Statistical application for DNA microarray data analysis

Bellina Teixeira, Joel Arrais, José Luís Oliveira

Resumo **– Os *microarrays* de ADN (ácido desoxirribonucleico) são usados para analisar a expressão de milhares de genes numa amostra ao mesmo tempo. Estas experiências geram muita informação que é preciso armazenar e processar com recurso a meios informáticos.**

Mind (Microarray Information Database) é uma aplicação web desenvolvida no grupo de bioinformática da Universidade de Aveiro que aborda o armazenamento e o processamento de dados de *microarrays*.

Neste artigo descreve-se o desenvolvimento de um módulo de análise de dados de experiências de *microarrays* de ADN para o Mind. Apresentam-se procedimentos e ferramentas de análise de dados, nomeadamente métodos estatísticos, descrevem-se aspectos próprios do desenvolvimento do módulo e o resultado final é exposto.

Abstract **– DNA (deoxyribonucleic acid) microarrays are used to analyze the expression of thousands of genes within a sample at the same time. These experiments generate a lot of information that is necessary to be stored and processed with the aid of computerized means.**

Mind (Microarray Information Database) is a web application developed at the bioinformatics group of the University of Aveiro that addresses the storage and processing of microarray data.

In this article the development of a DNA microarray data analysis module for Mind is described. Some methods and tools for microarray data analysis are presented, namely statistical methods, some aspect of the module development are described and the final result is exposed.

I. INTRODUCTION

DNA microarrays are a valuable tool in gene expression assessment as they can probe, at the same time, for the expression levels of thousands of genes within a sample. They have been used a lot to assess differential expression between two or more conditions under study. A typical example is drug testing: a sample of cells of an organism treated with a medication can be compared with an untreated sample to see what genes have different expression between both. Ideally, a drug should target the genes that cause the disease, while not interfering on the other ones. To learn about what genes are related to a disease, DNA can be used to compare a healthy with a diseased sample: the genes that are differentially expressed between both are likely to be due to the disease.

Mind (Microarray Information Database) is a web application project developed at the University of Aveiro to address storage and processing of microarray data [1]. It contains a LIMS (Laboratory Information Management System) for storage and a Data Analysis module for processing.

This project aims to integrate gene regulation assessment functionalities into Mind, in addition to the quality control functions it contained previously, to allow users to perform complete gene differential expression analysis within Mind, using their experimental data stored in the Mind LIMS as a starting point.

Also developed by the University of Aveiro bioinformatics group, GeneBrowser is a web application that, using a list of genes as input, is able to assist the user in their functional analysis [2]. As the new Data Analysis module generates lists of regulated genes, by providing a link to open these in GeneBrowser, in constitutes itself as a bridge between both web applications.

II. A DNA MICROARRAY EXPERIMENT

DNA microarrays measure mRNA present in the sample cells to infer about gene expression. The more expressed a gene is in a cell, at a given time, the more mRNA molecules are transcript of it for protein synthesis. In this technology, the mRNA is extracted from sample cells to infer about the expression of their genes, as these lead to their ultimate phenotype.

A microarray slide itself consists usually of a microscope-like glass slide with thousands of spots arranged in a grid-like fashion. Each spot contains many identical DNA sequences relative to a gene. A slide can contain spot replicates, meaning several spots that are relative to the same gene.

Microarray technology relies on hybridization between the mRNA of the samples and the DNA probes that were spotted or printed onto the slide, to measure gene expression in that sample. On a microarray slide, hybridization occurs between the mRNA and DNA sequences, which are complementary. On two-color arrays (Mind currently supports only two-color microarrays) two mRNA samples are labeled with different fluorescent dyes, usually green and red cyanines, and are applied over one same slide to hybridize. Then, the microarray is scanned as the light excites the fluorescent cyanines. Spots to which mostly the red sample has bound will glow red, spots to which mostly the green sample has bound will reveal a

green color, and spots with equal amounts of both will show a yellow color. Color intensities also vary, from bright colored (red, green or yellow) to black, depending on how much mRNA bound – genes are not only expressed or not, but rather more or less expressed. The pictures obtained are converted to a tab delimited text-file by a microarray image processing program. The text file contains as many rows as the spots of the microarray slide, and indicates, for each spot, the measured red and green foreground and background intensities (Figure 1). The raw data files, along with the array layout description file, which indicates the name of the gene each spot refers to, are the starting material for microarray data analysis.

