Stockholms universitet

Microarray Data Analysis of Dyslexia Candidate Genes

Sini Kilpeläinen

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2008:1
Matematisk statistik
December 2008

www.math.su.se/matstat

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2008:1**
http://www.math.su.se

Microarray Data Analysis of Dyslexia Candidate Genes

Sini Kilpeläinen*

December 2008

Abstract

The aim of this project is to identify downstream target genes and pathways of dyslexia candidate genes (DCGs) in a cell model. The microarray data is first controlled and preprocessed by means of quality control and normalisation of the arrays. Then a linear model is fitted to the log-intensity expression values and parameters are estimated for contrasts of the treated samples against the controls. The most significant genes are listed according to different statistics and expression measures and these are illustrated with different kinds of plots. The statistical analysis of the microarray data is performed using the statistical software R as implemented in the Bioconductor packages.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: sini.kilpelainen@gmail.com . Supervisor: Ola Hössjer.

Sammanfattning

Syftet med det här projektet är att identifiera nedströmsgener och signalvägar för kandidatgener för dyslexi (DCG) i en cellmodell.

Microarraydata är först kontrollerad och förbehandlad genom kvalitetskontroll och normalisering av arrayerna. Sedan anpassas en lineär modell till data, givna i form av log-intensiteter för genuttryck, och parametrarna för olika kontraster mellan behandling och kontroll skattas. De mest signifikanta generna listas med hjälp av olika statistiska mått och dessa illustreras med olika plottar.

Statistisk analys av microarraydata genomförs i R med hjälp av Bioconductorpaketet.

Preface

I have completed my degree thesis with a great enthusiasm and a growing interest for biostatistics. The person who gave me an opportunity to write this thesis was Juha Kere, a Professor in Molecular Genetics at the Department of Biosciences and Nutrition, who I contacted via University of Helsinki. I got a chance to visit Juha's research group in Novum, Karolinska Institute, where they work to find genes behind endemic diseases. I want to thank Juha for giving me a chance to write my thesis for them.

I am very grateful to my supervisors at Novum, MSc Kristiina Tammimies, and PhD Per Unneberg, who stand behind the design of the projekt and the data and articles concerning gene expression analysis. They have been a great support during the whole period. I also want to thank my supervisor at the Division of Mathematical Statistics, Ola Hössjer, and the system manager, Tomas Ericsson!

I think microarrays offer an interesting area for statisticians because it perfectly combines both biology and bioinformatics and by implementing statistical methods microarray experiments will result in interesting discoveries.

Working with this project has been truly interesting and I am glad that I have learned how to process a microarray analysis. Still, this area is very extensive, and my thesis will only give a simple overview of the microarray data analysis. Microarray analysis requires some understanding of biology which made this project even more challenging for me who doesn't have any biological background. I hope that this thesis will give an approach to microarray analysis from a statistical point of view but also that it gives the reader an understanding of the underlying biology.

Stockholm – 2008
Sini Kilpeläinen

Table of contents

1 Introduction

This first chapter will give a reader an overview of the thesis by introducing the background for microarray analysis and defining the aim of this project. The analysis of the microarray experiment is performed using the statistical software package R (http://www.r-project.org/), and by implementing the Bioconductor packages (http://www.bioconductor.org/). References will be marked with [ ] in the text, and R-commands are typeset in *italics*.

1.1 Biological background

1.1.1 Gene expression

A cell stores its genetic information in a DNA molecule, which contains genes. Depending on the function of the cell, it uses different genes to make proteins by copying the code of the gene into messenger RNA (mRNA) in a procedure called transcription. This is illustrated in Figure 1.

A transient transfection of specific genes into a cell line can be used to test the function of these particular genes. Transfection means an introduction of DNA into the cell line, where these specific genes are expressed only a short period of time.

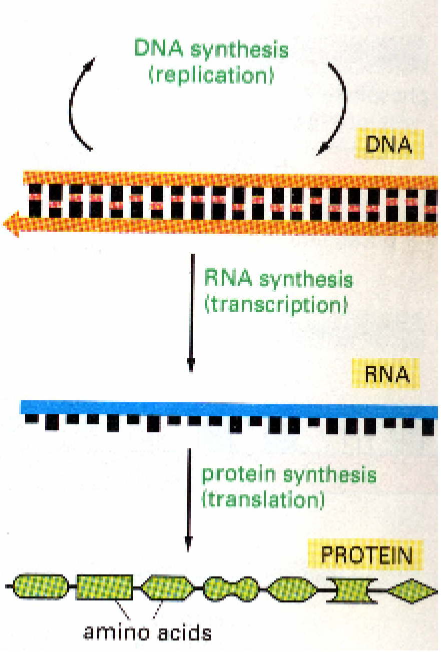


Figure 1. From DNA to protein. [I:9]

1.1.2 Dyslexia and genetics

Dyslexia is a complex disorder defined as unexpected difficulty in learning to read, despite adequate education, intelligence, social environment and normally functioning senses. Dyslexia, also called "word blindness", is one of the most common neuro-developmental disorders and it affects around 5-10% of the population, most often school-age children.

Genetic studies have identified chromosomes including 1, 2, 3, 6, 11, 15, 18 and X that are linked to dyslexia by using linkage and association studies. In these regions of linkage so far ten genes have been associated with dyslexia, *DYX1C1, DCDC2, KIAA0319, ROBO1, MRPL19, C20RF3, PCNT, DIP2A, S100B* and *PRMT2.* [II]

1.2 Gene expression microarray

Most of the cells in the human body contain the same genes, but not all of the genes are used in each cell. Some genes are active, or "expressed", when needed. To understand how the cells work and which genes are active or inactive in different kinds of cells, a gene expression microarray technology can be used. Microarrays allow to examine thousands of genes at the same time, and helping identify genes that are expressed in different cells and find relationships between individual genes.
In molecular biology and medical research, microarray technology is used to understand and learn more about different diseases.

DNA oligonucleotides that correspond to different genes are placed on a single microscope slide, which is called microarray. mRNA that is then extracted from cells synthesized to complementary DNA by the enzyme reverse transcriptase and labeled with fluorescent tags is then hybridized on the slides. The scanner measures the brightness of fluorescence of each sample on the slide, and this brightness will help determining if the genes are active in the cell. Genes that are expressed differentially between diseased and healthy human arrays may be involved in causing the disease.

1.2.1 Affymetrix expression arrays

Expression arrays are used to monitor the expression of thousands of genes simultaneously to study certain treatment, disease or developmental stage. The Affymetrix platform uses a oligonucleotide based system to detect expression.

Each gene is represented with several (1-30) probes and each probe is 25-60 base-pairs long . The probes spotted on the arrays are short oligonucleotides designed to match the known or predicted open reading frames (genes), see Figure 2.

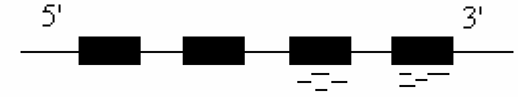


Figure 2. Boxes represent the part of transcript that is translated into protein. In the Human Genome U133A plus 2.0 array the probes are designed to match the 3' end of the open reading frame (lines below boxes).

1.2.2 Microarray Data Analysis

The analysis is started by describing the theory of microarray experiments, and then proceeded with the statistical part and how to process it in R. Furthermore the same experiment will be easily reprocessed.

The statistical analysis can be divided into several steps:

Visualization of the data is done first to get a better picture of what data looks like. Plotting the data helps identifying problem areas and to determine if data reduction is necessary. Plots will also help to follow and understand changes after each step of processing the data.

Quality control is an important part in the beginning of the process, where the quality of the data is controlled by different methods in the R-package called *affy*.

Normalization of data is necessary for every microarray experiment. It is done to correct for systematic technical differences and to reduce their variation.

Data analysis is done by fitting a linear model to data, formulating hypotheses of interesting contrasts and processing t-statistics and B-statistics implemented in a package called *limma*.

Identification of differentially expressed genes is probably the most interesting part of the project. Most significant genes are ranked from the above mentioned statistics and illustrated with different kinds of plots and diagrams.

Clustering will help to identify similar samples and group them into clusters to illustrate possible patterns between gene expression levels.

Classification and biological interpretation is the last step of the project for understanding the biological conclusion of the experiment.

The following packages must be loaded before the analysis can be carried out.

```
> library(affy)
> library(limma)
> library(simpleaffy)
> library(hgu133plus2)
> library(aroma)
> library(MASS)
```

1.3 Aim of the project

The aim of this project is to identify differentially expressed genes by comparing log intensity expression values of control cells and cells that received a treatment of overexpression of different dyslexia candidate genes.
The experiment is based on gene expression microarray data that will be preprocessed, normalized and analysed statistically. The high dimension of the data and a huge amount of genes with different variances will make the experiments statistically challenging.
Identifying differentially expressed genes will give a clue that the genes might have some kind of biological connection to the dyslexia candidate genes, which will lead to further studies and confirmation.

2 Data

2.1 Explaining data

The human neuroblatoma SH-SY5Y cell line was transiently transfected with plasmids containing three different dyslexia candidate genes, here Gene1, Gene2 and Gene3, and empty plasmid vectors as controls. The transfections were made at different times, Gene1 and the first control were taken in September 2007, and both Gene2 and Gene3 and the other control in March 2008. All samples are in triplicates (see Table 1).
These three genes have been chosen from the ten dyslexia candidate genes mentioned in 1.1.1 due to their significance in earlier studies.

Table 1. Measured samples 080403.

Sample	Array type	Condition	Replicates
Ctrl07	Human Gene U133 plus 2.0	Empty vector	3
Gene1	Human Gene U133 plus 2.0	Vector with gene	3
Ctrl08	Human Gene U133 plus 2.0	Empty vector	3
Gene2	Human Gene U133 plus 2.0	Vector with gene	3
Gene3	Human Gene U133 plus 2.0	Vector with gene	3

After the overexpression of the genes, RNA was extracted from the cells 24 hours after transfections. The RNA quality was controlled and the expression microarray experiments were done in the Bioinformatic and Expression analysis core facility (B.E.A) at the Karolinska Institute, from where the data was received in form of gene expression intensities from the arrays.

The data is very high dimensional. Each sample (controls and genes) is a 1164x1164 matrix, which means 1354896 elements in a data set. These elements represent the intensity of expression in the Human Genome U133 Plus 2.0 Array, the first and most comprehensive whole human genome expression array.

When estimating effects of a treatment it is important to find out which effect will occur *without* the treatment, but in the same circumstances. That is why we have control groups, i.e. empty vectors, to compare with the gene vector treatment.

Data is loaded into R by defining a phenodata object, and then reading celfiles with the ReadAffy function.

Data<-ReadAffy()

Data will include all 15 celfiles, 3xGene1, 3xGene2, 3xGene3, 3xCtrl07, 3xCtrl08.

2.2 Visualisation of data

Visualising microarray data helps identifying problem areas and to determine if transformations are necessary. It is important to spend time on this before starting the analysis, to make sure that there are no problems for example with hybridization. The biggest problem is the large amount of data to deal with. Each array may have up to 100,000 genes and we have many arrays in the experiment. This will produce millions of data points, as described above. To deal with this it is quite useful to break down data into smaller chunks and summarize as much as possible, though being careful not to lose information from data.

Plots of intensity levels are to check the quality of the arrays and identify the ones that differ from the rest. The data will be plotted as a whole, but also as smaller parts, to see more clearly which arrays are different or damaged, and which patterns might be found.

2.2.1 Boxplots and histograms

A boxplot contains the median, the upper and lower quartiles, the range, and individual extreme values. The central box in the plot represents the inter-quartile range (IQR), the difference between the upper and lower quartiles. The line in the middle of the box is the median. Extreme values, more than 1.5 IQR above the 75th percentile and 1.5 IQR below the 25th percentile, are plotted individually. Data points that fall outside the boxes' boundaries are considered outliers. Each box shows the distribution of expression values of one array, and these values will often range between 0 and 16. [I:1]
A histogram will show the intensity level in even intervals and the area of each pile corresponds to the number of measurement values with a given log intensity, and hence the total area of piles is the total number of observations. (See Figure 3)

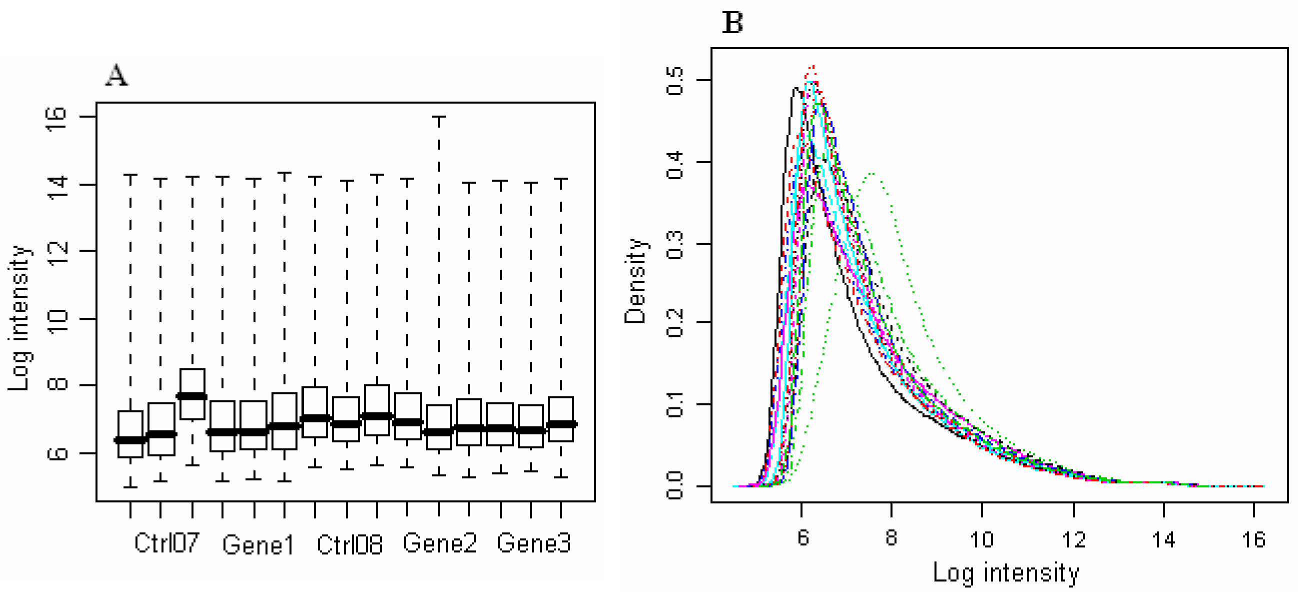


Figure 3.
Boxplot (A) shows the intensities, and histogram (B) the densities, for each sample of the raw data. These plots show some deviation for one of the replications of Ctrl07, but despite this I decided to include this replication in the experiment without reducing data.

