HÖGSKOLAN SKÖVDE

School of Humanities and Informatics
Dissertation in Bioinformatics 20p
Advanced level
Spring term 2006

The effect of normalization methods on the identification of differentially expressed genes in microarray data

Vilhelm Yngvi Kristinsson

The effect of normalization methods on the identification of differentially expressed genes in microarray data

Submitted by Vilhelm Yngvi Kristinsson to the University of Skövde as a dissertation towards the degree of B.Sc. by examination and dissertation in the School of Humanities and Informatics. This work has been supervised by Jane Synnergren.

2007-01-29

I hereby certify that all material in this dissertation which is not my own work has been identified and that no work is included for which a degree has already been conferred on me.

Signature: ____

The effect of normalization methods on the identification of differentially expressed genes in microarray data

Vilhelm Yngvi Kristinsson

Abstract

In this thesis the effect of normalization methods on the identification of differentially expressed genes is investigated. A zebrafish microarray dataset called Swirl was used in this thesis work. First the Swirl dataset was extracted and visualized to view if the robust spline and print tip loess normalization methods are appropriate to normalize this dataset. The dataset was then normalized with the two normalization methods and the differentially expressed genes were identified with the LimmaGUI program. The results were then evaluated by investigating which genes overlap after applying different normalization methods and which ones are identified uniquely after applying the different methods. The results showed that after the normalization methods were applied the differentially expressed genes that were identified by the LimmaGUI program did differ to some extent but the difference was not considered to be major. Thus the main conclusion is that the choice of normalization method does not have a major effect on the resulting list of differentially expressed genes.

Keywords: differentially expressed genes, normalization, microarray, MA-plot, box plot.

Acknowledgments

I would like to thank my supervisor Jane Synnergren at the University of Skövde for being extremely patient and supportive during this project. Without her ideas and comments I would have been lost. I would also like to thank Björn Olsson at the University of Skövde for comments and support during the project. Also I would like to thank my dear friend Samuel Genheden for his support and ideas when I needed them. And last but not least my girlfriend Bryndís and my family for their precious support. In the middle of this project I went through some hard times and without all of the understanding and support of all of you I would not have been able to finish this project.

Table of contents

1 Introduction

Over the years biological data for many organisms has been increasing exponentially. The need for effective methods to analyze these data to increase our understanding of the regulation and functions of genes is crucial. Techniques for analyzing gene expression data have been developed which facilitate this task. However, these techniques are not perfect and they need to be improved.

One of the main objectives in sequence analysis is functional genomics. Functional genomics entails understanding the effects and functions of genes: how and why genes behave in certain species and under specific conditions (Mehta, 2005). One of the ways to determine the behavior of a gene under specific conditions is to conduct a microarray experiment which will measure the gene expression of the condition and species in question. There are several techniques to measure the gene expression levels, e.g. RT-PCR and microarrays. RT-PCR is suitable for measuring the gene expression level for a limited number of genes while microarrays can measure the expression of thousands of genes simultaneously (Kerr and Churchill, 2001). With this advancement the amount of expression data has increased enormously which increases the demand for biologically relevant interpretation.

As described in Irizarry et al. (2002) variation between data that is caused by different gene expression is referred to as *interesting variation*. However, sometimes the variation is caused by the preparation of the microarray and the processing of the microarray (labeling, hybridization, and scanning). This is known as *obscuring variation* and can have many different effects on the data. By using normalization this obscuring variation can be reduced. Unless arrays are appropriately *normalized*, comparing data from different arrays can lead to misleading results.

There are many normalization methods based on different algorithms, and an investigation of the disparity between their results is important for further analysis of the microarray data. This is a rather complicated task since different normalization methods produce different results depending on what kind of datasets they are applied to and what kind of normalization that is needed.

The main objective in a microarray experiment is usually to identify differentially expressed genes. Identifying differentially expressed genes is to identify if two or more genes are differentially expressed under one or more conditions. It is very important that the data is properly normalized before this is done, otherwise the results can be misleading.

The aim of this thesis is to investigate the effect of using different normalization methods and how they affect the identification of differentially expressed genes. Since the normalization methods available at present differ in several ways, it is the hypothesis in this thesis that the choice of normalization method also has a major effect on the identification of differentially expressed genes. To investigate this possibility, two different normalization methods are applied to a set of microarray data. The methods used are "print tip loess" and "robust spline" which are both available in the LimmaGUI package. After the data has been normalized with the two normalization methods it is analyzed separately with linear models that are also available in the LimmaGUI package to identify differentially expressed genes. A list of differentially expressed genes is generated for each normalization method and they are then compared to examine the different effects the normalization methods have on the identification of differentially expressed genes. To examine the different effects the percentage of genes that overlap between the gene lists are analyzed and also the genes that are uniquely identified in each gene list. The genes that are identified uniquely in the two data sets normalized by two different normalization methods will show the different effects the normalization methods have on the identification of differentially expressed genes.

The rest of this report is structured as follows: Section 2 describes the background knowledge needed to understand the method used in this work. This is not required reading if the reader is familiar with the topic. Section 3 discusses problems relevant to this thesis and the aims and objectives are stated. Section 4 describes the used method and section 5 describes the results and analyses which were derived during the thesis project.

2 Background

2.1 Basic biology

The DNA is the genetic material which maintains the cellular and biochemical functions of an organism. The DNA is stationary in the nucleus of each cell in the organism. In most organisms the DNA is a double stranded polymer. This polymer has a sequence of units (nucleotides) and when double stranded the units on one strand are complementary to the other strand. The nucleotides are grouped together to produce a codon. Each codon has three nucleotides which are translated into one amino acid This process is called the central dogma of molecular biology and is illustrated in figure 1.

