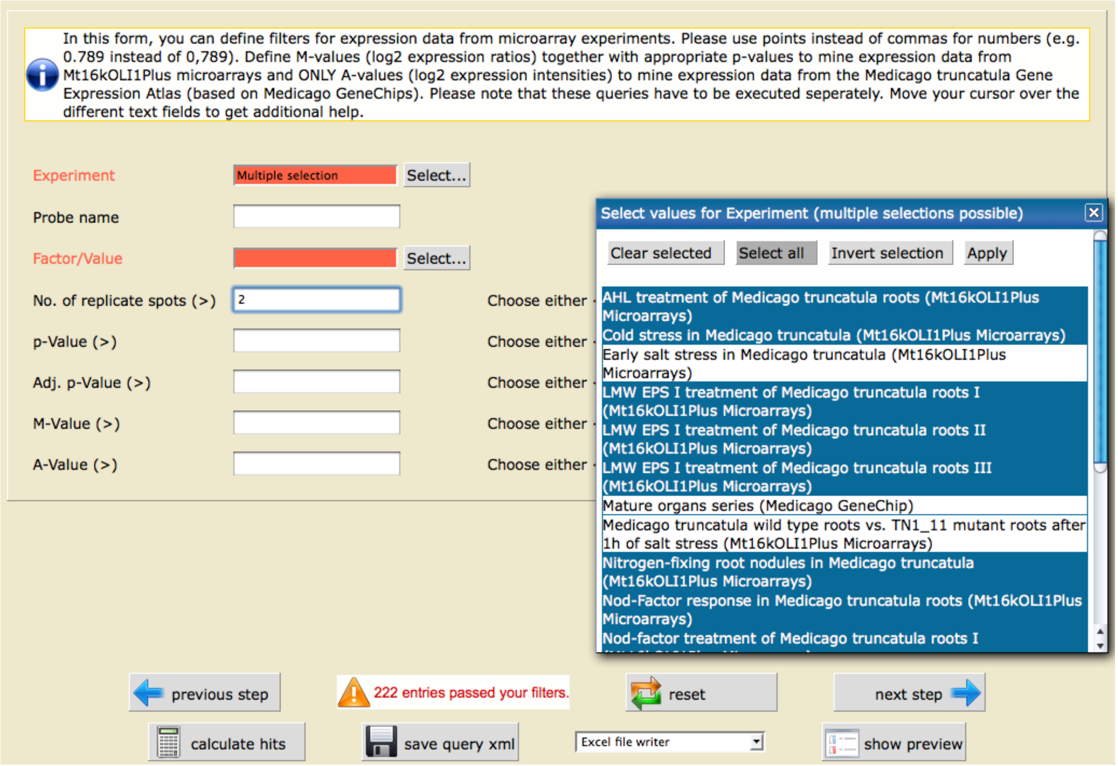
Visualization

TRUNCATULIX uses a web interface for user interactions. The interface is build using HTML, combined with the ECHO2 framework[4] and JAVA, based on the IgetDB data warehouse frontend.

As designed in Chapter four, three filter-pages are implemented in a clearly arranged way. One page for selecting the export options completes the basic search and export functionality. A progress bar at the top of each page shows the current filter step and allows to jump back and forth to one of the other filter and export panes. At the bottom of each page a counter informs the user how many genes remain according to the current filtering. The first filter concerns the genes and annotations. The user can filter for text or text fragments in the `ANNOTATION_DESCRIPTION`, for (a part) of the sequence, for the gene name, for one or more EC Numbers (or prefixes), and for gene products (see Figure 5.7). The `ANNOTATION_DESCRIPTION` is the most interesting filter option and thus marked in red. For the sake of clarity, more filter boxes for `GENE_ANNOTATION_MAIN` table attributes were left out.

The second filter panel is dedicated to the microarray gene expression datasets

[4] http://echo.nextapp.com



In this form, you can define filters for expression data from microarray experiments. Please use points instead of commas for numbers (e.g. 0.789 instead of 0,789). Define M-values (log2 expression ratios) together with appropriate p-values to mine expression data from Mt16kOLI1Plus microarrays and ONLY A-values (log2 expression intensities) to mine expression data from the Medicago truncatula Gene Expression Atlas (based on Medicago GeneChips). Please note that these queries have to be executed seperately. Move your cursor over the different text fields to get additional help.

Experiment: Multiple selection Select...
Probe name:
Factor/Value: Select...
No. of replicate spots (>): 2 Choose either
p-Value (>): Choose either
Adj. p-Value (>): Choose either
M-Value (>): Choose either
A-Value (>): Choose either

Select values for Experiment (multiple selections possible)

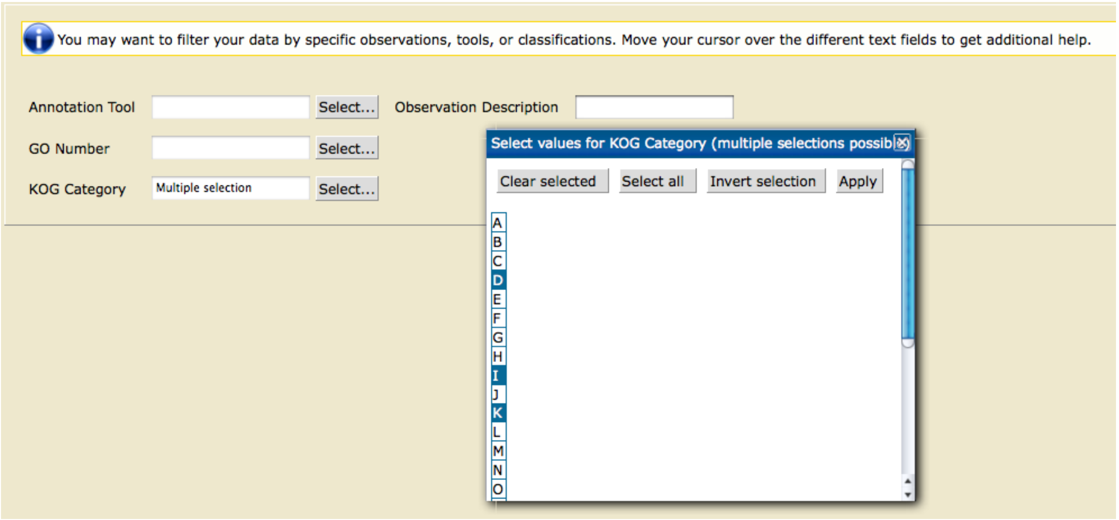
- AHL treatment of Medicago truncatula roots (Mt16kOLI1Plus Microarrays)
- Cold stress in Medicago truncatula (Mt16kOLI1Plus Microarrays)
- Early salt stress in Medicago truncatula (Mt16kOLI1Plus Microarrays)
- LMW EPS I treatment of Medicago truncatula roots I (Mt16kOLI1Plus Microarrays)
- LMW EPS I treatment of Medicago truncatula roots II (Mt16kOLI1Plus Microarrays)
- LMW EPS I treatment of Medicago truncatula roots III (Mt16kOLI1Plus Microarrays)
- Mature organs series (Medicago GeneChip)
- Medicago truncatula wild type roots vs. TN1_11 mutant roots after 1h of salt stress (Mt16kOLI1Plus Microarrays)
- Nitrogen-fixing root nodules in Medicago truncatula (Mt16kOLI1Plus Microarrays)
- Nod-Factor response in Medicago truncatula roots (Mt16kOLI1Plus Microarrays)
- Nod-factor treatment of Medicago truncatula roots I

222 entries passed your filters.

Figure 5.8: The screenshot shows the filter panel for the microarray gene expression datasets. The user can select which experiments he wants to query, the minimal number of replicates in the experiment, and of course minima and maxima for the expression values.

(Figure 5.8). The EXPERIMENT, as well as the FACTOR_VALUE can be selected in a multiple selection list in a separate window. The probe name can be specified as free text. For the "No. of replicate spots", as well as for the expression values ("p-Value, " Adj. p-Value", "M-Value", and "A-Value"), the user can specify if the searched genes should have values above or below a specified threshold.
The third filter covers observations, as well as GO numbers and COG categories (Figure 5.9). The filters for these three tables are located on one filter panel for the sake of clarity.
After these three filter steps, an export pane is displayed to select which attributes of the gene datasets should be exported. This covers all attributes stored in all five tables of the data warehouse (see Figure 5.10). The found datasets can be exported as csv file, as xls file (Microsoft Excel), or as HTML file including a table with the datasets.

In addition to the standard search interface, a quicksearch page is implemented allowing to search in a single google-like search box. The user can thus simply query for values in the attributes ANNOTATION_DESCRIPTION, ANNOTATED_GENE_PRODUCT, probe NAME, and GENE_NAME, which proved to be the mostly used search fields



You may want to filter your data by specific observations, tools, or classifications. Move your cursor over the different text fields to get additional help.

Annotation Tool

Observation Description

GO Number

KOG Category: Multiple selection

Select values for KOG Category (multiple selections possib

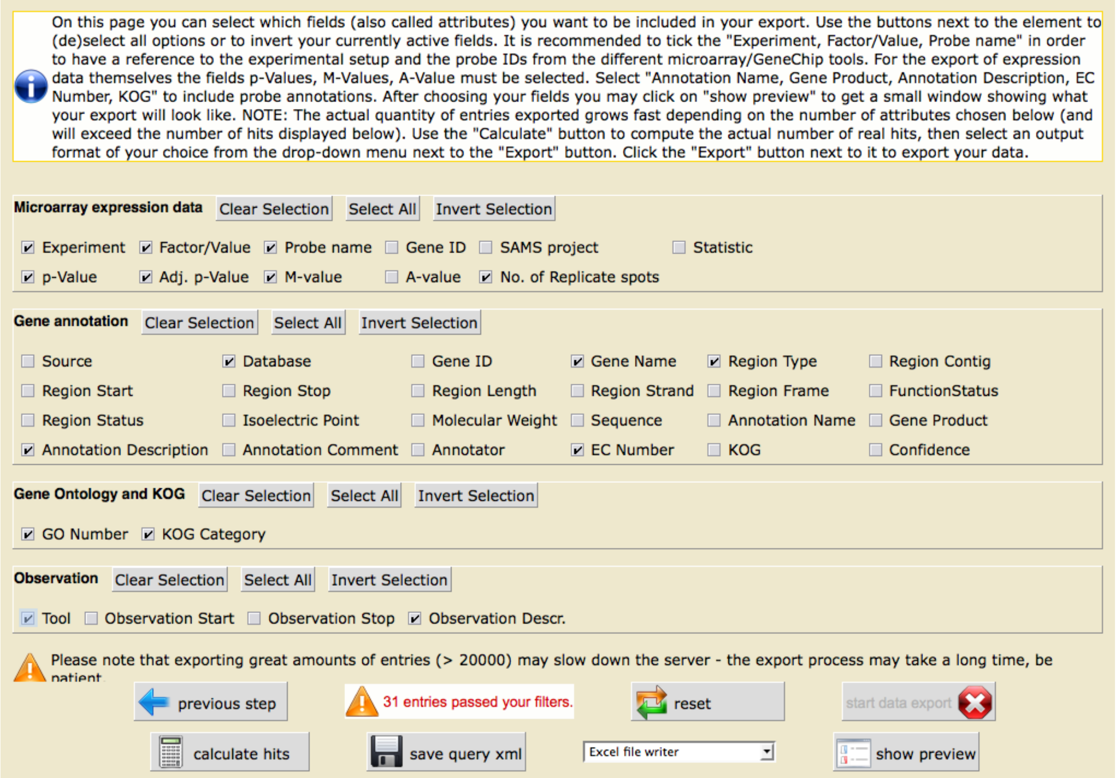
Figure 5.9: The filter panel of TRUNCATULIX for the observations, GO numbers and COG categories. The observations can be filtered for the used tool and for the observation description. GO numbers and COG categories can be selected in a new multiple selection window.

(see Figure 5.11). The query is implemented using the Lucene Java package[5], converting the query in "like % %" query strings. The textbox can also be used to query for other attributes, the syntax to be used is "attribute:[min_value TO max_value]" (e.g. "mean:[0 TO 2] " or pvalue:[0 TO 0.05]"). The query for a given text is processed and the user is directly forwarded to the same export panel as used for the complete query.

5.3 MediPlEx

MediPlEx (MEDIcago truncatula multiPLe EXpression tool) is designed to combine expression datasets from different types of transcriptome experiments (microarray gene expression experiments and EST experiments) and analyze these datasets together (see Section 4.3).
As a result, new insights into gene expression profiles could be gained and new candidates for further analyses could be found. The following sections describe the backend and frontend implementations of the four designed steps. As MediPlEx is implemented as an extension of SAMS (see Section 3.1) it is basically implemented in Perl. The SAMS web interface is using HTML and is created using Perl scrips and modules. For the integration of the MediPlEx functionality into SAMS, a new Perl module is created storing the described functions.

[5] http://lucene.apache.org/java/docs/index.html



On this page you can select which fields (also called attributes) you want to be included in your export. Use the buttons next to the element to (de)select all options or to invert your currently active fields. It is recommended to tick the "Experiment, Factor/Value, Probe name" in order to have a reference to the experimental setup and the probe IDs from the different microarray/GeneChip tools. For the export of expression data themselves the fields p-Values, M-Values, A-Value must be selected. Select "Annotation Name, Gene Product, Annotation Description, EC Number, KOG" to include probe annotations. After choosing your fields you may click on "show preview" to get a small window showing what your export will look like. NOTE: The actual quantity of entries exported grows fast depending on the number of attributes chosen below (and will exceed the number of hits displayed below). Use the "Calculate" button to compute the actual number of real hits, then select an output format of your choice from the drop-down menu next to the "Export" button. Click the "Export" button next to it to export your data.

Microarray expression data Clear Selection Select All Invert Selection

☑ Experiment ☑ Factor/Value ☑ Probe name □ Gene ID □ SAMS project □ Statistic
☑ p-Value ☑ Adj. p-Value ☑ M-value □ A-value ☑ No. of Replicate spots

Gene annotation Clear Selection Select All Invert Selection

☐ Source ☑ Database □ Gene ID ☑ Gene Name ☑ Region Type □ Region Contig
□ Region Start □ Region Stop □ Region Length □ Region Strand □ Region Frame □ FunctionStatus
□ Region Status □ Isoelectric Point □ Molecular Weight □ Sequence □ Annotation Name □ Gene Product
☑ Annotation Description □ Annotation Comment □ Annotator ☑ EC Number □ KOG □ Confidence

Gene Ontology and KOG Clear Selection Select All Invert Selection

☑ GO Number ☑ KOG Category

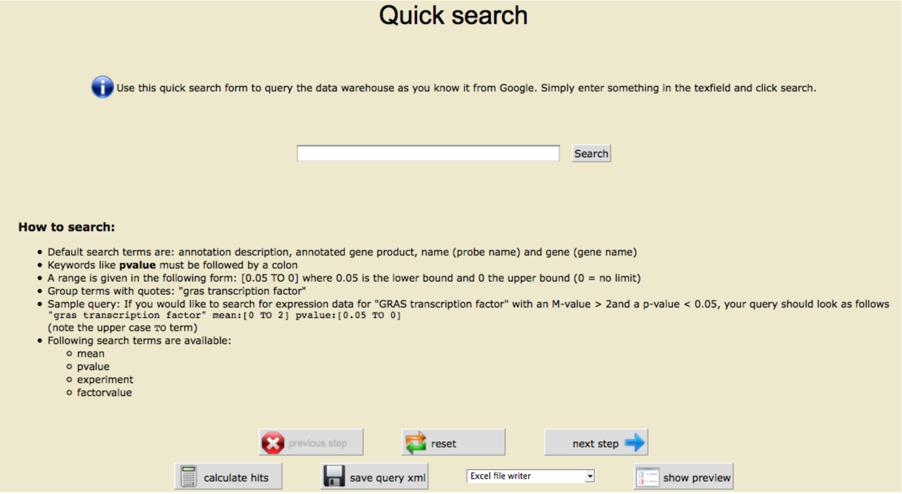
Observation Clear Selection Select All Invert Selection

☑ Tool □ Observation Start □ Observation Stop ☑ Observation Descr.

Please note that exporting great amounts of entries (> 20000) may slow down the server - the export process may take a long time, be patient

Figure 5.10: The screenshot show the export pane of TRUNCATULIX. The user can select each attribute stored in the database to be exported for the resulting genes. A preview option allows a sneak peak into the data. Various file formats are available to download the datasets.

.



Quick search

Use this quick search form to query the data warehouse as you know it from Google. Simply enter something in the texfield and click search.

How to search:

- Default search terms are: annotation description, annotated gene product, name (probe name) and gene (gene name)
- Keywords like **pvalue** must be followed by a colon
- A range is given in the following form: [0.05 TO 0] where 0.05 is the lower bound and 0 the upper bound (0 = no limit)
- Group terms with quotes: "gras transcription factor"
- Sample query: If you would like to search for expression data for "GRAS transcription factor" with an M-value > 2and a p-value < 0.05, your query should look as follows `"gras transcription factor" mean:[0 TO 2] pvalue:[0.05 TO 0]` (note the upper case TO term)
- Following search terms are available:
  - mean
  - pvalue
  - experiment
  - factorvalue

Figure 5.11: The quick search interface of TRUNCATULIX. The textbox offers a google-like search functionality, allowing fulltext search in the attributes ANNOTATION_DESCRIPTION, ANNOTATED_GENE_PRODUCT, probe NAME, and GENE_NAME. Other attributes can be queried using the notation "attribute:[min_value TO max_value]". The user will be forwarded to the export pane to select the attributes to be exported afterwards.

5.3.1 Gene selection

In the first step, genes that should be used for a combined analysis are selected. Therefore, the user selects under which conditions the genes should be expressed. This is done using the EST library and assembly information. The API of SAMS contains a module called "SteN" (STatistical Electronic Northern blot), which is able to filter genes/TCs according to the assembly of ESTs from different EST libraries[Küster *et al.* (2007)]. The module allows searching for genes expressed in different EST libraries, as well as a statistical evaluation (logarithmic likelihood ratio, see Section 2.1.4).

The search function allows to set one of three different states for each of the EST libraries of the assembly to filter the complete TC set for genes expressed under the selected EST library conditions. The three states are defined as followed:

- MUST contain ESTs:
  denotes that the TCs have to consist of at least one EST from this EST library.

- MUST NOT contain ESTs:
  means that the TCs are not allowed to contain one EST of this library.

- MAY contain ESTs:
  indicates that it does not matter if the TCs have ESTs assembled from this library or not. (The library will be ignored in the filter process.)

The states and libraries are connected via an ”OR” statement, so that genes are found that are expressed in only one of the libraries selected as ”MUST contain ESTs”. Thus, the query can be read as ”MUST contain ESTs from at least one of these libraries”.
All TCs of the assembly are scanned for their composition and the ones that do not match the query are sorted out. The logarithmic likelihood ratio (c.f. Section 2.1.4) is calculated on the basis of the selection of EST libraries, serving as an expression value for the TCs. The function to calculate the logarithmic likelihood ratio is implemented in the SteN module in SAMS. The calculation of the values is performed on a compute cluster. The results are used for further analysis.

Preselections of the EST libraries were adapted from the DFCI, more and finer adjusted preselection were created in collaboration with Helge Küster. The preselections are implemented as buttons with Javascript actions in the web interface and are documented in the Appendix (A.1).
The frontend for the selection of the EST libraries is shown in Figure 5.12.