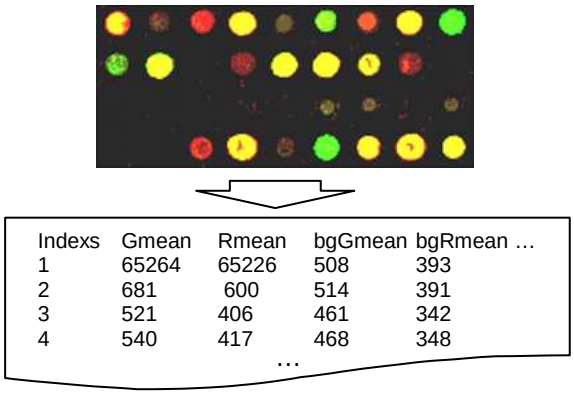

Figure 1 – From microarray image to raw data file

III. R LIBRARIES FOR MICROARRAY DATA ANALYSIS

The first step of microarray data analysis is quality control, which involves procedures to remove unwanted systematic biases from the data. DNA microarrays that analyze gene expression can be used to compare gene expression values between two or more samples or to learn about gene functions, how the work together and more; so quality control may be followed by gene regulation assessment analysis or by exploratory analysis. The objective of this project is the former – to add gene regulation analysis functions to Mind.

Both quality control and gene regulation assessment involve complex algorithms, which is why R libraries were used for the development of Mind. R is a statistical framework that allows data processing, generation of plots and more [3]. It has its own windows environment or it can be run from the operating system's command line. In both cases data processing and plot generation is done with the R scripting language. This makes R ideal to execute scripts from an application, which is what is done in Mind: the website site provides a GUI to the user, and using his input, processes the microarray data with R.

This software is open source and it can be downloaded from the Comprehensive R Archive Network website (CRAN). R base functions are the ones that come with the downloaded software and include arithmetic operations, plots, statistical methods and more. Additional libraries or packages can be downloaded and installed into R, such as libraries specific for the analysis of microarray data. Many bioinformatics related R libraries are released under the Bioconductor project. Of the R libraries used in Mind, that will now be introduced, Limma, Multtest and Maanova are part of the Bioconductor project, while Samr is not.

A. Limma library functions for quality control

The R library Limma, includes several functions to perform quality control on two-color array data [4]. For **background correction**, it offers the options: half, minimum, subtraction, norm exp, edwards and moving minimum. The experiment raw data files include, for each channel (red and green), a foreground and a background intensity. The background intensity measured is due to several factors such as such as natural reflection of the glass or optical noise from the scanner. The most simple and intuitive method is background subtraction, which subtracts the measured background intensities to the foreground intensities. However, this method may lead to negative spot intensities when the background is higher than the foreground, thus making it impossible to calculate the M-values for these spots. In two-color arrays, M-values are the expression values for each spot,

$$M - value = \log_2\left(\frac{R}{G}\right) = \log_2(R) - \log_2(G)$$

where R is the spot's red intensity and G the spot's green intensity. There are no logarithms of negative values, so M-values cannot be calculated if one of the intensities is negative. The half and minimum methods are similar to the subtractions methods, but handle specially spots that could otherwise lead to negative or zero intensities. The norm exp and edwards methods, more complex, take other factors into account.

Limma provides several options for **normalization**: median, robust spline, loess and print-tip loess. The main idea behind normalization is to approximate the M-values to zero. In an experiment, most genes must not be found differently expressed between the conditions under study. In a microarray slide, most spots should have their red and green intensities similar to each other, thus M-values close to zero. If spots in a slide are found to, for example, have expression values consistently lower or higher, these should be adjusted with normalization. Other factors can be taken into account, depending on the normalization method. These methods are explained further in the Limma User's Guide and Help manual. The result will be M-values that show most spots of a slide are not regulated – more spot M-values are closer to M = 0. Normalization affects the red and green intensities that lead to the M-values.