3 Preprocessing data

I will use two Affymetrix preprocessing functions called Robust Mean Average (RMA) and Microarray Suite (Mas5.0) to transform the expression intensities of arrays. The preprocessing of the data is done with background correction, pm correction and normalization.

3.1 Background and PM correction

Scanning and hybridization can cause background noise on microarrays, and it can never be eliminated completely. **Background correction** reduces this noise and spatial trends on the array.
The *rma*-method is one of the most common algorithms that use a statistical model for background correction, and it is estimated by taking the median of the values inside the indicated areas. PM probe intensities are corrected by using a model for the distribution of probe intensities. This method will not correct MM probes.
In the *mas*-method the array i*s* broken into a grid of 16 regions and for each of these regions the lowest 2% of probe intensities are used to compute a background value. Each probe is then adjusted based on a weighted average of the background for each region. The weights are calculated for the distances between the location of the probe and the centriods of these 16 regions. This method corrects both PM and MM probes.

PM correction can be done with the function *mas* by substracting the ideal mismatch from perfect match (PM-MM). For calculating the intensities the software selects means of probe pairs Perfect Match and Mismatch (PM-MM). MM will always be less than the corresponding PM so negative values are avoided. For RMA the function *pmonly* will not do any adjustment to the pm values. [I:1]

To see more clearly what these functions will do to the data, they can be processed separately.

> bgRMA<-bg.correct(Data, "rma")

3.2 Normalization

In microarray experiments there is both biological and technical variation. Individual genes can be quite variable between replicates, and this is called *biological variation,* a random error. It can be minimized by doing replicated measurements and then averaging the results (averages in this experiment are Ctrl07, Ctrl08, Gene1, Gene2 and Gene3).

Technical variation, the systematic error, is one of the sources of variation that we do have a control over, but replicates will not help to reduce this error. This variation can have a statistically significant effect on results and that is why it is important to control this variation by having a well-defined procedure of the preparation of arrays.[I:4] To correct this systematic bias between arrays and the differences in processing arrays we will normalize the data. Normalization is a process performed to compensate for systematic measurement error both between and within arrays. It should reduce or remove this variation while leaving the more interesting biological differences that may exist [I:2]. It is necessary to normalize the intensities before any analysis is carried out.

Examining the plots of raw probe intensities in 2.2.1 can often show the need for normalization. The data is normalized using a *quantiles*-function, where the goal is to normalize arrays so that each array has a common intensity distribution. Figure 4 shows a boxplot and a histogram of the normalized data.

> DataNormalized<-normalize(Data, “quantiles”)

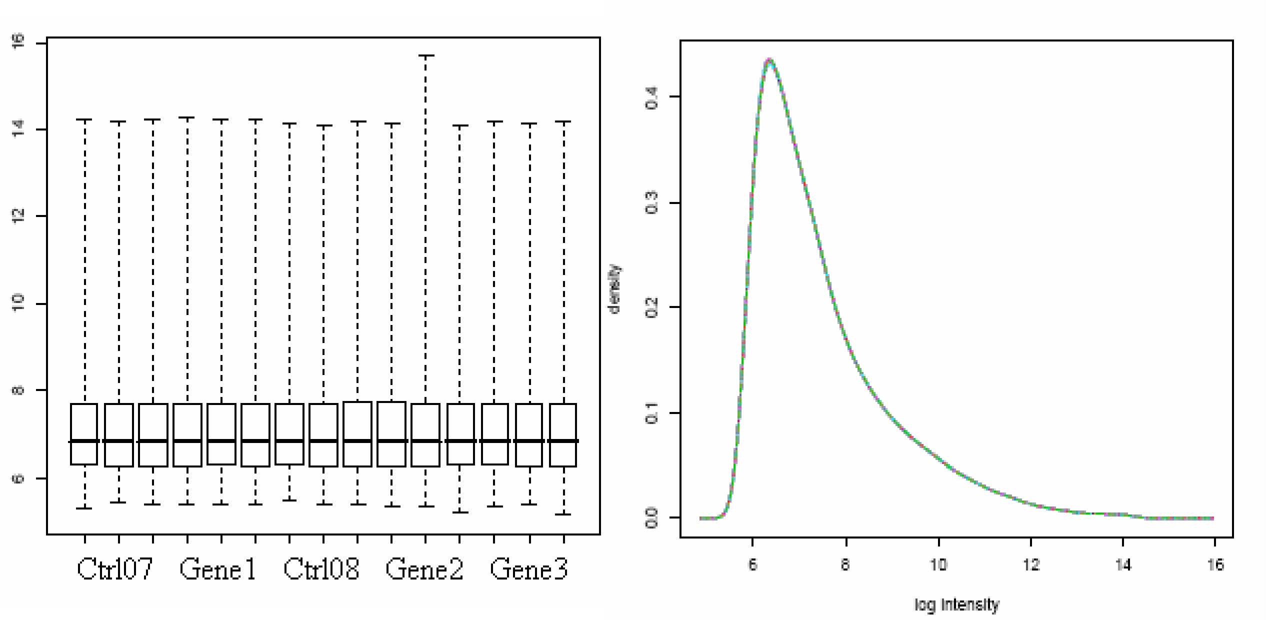


Figure 4. Boxplot and histogram after normalization. All arrays have the same distribution.

3.3 Expression values

To continue the process the probe level data is converted into summarized expression values. This includes background correction, normalization, probe specific background correction as just described, and a summary of probe set values into one expression measure for each probeset on the array (and a standard error for this summary statistic). RMA and Mas5.0 expression measures can be calculated from the affybatch object, Data, as follows:

eset_RMA <- expresso(Data, normalize.method = ”quantiles”, bgcorrect.method = “rma”, pmcorrect.method = ”pmonly”, summary.method = ”medianpolish”)

eset_Mas5 <- expresso(Data, bgcorrect.method = ”mas”, pmcorrect.method = “mas”, summary.method = ”mas”)

Medianpolish is the summary function used in the RMA expression summary, where a multichip linear model is fitted to data from each probeset.

$\log_2(PM_{ij}) = \alpha_i + \beta_j + \varepsilon_{ij}$

where α_i is a probe effect, β_j is the $\log_2$ expression value, and PM_{ij} the background adjusted, normalized PM intensity. [I:1]

The analysis will be continued with the RMA expression values, because *mas*-values are not normalized with this *affy*-method. RMA expression values are measured on a $\log_2$ scale and it is statistically better to use log transformated data because it turns the observations closer to a normal distribution, and decreases variance of summary statistics by reducing the influence of single measurements.

3.4 Quality control

It is important to assess the quality of the data obtained from any microarray experiment. There are several factors that might create variability in the quality of the results from each microarray, like variations in printing process, cDNA purity, RNA quality, hybridization and scanning. These may have a crucial impact on the results and lead to incorrect conclusions. Quality control is used to check that arrays have hybridised correctly and that sample quality is acceptable. Each observation that shows low quality must be discarded in the early phase of a microarray experiment.

3.4.1 Quality measures

There are several quality assessment metrics for identifying problematic arrays. These are Average Background, Scale Factor, Percent Present and 3'/5' ratios.

Quality measures can be extracted with the *qc* function and it is done for the Mas5.0 expression values by implementing a package called *simpleaffy*

```
Data_Exprs <- call.exprs(Data, “mas5”)
Data_QC <- qc(Data, Data_Exprs)
```

The Average Background is the average of the 16 background values which are calculated with RMA or Mas5.0 based on the lowest 2% of probe intensities, see Section 3.1. According to Affymetrix, the average background values should be comparable over arrays. The background intensities are used to correct the foreground intensities for local variation on the array surface. [I:1]

The Scale factor. Scaling is one of the normalization methods, where a baseline array is chosen, and the expression levels of all the other arrays are scaled to have the same mean intensity as the baseline array. After choosing the baseline array a linear regression is fitted between each array and the chosen array. The fitted regression line is then used for normalization. In mathematical terms, $E(Y_b)=\beta_i E(Y_i)$, where $E(Y_b)$ is the mean intensity for the baseline array, $E(Y_i)$ the mean intensity for array i, and β_i the scaling factor, estimated as $E(Y_b)/E(Y_i)$. The recommendation is that these values are within 3-fold of each other. [I:1]

The Percent present is the percentage of probe sets called ‘present’ on a chip. Replicated arrays should have similar values, and a higher percentage indicates better quality. [I:1]

See Table 2 in Appendix A for a summary of Average Background, Scale factor and Percent present –values.

3'/5' ratios. There are two quality control genes on a chip, β-actin and GADPH, and these are used to estimate RNA quality. They are represented by 3 probe sets; one from the 5' end, one from the middle and one from the 3' end of the transcript (see Figure 2 in 1.2.2). The ratio of the 3' expression to the 5' expression for these genes gives us a measure of RNA quality. These ratios are suggested not to be greater than 3. [I:1] See Table 3 in Appendix A for a summary of quality ratios.

Figure 5 in Appendix A shows a QC-plot and a somewhat more detailed explanation for the interpretation of these metrics.

3.4.2 Degradation plot

The RNA digestion plot illustrates the quality of RNA, the amount of RNA degradation that occured during the preparation of RNA, and how well second strand synthesis succeeded in preparation of samples.

The RNA degradation plot shows the mean expression from the 5' to the 3' end of the mRNA. Each sample is represented by a single line. In an ideal situation the lines should be flat, but that is usually not the case. If the lines are not flat, the slopes and profiles should be as similar as possible. Since RNA degradation typically starts from the 5' end of the molecule, we would expect probe intensities to be systematically lower at that end of a probeset compared to the 3' end. Figure 6 illustrates the degradation for the data.

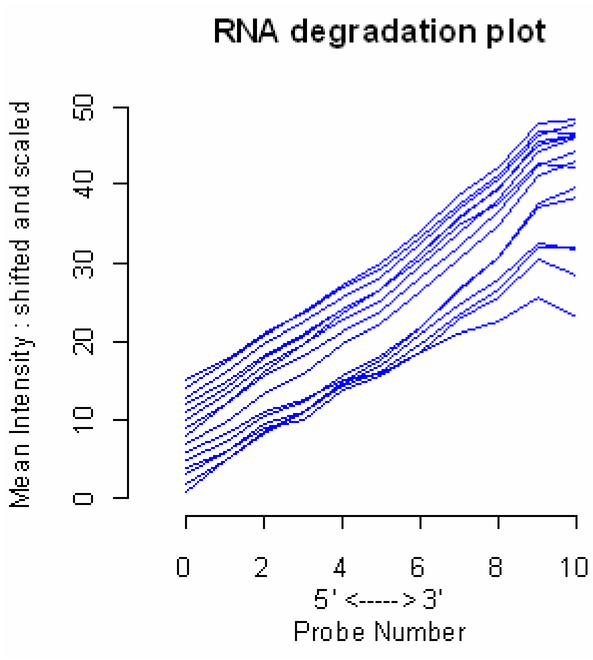


Figure 6. RNA degradation plot shows a good quality of arrays. A summary of degradation values is found in Table 4 in Appendix A.

4 Statistical analysis of the problem

4.1 Identifying Differentially Expressed Genes

After guaranteeing a good quality of the data we can begin identifying differentially expressed genes, which is the goal of microarray data analysis. The first step is to select a statistic which will rank the differentially expressed genes from strongest differential expression to weakest. Then a critical value can be chosen for the ranking statistic and every value above that is considered significant. The aim is to identify a number of candidate genes for confirmation and further study.

Affymetrix microarray data can be analysed with ordinary linear models, such as ANOVA. This experiment is done by fitting a regular linear model to data with a package called *limma.* Then the genes are ranked according to lowest p-values for the contrasts of interest.

The relation between gene expression values and array can be described using the following linear model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where **Y** is an nx1 response vector of expression values, **X** is an nxp design matrix, **β** is a px1 vector of regression parameters and **ε** is an nx1 vector of measurement noise.
Here, n=15 is the total number of samples and p=5 is the total number of groups that we investigate.

To run the analysis, we need to create a design matrix and a contrast matrix. Assume that the probe level intensities are converted to gene-wise expression levels with the RMA-method described in Chapter 3.

4.1.1 Design matrix

The design matrix defines the relations between RNA samples on each array. Each row of the design matrix corresponds to an array in the experiment and each column corresponds to one of the sample groups of Table 1.

The main purpose of this step is to estimate the variability in the data; hence the systematic part is modelled so it can be distinguished from random variation.

So the linear model **Y** = **Xβ** + **ε** will get the following componentwise form:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ X } \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$$

where columns of the design matrix **X** correspond to each group, Ctrl07, Gene1, Ctrl08, Gene2 and Gene3. Y_i corresponds to the expression level for a gene on array i, that is, one of the 15 samples after background correction, normalization and taking the logarithm of probe intensity levels.

Here, **Y** is assumed to be multivariate normally distributed, with mean **Xβ** and variance σ^2**I**. **ε** is normally distributed with mean **0** and variance σ^2**I**, where **0** is a nx1-vector of zeros and **I** is an identity matrix of order p.