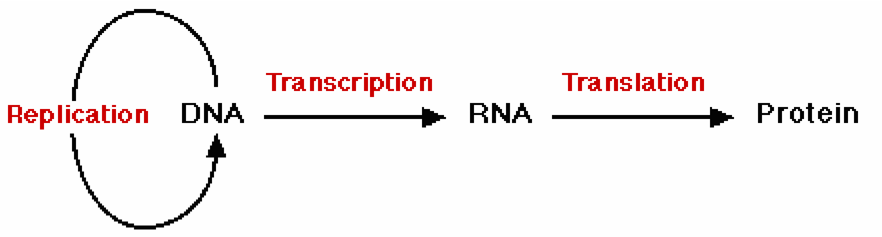


Figure 1 – A representation of the central dogma of molecular biology.

Proteins are produced by the genes that encode them. The genes are transcribed into mRNA which are then translated into amino acid polypeptides and those are later assembled into a protein. The structure of a protein is one of the factors that provide the protein with its function. The majority of amino acid sequence gets their structure by entering an enzyme called chaperon which “folds” the protein to its correct structure. Thus knowing the amino acid sequence of a protein is not enough to predict the function of that protein. There are other factors which can affect the structure of a protein which are for instance temperature and foreign chemicals. When predicting a function of a protein all these factors have to be taken into consideration.

2.2 Genomes and Genes

The definition of a gene in a molecular perspective is a specific nucleotide sequence that is transcribed into mRNA. A gene function is the function of the proteins that it encodes. In eukaryotes the DNA consist of introns (non-coding regions) and exons (coding region). Exons are the only part of the DNA that is transcribed into RNA and then later translated into a protein. In prokaryotes the DNA has no introns, only exons which are transcribed into RNA.

The DNA, with all the genes, is packed into chromosomes for storage until the time point when the gene is needed for transcription. The collection of all the chromosomes of the organism is called the genome of that particular organism.

2.2.1 Gene expression & Gene regulation

Gene expression is when the DNA sequence of a gene is converted into the functional product of that gene. Expressed genes include those that are transcribed into mRNA and translated into protein and also those that are just transcribed into RNA and not translated into proteins, for example ribosomal RNA (U.S. Dept. of Energy, 1997). Each cell of an organism has the same genome, but the cells are very different. This is because even though the cells have the same genetic material they use different parts of the genome. This is a very complex system which is not fully understood today. Thus, even though most of the genes from the organism are known it is not known in which cells and under which conditions these genes are expressed. Gene regulation is the control over gene expression, i.e. what amount of the functional product of each gene should be produced and at what time. This is very important because the slightest mishap can mean the death of the cell or the organism. Gene expression can be measured by measuring the mRNA concentration of a particular gene (Alberts et al., 2002). There are several techniques for measuring mRNA concentration where micorarrays are one of them.

2.3 Microarray technology

The DNA microarray is made out of a glass, plastic or silicon chip. This chip has many microscopic DNA spots on its surface which forms the array. The DNA spots are known as *probes,* because they are probing the sample which is hybridized to the chip. The sample which contains cDNA is called the *target* since the probes are looking to match these targets. Microarray experiments are typically made to compare two or more samples which represent two or more conditions, one being for example a cell that has mutated into a cancer cell and the other a normal cell (Lockhart and Winzeler, 2000).

There are two main types of microarrays; cDNA microarrays and oligonucleotide microarrays. The difference between these two microarrays is that in cDNA microarrays the cellular mRNA which is converted to cDNA hybridizes to a clone of a piece of DNA sequence but in oligonucleotide microarrays complementary mRNA (cRNA) hybridizes to short segments known as synthetic oligonucleotides. The probes used in the oligonucleotide microarray are much shorter than in the cDNA microarray. The cDNA microarray has advantages over the oligonucleotide microarray since it renders comparison of two conditions on a single chip, while oligonucleotide microarrays need one chip per condition (Sebastiani et al., 2003). These two types of microarrays will now be explained in more detail in the sections below.

2.3.1 cDNA microarrays

This section is based on information from Sebastiani et al. (2003).

The cDNA microarray or two channel microarray was developed at Stanford university.

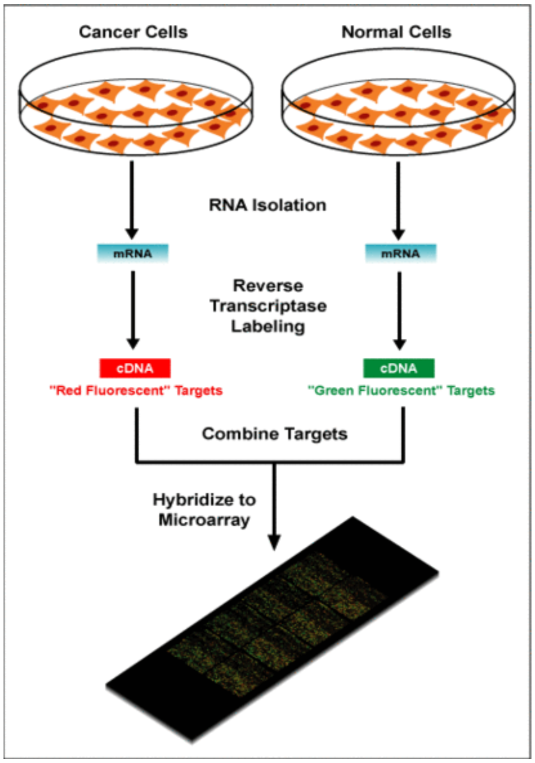


Figure 2 – A semantic representation of the expression microarray between a cancer cell and a normal cell. Non-copyrighted image from the Wikipedia Foundation (2005).