5.3.2 Selection of microarray expression datasets

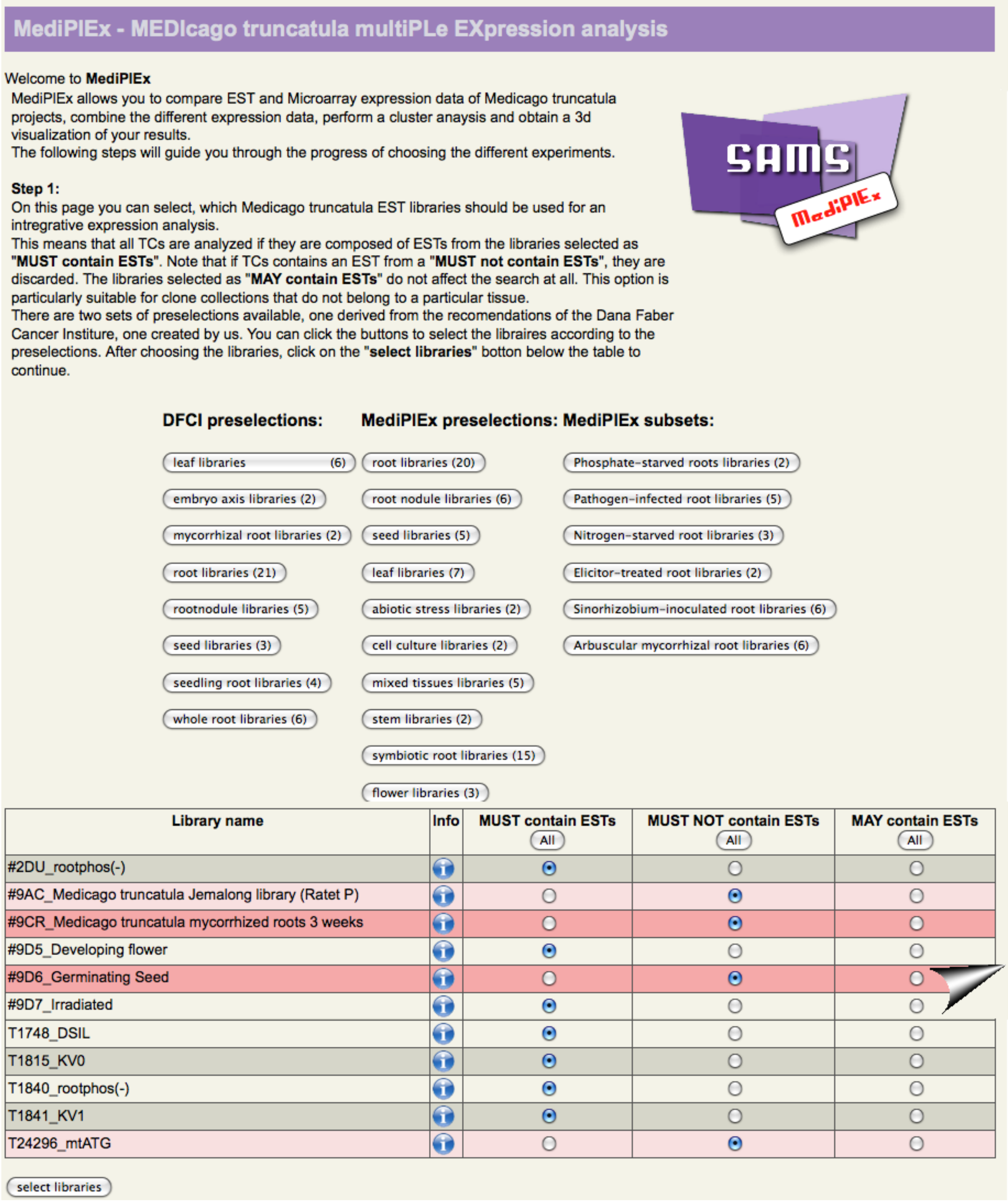
The second step in the expression analysis is to select microarray gene expression datasets. Therefore, a list of all microarray expression experiments stored in the TRUNCATULIX data warehouse is presented and the diverse experiments can be added to the expression analysis (see Figure 5.13). The expression values of the previously found genes are fetched from the TRUNCATULIX data warehouse via Perl API and combined with the calculated logarithmic likelihood ratio. A reimplemented TRUNCATULIX Perl API is used for the data retrieval. An overview of the implemented functions is documented in the Appendix A.2. Reimplementation was needed, because the advantages of a flexible API (BioMart Perl API) stood in contrast to a fast data retrieval.

5.3.3 Clustering of expression datasets

The expression datasets are subsequently transferred into the R software environment using the RSPerl package[6]. The complete datasets are clustered hierarchical using the hclust function and Ward's clustering algorithm.
Ward's clustering algorithm is chosen, because it tries to minimize the loss of information in each grouping.

The linkage function specifies a measure of the distance between two clusters. Ward's tries to minimize the ”error sum of squares” (ESS) after combining two

[6] http://www.omegahat.org/RSPerl/



MediPlEx - MEDIcago truncatula multiPLe EXpression analysis

Welcome to **MediPlEx**

MediPlEx allows you to compare EST and Microarray expression data of Medicago truncatula projects, combine the different expression data, perform a cluster anaysis and obtain a 3d visualization of your results.
The following steps will guide you through the progress of choosing the different experiments.

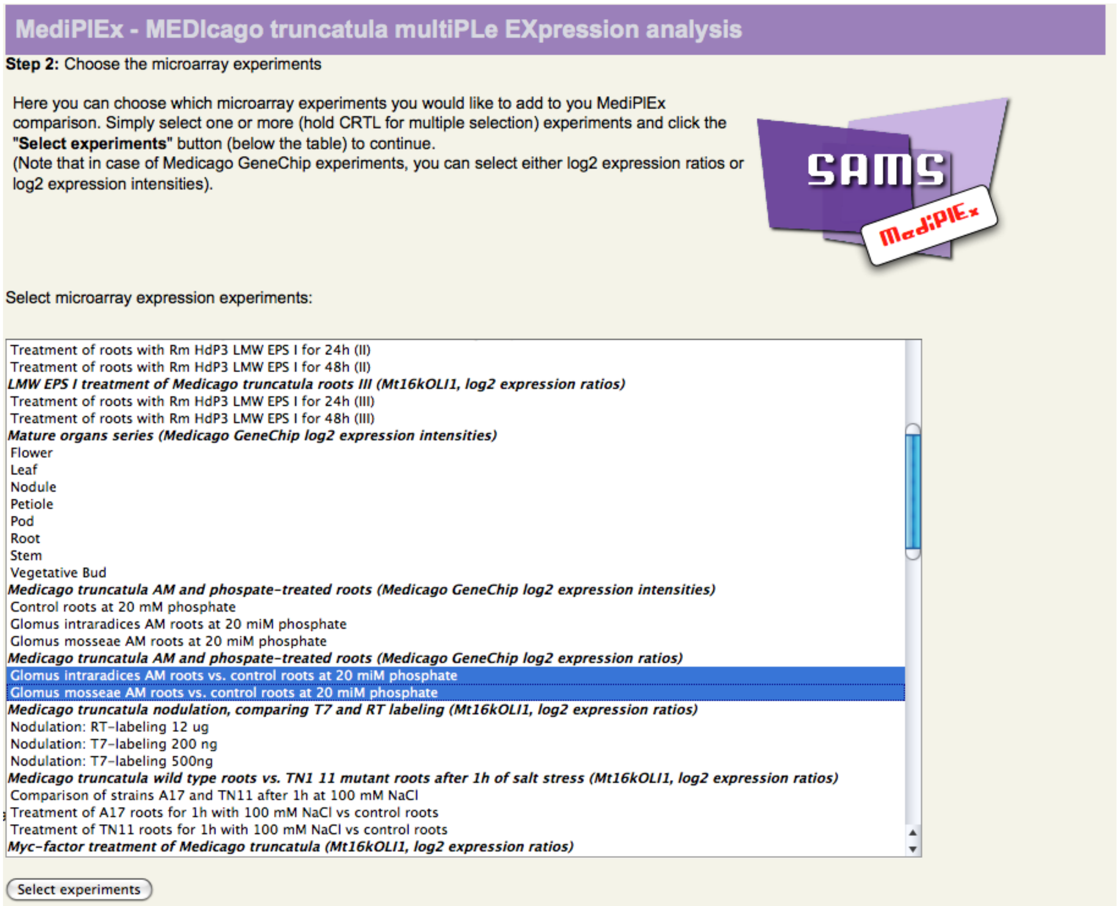
Step 1:
On this page you can select, which Medicago truncatula EST libraries should be used for an intregrative expression analysis.
This means that all TCs are analyzed if they are composed of ESTs from the libraries selected as **"MUST contain ESTs"**. Note that if TCs contains an EST from a **"MUST not contain ESTs"**, they are discarded. The libraries selected as **"MAY contain ESTs"** do not affect the search at all. This option is particularly suitable for clone collections that do not belong to a particular tissue.
There are two sets of preselections available, one derived from the recomendations of the Dana Faber Cancer Institure, one created by us. You can click the buttons to select the libraires according to the preselections. After choosing the libraries, click on the **"select libraries"** botton below the table to continue.

DFCI preselections:	**MediPlEx preselections:**	**MediPlEx subsets:**
leaf libraries (6)	root libraries (20)	Phosphate-starved roots libraries (2)
embryo axis libraries (2)	root nodule libraries (6)	Pathogen-infected root libraries (5)
mycorrhizal root libraries (2)	seed libraries (5)	Nitrogen-starved root libraries (3)
root libraries (21)	leaf libraries (7)	Elicitor-treated root libraries (2)
rootnodule libraries (5)	abiotic stress libraries (2)	Sinorhizobium-inoculated root libraries (6)
seed libraries (3)	cell culture libraries (2)	Arbuscular mycorrhizal root libraries (6)
seedling root libraries (4)	mixed tissues libraries (5)	
whole root libraries (6)	stem libraries (2)	
	symbiotic root libraries (15)	
	flower libraries (3)	

Library name	Info	MUST contain ESTs All	MUST NOT contain ESTs All	MAY contain ESTs All
#2DU_rootphos(-)	ⓘ	●	○	○
#9AC_Medicago truncatula Jemalong library (Ratet P)	ⓘ	○	●	○
#9CR_Medicago truncatula mycorrhized roots 3 weeks	ⓘ	○	●	○
#9D5_Developing flower	ⓘ	●	○	○
#9D6_Germinating Seed	ⓘ	○	●	○
#9D7_Irradiated	ⓘ	●	○	○
T1748_DSIL	ⓘ	●	○	○
T1815_KV0	ⓘ	●	○	○
T1840_rootphos(-)	ⓘ	●	○	○
T1841_KV1	ⓘ	●	○	○
T24296_mtATG	ⓘ	○	●	○

select libraries

Figure 5.12: The EST library selection and preselections in the MediPlEx web interface. Some preselections are adopted from the DFCI website, some are created in collaboration with Helge Küster. The library selection serves as filter for the genes. Only genes expressed under the special library conditions will be used in the expression analysis. The libraries can be selected manual in the table below the preselections. For each library, one other three states "MUST contain ESTs", "MUST NOT contain ESTs", or "MAY contain ESTs" has to be selected. The screenshot has been trimmed to fit the pagesize.



MediPlEx - MEDIcago truncatula multiPLe EXpression analysis

Step 2: Choose the microarray experiments

Here you can choose which microarray experiments you would like to add to you MediPlEx comparison. Simply select one or more (hold CRTL for multiple selection) experiments and click the **"Select experiments"** button (below the table) to continue.
(Note that in case of Medicago GeneChip experiments, you can select either log2 expression ratios or log2 expression intensities).

Select microarray expression experiments:

Treatment of roots with Rm HdP3 LMW EPS I for 24h (II)
Treatment of roots with Rm HdP3 LMW EPS I for 48h (II)
LMW EPS I treatment of Medicago truncatula roots III (Mt16kOLI1, log2 expression ratios)
Treatment of roots with Rm HdP3 LMW EPS I for 24h (III)
Treatment of roots with Rm HdP3 LMW EPS I for 48h (III)
Mature organs series (Medicago GeneChip log2 expression intensities)
Flower
Leaf
Nodule
Petiole
Pod
Root
Stem
Vegetative Bud
Medicago truncatula AM and phospate-treated roots (Medicago GeneChip log2 expression intensities)
Control roots at 20 mM phosphate
Glomus intraradices AM roots at 20 miM phosphate
Glomus mosseae AM roots at 20 miM phosphate
Medicago truncatula AM and phospate-treated roots (Medicago GeneChip log2 expression ratios)
Glomus intraradices AM roots vs. control roots at 20 miM phosphate
Glomus mosseae AM roots vs. control roots at 20 miM phosphate
Medicago truncatula nodulation, comparing T7 and RT labeling (Mt16kOLI1, log2 expression ratios)
Nodulation: RT-labeling 12 ug
Nodulation: T7-labeling 200 ng
Nodulation: T7-labeling 500ng
Medicago truncatula wild type roots vs. TN1 11 mutant roots after 1h of salt stress (Mt16kOLI1, log2 expression ratios)
Comparison of strains A17 and TN11 after 1h at 100 mM NaCl
Treatment of A17 roots for 1h with 100 mM NaCl vs control roots
Treatment of TN11 roots for 1h with 100 mM NaCl vs control roots
Myc-factor treatment of Medicago truncatula (Mt16kOLI1, log2 expression ratios)

Figure 5.13: The screenshot of MediPlEx show the microarray selection form. The microarray gene expression experiments stored in TRUNCATLIX are listed and can be selected for a combined analysis.

clusters into one single cluster. Each step in the clustering process tries to minimize the ESS increase.

The ESS of a set X of N_X values is the sum of squares of the deviations from the *mean* value or the *mean vector*. For a set X the ESS is described by the following expression:

$$ESS(X) = \sum_{i=1}^{N_x} |x_i - \frac{1}{N_X} \sum_{j=1}^{N_X} x_j|^2 \tag{5.1}$$

where $|\cdot|$ is the absolute value of a scalar value or the norm of a vector.

The linkage function is mathematically described as the distance between the clusters X and Y by the expression

$$D(X, Y) = ESS(XY) - [ESS(X) + ESS(Y)] \tag{5.2}$$

where XY is the combined cluster resulting from the fusion clusters X and Y; $ESS(\cdot)$ is the error sum of squares describes above.

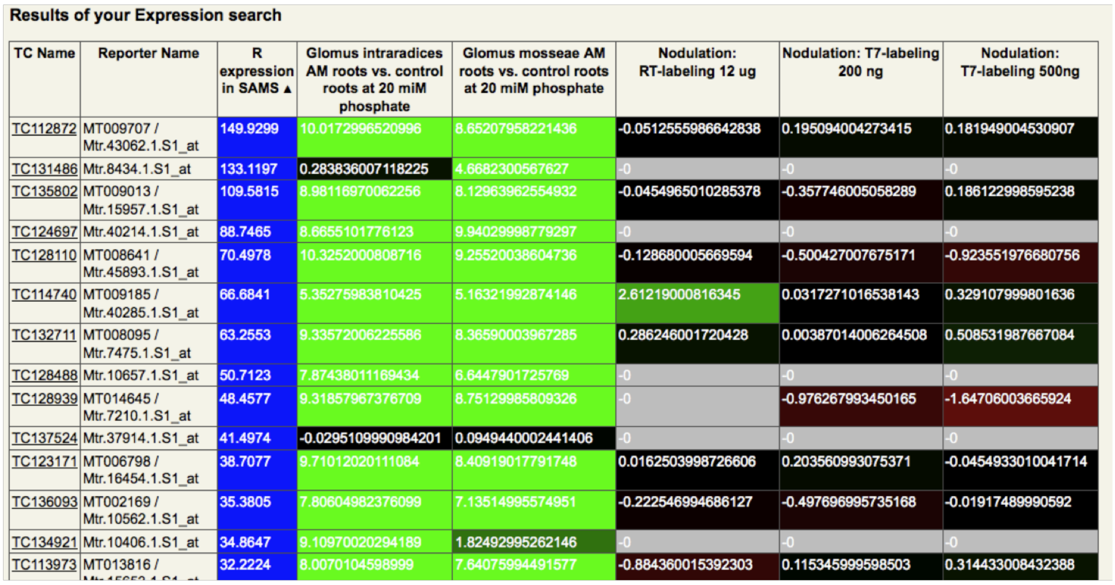
5.3.4 Visualization of results

As a result of the expression analysis, the genes that were found to be expressed under the selected EST library conditions are listed in a table (Figure 5.14). The gene names in the table store links to the genes in SAMS, so that the complete sequence and diverse tool results can be inspected. The table also lists the reporters spotted on the two microarrays Mt16kOliPlus (conventional oligonucleotide microarray for *Medicago truncatula*) and the Medicago GeneChip® that correlate to the found genes. The calculated logarithmic likelihood ratio is presented, as well as the microarray gene expression datasets. The complete table can be exported, adding the annotations of the genes to the exported csv file.

After the clustering progress a cluster dendrogram is created and presented in the web interface (see Figure 5.15).

A dropdown menu allows the user to select how many clusters are to be created.

A three dimensional visualization is implemented as a JAVA webstart application (see Figure 5.16). The application provides the user a better impression of the gene expression profiles. The assignment of the axis for the 3D visualization can be selected from the expression experiments. The clusters are stored as an XML file which is handed over to the 3D viewer. The 3D viewer reads the XML file (using the JAXP class (Java API for XML Processing)) containing the names for the axis of the coordinate system, the gene names, the expression values, the annotations, and the hyperlinks to the respective genes in SAMS. The 3D viewer is an interactive application which allows zooming and rotating the coordinate system using Java3D libraries. For each cluster a color is dedicated so that all genes can be represented in the 3D coordinate system at a time. The clusters can be selected and deselected, showing and hiding the genes of the clusters. Each gene is clickable, displaying the gene name, the annotation, the expression values, and a hyperlink to SAMS



Results of your Expression search

TC Name	Reporter Name	R expression in SAMS ▲	Glomus intraradices AM roots vs. control roots at 20 miM phosphate	Glomus mosseae AM roots vs. control roots at 20 miM phosphate	Nodulation: RT-labeling 12 ug	Nodulation: T7-labeling 200 ng	Nodulation: T7-labeling 500ng
TC112872	MT009707 / Mtr.43062.1.S1_at	149.9299	10.0172996520996	8.65207958221436	-0.0512555986642838	0.195094004273415	0.181949004530907
TC131486	Mtr.8434.1.S1_at	133.1197	0.283836007118225	4.6682300567627	-0	-0	-0
TC135802	MT009013 / Mtr.15957.1.S1_at	109.5815	8.98116970062256	8.12963962554932	-0.0454965010285378	-0.357746005058289	0.186122998595238
TC124697	Mtr.40214.1.S1_at	88.7465	8.6655101776123	9.94029998779297	-0	-0	-0
TC128110	MT008641 / Mtr.45893.1.S1_at	70.4978	10.3252000808716	9.25520038604736	-0.128680005669594	-0.500427007675171	-0.923551976680756
TC114740	MT009185 / Mtr.40285.1.S1_at	66.6841	5.35275983810425	5.16321992874146	2.61219000816345	0.0317271016538143	0.329107999801636
TC132711	MT008095 / Mtr.7475.1.S1_at	63.2553	9.33572006225586	8.36590003967285	0.286246001720428	0.00387014006264508	0.508531987667084
TC128488	Mtr.10657.1.S1_at	50.7123	7.87438011169434	6.6447901725769	-0	-0	-0
TC128939	MT014645 / Mtr.7210.1.S1_at	48.4577	9.31857967376709	8.75129985809326	-0	-0.976267993450165	-1.64706003665924
TC137524	Mtr.37914.1.S1_at	41.4974	-0.0295109990984201	0.0949440002441406	-0	-0	-0
TC123171	MT006798 / Mtr.16454.1.S1_at	38.7077	9.71012020111084	8.40919017791748	0.0162503998726606	0.203560993075371	-0.0454933010041714
TC136093	MT002169 / Mtr.10562.1.S1_at	35.3805	7.80604982376099	7.13514995574951	-0.222546994686127	-0.497696995735168	-0.01917489990592
TC134921	Mtr.10406.1.S1_at	34.8647	9.10970020294189	1.82492995262146	-0	-0	-0
TC113973	MT013816 /	32.2224	8.0070104598999	7.64075994491577	-0.884360015392303	0.115345999598503	0.314433008432388

Figure 5.14: The results of the combined gene expression analysis listed in a table. The genes found to be expressed in the selected libraries are listed by name, the associated reporters spotted on the two different microarray layouts are listed alongside. The expression values from the different selected microarray experiments are fetched from the TRUNCATULIX data warehouse and labeled according to their expression intensities.