The Maximum Likelihood estimate

$$\hat{\boldsymbol{\beta}} = (X^TX)^{-1}X^TY$$

of **β** is used as an expression measure for the gene for each one of the five sample groups.

In component form, the parameter estimates are

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} \end{pmatrix} \text{ X } \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{15} \end{pmatrix}$$

Less formally, we may write the estimates of β as the averages of the three replicates:

$\hat{\beta}_1$ = Average Control -07
$\hat{\beta}_2$ = Average Gene1
$\hat{\beta}_3$ = Average Control -08
$\hat{\beta}_4$ = Average Gene2
$\hat{\beta}_5$ = Average Gene3

4.1.2 Contrast matrix

To make it possible to compare the coefficients we will also define a contrast matrix which specifies comparisons that will be investigated between RNA samples. We are interested in sample combinations that were taken during the same period of time, that is, comparisons between Ctrl07 versus Gene1, Ctrl08 versus Gene2, and Ctrl08 versus Gene3.

The contrast matrix is created by

> contrast.matrix <- makeContrasts(gen1-ctrl07, gen2-ctrl08, gen3-ctrl08)

$$\mathbf{C} = \begin{pmatrix} -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & -1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

where the first column represents the contrast Gene1-Ctrl07, the second column Gene2-Ctrl08 and the third column Gene3-Ctrl08.

The three contrast parameters are contained in the 3x1 vector $\boldsymbol{\alpha} = \mathbf{C}^T\boldsymbol{\beta}$, and these are estimated by

$$\hat{\boldsymbol{\alpha}} = \mathbf{C}^T\hat{\boldsymbol{\beta}}$$

which can be written as

$$\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{pmatrix} \mathrm{X} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \end{pmatrix} = \begin{pmatrix} \hat{\beta}_2 - \hat{\beta}_1 \\ \hat{\beta}_4 - \hat{\beta}_3 \\ \hat{\beta}_5 - \hat{\beta}_3 \end{pmatrix}.$$

Here the three contrast estimates of α are of main interest and they are called Gene1-Ctrl07, Gene2-Ctrl08 and Gene3-Ctrl08 in plots and tables.

The linear model is fitted for RMA summarized data expression levels according to

```
> fit_RMA<- limFit(eset_RMA, design.matrix)
> fit2_RMA_contrast <- contrasts.fit(fit_RMA, contrast.matrix)
> fit2_RMA_eb <- eBayes(fit2_RMA_contrast)
```

4.2 Testing Hypotheses

Once the design and contrast matrices are created, we can run the analysis by computing t-statistics and log-odds of differential expressions and test against alternative hypotheses that there is difference in gene expression. We will set up a null hypothesis “No mean difference in the gene’s expression between the two groups” and the alternative hypotheses:

H_0: $\beta_i - \beta_j = 0$ versus H_1: $\beta_i - \beta_j \neq 0$

which we can form for our three contrasts:

H_{01}: $\beta_2-\beta_1 = 0$ versus H_{11}: $\beta_2-\beta_1 \neq 0$
H_{02}: $\beta_4-\beta_3 = 0$ H_{12}: $\beta_4-\beta_3 \neq 0$
H_{03}: $\beta_5-\beta_3 = 0$ H_{13}: $\beta_5-\beta_3 \neq 0$

Tests for t- and B-statistics are run with the package *limma*. Limma contains a function, *topTable*, that will list the probe sets which show biggest difference for each comparison defined by the contrast matrix C.

4.2.1 Expression measures and statistics

There are several measures for differential expression, and they can be listed as follows. (*topTable* will return a list of these values)

The **M-value (M)** corresponds to a log-fold difference in expression intensities. In this experiment, it is the log-fold difference in expression levels between gene and control arrays. High M-values indicate greater expression so we will concentrate on genes that have M-values genuinely different from zero. Since there are three replicates for each group of Table 1 and the Y_i are the logarithm expression levels, M can be expressed as follows:

$$M = \hat{\alpha}_j$$

for constrast j=1,2,3.

The **A-value (A)** is the average logarithm expression level for all arrays, or in this case, all six arrays included in the contrast. For instance, for contrast Ctrl07-Gene1,

$$A = ½ *(\hat{\beta}_1 + \hat{\beta}_2)$$

and similarly for the other two contrasts.

M is plotted against A for each contrast in Section 5.1.1.

The **p-value**. We will rank the genes according to their p-values, which is the probability of obtaining an estimated contrast at least as extreme as the observed one under the null hypotheses. It is the lowest level of significance where the null hypothesis can be rejected and low p-values corresponds to higher significance (note that our null hypotheses was β_i-$\beta_j = 0$ for the three contrasts).

The p-value is obtained from the distribution of the moderated t-statistic, usually after some form of adjustment for multiple testing. Given a set of p-values, the functions *topTable()* and *decideTests()* will also return *adjusted p-values* for multiple testing. The meaning of the adjusted p-value is that you can control the false discovery rate to be less than a risk level 0.05, and then select all genes that have adjusted p-value less than 0.05 as differentially expressed. See Section 4.2.2 for more details.

The moderated **t-statistic (t)** is the ratio of the M-value to its standard error. Here M is the difference in two samples of sizes n_1=n_2=3, with equal variances σ^2.
Hence, the formula for the t-statistic will be

$$t = \frac{M}{(s * \sqrt{(\frac{1}{n_1} + \frac{1}{n_2})})} = \frac{M}{(s * \sqrt{\frac{2}{3}})}$$

for contrast Gene1-Ctrl07, and similarly for the other two contrasts.

The estimated standard deviation, s, is

$$s = \frac{1}{f}\sqrt{\sum_{i=4}^{6}(Y_i-\hat{\beta}_2)^2+\sum_{i=1}^{3}(Y_i-\hat{\beta}_1)^2}$$

If $|t| > t_{0.025}(f)$ the null hypotheses is rejected at 5% level. Here, f is the number of degrees of freedom in a two-sample t-test with equal variances,

$f = (n_1-1)+(n_2-1) = 4$ (Reject H_0 if $|t| > t_{0.025}(4) = 2.78$)

The ordinary t-statistic is not ideal in this case because genes with small sample variances (n=3) can give a large t-statistic even if they are not differentially expressed. To reduce this problem we can instead use B-statistics. [I:8]

The **B statistic**, Bayes log posterior odds statistic, is the log-odds of differential expression, and is defined as

$$B = \log \frac{P(H_{1i}|M)}{P(H_{0i}|M)}$$

If this value is greater than 0 then there is more than a 50% chance that the gene is truly differentially expressed given its fitted M value. Say that we have B=1. The odds of differential expression is exp(1)=2.7183, almost three to one. The probability that the gene is differentially expressed is exp(1)/(1+exp(1))=0.7311, about 73%. B=0 corresponds to a 50% chance that the gene is differentially expressed. Negative B-values indicates that the gene is not differentially expressed. The statistic B avoids the problems of using averages or t-statistics mentioned above. Observations are weighted with prior assumptions. With fewer observations the prior assumptions will dominate, thus reducing random fluctuations.

The B-statistic is automatically adjusted for multiple testing by assuming that, for example, 1% of the genes are expected to be differentially expressed. The p-values and B-statistics usually rank the genes very similarly. All three measures, p-value, t– and B-statistic, are closely related, even though B-statistic depends on a prior guess for the number of differentially expressed genes. [I:1]
Our gene selection is based on the p-value. See Table 5 in Appendix A for the list of differentially expressed genes.

4.2.2 Multiple testing

We can compute a summary statistics by the eBayes() function for each contrast. A list of most significant probe sets in a specific contrast can be obtained in R by

> *topTable(fit2_RMA_eb, coef=1, number=100, adjust="BH", sort.by="P")*

where coef = 1 corresponds to comparison of Gene1-Ctrl07, coef = 2 to Gene2-Ctrl08 and coef=3 to Gene3-Ctrl08.

This function will do thousands of tests simultaneously and in this kind of multiple testing there are methods that correct for multiple comparisons. Bonferroni correction will control the chance of any false positives, and correct the cut-off significance (α) by multiplying it by the number of tests carried out. If this corrected value is still less than 0.05 the null hypotheses is rejected.

Instead of controlling the chance of any false positives (as Bonferroni), **false discovery rate (FDR)** controls the expected *proportion* of false positives (type I errors). A FDR threshold is determined from the observed p-value distribution. For example, if the algorithm returns 100 genes with a false discovery rate of 0.3, then we should expect 70 genes to be correctly classified as positives. This is quite useful when there are thousands of genes on the array most of them not being differentially expressed.

The p-values for the contrasts of interest are adjusted for multiple hypotheses testing by a call to adjust. The "BH" method, (Benjamini and Hochberg's method), that was used in this case, controls the expected false discovery rate (FDR) below a specified value. It is an adjustment method considered most appropriate for microarray studies. [I:1]

The *topTable* function will result in a table of M- and A-values, t-statistics, p-values and B-statistics. Here we are mostly interested in columns for M , p-values and B.

4.3 Clustering

Clustering is the most popular method currently used in analysing gene expression. It is very useful for large data sets because it reduces the dimensionality of the system and by this allows easier management of the data set. The goal of clustering is to group together samples with similar properties based on gene expression levels of the genes. These expression profiles can classify the different RNA sources into meaningful groups and in this way the cluster analysis will help to identify new pathway relationships and gene functions that may be critical to cellular control in health and disease. Clustering can even be used for quality control, by identifying outliers or experimental errors. Different clustering methods can have very different results and it is not yet clear which clustering methods are most useful for gene expression analysis.

We can study the gene expression values by comparing rows in the expression matrix. Then we may find similarities or differences between different genes and thus conclude about the correlation between the two genes. If we find that two rows are similar, we can hypothesize that the respective genes are co-regulated and possibly functionally related. We can also identify spatial or temporal expression patterns. Comparing columns in the expression matrix we can identify new classes of biological samples and detect experimental artifacts.

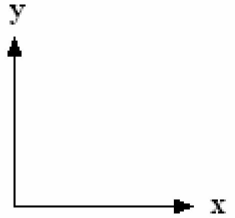
When trying to group together objects that are similar, we need a measure of similarity, a distance metric. There are different kinds of distance metrics and clustering is highly dependent upon the distance metric used.

4.3.1 Distance metric

There are some assumptions to apply when investigating the distance between two points, x and y, in an n-dimensional space $\mathbf{R}^n$. A distance metric *d* has the following properties:

a) The distance should be symmetric, $d(x,y) = d(y,x)$
b) The distance between any two points should be a real number greater than or equal to zero, $d(x,y) \geq 0$
c) The distance between two points x and y should be shorter than or equal to the sum of the distances from x to a third point z and from z to y: $d(x,y) \leq d(x,z) + d(z,y)$ [I:5]

The distance between two n-dimensional vectors $x = (x_1, x_2,..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ can be calculated with different methods. I will use the distance metric called Manhattan distance, which is measured along directions that are parallel to the x- and y-axes meaning that there is no diagonal direction.



Manhattan distance
$d_M(\mathbf{x},\mathbf{y}) = | x_1-y_1 | + | x_2-y_2 | +... +| x_n-y_n | = \sum | x_i-y_i |$

where $| x_i-y_i |$ represents the absolute value of the difference between x_i and y_i.

```
expression.matrix<-exprs(eset)
rma.dist1<-dist(t(expression.matrix), method=”man”)
rma.mds1<-isoMDS(rma.dist1)
```

Another often used method is the Euclidean distance, which takes into account both the direction and the magnitude of the vectors. This is measured along a diagonal direction between x and y.

$$d_E(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)2 + (x_2-y_2)2 + ... + (x_n-y_n)2} = \sqrt{\sum (x_i-y_i)2}$$

When comparing the Manhattan distance with the Euclidean, we can clearly see that Manhattan distance is greater because of the Pythagorean Theorem. Data clustered with Manhattan distance metric might also be less scattered and more robust regarding miscalculated data. (See reference 5 for more details about clustering methods)

5 Results

5.1 Illustrating significant genes

The statistical analysis resulted in tables of the 100 most significant genes for each contrast, and these genes were sorted both by p-values and M-values. Inspection of the tables revealed a few candidates that were significant in terms of p-values, t-statistics, M-values and B-statistics. No specific names of the genes are given due to the fact that the material of this work is unpublished.

Results are based on tests of the contrast estimates defined in 4.1.2.

5.1.1 Plots and diagrams

The quantile-quantile (Q-Q) plot helps to discover genes that have unusually large t-statistics, which could be an indication of differential expression. It plots the sample quantiles of the data against the theoretical quantiles of the t-distribution. Also a 45-degree reference line is plotted, and the points that are not approximately lying on the line shows the genes with large t-statistics, see Figure 7.
However, t-statistics are not perfectly suited for microarray analysis since a large t-statistic can be due to small standard deviations.

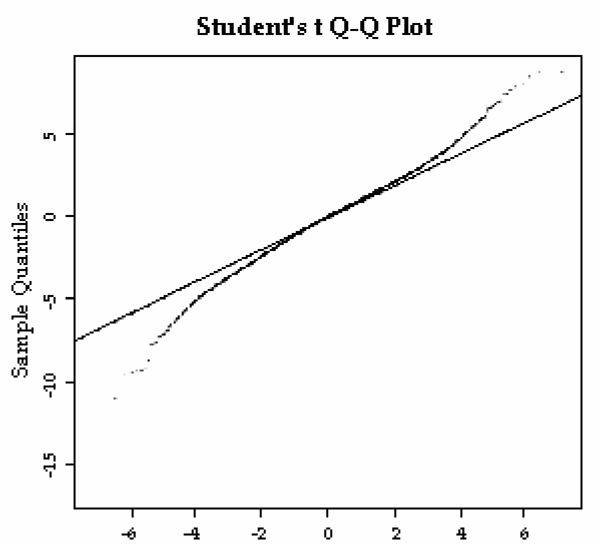


Figure 7. The Q-Q Plot clearly shows that there seems to be some good candidate genes that are differentially expressed. (The plotted line is the average for the 15 samples)

The MA-plots show the log-fold expression change, M, versus the average log intensity, A, for all genes on the array. A high log fold change and a low intensity is related to less reliable significance. The pairwise graphical comparison of intensity data can be useful when diagnosing problems in replicate sets of arrays. Comparisons for all three contrasts are done using the means of the three replicates. The MA-plot uses M as the y-axis and A as the x-axis, see Figure 8.