The first step in the process is to select cDNA probes which are going to be on the microarray. These probes are distributed on the array by a high-speed robot. Each probe corresponds to a gene which should be represented in the cDNA sample if the gene is expressed. The mRNA is usually extracted from two cell samples under different conditions, for example, tumor cells and healthy cells. The samples are then reverse transcribed into cDNA and labeled with fluorescent dyes, commonly Cy3 (green) and Cy5 (red). In figure 2 the normal sample has been labeled green and the cancer sample red. The samples are then mixed together and the mixture is hybridized to the probes on the glass slide. If the cDNA sequences in the samples find their complementary sequence on the glass slide they bind together.

After the hybridization the intensity for each color of each spot on the microarray is measured by a scanning microscope. Red color indicates that the gene is only expressed in the cancer cell sample; green color indicates that the gene is only expressed in the normal cell sample and yellow color indicates that the gene is expressed in both samples. Sometimes the color can be yellow-green and that indicates that the gene is expressed more in the normal cell, and vise versa if the color is yellow-red (brown/orange). If the color is black/gray it denotes that the gene is not expressed in either of the cell types.

There are two drawbacks with this method. One is that there is some risk of cross-hybridization and the second is that a large amount of total RNA is required to prepare the target (Duggan et al., 1999). Cross-hybridization is when a cDNA from the target sample binds to a similar probe in terms of codons but the probe is not completely complementary and this result in false detection values.

2.3.2 Oligonucleotide microarrays

This section is based on information gotten from Sebastiani et al. (2003).
Oligonucleotide microarrays are different from cDNA microarrays because their target sample is represented by a small cDNA fragment which is specific to a particular gene.

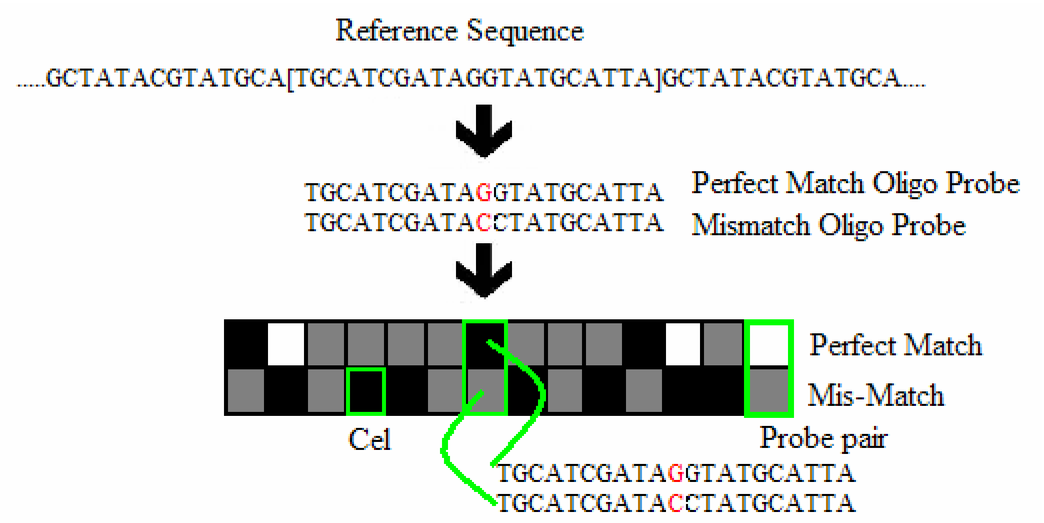


Figure 3 - A representation of how the PM probes and MM probes work together. See text for more details (Sebastiani et al., 2003)

One example of Oligonucleotide microarrays is Affymetrix GeneChip arrays. They use a short fragment made out of synthetic oligonucleotides that are later placed on a silicon chip. Affymetrix produces silicon chips that have already been prepared with probes. It is argued that this technology is better than the cDNA microarrays because the probes are represented by a set of well-chosen small segments of cDNA instead of a long subsequence of the gene. This reduces the chance that fragments of the target will randomly hybridize to the probes, thus reducing the likelihood of cross-hybridization. As can be seen in figure 3 each gene is not represented by its cDNA but by a set of fixed-length independent segments unique to the DNA of the gene.

Each gene is represented by 11-20 *probe pairs* and all together they are called a *probe set*. Figure 3 shows an overview of how the PM and MM works. A probe pair consists of a PM (perfect match) probe and a MM (mismatch) probe. The PM probe is chosen so that it will represent a unique part of a gene, so that the odds will increase that it will hybridize with high specificity. The PM probe is identical to the MM probe except for a single base in the central position which is replaced with a complementary base. The MM probe is used as a specificity control because if it hybridizes we know that it is due to some kind of cross-hybridization or background signal such as cell debris or salts that bind to the probes. A probe cell is a single square-shaped area on the array that contains many copies of a given 25-mer oligonucleotide or probe. Each probe cell of an Affymetrix oligonucleotide consists of millions of PM and MM probes. Probe cells which tag the same gene are scattered all over the matrix to avoid systematic bias.

The target is prepared by extracting the entire mRNA from the target cell. The mRNA is then reverse-transcribed into cDNA which is made double stranded. By using a transcription reaction the cDNA double helix is then converted to cRNA which fluorescently labels the target. The silicon chip is then allowed to hybridize to the target and then scanned with a laser scanner. The scanner generates an image that is organized so that the signal intensity of each probe cell can be measured (Sebastiani et al., 2003).