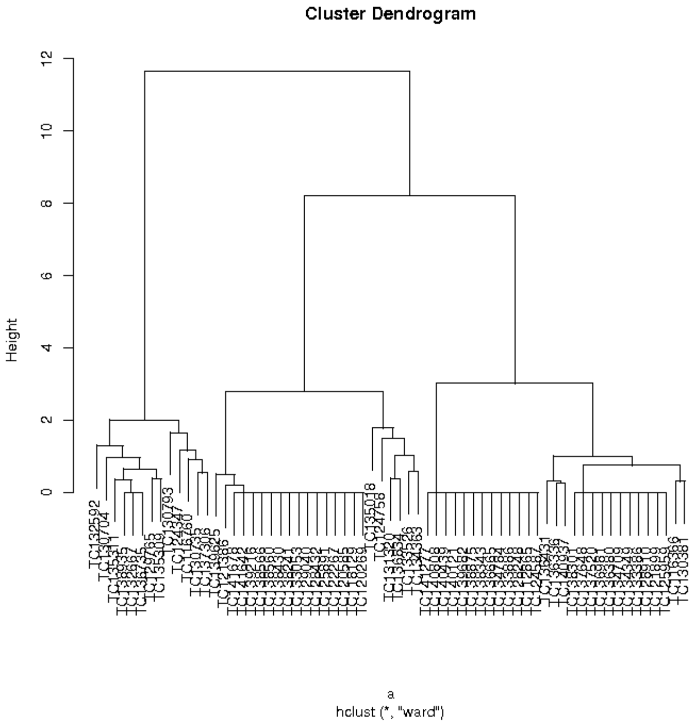
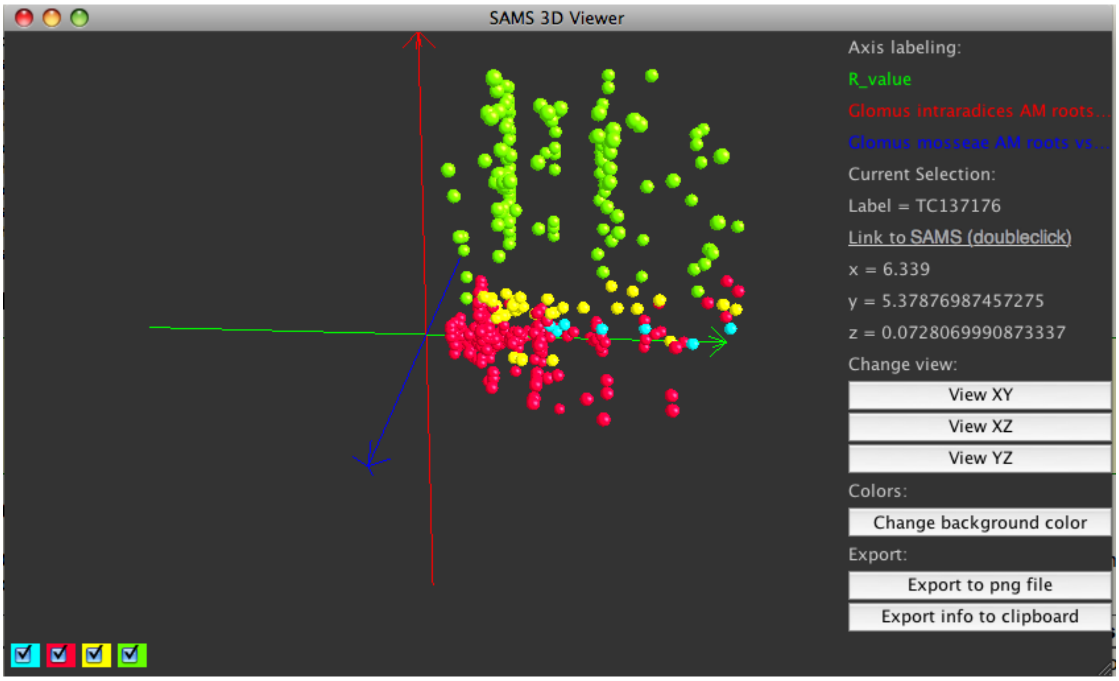


Figure 5.15: The cluster dendrogram created in the hierarchical clustering of the gene expression datasets. The x-axis lists the genes, the clustertree illustrates the similarities of the expression profiles of the genes.



Axis labeling:
R_value
Glomus intraradices AM roots...
Glomus mosseae AM roots vs...

Current Selection:
Label = TC137176
Link to SAMS (doubleclick)
x = 6.339
y = 5.37876987457275
z = 0.0728069990873337

Change view:

Colors:

Export:

Figure 5.16: The screenshot shows the interactive 3D visualization used to demonstrate the clustering of the genes. The spheres are colored according to the appropriate clusters. The clusters can be activated and deactivated on the bottom left. The information about the genes is displayed on the right when clicking a gene. A link to SAMS provides access to the gene sequence, observations and annotations. A screenshot of the clustering can be saved as png file.

on the right part of the application. The application is started by clicking on the button "Cut the clustertree and show the results in 3D". The created clusters can be exported as csv files storing the expression values, as well as the annotations of the genes.

CHAPTER 6

Results

This Chapter focuses on the results of the thesis. These are divided into the implemented tools and the results in using the tools on *Medicago truncatula* datasets.
In the first section, the extension of the EMMA2 software is pointed out. Afterwards, the EMMA2 projects using the newly implemented Affymetrix GeneChip® functions are described. The main focus of the section lies on the project concerning *Medicago truncatula* and the Medicago GeneChip®, as it is the largest and most active EMMA2 GeneChip® project. Additionally, the ATH1 GeneChip® layout was imported and used for an *Arabidopsis thaliana* gene expression analysis project, which is also presented.
In the second section, the newly implemented data warehouse TRUNCATULIX, storing gene sequences, annotations, and expression datasets from *Medicago truncatula* is outlined. Examples for the biological application of TRUNCATULIX are given.
The last section concentrates on the main part of the thesis, MediPlEx (MEDIcago truncatula multiPLe EXpression tool). MediPlEx offers an integrative analysis of EST expression datasets and microarray expression datasets (conventional oligonucleotide microarrays and Affymetrix GeneChips®). Combining these datasets for a hierarchical clustering and visualizing the results in an interactive 3D interface, MediPlEx offers new methods in gene expression analysis. The biological results of the analysis using MediPlEx are represented afterwards.

Figure 6.1 shows a scheme how the implemented tools interact.

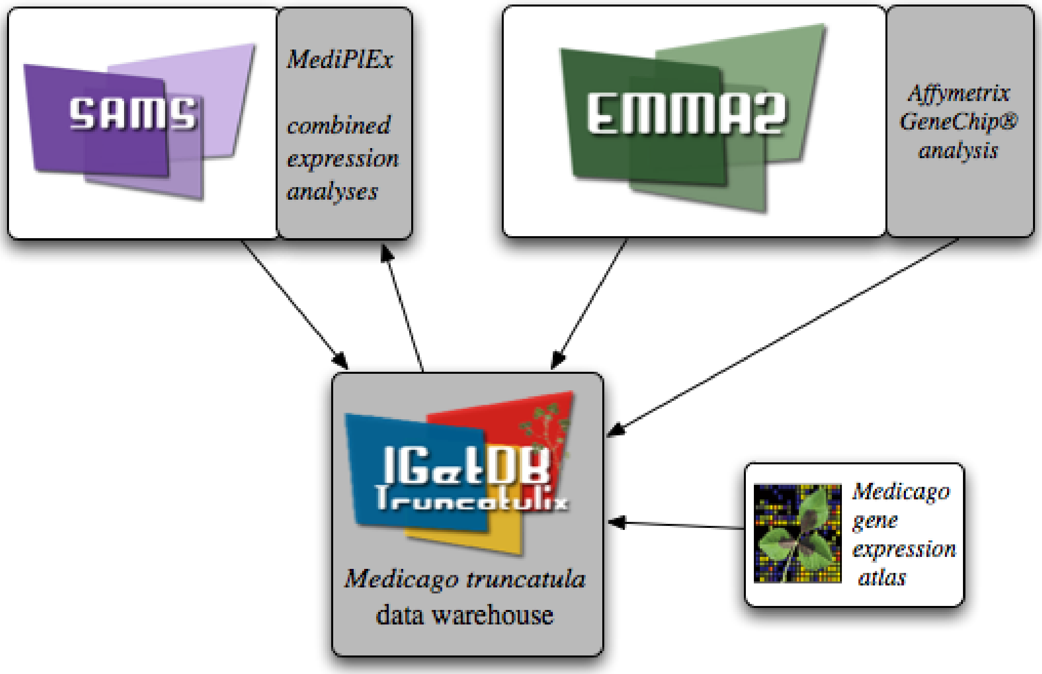

SAMS
MediPlEx
combined expression analyses
EMMA2
Affymetrix GeneChip® analysis
IGetDB Truncatulix
Medicago truncatula data warehouse
Medicago gene expression atlas

Figure 6.1: The scheme shows the implementations and extensions of the applications, as well as their interaction. EMMA2 was extended to analyze Affymetrix GeneChips®. Microarray datasets of GeneChip® and classical oligonucleotide microarrays were exported to the TRUNCATULIX data warehouse. Sequence and annotation datasets from SAMS were imported into the data warehouse. GeneChip® expression datasets stored in the Medicago gene expression atlas were also imported to the data warehouse. The combined expression analysis, as a part of SAMS, accesses the TRUNCATULIX data warehouse for a fast data retrieval.

6.1 Affymetrix GeneChip® analysis using EMMA2

During this thesis, the microarray analysis software EMMA2 was extended and adapted to load Affymetrix GeneChip® microarrays layouts and the corresponding GeneChip® expression datasets. The data can be normalized and analyzed using newly implemented normalization functions and adapted analysis functions in the pipeline based EMMA2 architecture. The web interface has been adapted to fit the needs of a GeneChip® microarray layout and experiment design.

Currently, there are two EMMA2 projects using Affymetrix GeneChip® technology, analyzing GeneChip® datasets of *Medicago truncatula* and *Arabidopsis thaliana*.

The *Medicago truncatula* project (**EMMA_Truncatulix**) is the largest EMMA2 GeneChip® project. There are 132 GeneChip® microarrays imported, 266 transformed datasets, five different created experiments and six cluster analyses. The five experiments cover:

- ***Medicago truncatula* response to supernatants from germinating Glomus spores.**
  This experiment studies gene expression in *Medicago truncatula* A17 wild type plants and in DMI3 mutants in response to supernatants from germinating *Glomus intraradices* spores. 24 GeneChips® were used for the analysis.

- **Gene expression in early stages of *Medicago truncatula* arbuscular mycorrhiza.**
  This experiment studies gene expression in *Medicago truncatula* J5 wild type as well as DMI1, DMI2, and DMI3 mutant roots after 5 dpi and 7 dpi with *Glomus intraradices*, respectively. As a control, non-inoculated roots of the same age were used. 24 GeneChips® were used for the analysis.

- ***Medicago truncatula* A17 and DMI3 responses to Glomus spores and supernatants.**
  This experiment studies gene expression in *Medicago truncatula* A17 wild type plants and in DMI3 mutants in response to supernatants from germinating *Glomus intraradices* spores and to contact with *Glomus intraradices* spores. 36 GeneChips® were used for the analysis.

- **Gene expression during infection of *Medicago truncatula* by AM fungi.**
  This experiment compares gene expression in early stages of arbuscular mycorrhiza. Gene expression in stage 1 infection areas containing appressoria

and PPAs is compared to control areas containing no appressoria and PPAs in A17 wild type plants at 5 days post inoculation with *Gigaspora margarita.* In a parallel approach, gene expression in stage 1 areas containing appressoria is compared to control areas containing no appressoria in DMI3 mutants at 5 days post inoculation with *Gigaspora margarita.* In DMI3 mutants, no PPAs are formed. The samples used were dissected manually, using GFP-labeled ER structures to identify PPA/appressoria regions. 12 GeneChips® were used for the analysis.

- ***Medicago truncatula* AM and phosphate-treated roots.**
  This experiment studies gene expression in *Medicago truncatula* roots colonized with the AM fungi *Glomus mosseae* and *Glomus intraradices* at 28 dpi. The plants were watered with 1/2 strength Hoaglands solution containing 20 μM phosphate. As a control, roots of comparable age watered with 1/2 strength Hoaglands solution containing 20 μM phosphate and 2 mM phosphate were used, to relate AM-specific and phosphate-induced gene expression. All expression profiles are based on whole roots preselected for high inductions of the phosphate transporter MtPT4 in the AM samples and the absence of nodulin genes in all samples. 12 GeneChips® were used for the analysis.

The experiments showed that in the set of differentially expressed genes, only a few transcription factors were modulated in the different mutants. Several genes implicated in primary metabolism, membrane transport or plant metabolism share a specific expression profile in wild-type and one mutant (Mtdmi1) and should be subject of a more detailed analysis. Another conspicuity is the activation of genes involved in cell wall synthesis or response to pathogens in two particular mutants (Mtdmi2/Mtsym2 and Mtdmi3/Mtsym13). The results underlined the complexity of gene expression in mycorrhizal roots colonized by AM fungi. The complete results of the analyses are published by Seddas *et al.* (2009).

The second EMMA2 project that uses the Affymetrix GeneChip® extension is the project **EMMA_Arabidopsis**, analyzing *Arabidopsis thaliana* gene expression. For the AtGenExpression Atlas, various international research labs study the gene expression of *Arabidopsis thaliana* using Affymetrix GeneChips®.
For the expression analysis of the developing flowers, over 120 GeneChips® were hybridized, two experiments were created using EMMA2 :

- **ATH1_ME00319_076-084+WT**
  The aim of this experiment was to compare different developmental silique and seed stages of the *Arabidopsis thaliana* wildtype Columbia-0. Therefore 24 GeneChips® were used in the analysis that is described

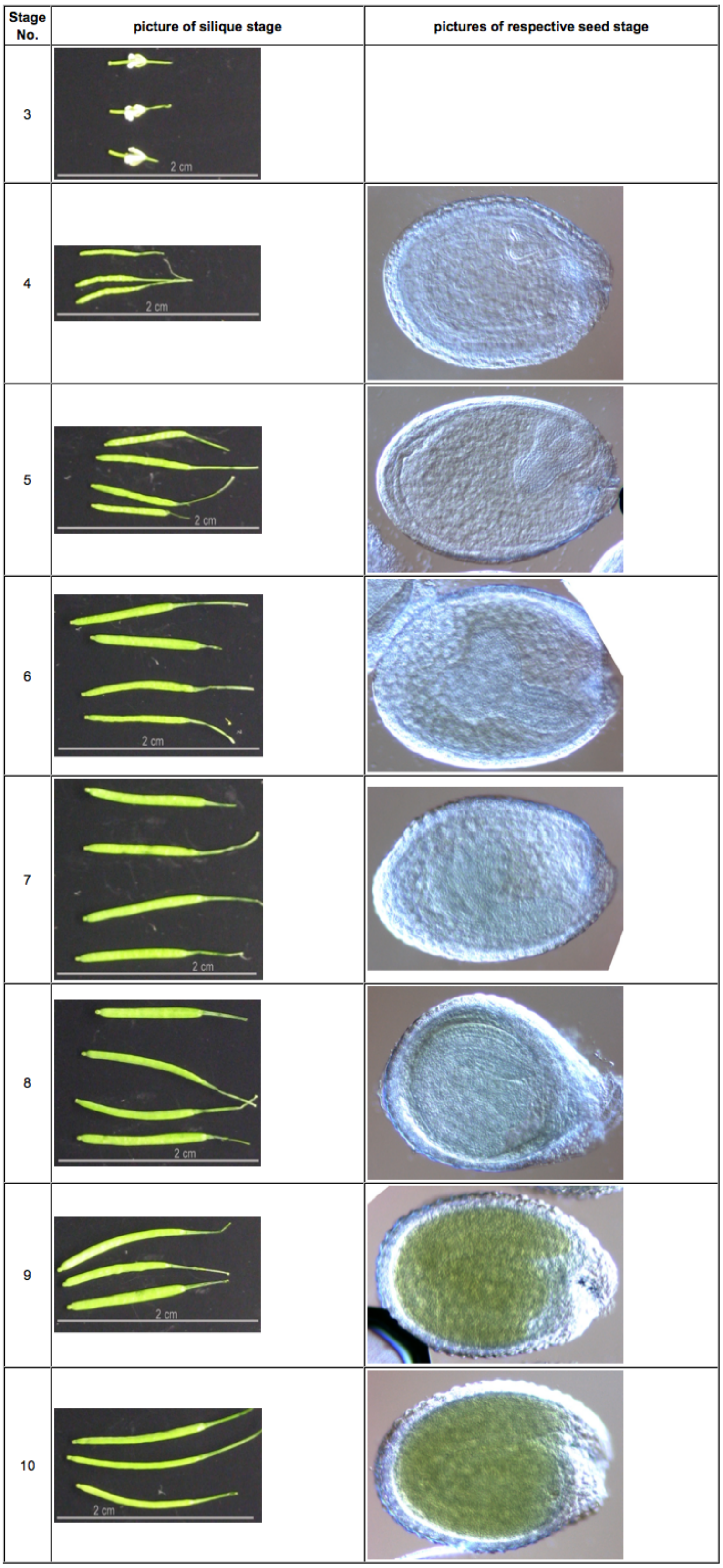


Figure 6.2: The picture shows the different silique and seed stages of *Arabidopsis thaliana* when extracting the mRNA for the gene expression hybridizations.

at *http://www.genomforschung.uni-bielefeld.de/GF-research/AtGenExpress-SeedsSiliques.html*. Figure 6.2 shows the different development stages that were used in the analysis (stages 3-10).

- **ME00319_FwtS**
  The aim of this experiment is a comparison of different parts of the flower, as well as developmental flower stages of *Arabidopsis thaliana* wildtype Columbia-0 and mutants. 25 GeneChips® were used for the analysis.

A corresponding publication is currently in preparation (Kleindt).

EMMA2 can be accessed at
https://www.cebitec.uni-bielefeld.de/groups/brf/software/emma/.

6.2 TRUNCATULIX

The TRUNCATULIX data warehouse stores information about gene sequences, annotations, and gene expression experiments for the model legume *Medicago truncatula*. Using the standard search dialogue, the user can search for specific genes and expression datasets in the complete database by filtering for annotations, special gene expression values, GO numbers, or COG clusters. A quick search additionally offers access and quick results to google-like typed queries. The results of the search can be exported in various file formats and be used for further analyses. TRUNCATULIX can freely be accessed at *http://cebitec.uni-bielefeld.de/truncatulix*.

The TRUNCATULIX data warehouse has been successfully used to find candidate genes for symbiotic signal transduction during nodulation in *Medicago truncatula* by Henckel *et al.* (2009). Therefore, a query was set up to find genes involved in symbiotic signal transduction: Starting with the knowledge about GRAS transcription factors, which are suggested to be activated during the nodulation of *Medicago truncatula*[Kaló *et al.* (2005); Smit *et al.* (2005); Limpens and Bisseling (2003)]. Using the different filter options, the data warehouse was scanned for genes that were automatically annotated as "GRAS transcription factor" (222 entries passed this filter criteria). Afterwards, the number of replicates was restricted to a minimum of 3 and the microarray experiments of interest were selected as the following ones: "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*

gene name	annotation
AJ499899	DELLA protein GAI (Gibberellic acid-insensitive mutant protein) (Restoration of growth on ammonia protein 2)
AL386880	DELLA protein GAI1 (VvGAI1) (Gibberellic acid-insensitive mutant protein 1)
AW559499	DELLA protein RGL1 (RGA-like protein 1) (RGA-like protein)
AW685610	DELLA protein GAI1 (VvGAI1) (Gibberellic acid-insensitive mutant protein 1)
TC114268	DELLA protein GAIP-B (GAIP-B) (CmGAIP-B) (Gibberellic acid-insensitive phloem protein B)
TC117900	DELLA protein GAI (Gibberellic acid-insensitive mutant protein) (Restoration of growth on ammonia protein 2)
TC117945	DELLA protein RGL1 (RGA-like protein 1) (RGA-like protein)
TC120850	DELLA protein GAI1 (VvGAI1) (Gibberellic acid-insensitive mutant protein 1)
TC122531	DELLA protein GAI1 (VvGAI1) (Gibberellic acid-insensitive mutant protein 1)
TC127458	DELLA protein RGL1 (RGA-like protein 1) (RGA-like protein)
TC130958	DELLA protein GAIP (GAIP) (CmGAIP) (Gibberellic acid-insensitive phloem protein)

Table 6.1: Results of the TRUNCATULIX search for GRAS transcription factors. Only genes with a detailed annotation and from the SAMS_Medicago_truncatula_DFCI9 project are listed.

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". The filters about GeneOntology numbers, COG categories, and observations were left blank. As a result, 100 entries were found to match the complete query and were exported. The complete list of the found genes can be viewed in the Appendix A.3.1.

The genes the were not only annotated as "GRAS transcription factors", but with a more precise function are listed in Table 6.1. Interestingly, these genes are all coding for RGA and it's near homologue GAI, which are both negative regulators of the gibberellin (GA) signal transduction[Dill and Sun (2001)].

Both are inducting signal transduction and are thus connected to the GRAS transcription.

6.3 MediPlEx

The application MediPlEx is created to combine gene expression datasets of different experiment types and analyze them together. With the use of MediPlEx, it is possible to calculate a logarithmic likelihood ratio for the expression of genes in certain EST libraries and combine this information with different microarray gene expression experiments (oligonucleotide microarrays and Affymetrix GeneChips®). The resulting data is clustered hierarchical and can be examined in an interactive 3D viewer. Export options with the corresponding annotation and expression datasets are implemented, as well as a sortable result-table. MediPlEx is freely accessible at *http://www.cebitec.uni-bielefeld.de/mediplex*.
The software is integrated in the current version of SAMS. MediPlEx has successfully been used by Henckel *et al.* (2010) to find correlations between oligonucleotide microarray experiments and GeneChip® expression analyses . Additionally, new candidate genes were found with the same expression profile for these experiments.