In Mind, quality control is performed on the data using R and Limma, with the parameters inputted by the user. Limma also generates different types of plots, personalized by user input (spot types information, filtering options and

choice of plots to generate). Mind uses these graphics to elaborate the quality control reports. For each raw data file, one normalized data file and one quality control report is produced.

Once the data is normalized, gene regulation assessment can begin. Unwanted unsystematic biases were minimized, so the biological variability and the truly regulated genes will stand out more clearly.

B. R libraries for gene regulation analysis

Gene regulation assessment should begin with **averaging spot replicates** in an array. Spots that refer to the same gene should have the average of their expression values taken. Otherwise, each spot will be treated as a distinct gene, even if they refer to the same one. In fact, simply taking the average may not be the best method, because it does not take into account variability among spot replicates, but at the moment it is the most commonly used method.

After spot replicate averaging, **filtering** may be performed, optionally. An intensity filter selects genes that show a minimum intensity value, on one or both channels, throughout a minimum number of arrays. A standard deviation filter selects genes based on a minimum standard deviation value throughout all the arrays. Filtering should not select too many or too little genes. On the group of genes that pass filtering, a statistical method is applied.

There are many statistical methods for gene regulation assessment, but a small selection is used on the vast majority of analysis done. As only a few methods could be chosen to implement in Mind at this time, a choice of commonly used methods was believed to be done: fold-change, t-test, ANOVA, Limma and SAM.

The **fold-change** method, when applied to a single array, compares the base two logarithms of the intensities of the control sample with the base two logarithms of the intensities of the experiment sample. In case of multiple arrays, the average of the log experiment intensities are compared with the average of the log control intensities. Genes that show a big difference between both are regulated. Figure 2 shows the fold-change method selecting genes with $\left|\log_2(E)-\log_2(C)\right|>2$

The **t-test** calculates a t-statistic and corresponding p-value for each gene, based on its experiment and control log intensities. If the t-statistics are calculated with the logged control and experiment intensities, the regulated genes are the same ones brought up by the fold-change method applied to multiple arrays, minus the ones that have higher p-values (Figure 3). However, in two-color array data, each spot's expression is more often given by its M-value. Each gene, that has spot replicates, has an M-value equal to the average of the spots' M-values. So in this case the t-test divides the array by the experiment and control groups and calculates a t-statistic for each gene based on the M-values it shows on each array (Figure 4).

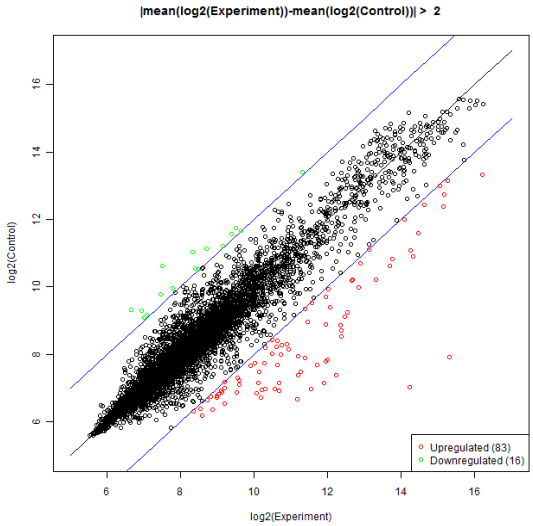

Figure 2 – Fold-change method illustrated with a scatter plot

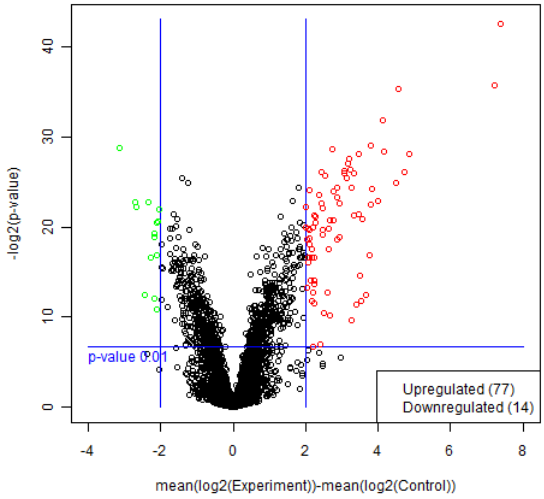

Figure 3- t-test illustrated with a volcano plot, t-statistics calculated based on logged control and experiment intensities