Although the Affymetrix GeneChip is less flexible than the cDNA arrays it does have some advantages. The amount of RNA needed to prepare the target is much smaller than for cDNA microarrays and the systematic bias is argued to be less because of the controls represented by the MM probes.

2.4 Analyzing microarray data

The process of analyzing data from a microarray experiment is as follows. First the data needs to be visualized to view if there is any obscuring variation that needs to be normalized. If there is any obscuring variation the data is normalized with the appropriate normalization method. Currently there are several normalization methods which are used to correct for different kinds of systematic bias in microarray data. Not all kinds of systematic variations and biases that exist will be discussed because that is beyond the scope of this thesis. The focus of this thesis is normalization methods that deal with intensity bias, spatial position bias and scale differences. After the normalization has been performed the data is visualized again to view if the normalization was a success. If the visualization results indicate that the normalization was a success the data can be analyzed. The analysis in this thesis is to identify differentially expressed genes.

2.4.1 Normalization

This section is based on information from Smyth and Speed (2003).

Normalization is the process to adjust for variation that is due to something else than a biological reason. In this thesis data from two channel cDNA microarrays are used, so the following description of the normalization process is assumed to be for data from a cDNA microarray experiment. In the process of conducting a two channel cDNA microarray experiment two different samples of mRNA which are colored differently are prepared and they are hybridized to the microarray. The hybridization results in green or red spots and each spot represents the expression of a particular gene. A scanner is used to measure the intensity of each of the two colors and to translate that to numbers that represent gene expression levels. Before comparing the two signals the intensity of the red and green colors must be normalized to reduce imbalance between the signals.

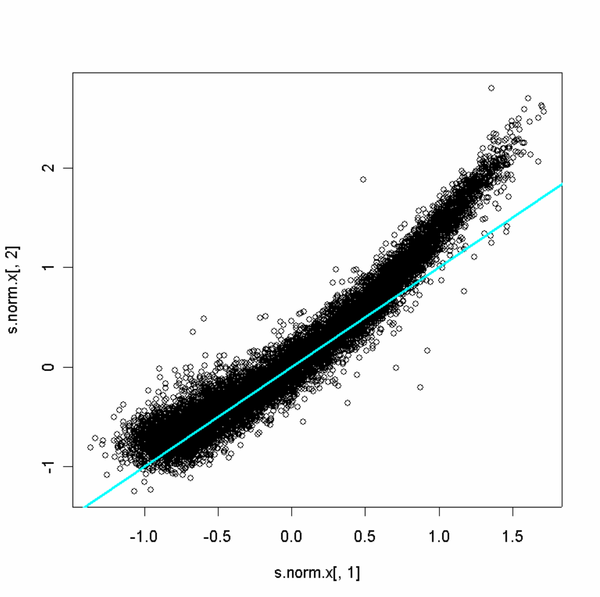


Figure 4 - A MA-plot which indicates a banana shaped plot which needs to be intensity normalized.

The imbalance can arise from different labeling efficiencies or scanning settings for the two fluorescent colors. Also they might fluorescent differently, which can produce variation between them because one signal might be stronger then the other. Often the imbalance is more complicated than what can be corrected by a simple scaling of one channel relative to the other. When this type of dye bias occurs, an intensity dependent normalization is needed. The dye bias can be between channels (between two mRNA samples) or it can be within one channel (between printed probes). When the bias is within a channel it is because the dyes may vary between spatial positions on the microarray. The spatial positions may differ because of, e.g. differences between the print-tips on the array printer. Variation between two channels can be because of, e.g. differences in print quality from differences in ambient conditions when the plates were processed or simply from changes in the scanner settings.

2.4.2 Visualization of Intensity and Spatial Trends

To analyse the need for normalization one can visualize the distribution of data in various plots e.g. MA-plots, which show the intensity and spatial trends. Smyth and Speed (2003) described that the R (red) and G (green) represent the expression for each gene and they are usually transformed into log ratios to reduce the expression range and make the data equally distributed around zero. On the y-axis is the M-value and its formula is M = Log R - Log G. On the x-axis is the A-value with the formula A = (Log G + Log R)/2. Figure 4 shows a banana-shaped plot, indicating that the data needs to be intensity normalized.

Another way to visualize bias in microarray data is to use a boxplot. Figure 5 shows a boxplot of two cDNA replicate microarrays. It is always good to use replicates of the arrays that you are going to compare. This way you can compare the replicates and see if there is any variation between them that is not due to biological variation. The green channel and the red channel are shown, and by comparing the red channels between these two replicates the quantiles should be lined up. They should line up because the distribution of values should be similar in the replicates. But as can be seen in the

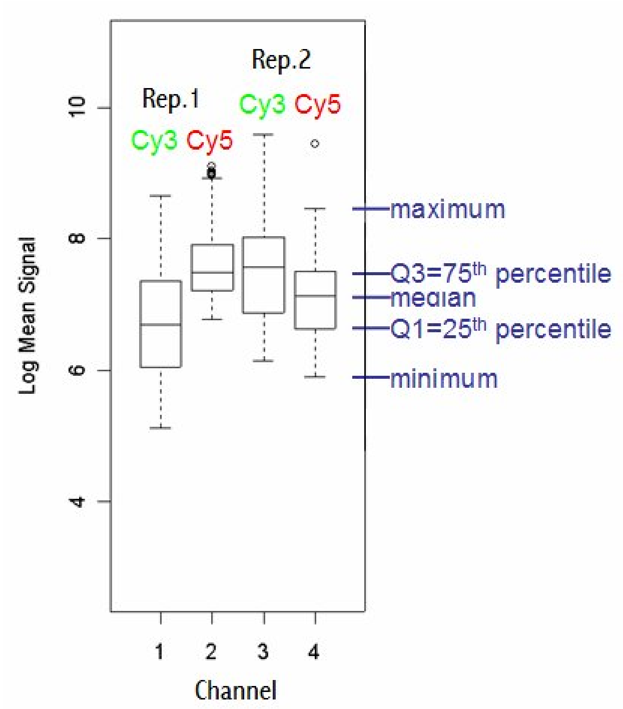


Figure 5 – A boxplot of two replicate cDNA microarrays that indicates that the two channels need to be normalized.

figure the replicates are not equal, which indicates that there is some variation between the microarrays that is not due to biological differences, and they need to be normalized to be comparable. This also applies for the green channel.