As an example of a comparison of expression based on a selection of different EST-libraries to GeneChip® analyses performed in the same biological background, arbuscular mycorrhiza (AM) symbiosis is focused. To identify AM-specific TCs and thus AM-specific genes, the preselection "Arbuscular mycorrhizal root libraries (6)" is selected, consisting of the following EST selection of libraries:

– MUST contain ESTs (using "OR" as concatenation):

#9CR	Medicago truncatula mycorrhized roots 3 weeks
#ARB	MTGIM
#ARE	MTAMP
#GFS	MHAM2
5520	MtBC
T1682	MHAM

– MUST NOT contain ESTs:

#2DU	rootphos(-)
#9AC	Medicago truncatula Jemalong library (Ratet P)
#9D5	Developing flower
#9D6	Germinating Seed
#9D7	Irradiated
#A8P	KVKC

#A8V	Phoma-infected
#ARC	MTFLOW
#ARD	MTPOSE
#CDE	MTAPHEU
#G7D	Medicago truncatula J5 roots
#G8F	MtSNF
#G8G	MtSC4
#G8H	MtSCF
#G8I	MtSN0
#G8J	MtSTW
#G8K	MtSTA
#G8L	MtSN4
#GFK	Virus-Infected Leaves
#GFL	Aphid-Infected Shoots
#GFM	Methyl Jasmonate-Elicited Root Cell Suspension Culture
#GOU	Medicago truncatula cv. J5 root
#IBH	MTY
#IPF	Subtracted medicago cDNA library specific for UV-B irradiation
#JBS	Medicago truncatula Clontech PCR select cDNA subtraction
#K5Q	MTOROCRE
#KAH	Medicago truncatula Subtractive PCR
#KB9	Medicago truncatula cv. 108-R Salt-stress SSH
#KBM	Medicago truncatula SSH 23 Hours
#KL5	Medicago truncatula SSH 6 Hours
#KOU	Medicago truncatula A17 glandular trichome
#L00	JCVI-MT1
#LLR	JCVI-MT3
1032	MtRHE
2764	Medicago truncatula R108Mt
2847	Medicago truncatula cDNA library
4046	Developing leaf
4047	Nodulated root
4048	Developing root
4049	Developing stem
5413	Drought
5414	Insect herbivory
5415	Phosphate starved leaf
5518	MtBA
5519	MtBB
7263	Elicited cell culture
T10014	MGHG
T10109	GVSN
T10110	DSLC
T10173	HOGA

T10493	GPOD
T10494	GESD
T11031	BNIR
T11127	GLSD
T12494	Leguminosins
T1510	KV2
T1581	DSIR
T1617	GVN
T1707	KV3
T1748	DSIL
T1815	KV0
T1840	rootphos(-)
T1841	KV1
T24296	mtATG

– MAY contain ESTs:

#IP8	NOLLY
T10174	kiloclone
T11958	MTUS
T12308	6KUG

The libraries set to "MAY contain ESTs" were not considered (ignored) since these represent clone libraries used for microarray construction and hence do not contain any information on tissue-specific gene expression.

The microarray expression datasets selected are the experiments "*Medicago truncatula* AM and phosphate-treated roots (Medicago GeneChip log2 expression ratios)": "*Glomus intraradices* AM roots vs. control roots at 20 miM phosphate" and "*Glomus mosseae* AM roots vs. control roots at 20 miM phosphate". Following the TC search, 843 TCs fulfilled the specified conditions of an AM-specific EST composition, and 829 of these were represented by reporters on the Affymetrix Medicago GeneChip®. The Top20 genes are listed in Table 6.2, the complete list can be looked at in the Appendix A.3.2. Sorting these TCs for the calculated logarithmic likelihood ratio identifies a range of AM-specific genes[Hohnjec *et al.* (2005, 2006); Küster *et al.* (2007)], as was suggested by the search. Remarkably, a TC encoding the mycorrhiza-specific phosphate transporter MtPt4, a key marker gene for an efficient AM symbiosis[Javot *et al.* (2007)], was identified as the top candidate. The identification of well-known AM-specific and AM-induced marker genes such as MtBcp1 (TC139394[Hohnjec *et al.* (2005)]), MtGlp1 (TC124054[Doll *et al.* (2003)]), MtGst1 (TC135802[Wulf *et al.* (2003)]), MtLec5 TC113973[Frenzel *et al.* (2005)]), MtMYBCC (TC117163[Küster *et al.* (2007)]), MtScp1 (TC114740[Liu *et al.* (2003)]), MtTi1 (TC123171[Grunwald *et al.* (2004)]) can be regarded as a proof-of-principle for the MediPlEx search strategy.

TC Name	Reporter Name	log likelihood ratio	Glomus intraradices AM roots vs. control roots at 20 miM phosphate	Glomus mosseae AM roots vs. control roots at 20 miM phosphate
TC112872	Mtr.43062.1.S1_at	149.9299	10.0172996520996	8.65207958221436
TC131486	Mtr.8434.1.S1_at	133.1197	0.283836007118225	4.6682300567627
TC135802	Mtr.15957.1.S1_at	109.5815	8.98116970062256	8.12963962554932
TC124697	Mtr.40214.1.S1_at	88.7465	8.6655101776123	9.94029998779297
TC128110	Mtr.45893.1.S1_at	70.4978	10.3252000808716	9.25520038604736
TC114740	Mtr.40285.1.S1_at	66.6841	5.35275983810425	5.16321992874146
TC132711	Mtr.7475.1.S1_at	63.2553	9.33572006225586	8.36590003967285
TC128488	Mtr.10657.1.S1_at	50.7123	7.87438011169434	6.6447901725769
TC128939	Mtr.7210.1.S1_at	48.4577	9.31857967376709	8.75129985809326
TC137524	Mtr.37914.1.S1_at	41.4974	-0.0295109990984201	0.0949440002441406
TC123171	Mtr.16454.1.S1_at	38.7077	9.71012020111084	8.40919017791748
TC136093	Mtr.10562.1.S1_at	35.3805	7.80604982376099	7.13514995574951
TC134921	Mtr.10406.1.S1_at	34.8647	9.10970020294189	1.82492995262146
TC113973	Mtr.15653.1.S1_at	32.2224	8.0070104598999	7.64075994491577
TC132245	Mtr.35424.1.S1_at	31.3864	9.55953979492188	0.705334007740021
TC129609	Mtr.10562.1.S1_at	26.8687	7.80604982376099	7.13514995574951
TC124054	Mtr.12500.1.S1_at	23.6306	7.93924999237061	7.79982995986938
TC130208	Mtr.44070.1.S1_at	22.1866	6.17437982559204	5.09914016723633
TC128493	Mtr.7489.1.S1_at	22.1866	-0.0920900031924248	10.6577997207642
TC126123	Mtr.8304.1.S1_at	21.6265	6.70066976547241	7.09717988967896

Table 6.2: The table lists the 20 genes with the highest logarithmic likelihood ratio calculated according to the query. The name of the spotted reporter on the GeneChip$^{®}$ is listed, as well as the expression values of the experiments.

CHAPTER 7

Summary, discussion and outlook

In the last Chapters three different tools were designed and implemented. The results of the application of these tools was presented in the previous chapter. The next section sums up the main aspects of the thesis. Afterwards a discussion, as well as an outlook to possible future improvements is presented.

7.1 Summary

The results of a gene expression analysis offer researchers the possibility to gain insights into the transcriptome of an organism under certain conditions. Being a sequencing based methods, EST library sequencing is one of the well established transcriptome analysis methods. In contrast to this, oligonucleotide microarrays provide cheaper and faster expression analysis results, nevertheless the sequences to be analyzed already have to be known. Newly developed microarrays, Affymetrix GeneChips®, provide even more spotted reporters and a more robust experimental setup. However, an application for the analysis of both types of microarrays did not exist. In the scope of this thesis, EMMA2, an application for the analysis of classical oligonucleotide microarrays has been extended to load, store, and analyze Affymetrix GeneChip® datasets. The performed analyses are as comparable as possible to classical oligonucleotide microarray analyses. EMMA2 and the Affymetrix GeneChip® extension are published by Dondrup *et al.* (2009a), the analyses performed with Affymetrix GeneChips® using EMMA2 were published by Seddas *et al.* (2009) and are in preparation (Kleindt). EMMA2 can be accessed at *https://www.cebitec.uni-bielefeld.de/groups/brf/software/emma/.*

The *Medicago truncatula* research community suffered from the fact of distributed data storage: Several databases stored different sequence datasets and various microarray expression experiment analyses. A query to find special genes and certain expression datasets needed a lot manual work in the different repositories. The TRUNCATULIX data warehouse host sequence and annotation datasets of five different sources combined with the results of over 20 microarray experiments (> 350 hybridizations). The datasets and query options are available in an easy to use web interface. Additionally, the database of TRUNCATULIX offers the stored datasets for further usage via a Perl and a Java API. TRUNCATULIX has been published (Henckel *et al.* (2009)) and is freely available at *http://www.cebitec.uni-bielefeld.de/truncatulix.*

The main goal of this thesis is the combination and analysis of the results of different gene expression analysis methods. Only one reference tool was found for a combined gene expression analysis, but practical appliance was limited. The combination of microarray gene expression datasets (oligonucleotide and GeneChip®) with the expression values (logarithmic likelihood ratio) of EST libraries only became possible because of these two previously accomplished tasks. Offering a user friendly web interface, MediPlEx is a great benefit for the *Medicago truncatula* research community. The 3D view of the clustered datasets offers a new approach to visualize correlated genes and find differences and agglomerations in expression profiles.
The results of the MediPlEx analyses help researchers in the filed of *Medicago truncatula* to find new candidates and gain new insights correlating gene expression. The tool MediPlEx has recently been submitted to BMC bioinformatics (Henckel *et al.* (2010)) and is freely available at *http://www.cebitec.uni-bielefeld.de/mediplex.*

7.2 Discussion and outlook

The presented work shows different achievements and implementations in the field of gene expression analysis and *Medicago truncatula* research.

The created Affymetrix GeneChip® analysis pipelines in the EMMA2 microarray analysis application offers a new scope for EMMA2. Other recently developed microarray formats could be supported in later EMMA2 releases. Obviously, other microarray analysis applications (3.2) also integrated

Affymetrix GeneChip® support, thus EMMA2 is not the only application for the analysis of GeneChips®.

The created data warehouse is designed to host datasets from the model legume *Medicago truncatula*, combining these to an extensive treasure of biological data. The backend and frontend components of the TRUNCATULIX data warehouse could be used to create a similar data warehouse for other model organisms like *Arabidopsis thaliana* or *Drosophila melanogaster*. The usability of the data warehouse could be increased with the implementation of a blast homology search, allowing to seek for homologue gene sequences in the database on the fly. As the amount of microarray hybridizations and datasets steadily grows, upcoming experimental data should (after a manual review and quality confirmation) also be integrated in the TRUNCATULIX data warehouse for a wider database and even more expression analysis possibilities.

MediPlEx proved to be a useful tool to combine gene expression analyses and to find new candidate genes. This is definitely based on the numerous EST libraries sequenced in the last decade and the huge amount of microarray experiments performed recently. The style of selecting genes according to their expression under different conditions (conditions of their EST library creation) is a big advance in contrast to a manual selection. Genes that are not yet found to be correlated might show up in the results and reveal their similar expression profiles.
The possibility to use the MediPlEx backend to create a similar tool for another organism (e.g. *Arabidopsis thaliana* → AraPlEx), should not induce any problems. Anyhow, to reach this task a data warehouse would be needed to provide a fast and effective data access, as well as some adaption need to be made (e.g. the manual created preselections of the EST libraries). This would imply a cooperation with a well trained expert in the field of (in this example) *Arabidopsis thaliana* research.

APPENDIX A

API documentation

A.1 MediPIEx preselections of the EST libraries

Only the libraries of the categories "MUST contain ESTs" and (if used) "MAY contain ESTs" are listed. The libraries are concatenated using 'OR'.

A.1.1 DFCI preselections

Leaf libraries

– MUST contain ESTs:
  #A8V Phoma-infected
  #IPF Subtracted medicago cDNA library specific for UV-B irradiation
  4046 Developing leaf
  5414 Insect herbivory
  5415 Phosphate starved leaf
  T1748 DSIL

Embryo axis libraries

– MUST contain ESTs:
  #A8V Phoma-infected

#IPF Subtracted medicago cDNA library specific for UV-B irradiation
4046 Developing leaf
5414 Insect herbivory

Mycorrhizal root libraries

- MUST contain ESTs:
  #KBM Medicago truncatula SSH 23 Hours
  #KL5 Medicago truncatula SSH 6 Hours

Root libraries

- MUST contain ESTs:
  #2DU rootphos(-)
  #9CR Medicago truncatula mycorrhized roots 3 weeks
  #CDE MTAPHEU
  #G7D Medicago truncatula J5 roots
  #JBS Medicago truncatula Clontech PCR select cDNA subtraction
  #K5Q MTOROCRE
  #KB9 Medicago truncatula cv. 108-R Salt-stress SSH
  1032 MtRHE
  4047 Nodulated root
  4048 Developing root
  5519 MtBB
  T10014 MGHG
  T10173 HOGA
  T11031 BNIR
  T1510 KV2
  T1581 DSIR
  T1682 MHAM
  T1707 KV3
  T1815 KV0
  T1840 rootphos(-)
  T1841 KV1

Rootnodule libraries

- MUST contain ESTs:
  #IP8 NOLLY
  2764 Medicago truncatula R108Mt

5519 MtBB
T10109 GVSN
T1617 GVN

Seed libraries

- MUST contain ESTs:
  #9D6 Germinating Seed
  T10494 GESD
  T11127 GLSD

Seedling root libraries

- MUST contain ESTs:
  T1510 KV2
  T1707 KV3
  T1815 KV0
  T1841 KV1

Whole root libraries

- MUST contain ESTs:
  #G8F MtSNF
  #G8G MtSC4
  #G8H MtSCF
  #G8I MtSN0
  #G8J MtSTW
  #G8K MtSTA

A.1.2 MediPlEx preselections

Root libraries

- MUST contain ESTs:
  #2DU rootphos(-)
  #CDE MTAPHEU
  #G8F MtSNF
  #G8G MtSC4

#G8H MtSCF
#G8I MtSN0
#GOU Medicago truncatula cv. J5 root
#JBS Medicago truncatula Clontech PCR select cDNA subtraction
#K5Q MTOROCRE
#KB9 Medicago truncatula cv. 108-R Salt-stress SSH
1032 MtRHE
4048 Developing root
5518 MtBA
T10014 MGHG
T10173 HOGA
T11031 BNIR
T1581 DSIR
T1748 DSIL
T1815 KV0
T1840 rootphos(-)

Root nodule libraries

- MUST contain ESTs:
  #G8L MtSN4
  #IP8 NOLLY
  2764 Medicago truncatula R108Mt
  5519 MtBB
  T10109 GVSN
  T1617 GVN

Seed libraries

- MUST contain ESTs:
  #9D6 Germinating Seed
  #ARD MTPOSE
  T10493 GPOD
  T10494 GESD
  T11127 GLSD

Leaf libraries

- MUST contain ESTs:
  #A8V Phoma-infected

#GFK Virus-Infected Leaves
#IPF Subtracted medicago cDNA library specific for UV-B irradiation
#KOU Medicago truncatula A17 glandular trichome
4046 Developing leaf
5414 Insect herbivory
5415 Phosphate starved leaf

Abiotic stress libraries

- MUST contain ESTs:
  #9D7 Irradiated
  #KAH Medicago truncatula Subtractive PCR

Cell culture libraries

- MUST contain ESTs:
  #GFM Methyl Jasmonate-Elicited Root Cell Suspension Culture
  7263 Elicited cell culture

Mixed tissues libraries

- MUST contain ESTs:
  #9AC Medicago truncatula Jemalong library (Ratet P)
  #L00 JCVI-MT1
  #LLR JCVI-MT3
  5413 Drought
  T10110 DSLC

Stem libraries

- MUST contain ESTs:
  #GFL Aphid-Infected Shoots
  4049 Developing stem

Symbiotic root libraries

- MUST contain ESTs:
  #9CR Medicago truncatula mycorrhized roots 3 weeks
  #A8P KVKC
  #ARB MTGIM
  #ARE MTAMP
  #G7D Medicago truncatula J5 roots
  #G8J MtSTW
  #G8K MtSTA
  #GFS MHAM2
  2847 Medicago truncatula cDNA library
  4047 Nodulated root
  5520 MtBC
  T1510 KV2
  T1682 MHAM
  T1707 KV3
  T1841 KV1

Flower libraries

- MUST contain ESTs:
  #9D5 Developing flower
  #ARC MTFLOW
  #IBH MTY

A.1.3 MediPIEx preselection subsets

These subsets of preselections are a refinement of some preselections.

Phosphate-starved roots libraries

- MUST contain ESTs:
  #2DU rootphos(-)
  T1840 rootphos(-)

Pathogen-infected root libraries

- MUST contain ESTs:
  #CDE MTAPHEU

#K5Q MTOROCRE
T11031 BNIR
T1581 DSIR
T1748 DSIL

Nitrogen-starved root libraries

- MUST contain ESTs:
  #G8G MtSC4
  #G8H MtSCF
  #G8I MtSN0

Elicitor-treated root libraries

- MUST contain ESTs:
  T10014 MGHG
  T10173 HOGA

Sinorhizobium-inoculated root libraries

- MUST contain ESTs:
  #G8J MtSTW
  #G8K MtSTA
  4047 Nodulated root
  T10014 MGHG
  T10173 HOGA
  T1510 KV2
  T1707 KV3
  T1841 KV1

Arbuscular mycorrhizal root libraries

- MUST contain ESTs:
  #9CR Medicago truncatula mycorrhized roots 3 weeks
  #ARB MTGIM
  #ARE MTAMP
  #GFS MHAM2
  5520 MtBC

T1682 MHAM

- MAY contain ESTs:
  #IP8 NOLLY
  T10174 kiloclone
  T11958 MTUS
  T12308 6KUG

A.2 TRUNCATULIX Perl API

fetch_experiments_or_factorvaluenames_from_truncatulix ($experiment_name)
The implemented function returns a list of all experiment names. If experimetn names are given, the function returns the corresponding factor_value_names.

search_for_gene_expression ($factor_value_name, $gene_names)
The implemented function fetches all expression values for a given factor_value_name and stores them in a hash. The list of gene_names is used to remove the unneeded expression values from the hash and only to return the searched values. This method performs faster than fetching large amounts of single datasets from the database.

search_for_gene_expression_by_gene ($gene_names)
The implemented function reveives a list of gene_names and fetches all expression values for all available factor_values from the TRUNCATULIX database.

_format_truncatulix_results ($results, $experiment_name,$gene_names)
The implemented function receives the results of one of the two functions to get gene expression values (search_for_gene_expression, search_for_gene_expression_by_gene) and reformats the output to a multi-nested-hash. %hash → experiment_name → gene_name → (a1)mean → value.

_cut_prefix ($experiment_name)
The implemented function removes the prefix from a given experiment_name. The prefix denotes the type of the experiment and is removed for the presentation to the user.

A.3 Results

A.3.1 TRUNCATULIX complete result table

The table lists the complete expression datasets found in the analysis in Chapter 6.2.