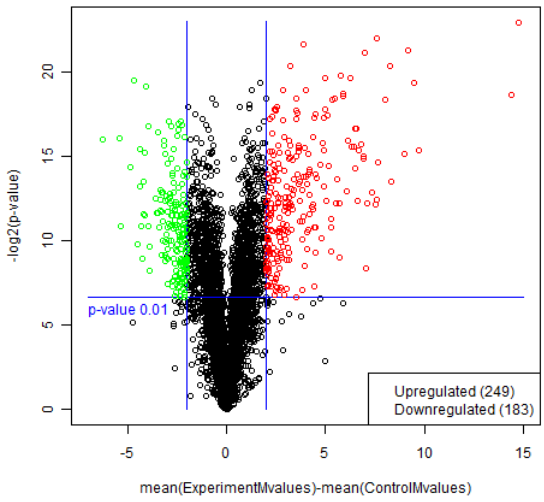

Figure 4 – t-test illustrated with a volcano plot, t-statistics calculated based on gene M-values in each array

The t-test is commonly used in its pair variant when each sample of the experiment group is related to a sample in the control group, for example, when studying a group of patients before and after being treated with a medication, the gene expression values of the control sample are paired with the values of the experiment sample.

The R library Multtest [5] calculates t-statistics to perform paired or unpaired t-tests. For unpaired t-tests, these can be performed assuming equal variances or not (Welch t-test). It also contains several multiple comparison procedures. When using the t-test to assess different assessment, one t-test is performed for each gene, to compare its expression in the two groups. Thousands of hypothesis tests are done for one experiment originating what is known in statistics as the multiple comparison issue. The Multtest is a library designed especially for microarray data analysis, and as so also includes multiple comparison correction procedures.

Just like the t-test is a classic statistical test for comparison of two groups, **ANOVA** or F-test is the counterpart for multiple group comparison. These tests have been found very useful for microarray data analysis, as long as the multi comparison issue is taken into account and correction procedures are applied to the raw p-values.

Both libraries Multtest and Maanova [6] offer functionalities to add the ANOVA test to Mind, including calculation of the F-statistics, the p-values and procedures to adjust the p-values (although the procedures offered by these libraries are different). As both libraries are appropriate for the functionality that was desired to add to Mind, Maaanova was selected.

Limma, although it is an R library, it is commonly referred to as being a methodology for gene regulation assessment, as one can say "perform gene expression analysis using Limma". Limma's approach involves using a design matrix to model the systematic part of the experimental data and two group comparison with empirical Bayes moderated t-statistics [7]. After the user defines the targets information of the data which describes, for each array, what mRNA sample was hybridized to each channel, Limma can compare the groups using the mRNA sample names and provide a top list of regulated genes.

SAM is also a favorite method for gene differential expression analysis and it was selected to be added to Mind in the two-class, two-class paired and multi-class variants, using the library Samr [8]. After defining the groups, the user can specify a delta value to select regulated genes.

Of these methods, the fold-change is the only one that is not truly statistical, because it does not take variability of gene expression values across arrays into account and it can be performed even when there is only one array. However, they can and are usually all referred to as "statistical methods" for practical reasons ("gene regulation assessment methods" is a name a bit too long for a menu item). After applying a statistical method, one can finally obtain a list of regulated genes. This list cannot however be considered without the opinion of the user. First a statistical method does not say for sure what genes are regulated and which ones are not: they just assign a confidence value for declaring each gene as regulated. Second, a statistical method may output a list of regulated genes even when the input does not makes sense. For example, it was tried to apply the ANOVA to three biologically identical groups, which should not show significant differences among them, and the method still selected 15 of 6,342 genes, each one having at a confidence of 99% or more (raw p-values lower than 0.01). The M-values for these genes for three groups ended up being all in the [-0,5 ; 0.5] range. Regardless of the test, the user should always have an intuitive look at the data involved and consider the biological conditions under study before obtaining a conclusion.

The quality control and gene regulation assessment methods described are common in many data analysis software, and recommended by the Mind users, so they constitute a selection of popularly used methods that to be offered in Mind.