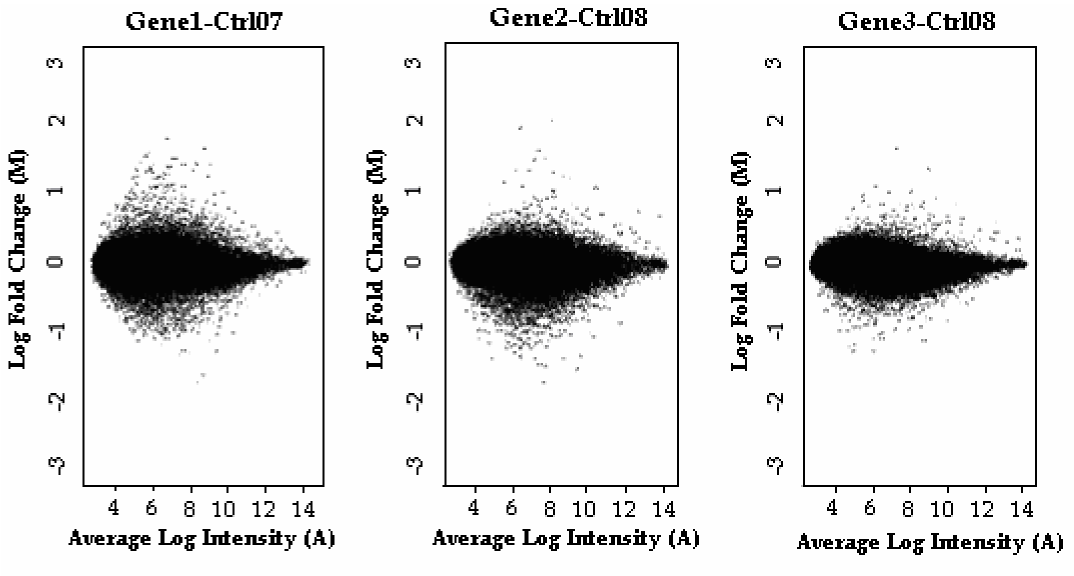


Figure 8. MA-plots of the three contrasts. The dots illustrate the difference in the log intensities for each probeset and we are most interested in those whose M-value is close to 2 or -2.

After finding the significant differentially expressed genes it can be interesting to plot these into the MA-plots, as shown in Figure 9.

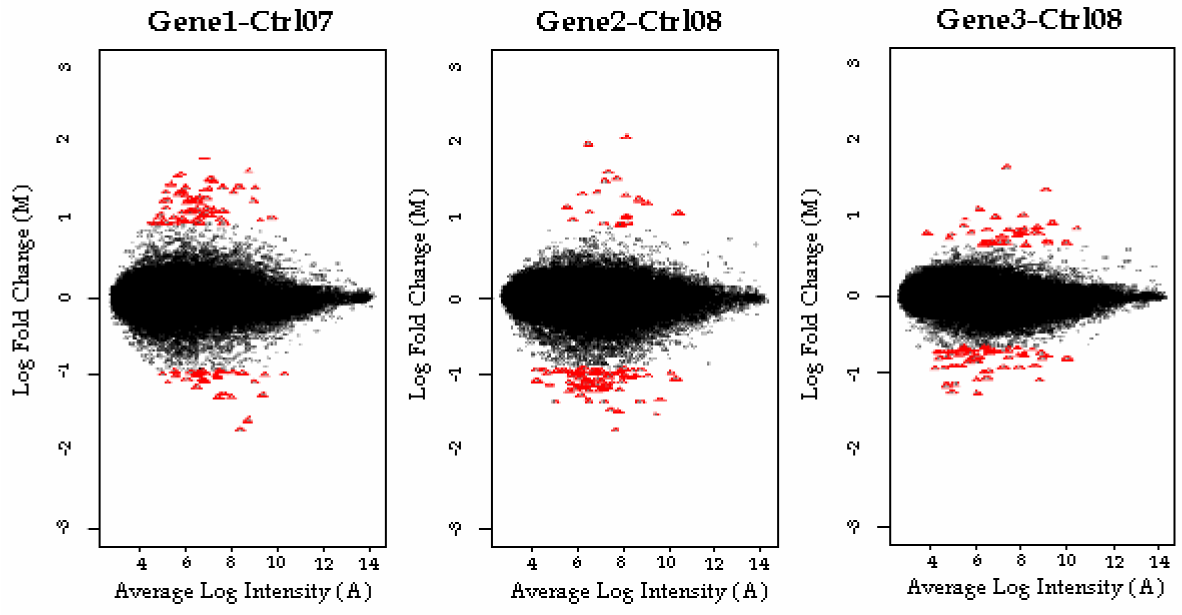
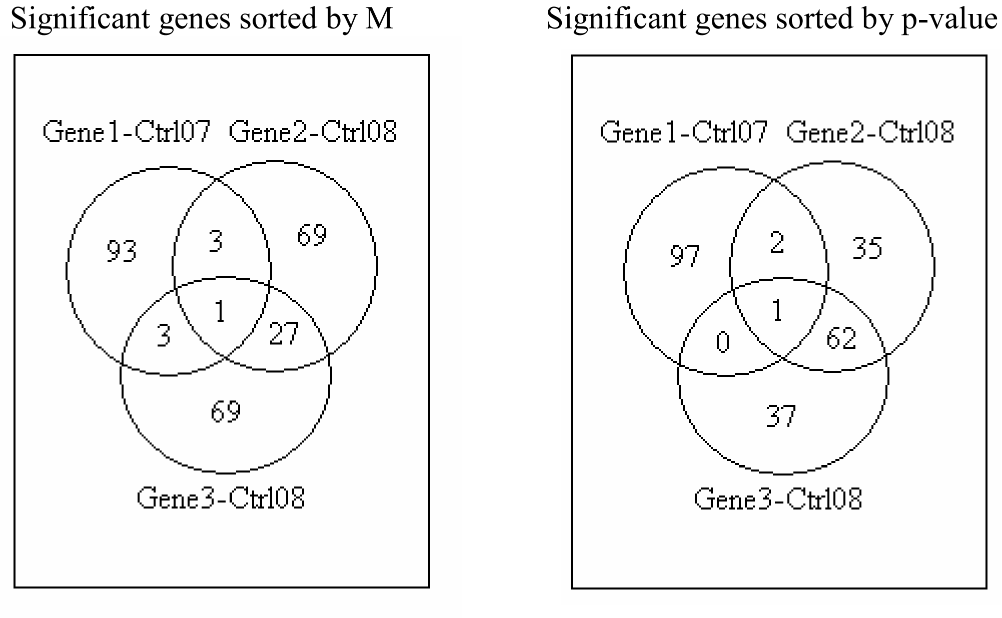


Figure 9. Red dots illustrates the most significant genes for each contrast.

When analyzing microarray data, **the Venn diagram** is a useful visualization tool to indicate differences or similarities between multiple experiments. It is comparing the gene lists for the three comparisons, and showing the number of genes that are significant in more than one comparison, as shown in Figure 10.

Venn diagram can be obtained with

> *vennDiagram(results)*



Significant genes sorted by M

Significant genes sorted by p-value

Figure 10. The Venn-diagrams are drawn for the 100 most significant genes in the three contrasts and it shows that for comparisons Gene2-Ctrl08 and Gene3-Ctrl08 there are most similar genes that are significantly differentially expressed (27 and 62).

The clustering plot in Figure 11 also shows the similarity of samples Gene2 and Gene3.

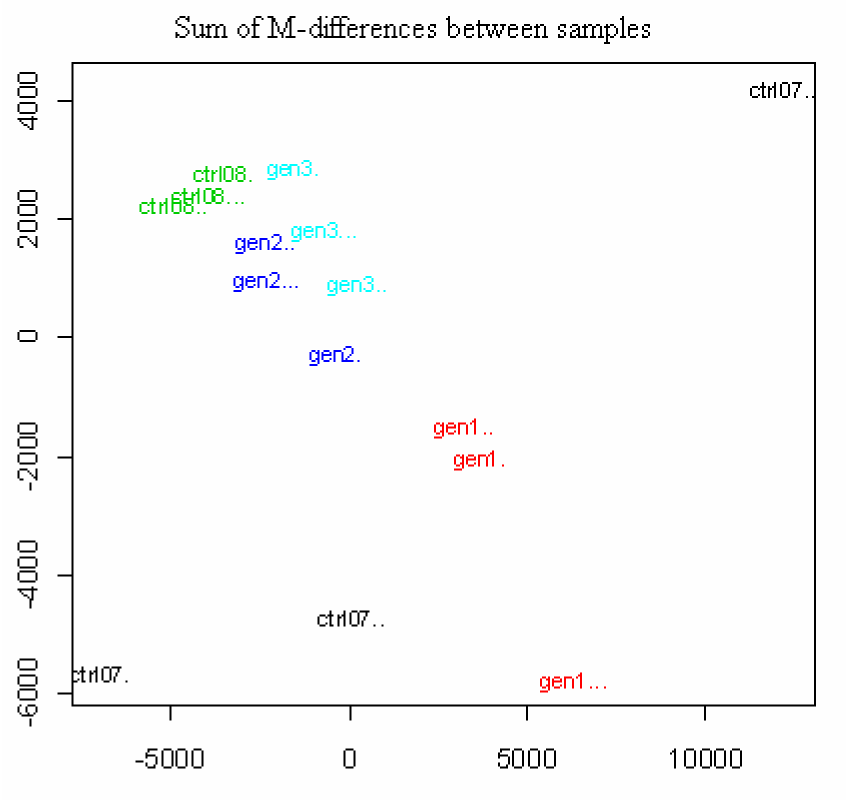


Figure 11. Clustering. Each sample is represented in the plot using multidimensional scaling with Manhattan distance. The most desirable thing to expect to see from this plot is a long distance between controls and genes when compared with each other, like here the distance between Gene1 and Ctrl07 –replicates. However we should be careful with Ctrl07..., because this replicate showed some deviation in the plots we draw before the analysis and it could have been removed from the data.
It can be established that the Gene2 and Gene3 samples show some similarities (as seen from the Venn-diagram) and therefore it is interesting to do som further tests on these two samples to see which genes are differing, see Table 6 in Appendix A.

A volcano plot is helpful in identifying significance change in differential expression of sets of genes between two conditions. It is an effective graph that displays the log-odds of differential expression (y-axis) against the log fold-change (x-axis). Volcano plots for each contrast are shown in Figure 12.

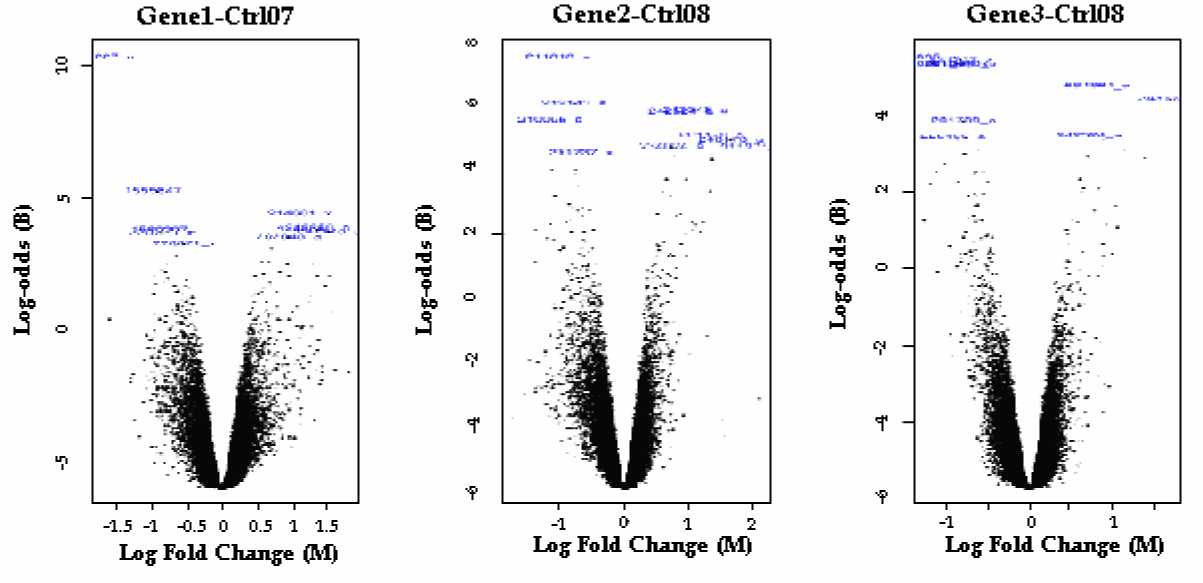


Figure 12. Volcano-plots. Genes with statistically significant differential expression according to B-statistics will lie above a horizontal threshold line. Genes with large fold-change (|M|>1) will lie outside a pair of vertical threshold lines. The blue dots are representing the genes with high effect and significance.

6 Discussion

Often the main reason for doing a microarray experiments is to identify genes that are differentially expressed between treatment and control groups and to find out which genes respond to the treatment. In this thesis, I have applied statistical methods to the analysis of gene expression microarrays. In particular, the microarray data was analysed within a linear model after quality control and normalization. t- and B-statistics were based on multiple testing between the contrasts taking the false discovery rate of prior assumptions on differential expression into consideration.

Biological interpretation is the understanding of the underlying biology after observing gene expression changes in an experiment. Many different measurements of differential expression are used, and the main purpose is to rank only a few of the thousands of genes included in the analysis.