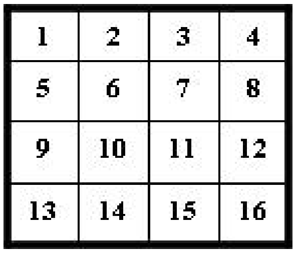
2.4.3 Loess Normalization

This section is based on information from Smyth and Speed (2003).

There are three types of loess normalization; print-tip normalization, composite normalization and global normalization. In short, print-tip loess normalization adjusts intensity and spatial trends that are within the array. The composite loess normalization on the other hand uses control spots that are known not to be differentially expressed. These control spots are then used to produce the loess curve and then the data is normalized according to that curve. Global normalization does not take print tip groups or control spots into consideration. In this thesis the print tip loess normalization method is the only one used.

The other ones are just described as background information on loess normalization.

Print-tip normalization normalizes the intensity and spatial bias within each print tip group. The probes are laid out on the microarray in terms of print tips. These print tips are on a so called array printer and to speed up the process of preparing the array there are many print tips grouped together. To demonstrate further lets consider an example.



1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Figure 6 – A representation of the arrangment of print tip groups.

There are 48 print tips in each print tip group and they have the arrangement of 4 x 4 probes as seen in figure 6. Each of these print tip groups can vary because of, e.g. differences in the sizes or openings of the print tips due to many hours of printing. We normalize this kind of variation by making a loess curve for each tip group and each M-value is made to fit this curve by subtracting the curve from the M-value. The normalized log-ratio *N* is the remainder and can be represented in the formula:

$$N = M - loess_i(A)$$

where loess_i (A) is the function of the print tip loess curve for tip group *i*. The loess curve is determined by a series of local regressions, one for each point in the MA-plot. The local regression curve is linear of degree 1. Each point on the curve is predicted by local regression of 40% of the points that are closest to it in terms of A-value. The local regressions are also estimated by using M estimation with Tukey’s biweight function. For a more detailed description on composite and global loess normalization please refer to Smyth and Speed (2003).

2.4.4 Robust spline normalization

The robust spline normalization method is used to deal with intensity dependent and spatial bias. This method is available only in the LimmaGUI package and is based on unpublished work of Gordon Smyth. A short description of the method is that the M values of a single microarray are normalized by using robustly fitted regression splines and empirical Bayes shrinkage. This method is very similar to the print tip loess method which is described in the section above but instead of using loess curves the robust spline method uses regression splines and empirical Bayes shrinkage to shrink the individual print-tip curves towards a familiar value. The advantage is that robust spline normalization results in a far more stable inference when the number of arrays is small (Smyth, 2004).

2.4.5 Identifying differentially expressed genes

There are several algorithms for identifying differentially expressed genes and some of them are available in the program LimmaGUI, which will be described in the next chapter. Simpler techniques such as fold change are also frequently used to identify differentially expressed genes. In this work the LimmaGUI is used and the false discovery rate (FDR) method is selected to identify potentially significant genes, ranked by p-value.

Identifying differentially expressed genes is a demanding problem to tackle and there are many difficulties in the field. Some genes that are not differentially expressed are identified as such by chance and this is a problem since we don’t want false positives to be included in the results. In the LimmaGUI program the false discovery rate (FDR) is estimated by analyzing the permutated measurements of each gene that is controlled.

2.4.6 The LimmaGUI program

LimmaGUI is a Graphical User Interface for Gordon Smyth's limma package (Linear Models for MicroArray data). The limma package (Smyth, 2006) is implemented for the R statistical program. As the name implies the limma package uses Empirical Bayes linear modeling which was proposed by Smyth (2004) to analyze microarray data. Linear models can be used in drawing regression lines through microarray data which needs to be normalized and have been found to be very efficient. The limmaGUI provides a point and click interface to the main functions of the limma package, which simplifies the use.

In summary, the program is used to normalize, find differentially expressed genes and perform diagnostic plots for microarray data. The normalization methods that are available within the package are median normalization, print tip loess, global loess, composite loess and robust spline. In this thesis the print tip loess and robust spline are compared. The diagnostic plots are used to view the microarray data and determine what kind of normalization is needed. The program offers many diagnostic plots and in this thesis the MA-plots and boxplots are mostly used. The print tip version of the MA-plots is also used to see if there are any differences that might be caused by the print tips. The program also offers some alternatives of how to find differentially expressed genes. In this thesis the option FDR (False discovery rate) is chosen and to rank the genes according to p-value. The FDR is chosen because it emphasizes the proportion of errors among the identified differentially expressed genes (Reiner et al., 2003).

The FDR is the expected percentage of false positives in a set of genes. For example if we get a false discovery rate of 40% in a set of 100 genes then we should expect that 60 of them are correctly identified as differentially expressed. By using the FDR to adjust the p-value the results about which genes are differentially expressed becomes more reliable.