IV. APPLICATION DEVELOPMENT AND OVERVIEW

The Mind web application is a Microsoft Visual Studio 2008 ASP .NET and C# web site project and the Mind data base a Microsoft SQL 2008 database (initially developed with the 2005 tools). Mind contains several modules, among them LIMS and Data Analysis, and the entire web application uses the database (Figure 5).

This projects works on the Data Analysis module, which in addition to using the database, requires some R scripts. The R scripts are part of the development, and contain R code to perform quality control and gene regulation assessment, using the user's input. The R scripts for quality control were already created as of the beginning of this project, to apply the Limma quality control functions to the data, and were found satisfactory by the Mind users. Additional scripts were written for spot replicate averaging, filtering and the statistical tests.

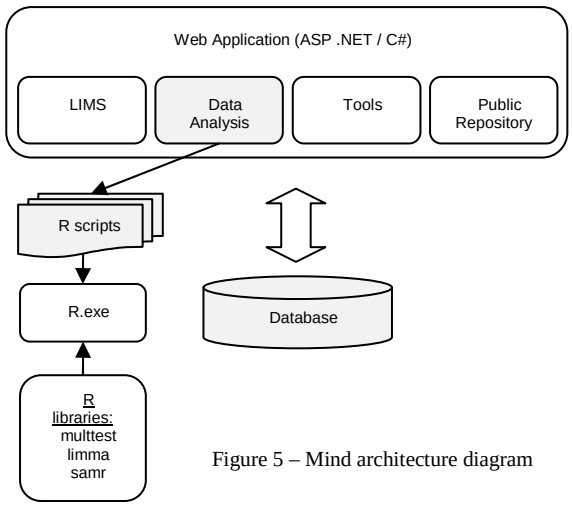

Figure 5 – Mind architecture diagram

The new Mind Data Analysis module proposed and implemented contains three sub-modules: Data Set, Quality Control and Gene Regulation (Figure 6).

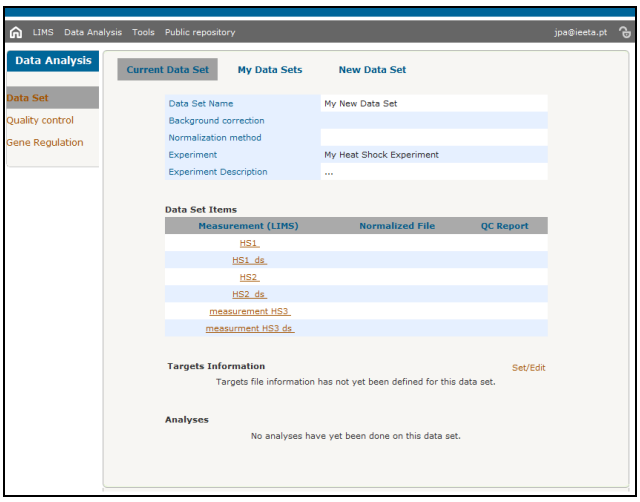

Figure 6 – Application Data Analysis module overview

In the Data Set sub-module, data sets can be created and managed. In microarray data analysis, a data set is commonly a selection of files on which quality control and/or gene regulation assessment analysis is performed. In Mind, it is not only a selection of data, but also an entity to which all analysis data and reports are associated to. Data sets can be created, viewed and deleted by the user in the Data Set sub-module.

Using the data analysis module requires the user to have his experiment stored in the Mind LIMS. If the user has no experiments saved, he will have to upload one to the Mind LIMS in order to create a data set necessary for data analysis. A data set is relative to an experiment and multiple data sets can be created based on one experiment. Each data set contains a selection of raw data files of that experiment. A data set can be defined as the Current Data Set, either upon creation, either choosing from the list of created data sets.

In the Quality Control sub-module, the user performs quality control on the Current Data Set. He can define the spot types (information about the colors to use in the Limma plots), the filtering, the background correction, the normalization and the plot generation options, or he can retrieve previously saved settings (Figure 7).