The 100 most significantly expressed genes were listed for each contrast according to the p-value. This was most relevant to do because the t- and B-statistics showed to be significant for all of these. Moreover, we are interested in minimized p-values, the lowest level of rejecting a true null hypotheses, in our case that no differential expression is observed. Thus, looking at the adjusted p-value smaller than 5% and when taking into consideration the log fold change (M) as well, the list of good candidates could be brought out.

Interpreting the significance and defining cutoffs can be difficult, but in this case it was most relevant to emphasize genes with $M>1$ and adjusted p-values <0.05. The ranked list of the differentially expressed genes according to p-value had some differences compared to the list according to fold change, so it was useful to test different ways and compare the results to receive the best possible information from the experiment. Genes that are significant for alternative statistics are less likely to be false positives. Clustering the data showed us some similarities between Gene2 and Gene3 and this was tested further by finding the most significantly differing genes. However, the B-statistics was negative and the adjusted p-value was high for all of these genes, so no significant difference could be found.

Even if we find some gene that show great significance in the toptables, this alone may not give us enough information of the biological connections with the other possible genes of interest. Different kinds of plots are a good way to illustrate differentially expressed genes and possible clusters and patterns of which genes are co-regulated or may be included in the same biological process.

Biological and statistical significance is not always easy to interpret since each model and test has its own assumptions and criteria. There are also many transformations and normalization methods when preprocessing the data, and the same results are not necessarily achieved with different methods.

Of course, the amount of replicates on an array also has an important role. It is useful to understand how many replicates it is necessary to run and how many replicates are possible to carry out in an experiment. In this case, three replicates were considered as a proper amount to finance. More than three replications on samples would have given more statistical power to the tests and more reliable results could have been obtained.

In other microarray studies where differentially expressed genes show more significance it can be interesting to pick up genes with M>2, for example, and make lists and (network) graphs that show which parts of the biological process the genes have a significant effect on.

After identifying differentially expressed genes, it might be interesting to visit www.geneontology.org to find some more information of the properties of a particular gene, so that genes with biological relevance can be chosen for further investigation.

7 References

I

1. Gentleman, Carey, Huber, Irizarry, Dudoit. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor.* 2005. Springer.

2. Bolstad, Collin, Simpson, Irizarry, Speed. *Experimental Design and Low-level Analysis of Microarray Data.* 2004. Elsevier Inc.

3. *Human Molecular Genetics*. 2006. Vol. 15, No. 19

4. Affymetrix's homepage for an overview of microarray analysis. Learning Center for Expression Data Analysis Series www.affymetrix.com

5. Korol, A.B. *Microarray Cluster Analysis and applications*. Institute of Evolution, University of Haifa.
   (http://www.science.co.il/enuka/Essays/Microarray-Review.pdf)

6. Dudoit, S. Gentleman, R. *Cluster Analysis in DNA Microarray Experiments*.
   Bioconductor short course. 2002.
   (http://www.bioconductor.org/workshops/2002/Seattle02/Cluster/cluster.pdf)

7. Smyth, G.K. *Linear models and empirical bayes methods for assessing differential expression in microarray experiments*. Stat Appl Genet Mol Biol, Article3, 2004.

8. Smyth, G.K. Yang, Y.H. Speed, T. *Statistical Issues in cDNA Microarray Data Analysis*.
   2002. Department of Statistics, University of California, Berkeley.
   (http://www.stat.berkeley.edu/users/terry/zarray/TechReport/mareview.pdf)

9. Alberts et al. *Molecular biology of the Cell*. 4th edition. 2002.

II

DYX1C1:

Taipale M, Kaminen N, Nopola-Hemmi J, Haltia T, Myllyluoma B, Lyytinen H, Muller K, Kaaranen M, Lindsberg PJ, Hannula-Jouppi K, Kere J.
A candidate gene for developmental dyslexia encodes a nuclear tetratricopeptide repeat domain protein dynamically regulated in brain.
Proc Natl Acad Sci U S A. 2003 Sep 30;100(20):11553-8. Epub 2003 Sep 3.
PMID: 12954984 [PubMed - indexed for MEDLINE]

DCDC2:

1: Schumacher J, Anthoni H, Dahdouh F, König IR, Hillmer AM, Kluck N, Manthey M, Plume E, Warnke A, Remschmidt H, Hülsmann J, Cichon S, Lindgren CM, Propping P, Zucchelli M, Ziegler A, Peyrard-Janvid M, Schulte-Körne G, Nöthen MM, Kere J.
Strong genetic evidence of DCDC2 as a susceptibility gene for dyslexia.
Am J Hum Genet. 2006 Jan;78(1):52-62. Epub 2005 Nov 17.
PMID: 16385449 [PubMed - indexed for MEDLINE]

2: Meng H, Smith SD, Hager K, Held M, Liu J, Olson RK, Pennington BF, DeFries JC, Gelernter J, O'Reilly-Pol T, Somlo S, Skudlarski P, Shaywitz SE, Shaywitz BA, Marchione K, Wang Y, Paramasivam M, LoTurco JJ, Page GP, Gruen JR.
DCDC2 is associated with reading disability and modulates neuronal development in the brain.
Proc Natl Acad Sci U S A. 2005 Nov 22;102(47):17053-8. Epub 2005 Nov 8. Erratum in: Proc Natl Acad Sci U S A. 2005 Dec 20;102(51):18763.
PMID: 16278297 [PubMed - indexed for MEDLINE]

KIAA0319:

1: Cope N, Harold D, Hill G, Moskvina V, Stevenson J, Holmans P, Owen MJ, O'Donovan MC, Williams J.
Strong evidence that KIAA0319 on chromosome 6p is a susceptibility gene for developmental dyslexia.
Am J Hum Genet. 2005 Apr;76(4):581-91. Epub 2005 Feb 16. Erratum in: Am J Hum Genet. 2005 Nov;77(5):898.
PMID: 15717286 [PubMed - indexed for MEDLINE]

ROBO1:

1: Hannula-Jouppi K, Kaminen-Ahola N, Taipale M, Eklund R, Nopola-Hemmi J, Kääriäinen H, Kere J.
The axon guidance receptor gene ROBO1 is a candidate gene for developmental dyslexia.
PLoS Genet. 2005 Oct;1(4):e50. Epub 2005 Oct 28.
PMID: 16254601 [PubMed - indexed for MEDLIN

C2ORF3 and MRPL19:

Anthoni H, Zucchelli M, Matsson H, Müller-Myhsok B, Fransson I, Schumacher J, Massinen S, Onkamo P, Warnke A, Griesemann H, Hoffmann P, Nopola-Hemmi J, Lyytinen H, Schulte-Körne G, Kere J, Nöthen MM, Peyrard-Janvid M.
A locus on 2p12 containing the co-regulated MRPL19 and C2ORF3 genes is associated to dyslexia.
Hum Mol Genet. 2007 Mar 15;16(6):667-77. Epub 2007 Feb 19.
PMID: 17309879 [PubMed - indexed for MEDLINE

PCNT, DIP2A, S100B and PRMT2:

1: Poelmans G, Engelen JJ, Van Lent-Albrechts J, Smeets HJ, Schoenmakers E,
Franke B, Buitelaar JK, Wuisman-Frerker M, Erens W, Steyaert J, Schrander-Stumpel C.
Identification of novel dyslexia candidate genes through the analysis of a
chromosomal deletion.
Am J Med Genet B Neuropsychiatr Genet. 2008 Jun 2. [Epub ahead of print]
PMID: 18521840 [PubMed - as supplied by publisher]

Appendix A

Table 2. Quality values over arrays.

Sample	Average Background	Scale Factor	Percent Present
Ctrl07.	40.0	0.8	49.3
Ctrl07..	41.2	0.8	47.3
Ctrl07...	66.4	0.7	42.3
Gene1.	41.3	0.9	43.7
Gene1..	43.7	0.9	45.1
Gene1...	42.9	0.7	46.5
Ctrl08.	54.2	0.6	46.6
Ctrl08..	56.1	0.7	46.9
Ctrl08...	50.5	0.8	46.1
Gene2.	54.8	0.8	44.7
Gene2..	48.3	0.9	45.6
Gene2...	50.0	0.9	46.0
Gene3.	59.0	0.7	46.1
Gene3..	51.6	0.9	45.1
Gene3...	53.6	0.8	45.3

* Values illustrated in the QC-plot

Table 3. Quality ratios over arrays.

Sample	β-actin 3'/5'	GADPH 3'/5'	β-actin 3'/M	GADPH 3'/M
Ctrl07.	0.837	-0.011	0.523	0.035
Ctrl07..	0.694	0.092	0.415	0.106
Ctrl07...	0.672	0.005	0.435	0.083
Gene1.	0.828	0.066	0.411	0.094
Gene1..	0.751	0.028	0.395	0.097
Gene1...	0.332	-0.050	0.166	0.061
Ctrl08.	0.921	0.041	0.629	0.144
Ctrl08..	0.889	-0.001	0.535	0.166
Ctrl08...	1.041	0.038	0.548	0.100
Gene2.	1.400	0.040	0.800	0.116
Gene2..	1.311	0.112	0.733	0.117
Gene2...	1.081	0.084	0.517	0.132
Gene3.	0.937	-0.037	0.610	0.157
Gene3..	1.145	0.093	0.599	0.113
Gene3...	0.967	0.039	0.560	0.062

* The data is presented as log 2 differences, so these values should be raised to power 2 to get the correct effect. β-actin should be within 3 and GADPH around 1. The differences are based on normalized expression values.

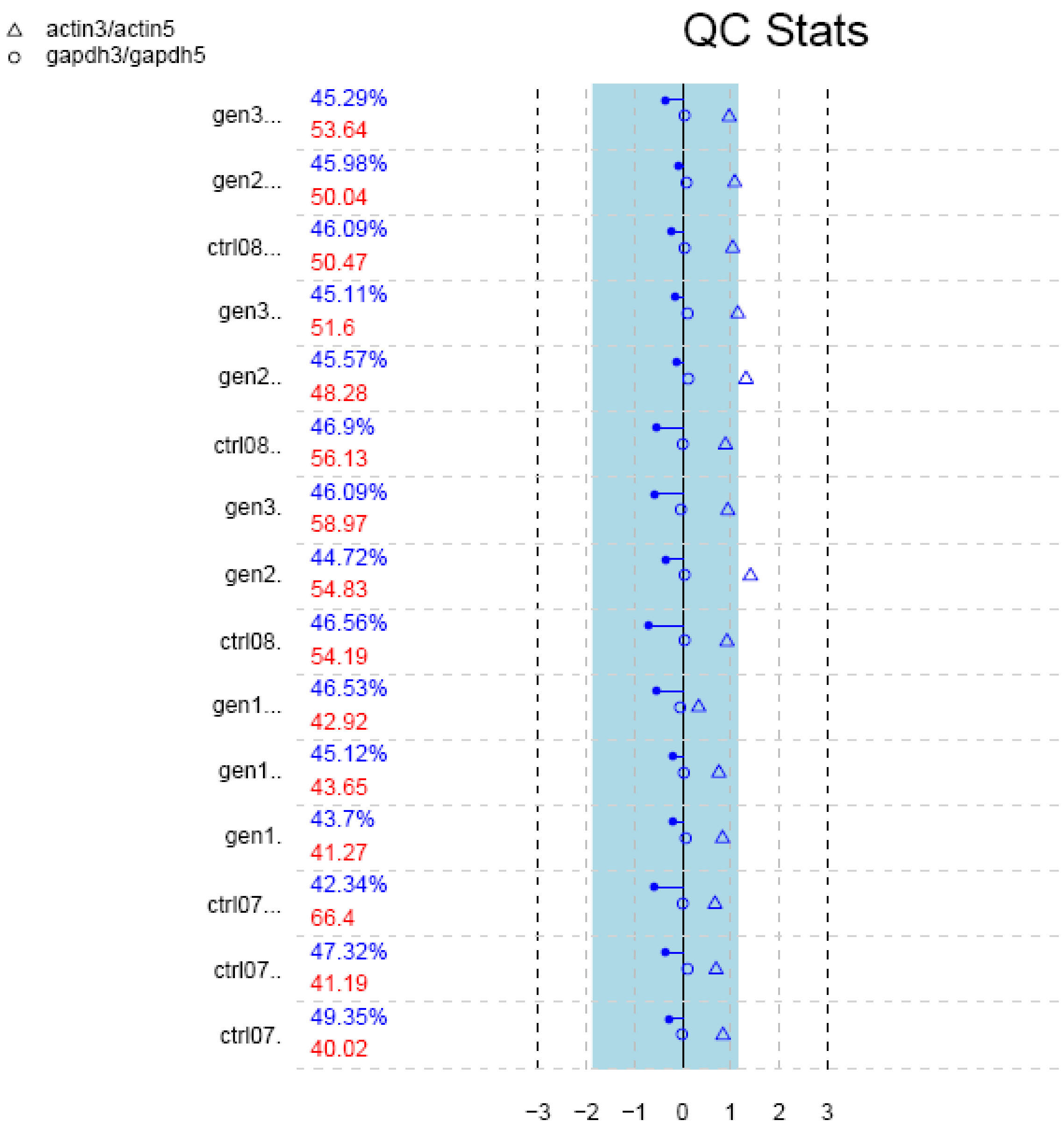


Figure 5. The QC-plot of data represents a summary of quality values for each array. All arrays have scale factors within the blue region, and its position can be found by calculating the mean scale factor for all arrays having the borders of 1.5 up and down. The scale factors for all arrays are the blue points plotted as a line from the central vertical 0-fold line. The figure shows also GAPDH and beta-actin 3'/5' ratios, which are represented as a pair of points.

All points plotted in the QC-plot should be blue, red indicates problems. So according to this plot, our arrays are of good quality.

On the left are the percentages of genes called present on each array and the average background. If the array has significantly different values (differing more than 10% in %-present or 20 units in background intensity), they are coloured red, otherwise blue. [I:1]

Note that there is a substantial spread in the background values, which is an indication of quality problems. (Recall the deviation of sample Ctrl07...; it has a much higher background than the rest.)