3 Problem description and statement

Microarrays give us the possibility to investigate a large number of genes under numerous conditions simultaneously. This is a promising technique for investigating biological processes and the relationship between genes. But the problem lies in analyzing this enormous amount of data. Often, non-biological variation is present in the microarray data which distorts the result when comparing microarrays. By normalizing the data the obscure variation is reduced so that biological differences are more easily detected. There are a number of methods available for normalizing microarray data. None of the methods give the exact same result which brings up the question if the choice of normalization method has a significant effect on the subsequent analysis of the data.

By using normalization methods we risk loosing information on genes that are truly differentially expressed. This thesis will focus on the normalization of cDNA microarrays that have intensity bias, spatial position bias and scale differences. Two different methods for reducing these types of biases are print tip loess and robust spline normalization and in this thesis the focus will be in investigating their effect on the subsequent process of detecting differentially expressed genes. Print tip loess normalization is used to remove intensity dependent dye bias and spatial position bias and in this thesis it will be combined with scale normalization as recommended by Yang, et al. (2002). Robust spline normalization is also used to remove intensity bias but the method is quite different. The robust spline method uses robustly fitted regression splines and empirical Bayes shrinkage to normalize the microarray data. Here the robust spline method is also combined with scale normalization. Both of these normalization methods normalize the same type of bias so it is interesting to see if the lists of differentially expressed genes that are derived will differ.

3.1 Aims and Objectives

The aim of this thesis is *to investigate the effect of normalization methods when identifying differentially expressed genes. The evaluation will mainly compare the genes* *that are classified as differentially expressed in the data set after applying these two normalization methods with the ones that are detected separately for each method.*

The genes that are identified as differentially expressed after the normalization methods have been applied are listed. In this thesis it is proposed that by comparing the lists of differentially expressed genes it can be determined if the normalization methods have a major effect on the identification of differentially expressed genes.

The objectives are:

- Extraction of the Swirl microarray data.
- Visualization of the Swirl data to determine if the print tip loess and robust spline methods are appropriate normalization methods for this data set.
- Normalization of data using both print tip loess and robust spline normalization separately.
- Detection of differentially expressed genes by using the LimmaGUI package for the print tip loess normalized data and for the robust spline normalized data.
- Evaluation of results – investigate which genes overlap after applying different normalization methods and which ones are identified uniquely after applying the different methods. This will indicate if the normalization methods have a major effect on the identification of differentially expressed genes in microarray data.

4 Method

4.1 Extraction of the microarray data from the Swirl Zebrafish experiment

The Swirl dataset is provided by Katrin Wuennenberg–Stapleton from the Ngai Lab at UC Berkeley. Yang and Dudoit (2006) explained that the Swirl experiment consists of a wild-type zebrafish and the swirl type which is a point mutant in the BMP2 gene that affects the dorsal/ventral body axis. The goal of the experiment is to detect differentially expressed genes between the swirl and wild type zebrafish. The dataset contains four double colored cDNA replicate microarray slides. Each slide has been hybridized with target cDNA from the swirl mutant and the wild type zebrafish. The swirl mutant cDNA was labeled with Cy5 and the cDNA from the wild type zebrafish was labeled with Cy3. The microarray used contains 8,448 cDNA probes, which include 768 control spots (e.g. negative, positive, and normalization control spots). The microarrays are printed with 4 x 4 print tips which are partitioned into 4 x 4 grid matrices. Each grid contains a 22 x 24 spot matrix that is printed with a single print tip.

Each of the slides produced a 16-bit image which was processed by the image analysis software Spot. The dataset consists of four output files which are swirl.1.spot, swirl.2.spot, swirl.3.spot, swirl.4.spot. Each spot file consists of 8,448 rows and 30 columns. Each row corresponds to a particular spot and the columns contain various statistics provided by the Spot image analysis program.

4.2 Visualization of the Swirl dataset

In order to normalize the data properly a visualization process is needed to determine what kind of obscuring variation might be present in the data. By using MA-plots and boxplots different types of variation can be visualized. Images are also used to identify spatial background variation. The limmaGUI package was used to derive the plots for this visualization process.

MA-plots are created to investigate intensity based bias. A “banana shaped” MA-plot is often due to intensity based dye bias. This type of bias can be corrected for with the print tip loess or robust spline normalization method. Print tip group boxplots are created to investigate if there is any obscuring variation that might be caused by the print tip groups. A boxplot for each replicate plate is created to investigate if there are any scale differences between replicate plates. If the boxes are uneven, scale normalization is preferred. Background images are created to investigate if the dyes are stronger in one particular area of the microarray. Print tip loess or robust spline normalization methods are both appropriate for correcting this bias.

4.3 Normalization of the Swirl data

The visualization of the swirl dataset showed intensity based dye bias which was normalized with the print tip loess and the robust spline method for comparisons. The visualization also showed spatial variance in the background and the boxplots indicated some differences within the print tips for each print tip group which can be reduced by applying print tip loess or robust spline. The boxplot of each replicate microarray showed some variance and thus scale normalization was also needed.

The LimmaGUI package was used to normalize the microarrays. First the Swirl microarrays were normalized with the print tip loess normalization method and after that scale normalization was applied. Second, the Swirl microarrays were normalized with the robust spline method followed by scale normalization. Although the print tip normalization is used to normalize obscuring variation that is the cause of print tip groups it is comparable to the robust spline method as they both normalize independent dye bias. The R documentation of the limma package states about the robust spline method (Smyth, 2005):

“This function implements an idea similar to print-tip loess normalization but uses regression splines in place of the loess curves and uses empirical Bayes ideas to shrink the individual print-tip curves towards a common value. This allows the technique to

introduce less noise into good quality arrays with little spatial variation while still giving good results on arrays with strong spatial variation."