After the parameters are defined, quality control is run, in the background and all measurements in parallel. Quality control handles each raw data file individually; therefore they can be normalized in parallel. Enabling parallel processing allows big gains relatively to perform the processing sequentially, and these gains are increased with the hardware characteristics of the server computer, namely the number of processing cores. Besides being parallel, processing is done in the background with the aid of AJAX, which is now embedded into the Visual Studio 2008 IDE. The Quality Control sub-module GUI consists of a panel with five tabs to define all the parameters or a list of saved parameters to select one. Each time quality control is run with defined parameters, these can be saved so they are available for future analyses. Once quality control processing is done, the user is notified that the normalized data and report files are available under the Current Data Set. The user can view them, re-run quality control if necessary, and finally proceed to gene regulation assessment.

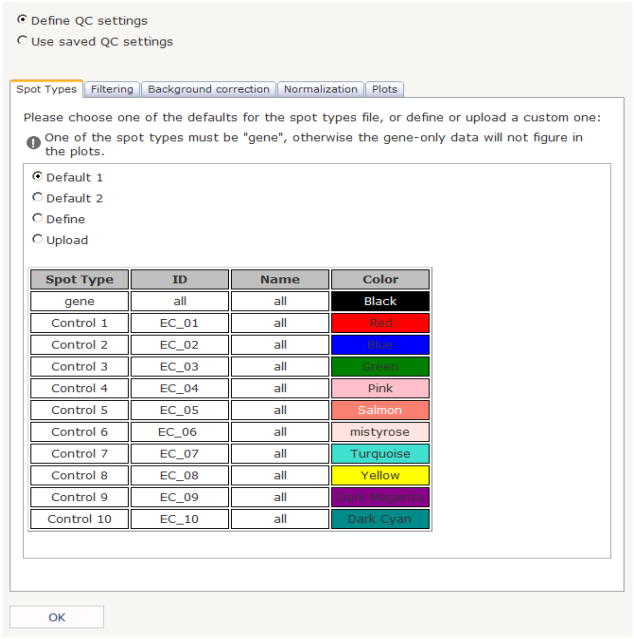

Figure 7 - Quality Control: define parameters

The gene-regulation sub-module starts by loading the normalized data files, of the Current Data Set, by averaging spot replicates or not. In a microarray slide, several spots can refer to the same gene, so spot replicate averaging, which averages the spot expression values according to the gene name (indicated by the experiment's array layout file stored in the LIMS).

Next, optionally, follows filtering to select a group of genes to run the statistical test on.

On the selected genes (or on all the genes, if no filtering was done) a statistical method is applied to output a list of regulated genes. The choices are: fold-change, t-test (paired or unpaired), Limma, SAM (two-class unpaired, two-class paired and multi-class) and one-way ANOVA. The unpaired t-test can be done considering the ratio log fold-change or the difference log fold-change. All these tests start by defining the groups, associating the measurements or bioassays of the data set to each group, except for Limma. Limma analyses the samples using the targets information, the names of the mRNA samples associated to the red and green channels of each bioassay. The targets information is defined for the data set, as it is not only mandatory for Limma, but helpful for group definition in the other tests. After the groups are defined (two groups for fold-change, t-test and SAM two-class, two paired groups for t-test paired and SAM two-class paired and several groups for ANOVA and SAM

multi-class), the parameters specific to the statistical test are defined and the test is run. Figure 8 shows the input insertion to perform the t-test, right after the groups have been defined. When the test is complete, a plot showing the number of the selected genes is shown, and the user can re-run the test with different parameters, or even choose a different test. Each time a test is performed, it will be on the filtered data. Once an adequate number of genes are declared regulated by a statistical test, the user can save the analysis, which will be available under the Current Data Set. For each gene regulation analysis run, the Current Data Set contains links to the data and report files, and a link to view the regulated genes in Gene Browser (Figure 9).