Table 4. RNA degradation.

RNA degradation

	Ctrl07.	Ctrl07..	Ctrl07...	Gene1.	Gene1..	Gene1...
slope	3.84e+00	3.19e+00	2.78e+00	2.96e+00	3.55e+00	1.93e+00
P-value	1.12e-10	1.61e-10	4.15e-09	8.83e-10	1.07e-08	3.63e-08
	Ctrl08.	**Ctrl08..**	**Ctrl08...**	**Gene2.**	**Gene2..**	**Gene2...**
slope	3.69e+00	3.67e+00	3.57e+00	3.94e+00	3.68e+00	3.58e+00
P-value	2.39e-10	1.45e-10	4.85e-11	6.17e-13	5.11e-11	2.72e-11
				Gene3.	**Gene3..**	**Gene3...**
				3.60e+00	3.22e+00	3.34e+00
				1.63e-10	8.80e-11	1.77e-10

* RNA is of good quality with very small P-values.

Table 5. List of differentially expressed genes for the three contrasts (RMA)

Gene1-Ctrl07

ID	M	A	t	P.Value	adj.P.Value	B
222227_at	-1,715	8,339	-16,736	0,00000000	0,00000476	10,317
1555847_a_at	-0,932	7,117	-8,349	0,00000072	0,01963507	5,264
214001_x_at	1,114	6,169	7,599	0,00000218	0,03978367	4,450
213350_at	1,371	8,044	7,109	0,00000469	0,04006111	3,869
1560297_at	-0,843	6,117	-7,078	0,00000493	0,04006111	3,832
216598_s_at	1,242	6,614	7,064	0,00000504	0,04006111	3,814
241713_s_at	1,529	7,198	6,997	0,00000562	0,04006111	3,731
220720_x_at	-0,875	6,846	-6,970	0,00000586	0,04006111	3,698
202648_at	0,957	6,426	6,840	0,00000724	0,04399651	3,534
220071_x_at	-0,542	7,820	-6,641	0,00001004	0,05490088	3,277
203468_at	-0,661	6,877	-6,579	0,00001113	0,05532042	3,196
243642_x_at	0,711	5,432	6,520	0,00001228	0,05593922	3,118
202605_at	-0,650	9,534	-6,316	0,00001733	0,07112873	2,844
219117_s_at	-0,681	8,622	-6,287	0,00001821	0,07112873	2,804
233810_x_at	0,679	5,813	6,115	0,00002446	0,07596797	2,566
206438_x_at	0,583	7,986	6,112	0,00002459	0,07596797	2,562
219138_at	1,296	6,826	6,097	0,00002523	0,07596797	2,541
216187_x_at	-0,789	8,309	-6,084	0,00002581	0,07596797	2,523
216246_at	0,945	6,908	6,071	0,00002640	0,07596797	2,505
235094_at	-0,869	7,274	-5,937	0,00003333	0,08703938	2,315
225934_at	0,835	6,041	5,889	0,00003628	0,08703938	2,246

232266_x_at	-0,799	8,925	-5,878	0,00003701	0,08703938	2,230
239959_x_at	0,599	5,964	5,842	0,00003940	0,08703938	2,178
1552287_s_at	-0,972	8,043	-5,811	0,00004166	0,08703938	2,133
224321_at	1,372	5,061	5,785	0,00004358	0,08703938	2,096
224023_s_at	0,962	5,915	5,782	0,00004384	0,08703938	2,091
234208_at	0,478	5,658	5,775	0,00004437	0,08703938	2,081
214041_x_at	1,542	7,005	5,767	0,00004504	0,08703938	2,069
242337_at	-0,747	5,855	-5,753	0,00004617	0,08703938	2,048
214016_s_at	-0,772	9,580	-5,694	0,00005126	0,09342059	1,962
226693_at	-0,503	7,221	-5,673	0,00005321	0,09384968	1,931
243537_at	-0,827	5,152	-5,639	0,00005653	0,09659023	1,881
213481_at	0,696	5,849	5,591	0,00006173	0,10122965	1,809
226334_s_at	-0,949	7,714	-5,573	0,00006370	0,10122965	1,783
200908_s_at	1,488	5,326	5,549	0,00006656	0,10122965	1,746
1554602_at	-0,678	6,488	-5,548	0,00006665	0,10122965	1,745
201505_at	-0,525	10,032	-5,527	0,00006921	0,10227333	1,714
217715_x_at	-0,460	5,638	-5,501	0,00007251	0,10244123	1,675
206180_x_at	0,599	7,396	5,497	0,00007307	0,10244123	1,669
228847_at	-0,615	5,605	-5,448	0,00007997	0,10685389	1,594
228318_s_at	-0,647	7,579	-5,438	0,00008135	0,10685389	1,580
213737_x_at	-0,634	8,030	-5,397	0,00008772	0,10685389	1,517
205584_at	-0,651	7,027	-5,393	0,00008838	0,10685389	1,510
1556316_s_at	0,579	5,644	5,384	0,00008982	0,10685389	1,497
221546_at	0,793	5,220	5,372	0,00009172	0,10685389	1,479
206057_x_at	0,471	5,837	5,362	0,00009349	0,10685389	1,464
1553909_x_at	-0,562	5,316	-5,358	0,00009426	0,10685389	1,457
222252_x_at	-0,494	6,194	-5,357	0,00009435	0,10685389	1,456
1552450_a_at	0,523	5,201	5,343	0,00009673	0,10685389	1,435
226318_at	-0,976	7,988	-5,338	0,00009772	0,10685389	1,427
236389_x_at	0,440	4,110	5,309	0,00010307	0,10941116	1,382
213813_x_at	1,274	6,371	5,304	0,00010406	0,10941116	1,374
218253_s_at	-0,425	9,049	-5,266	0,00011157	0,11448713	1,316
209460_at	-0,516	9,845	-5,243	0,00011637	0,11448713	1,280
226363_at	-0,604	7,998	-5,243	0,00011643	0,11448713	1,280
228674_s_at	-0,741	8,312	-5,239	0,00011726	0,11448713	1,274
229420_at	1,446	7,604	5,195	0,00012715	0,11732972	1,206
210231_x_at	0,553	11,326	5,194	0,00012754	0,11732972	1,203
214191_at	-0,523	5,117	-5,189	0,00012860	0,11732972	1,196
1563069_at	0,424	4,304	5,184	0,00012981	0,11732972	1,188
213034_at	-0,571	7,602	-5,179	0,00013090	0,11732972	1,181
216768_x_at	-0,489	5,592	-5,149	0,00013860	0,12108643	1,133
220227_at	0,451	4,454	5,137	0,00014162	0,12108643	1,115
212274_at	-0,478	10,269	-5,135	0,00014221	0,12108643	1,112
220682_s_at	1,256	4,936	5,128	0,00014395	0,12108643	1,101
242608_x_at	-0,515	8,306	-5,113	0,00014804	0,12111723	1,078
210422_x_at	0,646	6,157	5,112	0,00014842	0,12111723	1,076
219950_s_at	0,475	4,008	5,087	0,00015562	0,12338288	1,036
224565_at	-1,009	9,435	-5,083	0,00015660	0,12338288	1,030
218147_s_at	-0,551	8,979	-5,079	0,00015797	0,12338288	1,023
238470_at	-0,646	5,890	-5,060	0,00016357	0,12596179	0,994
235916_at	-0,470	4,980	-5,052	0,00016596	0,12602662	0,981
204377_s_at	-0,425	5,317	-5,038	0,00017038	0,12678061	0,959
239537_at	-0,642	7,448	-5,022	0,00017571	0,12678061	0,933
216554_s_at	1,077	6,788	5,014	0,00017816	0,12678061	0,921

229656_s_at	-0,560	8,472	-5,014	0,00017835	0,12678061	0,920
226036_x_at	0,800	6,300	5,013	0,00017855	0,12678061	0,919
206485_at	0,440	5,831	4,989	0,00018702	0,12816790	0,880
1557418_at	-0,636	4,243	-4,978	0,00019082	0,12816790	0,863
222999_s_at	-0,625	10,536	-4,969	0,00019386	0,12816790	0,850
226808_at	-0,710	7,594	-4,961	0,00019681	0,12816790	0,837
203243_s_at	-0,502	9,763	-4,960	0,00019743	0,12816790	0,834
220600_at	-0,509	5,123	-4,958	0,00019809	0,12816790	0,832
229165_at	0,651	6,448	4,952	0,00020025	0,12816790	0,822
240496_at	-0,457	4,423	-4,949	0,00020133	0,12816790	0,818
226967_at	0,779	7,610	4,943	0,00020383	0,12816790	0,807
221995_s_at	0,966	5,460	4,940	0,00020485	0,12816790	0,803
1563431_x_at	0,974	6,553	4,932	0,00020784	0,12816790	0,791
217383_at	1,582	5,613	4,930	0,00020901	0,12816790	0,786
207466_at	-0,709	5,429	-4,924	0,00021105	0,12816790	0,778
220725_x_at	0,616	7,280	4,919	0,00021332	0,12816790	0,769
214395_x_at	1,012	4,989	4,911	0,00021652	0,12867597	0,756
207365_x_at	-0,578	8,022	-4,889	0,00022562	0,13052480	0,721
205531_s_at	-0,355	5,430	-4,888	0,00022627	0,13052480	0,719
221223_x_at	0,408	7,061	4,885	0,00022744	0,13052480	0,714
223513_at	-0,625	8,669	-4,873	0,00023249	0,13052480	0,696
214057_at	0,727	6,248	4,872	0,00023311	0,13052480	0,693
226665_at	-0,967	6,714	-4,870	0,00023395	0,13052480	0,690
236825_at	0,394	5,398	4,845	0,00024544	0,13465917	0,650
1557996_at	-0,656	6,285	-4,836	0,00024940	0,13465917	0,636

Gene2-Ctrl08

ID	M	A	t	P.Value	adj.P.Value	B
211813_x_at	-1,032	5,814	-10,980	0,00000002	0,00129181	7,502
212143_s_at	-0,777	5,252	-9,171	0,00000023	0,00451604	6,050
213348_at	1,107	10,415	8,930	0,00000032	0,00451604	5,826
242524_at	0,877	9,090	8,901	0,00000033	0,00451604	5,799
210095_s_at	-1,149	6,076	-8,642	0,00000047	0,00518085	5,549
201170_s_at	1,352	6,159	8,193	0,00000090	0,00819686	5,093
243173_at	1,652	7,326	8,014	0,00000117	0,00913942	4,902
234024_at	0,740	8,743	7,841	0,00000151	0,00923036	4,713
211371_at	1,993	6,440	7,804	0,00000160	0,00923036	4,673
211737_x_at	-0,672	8,896	-7,681	0,00000193	0,00923036	4,535
206291_at	-0,962	5,093	-7,680	0,00000193	0,00923036	4,534
212558_at	-0,926	6,123	-7,611	0,00000214	0,00923036	4,456
231341_at	0,783	9,978	7,596	0,00000219	0,00923036	4,438
202504_at	1,371	6,861	7,526	0,00000244	0,00953697	4,359
222227_at	-0,741	8,339	-7,233	0,00000385	0,01318066	4,013
201893_x_at	-1,112	6,391	-7,233	0,00000386	0,01318066	4,013
213880_at	-0,752	8,827	-7,148	0,00000441	0,01418229	3,911
227566_at	-0,705	4,336	-7,020	0,00000541	0,01602523	3,753
1555338_s_at	0,660	6,854	6,971	0,00000585	0,01602523	3,693
229581_at	0,965	7,127	6,970	0,00000586	0,01602523	3,692
201739_at	-0,707	9,108	-6,801	0,00000772	0,02010589	3,477
1558795_at	1,334	8,095	6,683	0,00000937	0,02327551	3,326
200878_at	0,589	11,714	6,656	0,00000979	0,02327551	3,291
225238_at	-1,071	6,846	-6,576	0,00001118	0,02501331	3,187