This shows that these two normalization methods are comparable and normalize the same kind of variation which is present in the Swirl data.

4.4 Detection of differentially expressed genes

As described in section 2.4.6 the limmaGUI package uses linear models to fit the microarray data in order to find differentially expressed genes. There are several ways to rank which genes are differentially expressed and in this thesis the FDR (False discovery rate) and p-value is used. The FDR method is used to adjust the p-value for multiple testing. The threshold for differentially expressed genes was set to $p < 0.05$.

4.5 Comparisons of differentially expressed genes given by each method

The lists of differentially expressed genes after normalizing with the print tip loess and the robust spline methods were derived by the limmaGUI package. Appendix A includes the lists of differentially expressed genes after using each of these two normalization methods (table 1 for the print tip loess method and table 2 for the robust spline method). To compare the overlap between the two methods the R program was used. The code that was used is shown in Listing 1.

Listing 1 – A representation of the code that was used in the R program to compare the two normalization methods.

To compare the overlap between the two lists of DEGs the R program is used. First the gene lists are derived from the LimmaGUI program as a text file (robust.txt for the robust spline method and printip.txt for the print tip loess method) and then it is read into the R program as a dataframe with the following code:

```
Robust <- read.table("robust.txt")
Printtip <- read.table("printip.txt")
```

Then the two gene lists are compared by using the intersect command.

```
Overlapping <- intersect(Robust[,"id"],Printtip[,"id"])
```

5 Results and analysis

5.1 Results from normalization

As emphasized in chapter 4.2 it is important to visualize the data before normalization to determine what normalization method is appropriate. As can be seen on the MA-plot images in figure 7 the plot for the un-normalized data is “banana shaped” which indicates that there is independent dye bias which needs to be corrected for.

The background images can be seen in Appendix B (Figure 1-4). Replicate slide number 1 shows that there is some spatial variation in the green background image where there is stronger green intensity in the middle of the slide and also in the corners. The same pattern can be seen for the red background image but the intensity is stronger only in the middle of the slide. For replicate slide number 2 the background images for both green and red signal are stronger in the upper left corner and near the edges of the microarray. It seems to be similar for replicate slide number 3 where the intensity is stronger in the upper edge of the microarray and in some random spots over the array that are stronger. As for replicate slide number 4 there is some spatial variation which can be seen as stronger colored spots to the left in the upper half for the green image and in the upper half in general, and in particular at some spots to the left and right in the upper half for the red image.

In Appendix B (Figure 5-8) the MA-plots show “banana shaped” loess lines through all the print tip groups. This is an indication of print tip bias which needs to be compensated.

The final normalization needed was scale normalization. Boxplots of all replicate arrays side by side can be seen in Appendix B (Figure 9). The print tip loess and robust spline methods both normalize some of the scale differences as can be seen in each figure. The boxplot in figure 9a of the data that has been normalized with the print tip method shows some unequal distribution of the intensity signals between replicate plates, which indicates that there are some scale differences. The same applies for figure 9b where the robust spline method was used. Thus scale normalization was used after the print tip loess and robust spline normalization.