Gene Name	Gene Product	Database
MTYA330TF_tmp_266	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_454
AW686667	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI8
BE205257	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI8
BE326037	short-root transcription factor (SHR)	SAMS_Medicago_truncatula_DFCI8
BI262875	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI8
BM813961	DELLA protein GAI	SAMS_Medicago_truncatula_DFCI8
BQ136987	hypothetical protein predicted by Glimmer/Critica	SAMS_Medicago_truncatula_DFCI8
CX527591	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
CX531568	Nodulation-signaling pathway 1 protein.	SAMS_Medicago_truncatula_DFCI8
CX549343	scarecrow-like transcription factor 8 (SCL8)	SAMS_Medicago_truncatula_DFCI8
TC100497	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
TC102472	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
TC102836	scarecrow-like transcription factor 6 (SCL6)	SAMS_Medicago_truncatula_DFCI8
TC104252	scarecrow transcription factor family	SAMS_Medicago_truncatula_DFCI8
TC104740	DELLA protein GAIP-B	SAMS_Medicago_truncatula_DFCI8
TC105118	DELLA protein RGL1	SAMS_Medicago_truncatula_DFCI8
TC105615	DELLA protein RGL1	SAMS_Medicago_truncatula_DFCI8
TC106879	DELLA protein GAI1	SAMS_Medicago_truncatula_DFCI8
TC107253	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
TC108534	DELLA protein GAIP-B	SAMS_Medicago_truncatula_DFCI8
TC109336	DELLA protein GAI	SAMS_Medicago_truncatula_DFCI8
TC109615	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI8
TC109993	DELLA protein GAIP	SAMS_Medicago_truncatula_DFCI8
TC110367	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
TC110418	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI8
TC111546	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
TC112219	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
TC94843	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
TC95744	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI8
TC97928	DELLA protein RGL1	SAMS_Medicago_truncatula_DFCI8
TC98097	Nodulation-signaling pathway 2 protein.	SAMS_Medicago_truncatula_DFCI8
TC98320	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI8
TC98552	short-root transcription factor (SHR)	SAMS_Medicago_truncatula_DFCI8
TC99912	scarecrow transcription factor family	SAMS_Medicago_truncatula_DFCI8
AJ388937	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI9
AJ497361	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI9
AJ499899	DELLA protein GAI	SAMS_Medicago_truncatula_DFCI9
AL374023	contains EST AU094565(E11846)	SAMS_Medicago_truncatula_DFCI9
AL386879	DELLA protein DWARF8 (Protein dwarf-8).	SAMS_Medicago_truncatula_DFCI9
AL386880	DELLA protein GAI1	SAMS_Medicago_truncatula_DFCI9
AL388510	DELLA protein RGA2 (RGA-like protein 2) (BrRGA2).	SAMS_Medicago_truncatula_DFCI9
AW559499	DELLA protein RGL1	SAMS_Medicago_truncatula_DFCI9
AW586344	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI9
AW685610	DELLA protein GAI1	SAMS_Medicago_truncatula_DFCI9
BF518829	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
BG587215	Nodulation-signaling pathway 1 protein.	SAMS_Medicago_truncatula_DFCI9
BI308453	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
BM814126	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
CB894619	scarecrow transcription factor family	SAMS_Medicago_truncatula_DFCI9
CX542143	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
CX550676	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
DY617552	gibberellin response modulator-like protein	SAMS_Medicago_truncatula_DFCI9
TC112834	Nodulation-signaling pathway 1 protein.	SAMS_Medicago_truncatula_DFCI9
TC112920	Nodulation-signaling pathway 2 protein.	SAMS_Medicago_truncatula_DFCI9
TC114268	DELLA protein GAIP-B	SAMS_Medicago_truncatula_DFCI9
TC115452	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC115561	scarecrow transcription factor family protein	SAMS_Medicago_truncatula_DFCI9
TC116221	phytochrome A signal transduction 1 (PAT1)	SAMS_Medicago_truncatula_DFCI9
TC117409	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
TC117900	DELLA protein GAI	SAMS_Medicago_truncatula_DFCI9
TC117945	DELLA protein RGL1	SAMS_Medicago_truncatula_DFCI9
TC119390	Nodulation-signaling pathway 1 protein.	SAMS_Medicago_truncatula_DFCI9
TC120300	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC120726	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC120850	DELLA protein GAI1	SAMS_Medicago_truncatula_DFCI9
TC121570	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC122531	DELLA protein GAI1	SAMS_Medicago_truncatula_DFCI9
TC124034	Nodulation-signaling pathway 2 protein.	SAMS_Medicago_truncatula_DFCI9
TC125937	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC126429	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC127458	DELLA protein RGL1	SAMS_Medicago_truncatula_DFCI9

TC128758	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
TC129785	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC130218	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC130639	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
TC130958	DELLA protein GAIP	SAMS_Medicago_truncatula_DFCI9
TC132070	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
TC134925	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_DFCI9
TC135080	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
TC138569	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_DFCI9
AC121238_43.5	DELLA protein GAIP-B	SAMS_Medicago_truncatula_Genome_2.0
AC137079_38.5	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_Genome_2.0
AC137703_19.5	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_Genome_2.0
AC146554_29.5	Nodulation-signaling pathway 2 protein.	SAMS_Medicago_truncatula_Genome_2.0
AC148484_20.5	DELLA protein DWARF8 (Protein dwarf-8).	SAMS_Medicago_truncatula_Genome_2.0
AC153351_7.5	Protein MONOCULM 1.	SAMS_Medicago_truncatula_Genome_2.0
AC155890_4.5	DELLA protein GAI1	SAMS_Medicago_truncatula_Genome_2.0
AC162278_27.4	DELLA protein GAIP-B	SAMS_Medicago_truncatula_Genome_2.0
AC174290_1.4	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_Genome_2.0
AC174290_18.4	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_Genome_2.0
AC174290_8.4	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_Genome_2.0
AC183753_16.4	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_Genome_2.0
AC183753_21.4	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_Genome_2.0
AC183753_23.4	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_Genome_2.0
AC192072_11.3	Chitin-inducible gibberellin-responsive protein 1.	SAMS_Medicago_truncatula_Genome_2.0
AC202572_10.4	Protein MONOCULM 1.	SAMS_Medicago_truncatula_Genome_2.0
CR538722_9.4	Nodulation-signaling pathway 2 protein.	SAMS_Medicago_truncatula_Genome_2.0
CR955006_12.5	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_Genome_2.0
CT027662_14.5	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_Genome_2.0
CT961058_2.4	Chitin-inducible gibberellin-responsive protein 2.	SAMS_Medicago_truncatula_Genome_2.0

6.2

A.3.2 MediPIEx complete result talbe

The table lists the complete results of the expression analysis performed in Chapter 6.3.

TC/Gene name	Glomus intraradices AM roots vs. control roots at 20 miM phosphate	Glomus mosseae AM roots vs. control roots at 20 miM phosphate	log likelihood ratio
TC124213	0.0355979986488819	0.610783994197845	9.7133
TC126278	0.159541994333267	6.35716009140015	9.7133
TC130998	0.0856510028243065	7.78726005554199	9.7133
TC126387	-0.737824976444244	-0.018996000289917	9.6097
TC131394	-0.438883006572723	-0.380724012851715	9.5210
TC120736	0.277628004550934	0.322939991950989	9.5085
TC124092	-0.141607999801636	-0.169365003705025	9.5085
TC126731	2.33139991760254	1.53034996986389	9.5085
TC127419	4.35923004150391	3.75247001647949	9.5085
TC128013	-0.136836007237434	10.2652997970581	9.5085
TC130355	3.64326000213623	2.8582398891449	9.5085
TC133265	6.41361999511719	0.140417993068695	9.5085
TC136584	-0.276486992835999	0.483936995267868	9.5085
TC137521	4.44788980484009	4.53970003128052	9.5085
TC137525	7.49742984771729	0.0615790002048016	9.5085
TC129116	3.2717399597168	2.18694996833801	9.2637
TC129598	0.84383898973465	-0.37024798989296	9.2637
TC130943	3.31514000892639	2.29018998146057	9.2637
TC131324	5.98916006088257	7.10192012786865	9.2637
TC131851	7.08785009384155	0.594492018222809	9.2637
TC133358	0.84383898973465	-0.37024798989296	9.2637
TC134749	5.15231990814209	0.218233004212379	9.2637
TC136368	7.38171005249023	0.303952008485794	9.2637
TC141179	0.0430580005049706	0.25026598572731	9.2637
TC141361	0.722773015499115	0.0434169992804527	9.2637
TC132081	7.80604982376099	7.13514995574951	9.2314
TC117388	-0.127856001257896	-0.0805480033159256	9.1334
TC124697	8.6655101776123	9.94029998779297	88.7465
TC128357	-0.157627999782562	-0.616146981716156	8.9687
TC135135	-0.355266004800797	7.5527400970459	8.7719
TC137449	-0.0674550011754036	8.18527030944824	8.7719
TC133707	6.7458701133728	5.07664012908936	8.1859
TC136670	0.741454005241394	1.10337996482849	8.0565
TC128110	10.3252000808716	9.25520038604736	70.4978
TC129755	2.46492004394531	0.325302988290787	7.7198
TC130388	9.30231952667236	0.48351201415062	7.7198
TC131422	0.152826994657516	0.121142998337746	7.7198
TC135339	3.69413995742798	-0.0195359997451305	7.7198
TC135597	6.10416984558105	2.7468900680542	7.6032

TC123338	-4.68512010574341	-1.14790999889374	7.3622
TC132158	0.17383199930191	5.93111991882324	7.3622
TC130834	8.6708402633667	7.3244800567627	7.0509
TC117354	0.496277987957001	-0.222003996372223	7.0352
TC114740	5.35275983810425	5.16321992874146	66.6841
TC132711	9.33572006225586	8.36590003967285	63.2553
TC131492	8.79374980926514	7.34459018707275	6.8836
TC135622	-0.0263550002127886	5.83116006851196	6.8836
TC123846	0.285201013088226	0.669620990753174	6.8099
TC113920	-0.0498340018093586	-1.15924000740051	6.7882
TC117057	0.372525006532669	0.356023997068405	6.7882
TC125138	0.02653600089252	-0.0337390005588531	6.7882
TC128669	1.59142994880676	1.88297998905182	6.6569
TC130814	5.3818998336792	5.9020299911499	6.6544
TC114612	5.68821001052856	5.08143997192383	6.6261
TC123049	4.32232999801636	2.67528009414673	6.5622
TC116339	-0.545656025409698	0.0114339999854565	6.4022
TC120317	0.226659998297691	1.08144998550415	6.3474
TC120620	-0.200045004487038	-0.287283003330231	6.3474
TC120730	-0.306701987981796	0.0655459985136986	6.3474
TC120774	0.224738001823425	-0.00211600004695356	6.3474
TC120931	-0.616635978221893	-0.0316170006990433	6.3474
TC122445	0.385935992002487	-0.360747009515762	6.3474
TC134519	-0.0900940001010895	0.140229001641273	6.3474
TC134748	-0.625854015350342	-0.653864026069641	6.3474
TC134790	0.261391997337341	0.193657994270325	6.3474
TC136845	-0.92993301153183	-1.41022002696991	6.3474
TC139555	-0.0483390018343925	-0.120718002319336	6.3474
TC125146	8.96438980102539	0.77061802148819	6.3390
TC125513	-0.0352009981870651	0.00347699993290007	6.3390
TC127195	0.20058299601078	9.53728008270264	6.3390
TC127721	7.62004995346069	5.38644981384277	6.3390
TC127825	0.680931985378265	0.126497000455856	6.3390
TC127993	0.189133003354073	0.356694996356964	6.3390
TC128036	-0.122340999543667	5.91245985031128	6.3390
TC128140	3.10238003730774	2.07936000823975	6.3390
TC128266	0.139388993382454	1.89561998844147	6.3390
TC128629	8.58522033691406	7.63905000686646	6.3390
TC129968	-0.0084009999409318	0.0103669995442033	6.3390
TC130209	-0.0806320011615753	-0.104633003473282	6.3390
TC130452	-0.440066993236542	0.513522982597351	6.3390

TC130479	8.33716011047363	7.81049013137817	6.3390
TC130888	5.82740020751953	4.50173997879028	6.3390
TC132189	-0.810491025447845	-0.646420001983643	6.3390
TC132286	-0.31971201300621	-0.104997001588345	6.3390
TC132382	-1.10250997543335	-0.549826979637146	6.3390
TC132629	9.31441020965576	0.336347997188568	6.3390
TC132849	0.0679420009255409	5.40047979354858	6.3390
TC133273	6.89094018936157	6.09833002090454	6.3390
TC133730	0.12109299749136	2.35842990875244	6.3390
TC133935	0.265684992074966	1.11738002300262	6.3390
TC134268	-0.0876030027866364	-0.599584996700287	6.3390
TC134391	0.16465699672699	-0.136985003948212	6.3390
TC134499	-1.75822997093201	-0.910618007183075	6.3390
TC134566	4.07922983169556	0.0566289983689785	6.3390
TC134607	-0.0835350006818771	-0.0958670005202293	6.3390
TC134856	-0.267302006483078	-0.4339399933815	6.3390
TC134871	0.0752499997615814	8.64890956878662	6.3390
TC134961	-0.147490993142128	6.59903001785278	6.3390
TC135053	5.61820983886719	0.0492889992892742	6.3390
TC135862	7.43605995178223	2.48864006996155	6.3390
TC136063	6.69602012634277	5.2660698890686	6.3390
TC136204	-0.109440997242928	5.5505199432373	6.3390
TC136469	0.915547013282776	3.8585000038147	6.3390
TC136968	5.75323009490967	0.0096469996497035	6.3390
TC137176	5.37876987457275	0.0728069990873337	6.3390
TC137373	0.0361309982836246	6.88813018798828	6.3390
TC137495	8.58522033691406	7.63905000686646	6.3390
TC137868	1.81669998168945	0.543666005134583	6.3390
TC138125	7.23420000076294	5.83658981323242	6.3390
TC138575	-0.053821999579668	6.39849996566772	6.3390
TC138854	7.2407398223877	0.451945006847382	6.3390
TC139012	0.021789999678731	0.159143000841141	6.3390
TC139229	-0.288830995559692	0.0140359997749329	6.3390
TC139433	-0.281780004501343	-0.158706992864609	6.3390
TC139475	0.800897002220154	-0.14020200073719	6.3390
TC140708	0.43954399228096	0.153257995843887	6.3390
TC117751	-0.0993800014257431	-0.373928993940353	6.2583
TC129351	-0.199746996164322	-0.40973499417305	6.1758
TC130323	7.08597993850708	0.710295975208282	6.1758
TC130616	0.768235981464386	0.318423986434937	6.1758
TC130684	6.15528011322021	0.724072992801666	6.1758

TC130828	0.149309992790222	0.114991001784801	6.1758
TC131018	2.14073991775513	2.27550005912781	6.1758
TC131213	6.7312798500061	-0.087067998945713	6.1758
TC131273	1.47541999816895	0.0236740000545979	6.1758
TC132240	0.765021979808807	0.0878940001130104	6.1758
TC132277	-0.0710960030555725	0.127277001738548	6.1758
TC132310	0.0860129967331886	-0.00607499992474914	6.1758
TC132328	5.1009202003479	0.0826620012521744	6.1758
TC132447	8.38994979858398	0.434208989143372	6.1758
TC132926	6.57658004760742	0.506087005138397	6.1758
TC133436	0.05901899933815	-0.0592189989984035	6.1758
TC136070	6.9540901184082	2.99131989479065	6.1758
TC121747	0.0281310006976128	0.315631985664368	6.1348
TC128731	-0.402442008256912	-0.100566998124123	6.1348
TC128687	3.85447001457214	2.6078200340271	6.1265
TC136005	0.685630023479462	0.114459000527859	6.0752
TC128488	7.87438011169434	6.6447901725769	50.7123
TC125366	-1.33430004119873	-0.130215004086494	5.9734
TC133425	8.20551013946533	6.19806003570557	5.9734
TC135530	7.36121988296509	0.709051012992859	5.9734
TC132665	8.92354011535645	6.87591981887817	5.9697
TC123295	0.450740993022919	0.63374799489975	5.9586
TC133133	-0.0276120007038116	-0.351199001073837	5.5701
TC136240	0.223030000925064	-0.336134999990463	5.5701
TC128752	-0.0791879966855049	-0.208544000983238	5.3698
TC128310	0.219380006194115	0.28523001074791	5.2235
TC114256	-0.459800988435745	-0.613906025886536	5.0911
TC115197	-0.0381489992141724	0.0612270012497902	5.0911
TC118482	0.515552997589111	0.635698974132538	5.0911
TC118593	-0.562547028064728	0.24252100288868	5.0911
TC119105	-0.339284002780914	-0.574702024459839	5.0911
TC119794	-0.266469985246658	-0.94216400384903	5.0911
TC121639	-0.372725993394852	-0.532002985477448	5.0911
TC122606	-0.124544002115726	0.268444985151291	5.0911
TC137255	0.244363993406296	0.277024000883102	5.0911
TC128939	9.31857967376709	8.75129985809326	48.4577
TC137524	-0.0295109990984201	0.0949440002441406	41.4974
TC140637	0.319999992847443	0.266759008169174	4.9569
TC125949	9.06083965301514	6.91123008728027	4.9305
TC127510	-0.0864859968423843	-0.142601996660233	4.9082
TC128169	-0.145313993096352	-0.123608998954296	4.9082

TC131334	0.185197994112968	0.0114700002595782	4.9082
TC133589	0.113980002701283	-0.0364699997007847	4.9082
TC134402	0.0646779984235764	0.748109996318817	4.9082
TC136195	-0.0214350000023842	-0.0552040003240108	4.9082
TC138314	-2.18299007415771	-0.427325993776321	4.9082
TC118303	0.291985988616943	1.91650998592377	4.8719
TC118420	0.463036000728607	0.485213994979858	4.8719
TC125678	0.132789999246597	0.0962840020656586	4.8719
TC125303	0.273380994796753	0.475585997104645	4.6957
TC132191	-0.090378999710083	0.100489996373653	4.6957
TC132943	-0.00990999955683947	0.00654800003394485	4.6582
TC130339	8.44738006591797	4.71007013320923	4.6319
TC130474	6.77299022674561	0.54645299911499	4.6319
TC130487	-1.00589001178741	-2.43432998657227	4.6319
TC131678	8.19908046722412	1.39285004138947	4.6319
TC132138	-0.126175999641418	0.0155830001458526	4.6319
TC132471	4.48218011856079	0.726336002349854	4.6319
TC133085	0.136714994907379	0.187285006046295	4.6319
TC133212	7.01039981842041	0.790431976318359	4.6319
TC133857	6.17259979248047	0.153997004032135	4.6319
TC134076	4.34353017807007	0.110322996973991	4.6319
TC134450	-0.0610809996724129	0.359501004219055	4.6319
TC134734	0.139294996857643	0.0162649992853403	4.6319
TC136074	-0.0554479993879795	0.194864004850388	4.6319
TC137227	3.75874996185303	-0.184492006897926	4.6319
TC138306	4.21962022781372	0.496796995401382	4.6319
TC139394	9.11524963378906	8.23309993743896	4.6319
TC140914	7.42476987838745	-0.0048030000180006	4.6319
TC117714	0.0509480014443398	-0.0741230025887489	4.5846
TC139873	-0.252330988645554	-0.21712900698185	4.5657
TC118839	1.77471995353699	-0.357688009738922	4.5514
TC137164	-0.143941000103951	-0.174125000834465	4.3589
TC135789	-0.316498011350632	-0.276237010955811	4.3547
TC136173	-0.169813007116318	-0.336349993944168	4.3547
TC120397	0.454616010189056	0.329154014587402	4.3520
TC131033	8.99361991882324	5.69989013671875	4.3479
TC131193	7.6042799949646	6.96932983398438	4.3479
TC132207	6.84457015991211	0.760065019130707	4.3479
TC132544	1.01540994644165	0.984166026115417	4.3479
TC129517	-0.128502994775772	7.0262598991394	4.2373
TC136675	0.0133619997650385	6.17702007293701	4.2373

TC138853	-0.0859280005097389	-0.20465299487114	4.2373
TC139513	0.315079003572464	6.14657020568848	4.2373
TC141254	4.84353017807007	3.28605008125305	4.2373
TC117162	0.0228909999132156	0.0900650024414062	4.2170
TC115617	-0.639982998371124	-0.518877983093262	4.0546
TC119371	0.179563000798225	-0.104076996445656	4.0290
TC114456	1.3145500421524	1.02296996116638	4.0283
TC123171	9.71012020111084	8.40919017791748	38.7077
TC136093	7.80604982376099	7.13514995574951	35.3805
TC134921	9.10970020294189	1.82492995262146	34.8647
TC113973	8.0070104598999	7.64075994491577	32.2224
TC132245	9.55953979492188	0.705334007740021	31.3864
TC116635	-0.380053013563156	0.246383994817734	3.9130
TC126475	2.44143009185791	2.34827995300293	3.8717
TC117861	0.0302639994770288	0.527622997760773	3.8360
TC121340	-0.218992993235588	-0.293967008590698	3.8321
TC132843	0.00419300002977252	0.533088982105255	3.7499
TC117446	-0.856003999710083	-0.58830201625824	3.6981
TC123095	0.0598260015249252	-0.00615399982780218	3.6981
TC125806	0.0113479997962713	-0.0360930003225803	3.6981
TC127280	-0.213348999619484	-0.259222000837326	3.6981
TC130124	-0.223936006426811	-0.143953993916512	3.6936
TC130291	0.2516950070858	0.379599004983902	3.6325
TC132492	-0.549578011035919	0.0372189991176128	3.6325
TC139309	7.24459981918335	7.82483005523682	3.6325
TC115310	-0.44132798910141	-0.767705023288727	3.6068
TC117008	-0.109600000083447	0.366641014814377	3.4844
TC119381	-0.0160649996250868	-0.0935570001602173	3.4844
TC120300	-0.0671579986810684	0.39342999458313	3.4844
TC120882	0.341165989637375	0.754553020000458	3.4844
TC122102	-0.573674023151398	-0.548675000667572	3.4844
TC124835	-0.658591985702515	-0.0997850000858307	3.4844
TC125972	0.0976720005273819	0.681361019611359	3.4844
TC126644	-1.40411996841431	-0.278324007987976	3.4844
TC127128	0.0304300002753735	0.242770001292229	3.4844
TC127380	-0.175929993391037	-0.282963007688522	3.4844
TC127427	-0.415803998708725	-0.849673986434937	3.4844
TC128543	0.752049028873444	0.105984002351761	3.4844
TC129693	-1.1462299823761	-0.578158974647522	3.4844
TC130905	-0.732155978679657	-0.685375988483429	3.4844
TC131751	-0.121676996350288	-0.284112006425858	3.4844