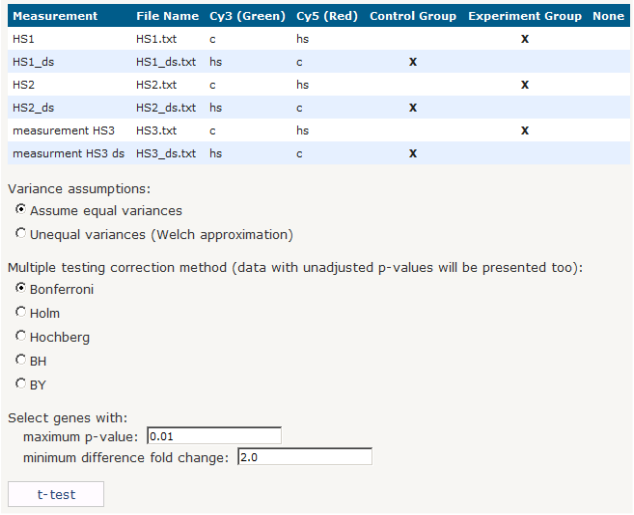

Figure 8 – Gene Regulation: insert t-test parameters

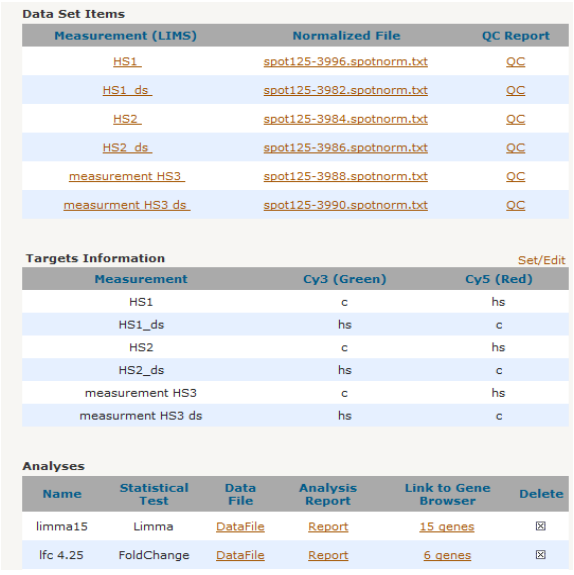

Figure 9 – Data Set: Current Data Set, with Quality Control and Gene Regulation results

Only the last quality control data is saved, so each analysis report also registers the background correction and normalization methods used to obtain the normalized data from the raw data files (spot types, filtering and plot choices do not affect the data). The saved gene regulation analysis of a data set can be deleted by the user when no longer needed.

Several tables were added to the database to support the new data set entity created and the association of gene regulation analysis and targets information to the data set. Allowing the user to save and retrieve quality control parameters also needs a new table, as does registering the user's Current Data Set.

V. CONCLUSION

The proposed Data Analysis module was developed, dedicated to two-array data processing, featuring several quality control functions, offered through the Limma R library, and several statistical tests for gene regulation analysis offered through several R libraries.

The Data Analysis module developed allows the user to perform gene regulation assessment analysis, from the raw data files and array description layout file to a list of regulation genes. With this module, the user can perform a complete analysis, within the Mind application, of his experimental data stored in the LIMS. Each analysis done and saved contains a link that opens GeneBrowser in a new browser tab or window, loaded with the regulated genes, so the user can perform subsequent functional analysis.

REFERENCES

1. Arrais, J.P., et al., *A Microarray Information Database*, in *Biocomputation, Bioinformatics, and Biomedical Technologies, 2008. BIOTECHNO '08. International Conference on*. 2008: Bucharest.
2. Arrais, J., et al., *GeneBrowser: an approach for integration and functional classification of genomic data*. Journal of Integrative Bioinformatics, 2007.
3. R Development Core Team *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. 2009, Vienna, Austria.
4. Smyth, G.K. and T. Speed, *Normalization of cDNA microarray data.* Methods, 2003. **31**(4): p. 265-73.
5. Pollard, K.S., et al., *multtest: Resampling-based multiple hypothesis testing.* R package version 2.1.1. http://CRAN.R-project.org/package=multtest.
6. Wu, H., et al., *Maanova: A software package for the analysis of spotted cdna microarray experiments.*, in *The Analysis of Gene Expression Data*. 2003, Springer London. p. 313-341.
7. Smyth, G.K., *Linear models and empirical bayes methods for assessing differential expression in microarray experiments.* Stat Appl Genet Mol Biol, 2004. **3**: p. Article3.
8. Tibshirani, R., et al., *samr: SAM: Significance Analysis of Microarrays.* R package version 1.26, http://www-stat.stanford.edu/~tibs/SAM.