218974_at	0,712	10,702	6,563	0,00001144	0,02501331	3,169
210393_at	-0,846	7,199	-6,397	0,00001509	0,03173736	2,949
228188_at	-0,806	5,024	-6,341	0,00001661	0,03364097	2,872
228728_at	-0,643	5,638	-6,314	0,00001739	0,03395360	2,836
230015_at	0,654	10,044	6,287	0,00001822	0,03434428	2,799
209897_s_at	-0,535	6,513	-6,248	0,00001945	0,03488577	2,746
228104_at	-0,715	5,915	-6,238	0,00001978	0,03488577	2,733
224707_at	0,521	8,416	6,185	0,00002170	0,03694871	2,658
213418_at	0,881	4,755	6,169	0,00002230	0,03694871	2,636
225728_at	0,813	8,295	6,131	0,00002377	0,03758884	2,585
218170_at	-0,646	11,096	-6,124	0,00002406	0,03758884	2,575
1558404_at	-0,673	3,513	-6,049	0,00002745	0,04168740	2,468
238877_at	-0,835	6,131	-6,003	0,00002970	0,04389400	2,404
230596_at	-0,884	6,395	-5,967	0,00003167	0,04416084	2,352
217057_s_at	1,234	9,044	5,965	0,00003175	0,04416084	2,350
216894_x_at	1,053	8,114	5,955	0,00003231	0,04416084	2,336
206577_at	-1,142	4,752	-5,908	0,00003508	0,04639553	2,269
209047_at	1,118	6,514	5,899	0,00003564	0,04639553	2,256
235368_at	-0,469	4,783	-5,840	0,00003956	0,05025374	2,171
209291_at	-1,361	4,962	-5,815	0,00004133	0,05025374	2,135
213182_x_at	1,073	8,133	5,815	0,00004136	0,05025374	2,135
204288_s_at	0,559	8,465	5,784	0,00004369	0,05125227	2,090
224997_x_at	1,189	5,524	5,760	0,00004560	0,05125227	2,055
209988_s_at	-0,815	10,025	-5,758	0,00004572	0,05125227	2,053
227662_at	-0,615	10,919	-5,756	0,00004593	0,05125227	2,049
1558796_a_at	0,957	7,793	5,707	0,00005013	0,05481658	1,977
242898_at	-1,198	6,456	-5,639	0,00005661	0,06069385	1,877
209466_x_at	-0,624	8,228	-5,604	0,00006025	0,06127473	1,825
236179_at	-1,087	4,200	-5,601	0,00006058	0,06127473	1,821
212171_x_at	0,779	8,618	5,593	0,00006145	0,06127473	1,809
224646_x_at	0,952	7,823	5,591	0,00006164	0,06127473	1,806
205736_at	0,601	5,483	5,578	0,00006312	0,06145698	1,787
213768_s_at	-0,601	8,011	-5,570	0,00006407	0,06145698	1,774
230695_s_at	0,630	5,402	5,545	0,00006699	0,06315157	1,737
238332_at	0,885	6,631	5,518	0,00007031	0,06515725	1,697
1560935_s_at	0,515	5,756	5,502	0,00007241	0,06598461	1,673
207414_s_at	0,535	7,625	5,445	0,00008028	0,07138047	1,587
237435_at	-0,945	8,761	-5,430	0,00008261	0,07138047	1,564
243940_at	-0,481	4,407	-5,425	0,00008326	0,07138047	1,557
1563658_a_at	-0,654	5,564	-5,423	0,00008355	0,07138047	1,554
219935_at	-0,988	4,651	-5,413	0,00008510	0,07142614	1,539
230859_at	0,540	7,614	5,406	0,00008622	0,07142614	1,528
218129_s_at	-0,536	8,235	-5,370	0,00009221	0,07359887	1,472
216235_s_at	-0,596	6,350	-5,366	0,00009287	0,07359887	1,466
239107_at	0,442	6,343	5,366	0,00009288	0,07359887	1,466
222162_s_at	-0,696	8,592	-5,321	0,00010073	0,07819175	1,398
210512_s_at	0,727	10,798	5,311	0,00010273	0,07819175	1,382
204284_at	0,516	7,120	5,309	0,00010297	0,07819175	1,380
220161_s_at	0,693	7,204	5,292	0,00010628	0,07947049	1,353
202998_s_at	0,761	7,498	5,286	0,00010756	0,07947049	1,343
219534_x_at	0,975	8,157	5,271	0,00011062	0,08022097	1,320
232119_at	-0,616	4,490	-5,266	0,00011151	0,08022097	1,313
220807_at	0,489	6,438	5,244	0,00011611	0,08244485	1,279
223588_at	-1,372	6,490	-5,122	0,00014579	0,10069727	1,088

229512_at	-0,421	7,207	-5,119	0,00014655	0,10069727	1,084
207074_s_at	0,505	9,958	5,116	0,00014734	0,10069727	1,079
244650_at	-0,544	6,904	-5,095	0,00015317	0,10313134	1,046
229357_at	-0,804	4,504	-5,090	0,00015467	0,10313134	1,038
223152_at	0,486	6,855	5,080	0,00015747	0,10373011	1,023
220600_at	0,520	5,123	5,068	0,00016110	0,10486100	1,004
210684_s_at	0,455	4,866	5,046	0,00016804	0,10705129	0,968
218566_s_at	-0,633	9,753	-5,044	0,00016838	0,10705129	0,967
225644_at	-0,570	9,023	-5,034	0,00017184	0,10799255	0,949
221245_s_at	0,807	6,676	5,010	0,00017973	0,11061452	0,912
223400_s_at	-0,740	6,442	-5,009	0,00018006	0,11061452	0,910
1560297_at	0,593	6,117	4,986	0,00018803	0,11345346	0,873
232071_at	-0,876	5,833	-4,980	0,00019015	0,11345346	0,864
243356_at	0,374	6,873	4,970	0,00019357	0,11345346	0,849
213802_at	-0,574	8,642	-4,966	0,00019512	0,11345346	0,842
211896_s_at	-0,665	5,634	-4,966	0,00019515	0,11345346	0,842
206089_at	0,444	8,651	4,961	0,00019713	0,11345346	0,833
206233_at	-0,855	6,815	-4,933	0,00020780	0,11834713	0,789
214157_at	0,681	8,093	4,922	0,00021210	0,11955052	0,771
233538_s_at	-0,466	4,091	-4,909	0,00021723	0,12051043	0,751
201617_x_at	-0,801	9,346	-4,907	0,00021821	0,12051043	0,747
209985_s_at	-0,834	6,450	-4,897	0,00022209	0,12142803	0,732

Gene3-Ctrl08

ID	M	A	t	P.Value	adj.P.Value	B
210095_s_at	-1,268	6,076	-9,541	0,00000014	0,00218521	5,490
211737_x_at	-0,824	8,896	-9,422	0,00000016	0,00218521	5,403
206291_at	-1,162	5,093	-9,279	0,00000020	0,00218521	5,297
212143_s_at	-0,786	5,252	-9,277	0,00000020	0,00218521	5,295
211813_x_at	-0,872	5,814	-9,273	0,00000020	0,00218521	5,292
234024_at	0,812	8,743	8,603	0,00000050	0,00455699	4,759
243173_at	1,685	7,326	8,172	0,00000093	0,00724992	4,383
201739_at	-0,789	9,108	-7,593	0,00000221	0,01507597	3,834
242524_at	0,712	9,090	7,231	0,00000387	0,01985848	3,465
228188_at	-0,914	5,024	-7,187	0,00000415	0,01985848	3,419
213880_at	-0,753	8,827	-7,166	0,00000429	0,01985848	3,397
212558_at	-0,865	6,123	-7,109	0,00000469	0,01985848	3,336
213348_at	0,881	10,415	7,105	0,00000472	0,01985848	3,332
222162_s_at	-0,918	8,592	-7,015	0,00000546	0,02131700	3,234
201170_s_at	1,135	6,159	6,883	0,00000675	0,02347532	3,090
243940_at	-0,609	4,407	-6,872	0,00000687	0,02347532	3,078
228104_at	-0,778	5,915	-6,786	0,00000791	0,02544328	2,981
231341_at	0,692	9,978	6,709	0,00000897	0,02647296	2,894
217057_s_at	1,385	9,044	6,694	0,00000920	0,02647296	2,877
201893_x_at	-1,005	6,391	-6,540	0,00001188	0,03246784	2,699
227566_at	-0,643	4,336	-6,404	0,00001492	0,03683851	2,539
237435_at	-1,111	8,761	-6,387	0,00001536	0,03683851	2,518
209897_s_at	-0,546	6,513	-6,382	0,00001550	0,03683851	2,512
232276_at	-0,649	5,424	-6,340	0,00001664	0,03790732	2,461
204284_at	0,605	7,120	6,218	0,00002048	0,04479221	2,313
204602_at	-0,496	11,354	-6,157	0,00002276	0,04666805	2,237
238877_at	-0,856	6,131	-6,150	0,00002305	0,04666805	2,228

206577_at	-1,180	4,752	-6,106	0,00002486	0,04855331	2,174
203726_s_at	0,649	5,175	6,015	0,00002910	0,05486158	2,060
212171_x_at	0,835	8,618	5,993	0,00003022	0,05507264	2,032
230015_at	0,620	10,044	5,952	0,00003249	0,05729863	1,980
200799_at	-0,547	10,641	-5,925	0,00003406	0,05784027	1,945
1558404_at	-0,657	3,513	-5,909	0,00003502	0,05784027	1,925
227662_at	-0,630	10,919	-5,894	0,00003597	0,05784027	1,905
232317_at	-0,732	4,330	-5,876	0,00003709	0,05794063	1,883
209988_s_at	-0,828	10,025	-5,850	0,00003884	0,05898628	1,849
213768_s_at	-0,615	8,011	-5,700	0,00005076	0,07259075	1,652
213802_at	-0,658	8,642	-5,690	0,00005163	0,07259075	1,639
202504_at	1,036	6,861	5,688	0,00005178	0,07259075	1,637
1554140_at	0,819	3,863	5,638	0,00005667	0,07632439	1,570
200878_at	0,499	11,714	5,633	0,00005723	0,07632439	1,563
229581_at	0,774	7,127	5,592	0,00006161	0,08020185	1,508
1563658_a_at	-0,665	5,564	-5,510	0,00007144	0,09083867	1,397
210393_at	-0,723	7,199	-5,469	0,00007689	0,09496486	1,342
203824_at	-0,729	4,946	-5,426	0,00008325	0,09496486	1,282
243356_at	0,408	6,873	5,415	0,00008485	0,09496486	1,268
209466_x_at	-0,602	8,228	-5,411	0,00008555	0,09496486	1,262
209465_x_at	-0,814	7,421	-5,403	0,00008673	0,09496486	1,251
225644_at	-0,611	9,023	-5,398	0,00008747	0,09496486	1,245
204288_s_at	0,522	8,465	5,398	0,00008761	0,09496486	1,244
218170_at	-0,568	11,096	-5,388	0,00008914	0,09496486	1,231
209291_at	-1,259	4,962	-5,381	0,00009032	0,09496486	1,221
235333_at	-0,643	7,472	-5,366	0,00009282	0,09575242	1,200
218974_at	0,581	10,702	5,354	0,00009487	0,09605701	1,184
204780_s_at	0,914	5,459	5,300	0,00010480	0,10418087	1,108
1558795_at	1,051	8,095	5,267	0,00011136	0,10872211	1,062
204339_s_at	0,449	12,558	5,244	0,00011611	0,11137386	1,030
219935_at	-0,953	4,651	-5,220	0,00012134	0,11185286	0,997
235368_at	-0,419	4,783	-5,217	0,00012210	0,11185286	0,992
1555338_s_at	0,494	6,854	5,214	0,00012275	0,11185286	0,988
209987_s_at	-0,764	8,146	-5,174	0,00013222	0,11851363	0,931
205311_at	0,458	12,672	5,141	0,00014054	0,12344733	0,885
226080_at	-0,495	8,373	-5,135	0,00014224	0,12344733	0,875
209170_s_at	-0,510	8,118	-5,096	0,00015280	0,12971607	0,821
228323_at	-0,445	8,890	-5,090	0,00015471	0,12971607	0,811
228115_at	-0,788	8,151	-5,083	0,00015658	0,12971607	0,802
221760_at	-0,582	8,838	-5,067	0,00016153	0,13181164	0,778
207542_s_at	0,695	6,524	5,029	0,00017331	0,13934851	0,724
211896_s_at	-0,672	5,634	-5,014	0,00017839	0,14135760	0,702
237322_at	0,470	4,536	4,982	0,00018918	0,14776621	0,656
216894_x_at	0,878	8,114	4,964	0,00019592	0,15087148	0,629
236179_at	-0,954	4,200	-4,917	0,00021410	0,16257949	0,561
239437_at	-0,467	5,691	-4,894	0,00022348	0,16738003	0,527
219230_at	-0,642	8,661	-4,872	0,00023318	0,17102368	0,494
211421_s_at	-0,701	8,489	-4,868	0,00023460	0,17102368	0,490
243013_at	-0,472	4,138	-4,860	0,00023836	0,17147897	0,477
225728_at	0,642	8,295	4,844	0,00024589	0,17460024	0,453
202747_s_at	-0,768	8,123	-4,811	0,00026139	0,17933529	0,406
224997_x_at	0,993	5,524	4,809	0,00026281	0,17933529	0,401
244551_at	-0,825	4,363	-4,801	0,00026673	0,17933529	0,390
227443_at	-0,428	6,472	-4,798	0,00026823	0,17933529	0,385

228728_at	-0,489	5,638	-4,796	0,00026896	0,17933529	0,383
234986_at	-0,693	5,318	-4,761	0,00028791	0,18280695	0,330
213182_x_at	0,879	8,133	4,760	0,00028827	0,18280695	0,329
228971_at	-0,479	8,578	-4,752	0,00029268	0,18280695	0,317
210512_s_at	0,650	10,798	4,749	0,00029449	0,18280695	0,313
229580_at	-0,622	8,120	-4,741	0,00029919	0,18280695	0,300
235616_at	-0,701	5,645	-4,735	0,00030253	0,18280695	0,292
207173_x_at	-0,940	7,919	-4,732	0,00030410	0,18280695	0,288
224707_at	0,399	8,416	4,730	0,00030528	0,18280695	0,285
229512_at	-0,389	7,207	-4,729	0,00030593	0,18280695	0,283
210729_at	-0,988	6,567	-4,722	0,00031027	0,18280695	0,272
203925_at	-0,502	8,361	-4,720	0,00031095	0,18280695	0,270
224894_at	-0,812	4,719	-4,713	0,00031555	0,18353784	0,259
222227_at	-0,481	8,339	-4,698	0,00032483	0,18694610	0,236
217853_at	-0,626	7,954	-4,675	0,00033953	0,19337016	0,201
207414_s_at	0,458	7,625	4,666	0,00034493	0,19406308	0,189
210430_x_at	0,439	5,540	4,660	0,00034908	0,19406308	0,180
230596_at	-0,689	6,395	-4,653	0,00035374	0,19406308	0,169
232071_at	-0,818	5,833	-4,650	0,00035592	0,19406308	0,164

* Note that even if the genes are sorted by the p-value here, we are also interested in high M-values (log-fold change between gene-control)

Table 6. List of the 10 most different genes for the contrast Gene2-Gene3.

Gen2-Gen3

ID	M	A	t	P.Value	adj.P.Value	B
204780_s_at	-1,084	5,459	-6,284	0,00001829	0,41147726	-1,942
211371_at	1,580	6,440	6,186	0,00002165	0,41147726	-1,972
213418_at	0,880	4,755	6,161	0,00002258	0,41147726	-1,979
1554140_at	-0,865	3,863	-5,960	0,00003201	0,43754024	-2,044
208180_s_at	-0,636	4,128	-4,770	0,00028284	0,99998961	-2,506
229069_at	-0,505	6,651	-4,759	0,00028913	0,99998961	-2,511
225238_at	-0,770	6,846	-4,726	0,00030782	0,99998961	-2,526
204566_at	-0,509	10,115	-4,717	0,00031287	0,99998961	-2,530
208048_at	0,345	4,556	4,453	0,00052063	0,99998961	-2,654
1554141_s_at	-0,631	4,075	-4,444	0,00052990	0,99998961	-2,659

* Note that here the B-statistic is negative, which indicates that none of these genes is differentially expressed.