TC133384	0.128756001591682	0.14903299510479	3.4844
TC133714	-1.17549002170563	-0.483310014009476	3.4844
TC135004	-0.0829190015792847	-0.0187480002641678	3.4844
TC135919	0.685064971446991	0.710372984409332	3.4844
TC136636	0.294703990221024	-0.305842012166977	3.4844
TC138653	0.00212000007741153	0.0397480018436909	3.4844
TC140757	1.0424599647522	0.54451197385788	3.4844
TC141020	-0.0600529983639717	-0.014151000417769	3.4844
TC130589	7.54164981842041	5.42484998703003	3.4803
TC136597	5.3825798034668	3.05013990402222	3.4803
TC116715	0.12635999917984	-0.0217940006405115	3.4343
TC131700	-0.0493699982762337	-0.0442419983446598	3.4331
TC135587	-0.314099013805389	-0.286810994148254	3.4290
TC114318	-0.0262289997190237	-0.0870449990034103	3.3941
TC114443	0.211918994784355	0.206212997436523	3.3941
TC114553	-0.480771988630295	0.820605993270874	3.3941
TC114916	0.0164020005613565	0.0527490004897118	3.3941
TC115037	-0.855708003044128	-0.0681070014834404	3.3941
TC115041	0.943168997764587	-0.243168994784355	3.3941
TC116445	-0.321202009916306	-0.451595991849899	3.3941
TC116669	-0.59782999753952	-0.126758992671967	3.3941
TC117272	-0.136289998888969	-0.0708549991250038	3.3941
TC118261	0.116094999015331	-0.025305999442935	3.3941
TC118767	0.111718997359276	0.134709998965263	3.3941
TC118905	0.308777004480362	0.0992330014705658	3.3941
TC119405	-0.00139999995008111	0.0380610004067421	3.3941
TC119589	-0.507748007774353	-0.234356999397278	3.3941
TC119974	-0.150969997048378	-0.071942001581192	3.3941
TC120720	-0.612821996212006	-0.69439297914505	3.3941
TC121772	-0.0358339995145798	-0.251134991645813	3.3941
TC121911	0.095160998404026	0.0788130015134811	3.3941
TC123190	-0.0432710014283657	0.00377799989655614	3.3941
TC123588	-1.21164000034332	-0.746321976184845	3.3941
TC123687	0.352048993110657	-0.0503080002963543	3.3941
TC124018	-0.659166991710663	-0.371506989002228	3.3941
TC124376	0.211942002177238	-0.00887000001966953	3.3941
TC124440	0.102174997329712	-0.0945110023021698	3.3941
TC124461	-0.103373996913433	0.0431209988892078	3.3941
TC125530	-2.70064997673035	-0.107179999351501	3.3941
TC126569	0.510851979255676	-0.294510006904602	3.3941
TC127115	-0.62975001335144	-0.853721976280212	3.3941

TC127436	-0.584532022476196	0.418760985136032	3.3941
TC127999	0.0848829969763756	0.116074003279209	3.3941
TC129050	-0.761977970600128	-0.656714975833893	3.3941
TC129317	0.0868650004267693	0.0372859984636307	3.3941
TC130091	-0.0613319985568523	-0.361678004264832	3.3941
TC130960	0.196447998285294	0.0221599992364645	3.3941
TC131031	0.020796999335289	0.0496959984302521	3.3941
TC131402	0.183270007371902	0.173950999975204	3.3941
TC132229	0.240707993507385	-0.0183739997446537	3.3941
TC132674	-0.182194992899895	-0.318789005279541	3.3941
TC133161	-0.0840110033750534	0.100185997784138	3.3941
TC133596	-0.104482002556324	-0.237232998013496	3.3941
TC133800	0.121935002505779	0.222800001502037	3.3941
TC134907	-0.389982014894485	-0.462475001811981	3.3941
TC134963	0.103967003524303	0.097972996532917	3.3941
TC136424	0.170487001538277	0.793542981147766	3.3941
TC137390	-0.00905099976807833	0.145780995488167	3.3941
TC138172	-0.316480994224548	-0.290710002183914	3.3941
TC139340	-0.072382003068924	-0.0601060017943382	3.3941
TC139494	-0.020949000492692	0.624387979507446	3.3941
TC139804	-0.117216996848583	0.0550480000674725	3.3941
TC141111	7.0408501625061	5.97206020355225	3.3941
TC135382	0.688766002655029	0.623048007488251	3.3313
TC139259	-0.0246549993753433	-0.153327003121376	3.3313
TC140926	-0.267955005168915	-0.429343998432159	3.3313
TC131603	0.58984100818634	0.6581130027771	3.3272
TC131889	6.73935985565186	5.87590980529785	3.3272
TC132339	7.96802997589111	7.06088018417358	3.3272
TC133182	6.903399944430542	6.19734001159668	3.3272
TC137154	9.12110042572021	7.86159992218018	3.3272
TC138044	10.1085996627808	9.28083992004395	3.3272
TC139113	5.20493984222412	3.93712997436523	3.3272
TC141641	0.213467001914978	-0.239201992750168	3.3272
TC124794	-0.00253699999302626	0.0428979992866516	3.2812
TC126061	-0.0718979984521866	-0.549755990505219	3.2164
TC120629	0.115979000926018	0.0470080003142357	3.1737
TC121310	0.415295988321304	-0.0882140025496483	3.1737
TC122952	-0.140095993876457	-0.204181000590324	3.1737
TC124141	0.00563200004398823	0.804037988185883	3.1737
TC126234	-0.148121997714043	0.0333699993789196	3.1737
TC126877	0.0432629995048046	0.188448995351791	3.1737

TC133953	-0.464552998542786	-0.420691013336182	3.1737
TC137578	-0.0673770010471344	-0.0755200013518333	3.1737
TC113011	6.94076013565063	5.31070995330811	3.1695
TC113066	-0.389236986637115	-0.703218996524811	3.1695
TC132225	-0.634553015232086	-0.86260998249054	3.1695
TC133144	7.09320020675659	6.02313995361328	3.1695
TC133229	6.33626985549927	4.17204999923706	3.1695
TC133971	-0.425148010253906	0.145858004689217	3.1695
TC134663	7.03036022186279	5.53994989395142	3.1695
TC135549	0.822798013687134	1.70363998413086	3.1695
TC140212	-0.46398600935936	-0.670060992240906	3.1695
TC141246	9.01764011383057	8.29601001739502	3.1695
TC115049	-0.00901699997484684	0.289929986000061	3.1303
TC115389	-0.158141002058983	-0.09740199893713	3.1303
TC115869	3.0665500164032	1.90565001964569	3.1303
TC115925	-0.487922012805939	-0.944101989269257	3.1303
TC116912	0.264398008584976	-0.0268970001488924	3.1303
TC117928	0.23971800506115	0.112282998859882	3.1303
TC119123	0.961302995681763	-1.84018003940582	3.1303
TC120782	-0.0668049976229668	-0.0758860036730766	3.1303
TC131690	-0.0432050004601479	-0.22181299328804	3.1303
TC112630	-0.789043009281158	-0.307709991931915	3.0879
TC115183	-0.779689013957977	-0.634657979011536	3.0879
TC129019	1.79232001304626	1.91459000110626	3.0879
TC129389	5.90803003311157	0.108134999871254	3.0879
TC129535	0.676886975765228	0.824001014232635	3.0879
TC129587	8.54706001281738	0.567373991012573	3.0879
TC129713	0.0516519993543625	-0.00486099999397993	3.0879
TC130026	2.77238011360168	0.00621900008991361	3.0879
TC130161	-0.177204996347427	0.0563579984009266	3.0879
TC130199	-0	-0	3.0879
TC130344	-0.103413999080658	-0.0310960002243519	3.0879
TC130379	-0.0800540000200272	-0.221182003617287	3.0879
TC130386	8.28339958190918	1.17797005176544	3.0879
TC130393	8.40773963928223	1.00039994716644	3.0879
TC130593	-1.39839994907379	0.506515979766846	3.0879
TC130830	-1.4763799905777	-1.13052999973297	3.0879
TC130844	0.129216000437737	0.0945099964737892	3.0879
TC130845	0.340131998062134	0.202079996466637	3.0879
TC130856	6.23939990997314	3.51793003082275	3.0879
TC130903	-0.136937007308006	-0.24346399307251	3.0879

TC130984	1.68604004383087	0.209050998091698	3.0879
TC131088	-0.00910000037401915	0.601242005825043	3.0879
TC131188	0.427345007658005	0.339621007442474	3.0879
TC131202	0.166629999876022	-0.0372769981622696	3.0879
TC131250	0.142040997743607	0.125524997711182	3.0879
TC131308	-0.0117669999599457	-0.0851060003042221	3.0879
TC131344	7.91173982620239	1.12321996688843	3.0879
TC131381	0.165686994791031	-0.10606499761343	3.0879
TC131425	0.236488997936249	0.100041002035141	3.0879
TC131429	1.72254002094269	0.159795001149178	3.0879
TC131445	0.137135997414589	0.0447599999606609	3.0879
TC131484	-0.27869701385498	-0.489235997200012	3.0879
TC131718	0.201791003346443	-0.865405023097992	3.0879
TC131734	-0.122688002884388	-0.0745330005884171	3.0879
TC131765	0.475596010684967	0.213072001934052	3.0879
TC131769	2.59754991531372	-0.146941006183624	3.0879
TC131776	-0.0177960004657507	-0.18715900182724	3.0879
TC131799	2.16829991340637	-0.29236900806427	3.0879
TC131847	4.04807996749878	-0.0661270022392273	3.0879
TC131921	5.42489004135132	0.385008990764618	3.0879
TC131969	3.06434988975525	0.170861005783081	3.0879
TC132174	0.353363007307053	0.26468101143837	3.0879
TC132180	0.0237109996378422	0.014465999789536	3.0879
TC132209	-0.179366007447243	-0.140772998332977	3.0879
TC132314	0.406901001930237	0.930249989032745	3.0879
TC132317	0.104313001036644	0.128911003470421	3.0879
TC132391	5.40302991867065	5.15984010696411	3.0879
TC132400	4.71823978424072	0.49268901348114	3.0879
TC132437	8.18875026702881	1.08534002304077	3.0879
TC132737	-0.318210989236832	-0.530921995639801	3.0879
TC132782	2.30475997924805	0.139816999435425	3.0879
TC132786	0.2603600025177	0.0359780006110668	3.0879
TC132869	2.3948700428009	0.0452260002493858	3.0879
TC132886	-0.0054109999909997	0.099992997944355	3.0879
TC132930	5.82582998275757	0.332073986530304	3.0879
TC132933	5.47037982940674	0.261227011680603	3.0879
TC132986	4.40354013442993	0.298866003751755	3.0879
TC132991	0.559548020362854	0.0074880002066493	3.0879
TC133041	6.74808979034424	1.47044003009796	3.0879
TC133117	-0.388078987598419	-0.0655599981546402	3.0879
TC133119	0.265511989593506	0.0470400005578995	3.0879

TC133122	9.00481033325195	0.826506972312927	3.0879
TC133129	1.46352994441986	0.227052003145218	3.0879
TC133162	7.2550802230835	0.715076982975006	3.0879
TC133215	-0.05867899954319	-0.10097000002861	3.0879
TC133224	-0.031640999019146	0.300909012556076	3.0879
TC133329	1.7966400384903	1.01121997833252	3.0879
TC133333	0.105434000492096	-0.284260988235474	3.0879
TC133351	9.11524963378906	8.23309993743896	3.0879
TC133398	1.1950900554657	0.232730001211166	3.0879
TC133401	7.20753002166748	-0.0630149990320206	3.0879
TC133433	-0.73029500246048	-1.46659994125366	3.0879
TC133438	-0.719202995300293	-0.690239012241364	3.0879
TC133563	9.39739036560059	5.45682001113892	3.0879
TC133567	0.356274992227554	0.216481000185013	3.0879
TC133575	5.60678005218506	4.37236022949219	3.0879
TC133588	0.520021975040436	-0.0864389985799789	3.0879
TC133611	-0.517444014549255	-0.440658986568451	3.0879
TC133632	0.234354004263878	0.00240399991162121	3.0879
TC133665	-0	-0	3.0879
TC133673	-0.601059973239899	-0.600480020046234	3.0879
TC133705	8.09700012207031	0.502604007720947	3.0879
TC133846	2.35535001754761	-0.154221996665001	3.0879
TC133861	6.85706996917725	0.941088974475861	3.0879
TC133899	0.00903599988669157	-0.0131449997425079	3.0879
TC133942	2.63479995727539	0.162734001874924	3.0879
TC134000	-0.369522005319595	0.621553003787994	3.0879
TC134004	4.31312990188599	-0.0850540027022362	3.0879
TC134005	0.485715985298157	-0.0292780008167028	3.0879
TC134009	8.35208988189697	1.41840994358063	3.0879
TC134017	3.6010000705719	0.448624014854431	3.0879
TC134019	0.013136999681592	-0.726688981056213	3.0879
TC134035	3.72350001335144	0.060139998793602	3.0879
TC134046	0.245725005865097	0.0952849984169006	3.0879
TC134080	0.337913990020752	-0.153843998908997	3.0879
TC134082	-0.461425989866257	-0.225948005914688	3.0879
TC134139	-0.252624988555908	-0.270047008991241	3.0879
TC134144	1.29307997226715	0.89444500207901	3.0879
TC134153	-0.64776599407196	-0.563404023647308	3.0879
TC134175	0.22197400033474	0.261148005723953	3.0879
TC134207	-0	-0	3.0879
TC134280	2.86207008361816	-0.237316995859146	3.0879

TC134321	-1.21553003787994	0.214230000972748	3.0879
TC134350	0.0544359982013702	0.0624609999358654	3.0879
TC134359	5.53685998916626	0.0636079981923103	3.0879
TC134380	-0.0467009991407394	-0.144012004137039	3.0879
TC134445	9.39739036560059	5.45682001113892	3.0879
TC134462	-0.01004199963063	0.0332910008728504	3.0879
TC134500	-0.449730008840561	0.0594779998064041	3.0879
TC134532	3.21543002128601	0.283183008432388	3.0879
TC134807	1.83326995372772	0.128448992967606	3.0879
TC134837	-0.471843987703323	-0.321794003248215	3.0879
TC134869	0.00937599968165159	0.0759899988770485	3.0879
TC134903	-0.337736010551453	-0.271495997905731	3.0879
TC134969	0.127478003501892	0.169293999671936	3.0879
TC134971	0.11583100259304	-0.0333929993212223	3.0879
TC135124	0.208363994956017	0.5207279920578	3.0879
TC135206	0.177175998687744	0.104438997805119	3.0879
TC135245	-0.200903996825218	-0.143711999058723	3.0879
TC135254	5.85739994049072	0.170471996068954	3.0879
TC135337	-0.13773900270462	-0.0768100023269653	3.0879
TC135343	2.44929003715515	0.268572986125946	3.0879
TC135375	4.63539981842041	0.651317000389099	3.0879
TC135412	-0	-0	3.0879
TC135426	-0.149128004908562	-0.168341994285583	3.0879
TC135432	1.03741002082825	0.140560999512672	3.0879
TC135448	0.34476900100708	0.165106996893883	3.0879
TC135505	0.0433459989726543	0.0160060003399849	3.0879
TC135604	-0.0919959992170334	-0.450421005487442	3.0879
TC135640	-0	-0	3.0879
TC135641	4.4410400390625	0.0492810010910034	3.0879
TC135648	5.1012601852417	0.152484998106956	3.0879
TC135697	2.44199991226196	0.575094997882843	3.0879
TC135714	-0	-0	3.0879
TC135741	5.09264993667603	0.696120023727417	3.0879
TC135743	-0.171462997794151	0.0311960000544786	3.0879
TC135762	5.90696001052856	5.24793004989624	3.0879
TC135814	8.12154006958008	0.271479994058609	3.0879
TC135816	4.52717018127441	0.0360630005598068	3.0879
TC135833	0.162200003862381	-0.107148997485638	3.0879
TC135909	-0.292394995689392	-0.303871005773544	3.0879
TC135928	-0	-0	3.0879
TC136079	0.491928994655609	0.203951999545097	3.0879

TC136126	3.24327993392944	0.351164013147354	3.0879
TC136158	-0.148099005222321	0.155320003628731	3.0879
TC136176	1.60956001281738	-0.30986499786377	3.0879
TC136365	6.64804983139038	0.570436000823975	3.0879
TC136387	0.798852026462555	0.114817000925541	3.0879
TC136422	0.252692013978958	-0.183069005608559	3.0879
TC136444	5.28011989593506	0.149038001894951	3.0879
TC136479	2.24161005020142	0.343950986862183	3.0879
TC136568	0.0552960000932217	0.00533200008794665	3.0879
TC136576	1.8145500421524	-0.155294999480247	3.0879
TC136578	0.358175992965698	-0.117930002510548	3.0879
TC136591	-0.0174669995903969	-0.118143998086452	3.0879
TC136683	0.246043995022774	0.150593996047974	3.0879
TC136713	-0.307466000318527	0.180825993418694	3.0879
TC136717	4.33283996582031	4.81598997116089	3.0879
TC136758	0.793385028839111	0.0226690005511045	3.0879
TC136762	-0.566516995429993	-0.215992003679276	3.0879
TC136768	-0	-0	3.0879
TC136842	-0	-0	3.0879
TC136859	1.20717000961304	0.0225469991564751	3.0879
TC136865	6.2071099281311	0.352243989706039	3.0879
TC136956	-0.0619329996407032	-0.278261989355087	3.0879
TC137056	-0.473605006933212	-0.663591980934143	3.0879
TC137084	1.92366003990173	1.34372997283936	3.0879
TC137102	2.15075993537903	-0.0381270013749599	3.0879
TC137120	-0	-0	3.0879
TC137148	-0.0617480017244816	-0.0185480006039143	3.0879
TC137192	-0	-0	3.0879
TC137316	-0.900831997394562	-0.789967000484467	3.0879
TC137326	4.51968002319336	0.125780001282692	3.0879
TC137327	-0.365648001432419	-0.104819998145103	3.0879
TC137410	-0.385612010955811	0.185170993208885	3.0879
TC137451	-0.290297001600266	-0.351958990097046	3.0879
TC137462	-0.098531000316143	-0.054552998393774	3.0879
TC137513	5.55813980102539	0.606168985366821	3.0879
TC137556	-0.599591016769409	-0.83184802532196	3.0879
TC137568	1.00625002384186	0.194921001791954	3.0879
TC137576	1.71165001392365	0.0215629991143942	3.0879
TC137743	4.33082008361816	0.0686699971556664	3.0879
TC137858	-0	-0	3.0879
TC137871	-0.502514004707336	-0.126118004322052	3.0879

TC137872	6.19129991531372	-0.100391998887062	3.0879
TC137873	5.58888006210327	0.157056003808975	3.0879
TC137967	0.0254089999943972	-0.0159859992563725	3.0879
TC138034	2.54998993873596	2.05941009521484	3.0879
TC138114	5.84568023681641	0.29763600230217	3.0879
TC138220	-0.764618992805481	-0.673291027545929	3.0879
TC138268	-0.428537994623184	-0.10598199814558	3.0879
TC138312	6.88439989089966	6.08643007278442	3.0879
TC138346	0.456757009029388	0.222185000777245	3.0879
TC138381	7.5770001411438	0.31768599152565	3.0879
TC138415	9.19890022277832	2.22947001457214	3.0879
TC138446	0.213467001914978	-0.239201992750168	3.0879
TC138541	-0.595984995365143	-1.08002996444702	3.0879
TC138582	5.62848997116089	0.204051002860069	3.0879
TC138583	2.059730052948	1.61466002464294	3.0879
TC138621	2.70958995819092	0.103965997695923	3.0879
TC138649	0.250721991062164	0.39172700047493	3.0879
TC138659	3.01270008087158	0.463393986225128	3.0879
TC138674	0.20567199587822	-0.203492999076843	3.0879
TC138688	0.342900007963181	0.11590900272131	3.0879
TC138932	0.399392992258072	0.0554450005292892	3.0879
TC139014	-0.373701989650726	-0.333727985620499	3.0879
TC139078	1.65004003047943	1.51363003253937	3.0879
TC139118	0.898591995239258	-0.16168899834156	3.0879
TC139156	-0.557318985462189	-0.360049992799759	3.0879
TC139182	0.499173998832703	0.00377100007608533	3.0879
TC139190	0.508316993713379	0.032125998288393	3.0879
TC139234	3.50600004196167	0.0646649971604347	3.0879
TC139255	-0	-0	3.0879
TC139275	0.39826700091362	0.455904006958008	3.0879
TC139312	0.117362998425961	-0.18928100168705	3.0879
TC139335	-0.452093005180359	-0.519997000694275	3.0879
TC139408	-0.386729001998901	-0.356382012367249	3.0879
TC139482	0.313024997711182	0.0116959996521473	3.0879
TC139504	5.35894012451172	0.580877006053925	3.0879
TC139725	0.0550450012087822	0.199775993824005	3.0879
TC139755	-1.40962994098663	-0.490644007921219	3.0879
TC139795	0.0712350010871887	0.149743005633354	3.0879
TC139866	2.85160994529724	-0.161577999591827	3.0879
TC139878	4.02458000183105	0.315719991922379	3.0879
TC139915	1.76217997074127	1.22032999992371	3.0879

TC139952	0.109491996467113	0.129176005721092	3.0879
TC140088	5.12565994262695	-0.000123999998322688	3.0879
TC140166	1.8486499786377	1.08580994606018	3.0879
TC140275	-0.0778850018978119	0.0108160004019737	3.0879
TC140345	0.272056996822357	0.428011000156403	3.0879
TC140404	4.3534197807312	-0.038077998906374	3.0879
TC140421	0.801813006401062	-2.54136991500854	3.0879
TC140476	0.168320000171661	0.069242000579834	3.0879
TC140526	0.915144979953766	0.171964004635811	3.0879
TC140558	-0.0966150015592575	-0.00718900002539158	3.0879
TC140569	1.67416000366211	0.106504999101162	3.0879
TC140599	0.164848998188972	-0.0846939980983734	3.0879
TC140656	2.56133008003235	-0.293787986040115	3.0879
TC140747	5.68404006958008	0.0142029998824	3.0879
TC140772	-0.281197994947433	-0.213784992694855	3.0879
TC140819	5.41776990890503	2.85824990272522	3.0879
TC140835	2.9488799571991	-0.0731360018253326	3.0879
TC140869	-0.13510499894619	0.0698510035872459	3.0879
TC140870	1.78894996643066	0.955303013324738	3.0879
TC140887	8.67564964294434	1.17155003547668	3.0879
TC140911	5.39917993545532	0.208515003323555	3.0879
TC140982	-0.217231005430222	-0.39360100030899	3.0879
TC141016	-0.198080003261566	-0.184448003768921	3.0879
TC141032	3.86401009559631	0.0911310017108917	3.0879
TC141123	0.421654999256134	0.0393289998173714	3.0879
TC141151	-0.172888994216919	0.00364800007082522	3.0879
TC141168	8.64713954925537	2.3090500831604	3.0879
TC141194	2.55707001686096	0.0468229986727238	3.0879
TC141206	-0.0370149984955788	-0.122188001871109	3.0879
TC141329	-0.0786309987306595	-0.000472000014269724	3.0879
TC141334	2.94886994361877	-0.161542996764183	3.0879
TC141338	1.17613995075226	0.0891589969396591	3.0879
TC141360	-0.181737005710602	-0.01369300018996	3.0879
TC141395	1.12343001365662	-0.0271039996296167	3.0879
TC141398	0.208582997322083	0.0157190002501011	3.0879
TC141495	-0.242155000567436	-0.0144980000331998	3.0879
TC141533	-0.187325000762939	-0.426813989877701	3.0879
TC141586	7.89462995529175	2.16901993751526	3.0879
TC141590	-0.182081997394562	-0.509351015090942	3.0879
TC141603	7.50578022003174	0.200914993882179	3.0879
TC141632	-0.0483979992568493	0.0481979995965958	3.0879

TC141701	4.61406993865967	-0.151798993349075	3.0879
TC141732	-0.146377995610237	-0.0594919994473457	3.0879
TC129609	7.80604982376099	7.13514995574951	26.8687
TC124054	7.93924999237061	7.79982995986938	23.6306
TC128493	-0.0920900031924248	10.6577997207642	22.1866
TC130208	6.17437982559204	5.09914016723633	22.1866
TC126123	6.70066976547241	7.09717988967896	21.6265
TC133696	9.5042896270752	8.2852201461792	21.1521
TC130894	0.055417999625206	-0.227484002709389	2.9259
TC117459	-0.471354991197586	-0.696443021297455	2.8822
TC132698	-0.523559987545013	-0.363216012716293	2.8822
TC138275	0.335956007242203	-0.234451994299889	2.8822
TC140243	-0.147383004426956	-0.170837000012398	2.8822
TC141427	-0.377815991640091	0.244065999984741	2.8822
TC131309	0.12024699896574	0.504437029361725	2.8754
TC131622	-0.244610995054245	0.215131998062134	2.8241
TC125792	-0.23157599568367	-0.110128998756409	2.7648
TC129398	-0.250308007001877	-0.294872999191284	2.7648
TC132105	-0.570378005504608	-0.281500995159149	2.7648
TC132571	0.700655996799469	0.657922029495239	2.7648
TC135001	-1.58307003974915	-0.511767029762268	2.7648
TC113542	7.39908981323242	5.49006986618042	2.7292
TC129483	-0.163708999752998	-0.227321997284889	2.7292
TC135031	0.0579590015113354	-0.277054011821747	2.7292
TC136092	0.407481014728546	0.303362995386124	2.7292
TC138452	-0.0516370013356209	0.10215300321579	2.7292
TC139987	0.169569000601768	-0.367987006902695	2.7292
TC141212	-0.0107429996132851	0.0151159996166825	2.7292
TC130559	0.239533007144928	0.280140995979309	2.6117
TC131517	-0.735634028911591	-0.260244995355606	2.6117
TC138999	-0.869071006774902	-0.597701013088226	2.6117
TC141011	-0.269870012998581	-0.60004198551178	2.6117
TC113283	-1.66603004932404	-1.05982005596161	2.4541
TC118335	-0.846540987491608	-0.548124015331268	2.4541
TC123559	-0.246983006596565	-0.335283994674683	2.4541
TC125580	-0.150942996144295	0.0653280019760132	2.4541
TC126397	0.20924599468708	0.219494000077248	2.4541
TC130911	-0.0382679998874664	0.0344889983534813	2.4541
TC131259	-0	-0	2.4541
TC131737	-0.280016988515854	-0.121693000197411	2.4541
TC131927	-0.0643080025911331	0.0990530028939247	2.4541

TC132678	-0.120784997940063	-0.0100600002333522	2.4541
TC134278	0.0266970004886389	-0.0749299973249435	2.4541
TC135155	-0.161733001470566	0.0756089985370636	2.4541
TC135760	0.34451100230217	0.786458015441895	2.4541
TC135878	0.0853269994258881	0.287463009357452	2.4541
TC136306	-0.93249100446701	-0.0581150017678738	2.4541
TC137101	0.0226850006729364	-0.048105001449585	2.4541
TC137353	-0.741744995117188	0.280436009168625	2.4541
TC137625	-0.00552299991250038	-0.354090988636017	2.4541
TC137796	-0.442294001579285	-0.477160006761551	2.4541
TC139221	-0.0252000000327826	0.0509970001876354	2.4541
TC139872	0.0113760000094771	-0.0199820008128881	2.4541
TC140019	-0.944253027439117	-0.567523002624512	2.4541
TC140335	-0.195460006594658	0.0264240000396967	2.4541
TC141647	-0.740091979503632	0.833779990673065	2.4541
TC128385	-0.371688991785049	0.0383299998939037	2.2718
TC134265	5.95384979248047	5.16078996658325	19.0171
TC130892	-0.259214013814926	10.5249996185303	17.1648
TC132766	-0.310658007860184	9.75214004516602	17.1648
TC137189	7.93924999237061	7.79982995986938	17.1648
TC117163	9.90200042724609	8.99077987670898	16.7988
TC125206	7.33025979995728	5.95730018615723	16.3179
TC131211	3.55378007888794	2.41005992889404	16.1016
TC127081	6.34950017929077	4.99368000030518	15.8476
TC133852	6.18787002563477	4.46468019485474	15.8476
TC134177	8.82112979888916	7.48449993133545	15.8476
TC134597	-0.142115995287895	9.70014953613281	15.8476
TC135484	1.66612994670868	0.804161012172699	15.8476
TC136480	3.39989995956421	0.164295002818108	15.8476
TC121095	8.34261989593506	8.54428005218506	15.5309
TC112872	10.0172996520996	8.65207958221436	149.9299
TC133466	2.88862991333008	2.32395005226135	14.6882
TC136824	9.39739036560059	5.45682001113892	14.3148
TC128569	5.10055017471313	3.79190993309021	14.0756
TC131486	0.283836007118225	4.6682300567627	133.1197
TC130256	-0.0794510021805763	-0.0532420016825199	13.8956
TC118935	5.13018989562988	4.15883016586304	13.6739
TC126200	2.56881999969482	8.08203983306885	12.6781
TC126222	6.69082021713257	5.65080976486206	12.6781
TC126931	5.13926982879639	5.5586199760437	12.6781
TC130767	7.32429981231689	5.4987998008728	12.6781

TC131682	0.428438007831573	-0.197053000330925	12.6781
TC135839	-0.0664039999246597	7.42482995986938	12.6781
TC134267	-0.224659994244576	8.81033992767334	12.6301
TC134022	1.8909900188446	-0.604825019836426	12.5296
TC136513	0.394811004400253	0.159302994608879	12.3517
TC122974	9.03695011138916	8.57763004302979	12.2531
TC129977	9.04615020751953	8.05712032318115	11.9469
TC131060	8.91236972808838	6.3719801902771	11.9469
TC118382	-0.719228982925415	-0.862945020198822	11.8793
TC121728	0.0166410002857447	-0.0666920021176338	11.8793
TC129607	1.34563004970551	0.722786009311676	11.7200
TC127749	9.95306015014648	9.24635982513428	11.1287
TC135802	8.98116970062256	8.12963962554932	109.5815
TC117435	9.90200042724609	8.99077987670898	10.8654
TC130445	4.31778001785278	0.47885400056839	10.8077
TC115823	0.899811029434204	0.73241001367569	1.9565
TC118927	0.09391900151968	0.328734010457993	1.9565
TC119115	2.65768003463745	3.85873007774353	1.9565
TC122412	-0.105084002017975	0.147769004106522	1.9565
TC124907	-0.103638000786304	0.0660099983215332	1.9565
TC127295	0.60030597448349	0.513886988162994	1.9565
TC127985	1.14181005954742	0.422226011753082	1.9565
TC129348	-0.121317997574806	-0.418927997350693	1.9565
TC130079	-0.0728230029344559	-0.172709003090858	1.9565
TC136133	0.168162003159523	0.0603320002555847	1.9565
TC138645	-0.19964300096035	-0.576839029788971	1.9565
TC141101	-0.631487011909485	-0.523413002490997	1.9565
TC141219	0.617069005966187	0.414956003427505	1.9565
TC114845	3.06169009208679	2.81160998344421	1.8547
TC129388	0.14274999499321	-0.110895998775959	1.8547
TC132811	0.0821250006556511	-0.959578990936279	1.8547
TC138023	-0.00664900010451674	-0.46263799071312	1.8547
TC116874	-0.14342400431633	0.214157000184059	1.8034
TC128948	0.531081020832062	0.606121003627777	1.8034
TC133050	-0.608843982219696	-1.53084003925323	1.8034
TC133698	-0.327127993106842	0.0311719998717308	1.8034
TC138969	-0.555625021457672	-0.66898500919342	1.8034
TC141577	-0.328119993209839	-0.502569973468781	1.8034
TC112785	-0.237756997346878	-0.0343349985778332	1.6970
TC112960	0.152493998408318	0.0813449993729591	1.6970
TC113482	-0.00991600006818771	0.034230001270771	1.6970

TC114044	0.422591000795364	0.241629004478455	1.6970
TC115075	-0.0529769994318485	-0.0498889982700348	1.6970
TC115472	0.167162001132965	0.303555995225906	1.6970
TC116286	-0.273131012916565	-0.0969650000333786	1.6970
TC116541	-0.543470025062561	-0.374509990215302	1.6970
TC116814	-0.0679690018296242	0.0158169995993376	1.6970
TC117054	-0.228708997368813	-0.43809500336647	1.6970
TC117127	-0.129130005836487	-0.267879009246826	1.6970
TC117337	-0.269459992647171	-0.293190002441406	1.6970
TC118149	0.888673007488251	0.57039201259613	1.6970
TC118451	-0.00184200005605817	0.0286110006272793	1.6970
TC118463	0.0556330010294914	0.341866999864578	1.6970
TC118943	-0.264764994382858	0.168254002928734	1.6970
TC119067	-0.681151986122131	-0.779031991958618	1.6970
TC119609	4.61116981506348	3.56291007995605	1.6970
TC119808	-0.256231993436813	0.0348850004374981	1.6970
TC120021	0.481265991926193	0.227340996265411	1.6970
TC121203	-0.353837996721268	-0.39614799618721	1.6970
TC122092	-0.00867199990898371	0.0239090006798506	1.6970
TC122470	6.17319011688232	4.68316984176636	1.6970
TC122825	-0.123025000095367	0.157957002520561	1.6970
TC123649	-0.144305005669594	-0.324003010988235	1.6970
TC123866	0.0401419997215271	-0.0657899975776672	1.6970
TC123871	-0.670413970947266	1.09888994693756	1.6970
TC123883	-0.568971991539001	-0.315488994121552	1.6970
TC125169	0.021671000868082	0.123116999864578	1.6970
TC125177	-0.224668994545937	-0.170341998338699	1.6970
TC125272	-0.184123992919922	-0.138310998678207	1.6970
TC125572	-0.687541007995605	-0.619231998920441	1.6970
TC126920	7.18803977966309	5.83605003356934	1.6970
TC127094	0.00337600009515882	-0.133006006479263	1.6970
TC127518	-0.167576998472214	0.234680995345116	1.6970
TC129478	0.179207995533943	0.285151988267899	1.6970
TC129856	0.779439985752106	0.728404998779297	1.6970
TC129924	-1.0058399438858	-1.14599001407623	1.6970
TC130786	-0.232475996017456	0.277337998151779	1.6970
TC130835	0.162222996354103	0.521948993206024	1.6970
TC131040	0.972733974456787	0.749477028846741	1.6970
TC131283	0.101709999144077	0.0272739995270967	1.6970
TC131634	-0.0860200002789497	-0.332462996244431	1.6970
TC132361	0.397433996200562	0.681433022022247	1.6970

TC133006	0.222839996218681	-0.193653002381325	1.6970
TC133762	-0.602367997169495	0.406661003828049	1.6970
TC135069	0.0521479994058609	-0.096949003636837	1.6970
TC135192	0.0963369980454445	0.371502012014389	1.6970
TC136039	0.34315899014473	0.190238997340202	1.6970
TC137799	-0.184380993247032	1.03267002105713	1.6970
TC137800	-0.217170998454094	-0.0032619999255985	1.6970
TC138137	-0.0294899996370077	-0.0969069972634315	1.6970
TC138370	4.33820009231567	2.40876007080078	1.6970
TC138713	-0.432170987129211	0.441219002008438	1.6970
TC139172	-0.43435001373291	0.103552997112274	1.6970
TC140649	-0.60513699054718	0.0144530003890395	1.6970
TC140728	0.0869249999523163	0.133495002985001	1.6970
TC141024	0.093359999358654	-0.203913003206253	1.6970
TC141120	0.0452789999544621	0.213914006948471	1.6970
TC141185	-0.11056499928236	-0.126112997531891	1.6970
TC141455	-0.150696992874146	-0.14716100692749	1.6970
TC130181	0.490298002958298	0.700892984867096	1.5909
TC112558	0.246463999152184	0.508777022361755	1.5440
TC122595	0.747371017932892	1.31175005435944	1.5440
TC126238	0.137624993920326	0.252853989601135	1.5440
TC128471	-0.0682839974761009	0.127344995737076	1.5440
TC129373	-0.250800997018814	-0.30647400021553	1.5440
TC129589	-0.613614976406097	-0.167823001742363	1.5440
TC129985	0.016780000180006	0.361005008220673	1.5440
TC130329	-0.164176002144814	-0.361119985580444	1.5440
TC130941	-0.091899998486042	0.125916004180908	1.5440
TC133889	0.610831022262573	0.28117299079895	1.5440
TC134003	-0.186297997832298	0.118601001799107	1.5440
TC134698	-0.383343994617462	-0.415921002626419	1.5440
TC134767	-0.466172009706497	-0.538863003253937	1.5440
TC134819	0.976388990879059	1.25479996204376	1.5440
TC135079	-0.225475996732712	0.0589569993317127	1.5440
TC136692	0.276547998189926	0.0699969977140427	1.5440
TC136782	-0.249258995056152	-0.464897006750107	1.5440
TC136866	-0.480601012706757	-0.52266800403595	1.5440
TC137216	0.179914996027946	0.272186011075974	1.5440
TC137463	0.400317996740341	0.413881003856659	1.5440
TC137626	-1.12957000732422	-0.85391902923584	1.5440
TC137760	-0.795026004314423	-1.0799800157547	1.5440
TC137762	0.16074800491333	0.0109590003266931	1.5440

TC138130	0.296115010976791	0.683948993682861	1.5440
TC138180	3.46795010566711	0.878309011459351	1.5440
TC139079	-0.289467006921768	-0.2049939930439	1.5440
TC139674	-0.361272007226944	0.208288997411728	1.5440
TC140643	-0.194949999451637	-0.307305991649628	1.5440
TC140994	-0.464724987745285	-0.194327995181084	1.5440
TC141225	-0.0906530022621155	-0.110950998961926	1.5440
TC141453	-0.0780669972300529	0.256475001573563	1.5440
TC141627	-0.350226014852524	-0.483747005462646	1.5440
TC141669	0.0493949986994267	0.0243999995291233	1.5440
TC141800	0.379642009735107	1.3094300031662	1.5440

Bibliography

Aitchison J.: *The Statistical Analysis of Compositional Data.*, chap. Monographs on Statistics and Applied Probability. Chapman & Hall, London (1988).

Aitchison J.: *Algebraic Methods in Statistics and Probability: Contemporary Mathematics Series*, pages 1–22. No. 287 in Contemporary Mathematics Series, American Mathematical Society, Providence, Rhode Island (2001).

Altschul S. F., Gish W., Miller W., Myers E. W., Lipman D. J.: Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, (1990).

Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H., Cherry J. M., Davis A. P., Dolinski K., Dwight S. S., Eppig J. T., Harris M. A., Hill D. P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J. C., Richardson J. E., Ringwald M., Rubin G. M., Sherlock G.: Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–9, (2000).

Attoor S., Dougherty E. R., Chen Y., Bittner M. L., Trent J. M.: Which is better for cdna-microarray-based classification: ratios or direct intensities. *Bioinformatics*, 20(16):2513–20, (2004).

Audic S., Claverie J. M.: The significance of digital gene expression profiles. *Genome Res*, 7(10):986–95, (1997).

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*, 145(4):1600–1618, (2007).

Bainbridge M. N., Warren R. L., Hirst M., Romanuik T., Zeng T., Go A., Delaney A., Griffith M., Hickenbotham M., Magrini V., Mardis E. R., Sadar M. D., Siddiqui A. S., Marra M. A., Jones S. J. M.: Analysis of the prostate cancer cell line lncap transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7:246, (2006).

Baldi P., Long A. D.: A bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–19, (2001).

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*, 19(9):998–1013, (2006).

Bekel T., Henckel K., Kuster H., Meyer F., Mittard Runte V., Neuweger H., Paarmann D., Rupp O., Zakrzewski M., Puhler 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*, 140(1-2):3–12, (2009).

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 journal*, April(12).