Appendix B

R-code

```
> biocLite()
# This will download all basic packages for Bioconductor
> library(affy)
> Data<-ReadAffy()
# Data includes Ctrl07., Ctrl07.., Ctrl07..., Gen1., Gen1.., Gen1..., Ctrl08., Gen2., Gen3.,
Ctrl08.., Gen2.., Gen3.., Ctrl08..., Gen2..., Gen3...
>cdfName(Data)
[1] “HG-U133_Plus_2”

> sampleNames(Data)<-c("ctrl07.","ctrl07..", "ctrl07...","gen1.","gen1..","gen1...",
"ctrl08.","gen2.","gen3.","ctrl08..","gen2..","gen3..","ctrl08...","gen2...","gen3...")
> sampleNames(Data)
[1] "ctrl07." "ctrl07.." "ctrl07..." "gen1." "gen1.." "gen1..."
[7] "ctrl08." "gen2." "gen3." "ctrl08.." "gen2.." "gen3.."
[13] "ctrl08..." "gen2..." "gen3..."

> ctrl07.
AffyBatch object
size of arrays=1164x1164 features (7 kb)
cdf=HG-U133_Plus_2 (54675 affyids)
number of samples=1
number of genes=54675
annotation=hgu133plus2
notes=


> hist(Data)
> boxplot(Data)
> DataNormalized<-normalize(Data, “quantiles”)
> hist(DataNormalized)
> boxplot(DataNormalized)
```

Quality control

```
> library(simpleaffy)
> exprs_Mas5<-call.exprs(Data, "mas5")
> qc_Mas5<-qc(Data, exprs_Mas5)
> ratios(qc_Mas5)
```

RNA degradation

```
> Datadeg<-AffyRNAdeg(Data)
> names(Datadeg)
[1] "N" "sample.names" "means.by.number" "ses"
[5] "slope" "pvalue"

> summaryAffyRNAdeg(Datadeg)
> plotAffyRNAdeg(Datadeg)
```

Expression values

```
> library(affy)
> eset<-rma(Data)
> eset_mas5<-mas5(Data)

> eset_Mas5<-expresso(Data, bgcorrect.method="mas",pmcorrect.method="mas",
summary.method="mas")

> eset_RMA<-expresso(Data, normalize.method="quantiles", bgcorrect.method="rma",
pmcorrect.method="pmonly", summary.method="medianpolish")

>bgRMA<-bg.correct(Data, “rma”)

design.matrix<-
matrix(data=c(1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,
1,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,1,0,0,1,0,0,1),nrow=15,ncol=5)
>colnames(design.matrix)<-c(“ctrl07”,”gen1”,”ctrl08”,”gen2”,”gen3”)

>contrast.matrix<-makeContrasts(gen1-ctrl07, gen2-ctrl08, gen3-ctrl08, levels=design.matrix)
```

Fitting a linear model

```
>library(limma)
>fit<-lmFit(eset, design.matrix)
OR
>fit_Mas5<-lmFit(eset_Mas5, design.matrix)
>fit_RMA<-lmFit(eset_RMA, design.matrix)

>fit2_RMA_contrast<-contrasts.fit(fit_RMA, contrast.matrix)
>fit2_RMA_eb<-eBayes(fit2_RMA_contrast)

>results<-decideTests(fit2_RMA_eb)
> summary(results)
   gen1 - ctrl07 gen2 - ctrl08 gen3 - ctrl08
-1          4          22          20
0        54666       54633       54647
1           5          20           8
```

TopTable sorted by M

```
>library(limma)
> topTable1<-topTable(fit2_RMA_eb, coef=1,adjust="BH",sort.by="logFC")   #n=10
> topTable2<-topTable(fit2_RMA_eb, coef=2,adjust="BH",sort.by="logFC")
> topTable3<-topTable(fit2_RMA_eb, coef=3,adjust="BH",sort.by="logFC")

> topRMA1<-topTable(fit2_RMA_eb, coef=1, number=100, adjust="BH", sort.by="logFC")
write.table(topRMA1, file="topRMA1.txt", quote=F, sep="\t")
> topRMA2<-topTable(fit2_RMA_eb, coef=2, number=100, adjust="BH", sort.by="logFC")
write.table(topRMA2, file="topRMA2.txt", quote=F, sep="\t")
```

```
> topRMA3<-topTable(fit2_RMA_eb, coef=3, number=100, adjust="BH", sort.by="logFC")
write.table(topRMA3, file="topRMA3.txt", quote=F, sep="\t")


> prob1.tt<-topTable1[,1]  # ID-column for n=10 Gene1-Ctrl07
> prob2.tt<-topTable2[,1]
> prob3.tt<-topTable3[,1]

> probes1.tt<-topRMA1[,1]  #ID-column for n=100  Gene1-Ctrl07
> probes2.tt<-topRMA2[,1]
> probes3.tt<-topRMA3[,1]
> tt.idx.1<-as.numeric(rownames(topRMA1))   # rownames for n=100  Cene1-Ctrl07
> tt.idx.2<-as.numeric(rownames(topRMA2))
> tt.idx.3<-as.numeric(rownames(topRMA3))


> library(KTH)
> venn.100<-vennDia(A=tt.idx.1,B=tt.idx.2,C=tt.idx.3, names=c("Gene1-Ctrl07","Gene2-Ctrl08","Gene3-Ctrl08"), cex=0.8)

>biocLite("hgu133plus2")
>library(hgu133plus2)

> tt1.symbol<-mget(probes1.tt, hgu133plus2SYMBOL)
> tt1.accnum<-mget(probes1.tt, hgu133plus2ACCNUM)
> tt1.genename<-mget(probes1.tt, hgu133plus2GENENAME)
> tt2.symbol<-mget(probes2.tt, hgu133plus2SYMBOL)
> tt2.accnum<-mget(probes2.tt, hgu133plus2ACCNUM)
> tt2.genename<-mget(probes2.tt, hgu133plus2GENENAME)
> tt3.symbol<-mget(probes3.tt, hgu133plus2SYMBOL)
> tt3.accnum<-mget(probes3.tt, hgu133plus2ACCNUM)
> tt3.genename<-mget(probes3.tt, hgu133plus2GENENAME)

> annotate.tt1<-cbind(tt1.symbol, tt1.accnum, tt1.genename)
> annotate.tt2<-cbind(tt2.symbol, tt2.accnum, tt2.genename)
> annotate.tt3<-cbind(tt3.symbol, tt3.accnum, tt3.genename)

> topRMA1_summary<-cbind(annotate.tt1, topRMA1)
> topRMA2_summary<-cbind(annotate.tt2, topRMA2)
> topRMA3_summary<-cbind(annotate.tt3, topRMA3)

>write.table(topRMA1_summary, file=”topRMA1_summary.txt”)
```

TopTable sorted by P-value

```
> topRMA1P<-topTable(fit2_RMA_eb, coef=1, number=100, adjust="BH",sort.by="P")
> write.table(topRMA1P, file="topRMA1P.txt", quote=F, sep="\t")
> topRMA2P<-topTable(fit2_RMA_eb, coef=2, number=100, adjust="BH", sort.by="P")
> write.table(topRMA2P, file="topRMA2P.txt", quote=F, sep="\t")
> topRMA3P<-topTable(fit2_RMA_eb, coef=3, number=100, adjust="BH", sort.by="P")
> write.table(topRMA3P, file="topRMA3P.txt", quote=F, sep="\t")

> probes1P<-topRMA1P[,1]
> probes2P<-topRMA2P[,1]
> probes3P<-topRMA3P[,1]
> idx1P<-as.numeric(rownames(topRMA1P))
> idx2P<-as.numeric(rownames(topRMA2P))
> idx3P<-as.numeric(rownames(topRMA3P))

> symbol1P<-mget(probes1P, hgu133plus2SYMBOL)
> accnum1P<-mget(probes1P, hgu133plus2ACCNUM)
> genename1P<-mget(probes1P, hgu133plus2GENENAME)
> symbol2P<-mget(probes2P, hgu133plus2SYMBOL)
> accnum2P<-mget(probes2P, hgu133plus2ACCNUM)
> genename2P<-mget(probes2P, hgu133plus2GENENAME)
> symbol3P<-mget(probes3P, hgu133plus2SYMBOL)
> accnum3P<-mget(probes3P, hgu133plus2ACCNUM)
> genename3P<-mget(probes3P, hgu133plus2GENENAME)

> annotate1P<-cbind(symbol1P, accnum1P, genename1P)
> annotate2P<-cbind(symbol2P, accnum2P, genename2P)
> annotate3P<-cbind(symbol3P, accnum3P, genename3P)
> write.table(annotate1P, file="summary1P.txt")
> write.table(annotate2P, file="summary2P.txt")
> write.table(annotate3P, file="summary3P.txt")
```

MA-plots for the three contrasts

```
> plot(x=fit2eb$Amean, y=fit2eb$coef[,1], main="Gene1-Ctrl07", xlab="Log2 mean
intensity", ylab="Log2 FC", ylim=c(-3,3), cex=0.5, cex.main=1.5)

> plot(x=fit2eb$Amean, y=fit2eb$coef[,2], main="Gene2-Ctrl08", xlab="Log2 mean
intensity", ylab="Log2 FC", ylim=c(-3,3), cex=0.5, cex.main=1.5)

> plot(x=fit2eb$Amean, y=fit2eb$coef[,3], main="Gene3-Ctrl08", xlab="Log2 mean
intensity", ylab="Log2 FC", ylim=c(-3,3), cex=0.5, cex.main=1.5)
```

MA-plots with red points

```
> plot(x=fit2_RMA_eb$Amean, y=fit2_RMA_eb$coef [,1],main="Gene1-Ctrl07",
xlab="Log2 mean intensity", ylab="Log2 FC", ylim=c(-3,3),cex=0.5,cex.main=1.5)
> points(x=fit2_RMA_eb$Amean[prob1.tt],y=fit2_RMA_eb$coef[,1][prob1.tt], pch=2,
col="red",cex=0.8)

> plot(x=fit2_RMA_eb$Amean, y=fit2_RMA_eb$coef [,2],main="Gene2-Ctrl08",
xlab="Log2 mean intensity", ylab="Log2 FC", ylim=c(-3,3),cex=0.5,cex.main=1.5)
> points(x=fit2_RMA_eb$Amean[prob2.tt], y=fit2_RMA_eb$coef[,2][prob2.tt], pch=2,
col="red",cex=0.8)
```

```
> plot(x=fit2_RMA_eb$Amean, y=fit2_RMA_eb$coef [,3],main="Gene3-Ctrl08",
xlab="Log2 mean intensity", ylab="Log2 FC", ylim=c(-3,3),cex=0.5,cex.main=1.5)
> points(x=fit2_RMA_eb$Amean[prob3.tt], y=fit2_RMA_eb$coef[,3][prob3.tt], pch=2,
col="red",cex=0.8)
```

Clustering

```
>library(MASS)
>expression.matrix<-exprs(eset)
>dist<-dist(t(expression.matrix), method=”man”)    #  “man” = Manhattan –method
> mds<-isoMDS(dist)
initial value 22.950836
final value 22.950836
converged

>plot(mds$points, type=”n”, ann=F, axes=T, main=”Sum of M-differences between
samples”)
>text(mds$points, labels=colnames(expression.matrix), col=c(1,1,1,2,2,2,3,4,5,3,4,5,3,4,5),
cex=0.8)


> volcanoplot(fit2_RMA_eb, coef=1, highlight=10)
> volcanoplot(fit2_RMA_eb, coef=2, highlight=10)
> volcanoplot(fit2_RMA_eb, coef=3, highlight=10)
> head(topTable(fit2_RMA_eb, coef=1))

> qqt(fit2_RMA_eb$t, df=fit2_RMA_eb$df.prior+fit2_RMA_eb$df.residual,pch=16,
cex=0.2)  # Student’s t Q-Q Plot
> abline(0,1)
```

MM and PM index

Returns the location of mis/perfect matches in the intensity matrix

```
>index<-pmindex(ctrl1)
>names(index)[1:2]
[1] “1007_s_at” “1053_at”
>index[1:2]

#the locations are ordered from 5’ to 3’ on the target transcript

>index<-mmindex(ctrl1)
>names(index)[1:2]
[1] “1007_s_at” “1053_at”
>index[1:2]
```

Contrast Gene2-Gene3

```
> fit<-lmFit(eset_RMA, design.matrix)
> fit_RMA_contrast2<-contrasts.fit(fit, contrast2)
> contrast2
      Contrasts
```

```
Levels   gen2-gen3
  ctrl07        0
  gen1          0
  ctrl08        0
  gen2          1
  gen3         -1
> fit_RMA_eb2<-eBayes(fit_RMA_contrast2)

> topGen2_Gen3<-topTable(fit_RMA_eb2,number=100, adjust="BH",sort.by="P")
> write.table(topGen2_Gen3, file="topGen2_Gen3.txt", quote=F,sep="\t")

> probesGG<-topGen2_Gen3[,1]
> idxGG<-as.numeric(rownames(topGen2_Gen3))

> symbolGG<-mget(probesGG,hgu133plus2SYMBOL)
> accnumGG<-mget(probesGG,hgu133plus2ACCNUM)
> genenameGG<-mget(probesGG,hgu133plus2GENENAME)
> annotateGG<-cbind(symbolGG,accnumGG,genenameGG)
> write.table(annotateGG,file="summaryGen2_Gen3.txt", quote=F, sep="\t")
```