Bieri T., Blasiar D., Ozersky P., Antoshechkin I., Bastiani C., Canaran P., Chan J., Chen N., Chen W. J., Davis P., Fiedler T. J., Girard L., Han M., Harris T. W., Kishore R., Lee R., McKay S., Müller H.-M., Nakamura C., Petcherski A., Rangarajan A., Rogers A., Schindelman G., Schwarz E. M., Spooner W., Tuli M. A., Van Auken K., Wang D., Wang X., Williams G., Durbin R., Stein L. D., Sternberg P. W., Spieth J.: Wormbase: new content and better access. *Nucleic Acids Res*, 35(Database issue):D506–10, (2007).

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M. J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M.: The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic Acids Res*, 31(1):365–70, (2003).

Bolstad B. M., Irizarry R. A., Astrand M., Speed T. P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–93, (2003).

Brazma A., Hingamp P., Quackenbush J., Sherlock G., Spellman P., Stoeckert C., Aach J., Ansorge W., Ball C. A., Causton H. C., Gaasterland T., Glenisson P., Holstege F. C., Kim I. F., Markowitz V., Matese J. C., 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. *Nat Genet*, 29(4):365–71, (2001).

Brenner S., Johnson M., Bridgham J., Golda G., Lloyd D. H., Johnson D., Luo S., McCurdy S., Foy M., Ewan M., Roth R., George D., Eletr S., Albrecht G., Vermaas E., Williams S. R., Moon K., Burcham T., Pallas M., DuBridge R. B., Kirchner J., Fearon K., Mao J., Corcoran K.: Gene expression analysis by massively parallel signature sequencing (mpss) on microbead arrays. *Nat Biotechnol*, 18(6):630–4, (2000).

Cannon S. B., 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 S. J., Schoof H., Mayer K. F. X., Rogers J., Quetier F., Oldroyd G. E., Debelle F., Cook D. R., Retzel E. F., Roe B. A., Town C. D., Tabata S., Van de Peer Y., Young N. D.: Legume genome evolution viewed through the medicago truncatula and lotus japonicus genomes. *Proc Natl Acad Sci U S A*, 103(40):14959–14964, (2006).

Chen Y., Dougherty E., Bittner M.: Ratio-based decisions and the quantitative analysis of cdna microarray images. *Journal of Biomedical Optics*, 2:364–373, (1997).

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*, 7(272).

Chisholm R. L., Gaudet P., Just E. M., Pilcher K. E., Fey P., Merchant S. N., Kibbe W. A.: dictybase, the model organism database for dictyostelium discoideum. *Nucleic Acids Res*, 34(Database issue):D423–7, (2006).

Clausen J.: Persistent objects with o2dbi. Tech. Rep. 2002-01, Technische Fakultät, Universität Bielefeld (2002).

Cleveland W., Devlin S.: Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, (1988).

Dietzsch J., Gehlenborg N., Nieselt K.: Mayday–a microarray data analysis workbench. *Bioinformatics*, 22(8):1010–2, (2006).

Dill A., Sun T.: Synergistic derepression of gibberellin signaling by removing rga and gai function in arabidopsis thaliana. *Genetics*, 159(2):777–85, (2001).

Doll J., Hause B., Demchenko K., Pawlowski K., Krajinski F.: A member of the germin-like protein family is a highly conserved mycorrhiza-specific induced gene. *Plant Cell Physiol*, 44(11):1208–14, (2003).

Dondrup M., Albaum S. P., Griebel T., Henckel K., Jünemann S., Kahlke T., Kleindt C. K., Küster H., Linke B., Mertens D., Mittard-Runte V., Neuweger H., Runte K. J., 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*, 10:50, (2009a).

Dondrup M., Huser A. T., Mertens D., Goesmann A.: An evaluation framework for statistical tests on microarray data. *J Biotechnol*, 140(1-2):18–26, (2009b).

Dudoit S., Yang Y. H., Callow M. J., Speed T. P.: Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. statistia sinica, 12:111–139, (2002). *Statistia Sinica*, 12:111–139, (2002).

Durinck S., Moreau Y., Kasprzyk A., Davis S., Moor B. D., Brazma A., Huber W.: Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–40, (2005).

Dwinell M. R., Worthey E. A., Shimoyama M., Bakir-Gungor B., DePons J., Laulederkind S., Lowry T., Nigram R., Petri V., Smith J., Stoddard A., Twigger S. N., Jacob H. J., RGD Team: The rat genome database 2009: variation, ontologies and pathways. *Nucleic Acids Res*, 37(Database issue):D744–9, (2009).

Eddy S. R.: Profile hidden markov models. *Bioinformatics*, 14(9):755–63, (1998).

Ewing B., Green P.: Base-calling of automated sequencer traces using phred. ii. error probabilities. *Genome Res*, 8(3):186–94, (1998).

Ewing B., Hillier L., Wendl M. C., Green P.: Base-calling of automated sequencer traces using phred. i. accuracy assessment. *Genome Res*, 8(3):175–185, (1998).

Fey P., Gaudet P., Curk T., Zupan B., Just E. M., Basu S., Merchant S. N., Bushmanova Y. A., Shaulsky G., Kibbe W. A., Chisholm R. L.: dictybase–a dictyostelium bioinformatics resource update. *Nucleic Acids Res*, 37(Database issue):D515–9, (2009).

Fey P., Gaudet P., Pilcher K. E., Franke J., Chisholm R. L.: dictybase and the dicty stock center. *Methods Mol Biol*, 346:51–74, (2006).

Forment J., Gilabert F., Robles A., Conejero V., Nuez F., Blanca J. M.: Est2uni: an open, parallel tool for automated est analysis and database creation, with a data mining web interface and microarray expression data integration. *BMC Bioinformatics*, 9:5, (2008).

Frenzel A., Manthey K., Perlick A. M., Meyer F., Pühler A., Küster H., Krajinski F.: Combined transcriptome profiling reveals a novel family of arbuscular mycorrhizal-specific medicago truncatula lectin genes. *Mol Plant Microbe Interact*, 18(8):771–82, (2005).

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. *Molecular & Cellular Proteomics*, 6(12):2165–2179, (2007).

Gautier L., Cope L., Bolstad B. M., Irizarry R. A.: affy–analysis of affymetrix genechip data at the probe level. *Bioinformatics*, 20(3):307–15, (2004).

Girard L. R., Fiedler T. J., Harris T. W., Carvalho F., Antoshechkin I., Han M., Sternberg P. W., Stein L. D., Chalfie M.: Wormbook: the online review of caenorhabditis elegans biology. *Nucleic Acids Res*, 35(Database issue):D472–5, (2007).

Grennan A. K.: Genevestigator. facilitating web-based gene-expression analysis. *Plant Physiol*, 141(4):1164–6, (2006).

Grunwald U., Nyamsuren O., Tamasloukht M., Lapopin L., Becker A., Mann P., Gianinazzi-Pearson V., Krajinski F., Franken P.: Identification of mycorrhiza-regulated genes with arbuscule development-related expression profile. *Plant Mol Biol*, 55(4):553–66, (2004).

Harris T. W., Antoshechkin I., Bieri T., Blasiar D., Chan J., Chen W. J., De La Cruz N., Davis P., Duesbury M., Fang R., Fernandes J., Han M., Kishore R., Lee R., Müller H.-M., Nakamura C., Ozersky P., Petcherski A., Rangarajan A., Rogers A., Schindelman G., Schwarz E. M., Tuli M. A., Van Auken K., Wang D., Wang X., Williams G., Yook K., Durbin R., Stein L. D., Spieth J., Sternberg P. W.: Wormbase: a comprehensive resource for nematode research. *Nucleic Acids Res*.

Harris T. W., Stein L. D.: Wormbase: methods for data mining and comparative genomics. *Methods Mol Biol*, 351:31–50, (2006).

Henckel K., Küster H., Stutz L., Goesmann A.: Mediplex, a tool to combine in silico & experimental gene expression profiles in the model legume medicago truncatula (2010), submitted.

Henckel K., Runte K. J., Bekel T., Dondrup M., Jakobi T., Kuster H., Goesmann A.: Truncatulix–a data warehouse for the legume community. *BMC Plant Biol*, 9:19, (2009).

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*, 33(8):737–748, (2006).

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.*, 137:1283–1301, (2005).

Huang X., Madan A.: Cap3: A dna sequence assembly program. *Genome Res*, 9(9):868–877, (1999).

International HapMap Consortium: The international hapmap project. *Nature*, 426(6968):789–96, (2003).

International HapMap Consortium: Integrating ethics and science in the international hapmap project. *Nat Rev Genet*, 5(6):467–75, (2004).

International HapMap Consortium: A haplotype map of the human genome. *Nature*, 437(7063):1299–320, (2005).

International HapMap Consortium: A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–61, (2007).

Irizarry R. A., Hobbs B., Collin F., Beazer-Barclay Y. D., Antonellis K. J., Scherf U., Speed T. P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–64, (2003).

Javot H., Penmetsa R. V., Terzaghi N., Cook D. R., Harrison M. J.: A medicago truncatula phosphate transporter indispensable for the arbuscular mycorrhizal symbiosis. *Proc Natl Acad Sci U S A*, 104(5):1720–5, (2007).

Journet E.-P., van Tuinen D., Gouzy J., Crespeau H., Carreau V., Farmer M.-J., Niebel A., Schiex T., Jaillon O., Chatagnier O., Godiard L., Micheli F., Kahn D., Gianinazzi-Pearson V., Gamas P.: Exploring root symbiotic programs in the model legume medicago truncatula using est analysis. *Nucleic Acids Res*, 30(24):5579–5592, (2002).

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*, 308:1786–1789, (2005).

Kanehisa M., Goto S.: Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, (2000).

Kimball R., Caserta J.: *Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data*. 0-7645-6757-8, John Wiley (2004).

Kimball R., Margy R.: *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling.*. ISBN 0471200247, John Wiley & Sons Inc (2002).

Knippers R.: *Molekulare Genetik*, vol. 9. Thieme, Stuttgart (2006).

Kreppel L., Fey P., Gaudet P., Just E., Kibbe W. A., Chisholm R. L., Kimmel A. R.: dictybase: a new dictyostelium discoideum genome database. *Nucleic Acids Res*, 32(Database issue):D332–3, (2004).

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.*, 68(1):19–32, (2007).

Laule O., Hirsch-Hoffmann M., Hruz T., Gruissem W., Zimmermann P.: Web-based analysis of the mouse transcriptome using genevestigator. *BMC Bioinformatics*, 7:311, (2006).

Li C., Wong W. H.: Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*, 98(1):31–6, (2001).

Limpens E., Bisseling T.: Signaling in symbiosis. *Current Opinion in Plant Biology*, 6(4):343–350, (2003).

Lipman D. J., Pearson W. R.: Rapid and sensitive protein similarity searches. *Science*, 227(4693):1435–1441, (1985).

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 Research*, 31(1).

Mulder N., Apweiler R.: Interpro and interproscan: tools for protein sequence classification and comparison. *Methods Mol Biol.*, 396:59–70, (2007).

Nagaraj S. H., Deshpande N., Gasser R. B., Ranganathan S.: Estexplorer: an expressed sequence tag (est) assembly and annotation platform. *Nucleic Acids Res*, 35(Web Server issue):W143–7, (2007).

O'Connell K.: There's no place like wormbase: an indispensable resource for caenorhabditis elegans researchers. *Biol Cell*, 97(11):867–72, (2005).

Okubo K., Hori N., Matoba R., Niiyama T., Fukushima A., Kojima Y., Matsubara K.: Large scale cdna sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genetics*, 2(3):173–179, (1992).

Parkinson H., Kapushesky M., Kolesnikov N., Rustici G., Shojatalab M., Abeygunawardena N., Berube H., Dylag M., Emam I., Farne A., Holloway E., Lukk M., Malone J., Mani R., Pilicheva E., Rayner T. F., Rezwan F., Sharma A., Williams E., Bradley X. Z., Adamusiak T., Brandizi M., Burdett T., Coulson R., Krestyaninova M., Kurnosov P., Maguire E., Neogi S. G., Rocca-Serra P., Sansone S.-A., Sklyar N., Zhao M., Sarkans U., Brazma A.: Arrayexpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res*, 37(Database issue):D868–72, (2009).

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.*, 35:747–50, (2007).

Parkinson H., Sarkans U., Shojatalab M., Abeygunawardena N., Contrino S., Coulson R., Farne A., Lara G. G., Holloway E., Kapushesky M., Lilja P., Mukherjee G., Oezcimen A., Rayner T., Rocca-Serra P., Sharma A., Sansone S., Brazma A.: Arrayexpress–a public repository for microarray gene expression data at the ebi. *Nucleic Acids Res*, 33(Database issue):D553–5, (2005).

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*, 22(19(5)):651–2, (2003).

Quackenbush J.: Microarray data normalization and transformation. *Nat Genet*, 32 Suppl:496–501, (2002).

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.*, 29:159–164, (2001).

R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2008), ISBN 3-900051-07-0.

Rayner T. F., Rocca-Serra P., Spellman P. T., Causton H. C., Farne A., Holloway E., Irizarry R. A., Liu J., Maier D. S., Miller M., Petersen K., Quackenbush J., Sherlock G., Stoeckert C. J., Jr, White J., Whetzel P. L., Wymore F., Parkinson H., Sarkans U., Ball C. A., Brazma A.: A simple spreadsheet-based, miame-supportive format for microarray data: Mage-tab. *BMC Bioinformatics*, 7:489, (2006).

Runte K.: Sophia. *in preparation.*

Sásik R., Calvo E., Corbeil J.: Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. *Bioinformatics*, 18(12):1633–40, (2002).

Schena M., ed.: *Microarray Analysis*, vol. 1. Wiley & Sons (2002).

Schena M., Davis R. W.: Hd-zip proteins: members of an arabidopsis homeodomain protein superfamily. *Proc Natl Acad Sci U S A*, 89(9):3894–8, (1992).

Schena M., Heller R. A., Theriault T. P., Konrad K., Lachenmeier E., Davis R. W.: Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol*, 16(7):301–6, (1998).

Schwarz E. M., Antoshechkin I., Bastiani C., Bieri T., Blasiar D., Canaran P., Chan J., Chen N., Chen W. J., Davis P., Fiedler T. J., Girard L., Harris T. W., Kenny E. E., Kishore R., Lawson D., Lee R., Müller H.-M., Nakamura C., Ozersky P., Petcherski A., Rogers A., Spooner W., Tuli M. A., Van Auken K., Wang D., Durbin R., Spieth J., Stein L. D., Sternberg P. W.: Wormbase: better software, richer content. *Nucleic Acids Res*, 34(Database issue):D475–8, (2006).

Seddas P., Gianinazzi-Pearson V., Schoefs B., Küster H., Wipf D.: *Plant-Environment Interactions, Signaling and Communication in Plants*, chap. Communication and Signaling in the Plant-Fungus Symbiosis: The Mycorrhiza., pages 45–71. Springer-Verlag, Berlin Heidelberg (2009).

Siegel S.: *Nonparametric statistics for the behavioral sciences.*. McGraw-Hill (1956).

Smedley D., Haider S., Ballester B., Holland R., London D., Thorisson G., Kasprzyk A.: Biomart–biological queries made easy. *BMC Genomics*, 10:22, (2009).

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*, 308:1789–1791, (2005).

Smyth G. K.: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1):Article 3, (2004).

Smyth G. K.: *Bioinformatics and Computational Biology Solutions using R and Bioconductor, chap. Limma: linear models for microarray data, pages 397-420*. Springer, New York (2005).

Smyth G. K., Speed T.: Normalization of cdna microarray data. *Methods*, 31(4):265–73, (2003).

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.*, 6:175–82, (1998).

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 Jr C., Brazma A.: Design and implementation of microarray gene expression markup language (mage-ml). *Genome Biology*, 3:research0046.1–0046.9, (2002).

Stekel D. J., Git Y., Falciani F.: The comparison of gene expression from multiple cdna libraries. *Genome Res*, 10(12):2055–2061, (2000).

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*, 11(4):41, (2003).

Thompson R., Ratet P., Küster H.: Identification of gene functions by applying tilling and insertional mutagenesis strategies on microarray-based expression data. *Grain Legumes*, 41:20–22, (2005).

Tusher V. G., Tibshirani R., Chu G.: Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, 98(9):5116–21, (2001).

Twigger S. N., S Smith J., Zuniga-Meyer A., Bromberg S. K.: Exploring phenotypic data at the rat genome database. *Curr Protoc Bioinformatics*, Chapter 1:Unit 1.14, (2006).

Velculescu V. E., Zhang L., Vogelstein B., Kinzler K. W.: Serial analysis of gene expression. *Science*, 270(5235):484–7, (1995).

Vencio R. Z. N., Varuzza L., de B Pereira C. A., Brentani H., Shmulevich I.: Simcluster: clustering enumeration gene expression data on the simplex space. *BMC Bioinformatics*, 8:246, (2007).

Whetzel P. L., Parkinson H., Causton H. C., Fan L., Fostel J., Fragoso G., Game L., Heiskanen M., Morrison N., Rocca-Serra P., Sansone S.-A., Taylor C., White J., Stoeckert C. J., Jr: The mged ontology: a resource for semantics-based description of microarray experiments. *Bioinformatics*, 22(7):866–73, (2006).

Wu Z., Irizarry R., Gentleman R., Murillo F., Spencer F.: A model based background adjustment for oligonucleotide expression arrays. *John Hopkins University, Department of Biostatistics Working Papers, Baltimore*, (1).

Wulf A., Manthey K., Doll J., Perlick A. M., Linke B., Bekel T., Meyer F., Franken P., Küster H., Krajinski F.: Transcriptional changes in response to arbuscular mycorrhiza development in the model plant medicago truncatula. *Mol Plant Microbe Interact*, 16(4):306–14, (2003).

Yang Y., Buckley M., Speed T.: Analysis of cdna microarray images. *Brief Bioinform*, 2(4):579–588, (2001).

Yang Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J., Speed T. P.: Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):e15, (2002).

Yin W., Chen T., Zhou S. X., Chakraborty A.: Background correction for cdna microarray images using the tv+l1 model. *Bioinformatics*, 21(10):2410–6, (2005).

Young N. D., Cannon S. B., Sato S., Kim D., Cook D. R., Town C. D., Roe B. A., Tabata S.: Sequencing the genespaces of medicago truncatula and lotus japonicus. *Plant Physiol*, 137(4):1174–1181, (2005).

Zimmermann P., Hennig L., Gruissem W.: Gene expression analysis and network discovery using genevestigator. *Trends in Plant Science*, 9(10):407–409, (2005).

Zimmermann P., Hirsch-Hoffmann M., Hennig L., Gruissem W.: Genevestigator. arabidopsis microarray database and analysis toolbox. *Plant Physiol*, 136(1):2621–32, (2004).

Zimmermann P., Laule O., Schmitz J., Hruz T., Bleuler S., Gruissem W.: Genevestigator transcriptome meta-analysis and biomarker search using rice and barley gene expression databases. *Mol Plant*, 1(5):851–7, (2008).

Acknowledgments

First of all, I wish to thank Prof. Dr. Helge Küster and Prof. Dr. Ralf Hofestädt for the opportunity to do my PhD thesis under their supervision. Their constant support and knowledge helped me to finish my project and this thesis.

I also thank Dr. Alexander Goesmann for help and advice in diverse bioinformatic questions. I thank all the people of the Computational Genomics group and BRF for the assistance in the different projects and the collaboration. Special thanks go to my office colleague Thomas Bekel who made long times short.

I would like to acknowledge the International Graduate School in Bioinformatics and Genome Research for funding the PhD project, the executive directors Dr. Dirk Evers and Dr. Susanne Schneiker-Bekel, and all PhD students of the graduate school for their support.

Thanks go to Tobias Jakobi, it was a pleasure to supervise your Bachelor thesis about TRUNCATULIX. For the help with the Sophia data warehouse backend and the IgetDB frontend I thank Kai Runte.

I wish to thank all friends I made at my time at University, especially Sebastian Burgemeister, Jochen Blom, Alexander Bunkowski, Christian Loyek, Johannes Huth, Bernhard Witte, Patrice Weisemann, and Sascha Winter. The times would have been less fun and filled up with too much work without them.

I am very grateful to my parents for their love and trust and the support throughout my studies.

Especially I would like to thank my wife Kathrin for her support over the whole time. Her constant assistent gave me the will to finish this work. Charlotte and Kian always put a smile on my face, even if sleep was short and time was running out.

Bielefeld, January 2010

Kolja Henckel

ERKLÄRUNG

Ich, Kolja Henckel, erkläre hiermit, dass ich die Dissertation selbständig erarbeitet und keine anderen als die in der Dissertation angegebenen Hilfsmittel benutzt habe.

Bielefeld, den 19. Januar 2010

Kolja Henckel