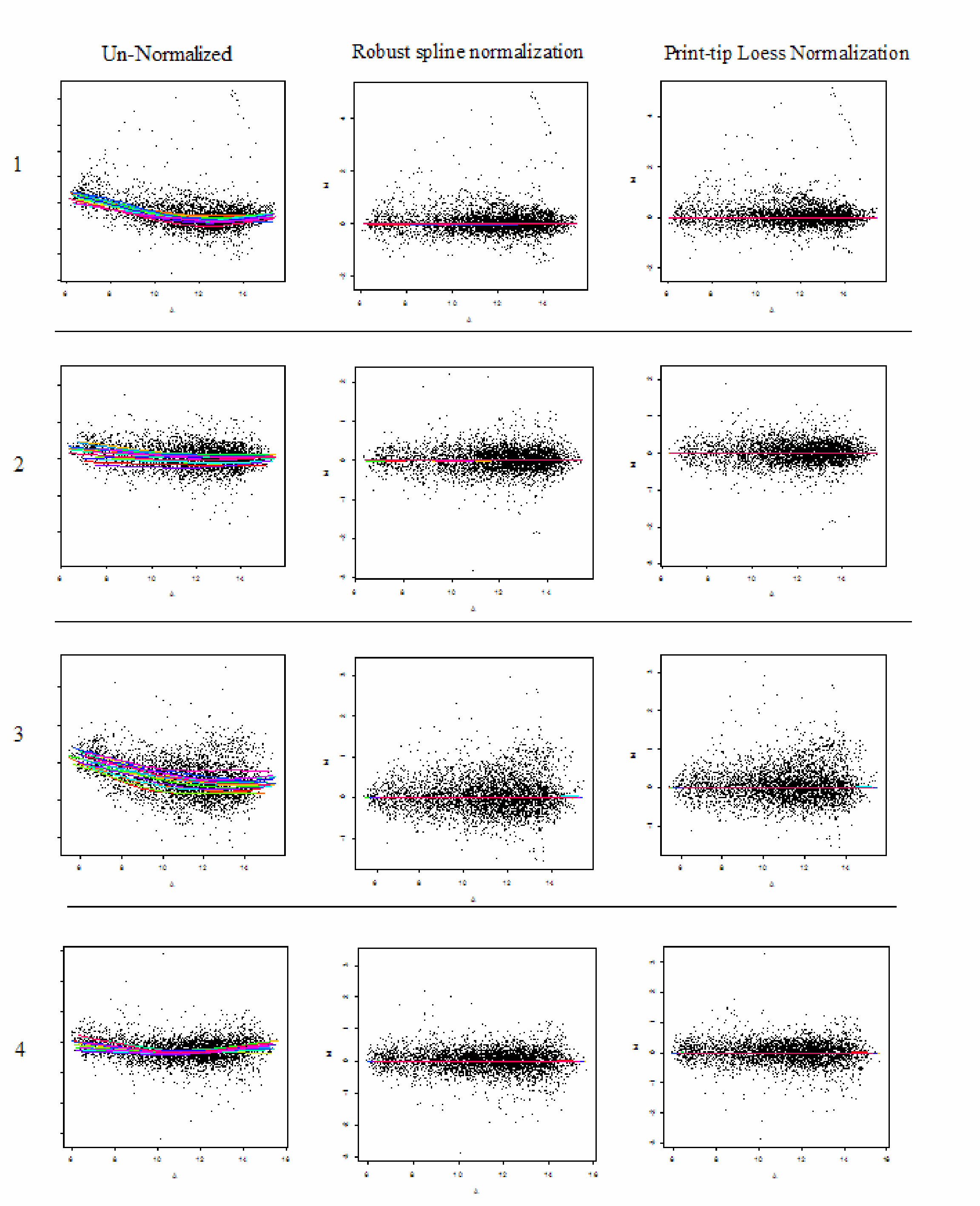


Figure 7 – MA plots of replicate slides 1 – 4 from the swirl dataset. Each column indicates a different normalization state. Column 1 represents un-normalized data. Column 2 represents Robust spline normalization and column 3 represents Print-tip Loess normalization.

To be sure that the normalization was efficient for both of the normalization methods the data was visualized after the normalization to evaluate the effect of the normalization. The MA-plots for each method, before and after normalization, is shown in figure 7. The figure shows a MA plot of each replicate slide and it can be seen clearly that the normalization has successfully decreased the independent dye bias because the plot is no longer “banana shaped” which indicates that the intensity is equal throughout the plot. This applies for all the replicate slides.

The limmaGUI package lacked the feature to generate background images and print tip group MA plots after the normalization process was completed. But since the MA-plots in figure 7 showed that the loess line is straight now and not curved as before it is assumed that the slides are ready to be analyzed.

In figure 1 Appendix C a boxplot of each replicate slide is shown before and after the scale normalization for each method. As can be seen in both figure 1a and 1b the boxes in the plot are more equal after the scale normalization which shows that the normalization has reduced the scale differences.

5.3 Results from the comparison of differentially expressed genes between methods

The differentially expressed genes (DEGs) identified after applying the normalization methods were compared using the R software as described in section 4.5.

To determine if a gene is differentially expressed or not a cutoff p-value of 0.05 was used. After the print tip loess method was applied 116 genes were identified with $p < 0.05$ and thus classified as differentially expressed. The identified DEGs are reported in Appendix A table 1. After the robust spline method was applied 120 genes were identified as differentially expressed. The identified DEGs are reported in Appendix A table 2.

The first part of the comparison was to calculate the number of genes that were detected as differentially expressed in both data sets after normalizing with these two methods respectively. There were 114 genes that overlapped between the two methods and the overlapping gene list can be seen in Appendix D table 1. To achieve measurements of overlap in percentage the following formula was used:

$$\frac{\mathit{Overlap}}{(\mathit{RobustSpline} + \Pr\mathrm{int}\,\mathit{Tip})/2} \tag{1}$$

Formula 1 calculates the number of DEGs that overlap between the two methods divided by the mean of uniquely identified DEGs from the robust spline and print tip loess method. The result from this was 96.6%. Table 1 below gives an overview of the statistics described in this section.

Table 1 - Statistics for the genes that were identified as differentially expressed after applying the two normalization methods. The columns represent: 1) the number of DEGs in the robust spline method, 2) the number of DEGs in the print tip loess method, 3) the number of DEGs that were identified in both methods, 4) the percentage between the robust spline and the print tip loess method (calculated with formula 1), and 5) the number of genes uniquely identified by the Robust spline method/Print tip loess method.

Nr. of DEGs in robust spline	Nr. of DEGs in Print tip loess	Nr. of DEGs in both methods.	Percentage (See formula 1)	Robust spline / Print tip loess
120	116	114	96.6%	6/2

To investigate whether the choice of p-value threshold affects the resulting percentage of overlapping DEGs, a number of different p-value cut off values were tested. The cut off values can be viewed in Appendix A table 1 for the print tip loess method and table 2 for the robust spline method.

Table 2 below shows the variation between different p-value thresholds. The p-value 0.05 which is used above is also kept in the table for comparisons. As can be seen in this table the different p-value thresholds only have minor effects on the percentage of overlapping DEGs. The only p-value that differs to some extent from the rest of the p-values is 0.02 and its percentage is 93.2% while the others range between 96 and 97%.

Table 2 - Comparisons of statistics for different p-value cut-offs. The first column represents the p-value threshold for identifying DEGs. Columns 2-6 are the same as columns 1-5 in Table 1 and are described in that caption.

P-value threshold	Nr. of DEGs in robust spline	Nr. of DEGs in Print tip loess	Nr. of DEGs in both methods.	Percentage (See formula 1)	Robust spline / Print tip loess
0.01	39	40	38	96.2%	1/2
0.02	70	65	63	93.3%	7/2
0.03	87	89	85	96.6%	2/4
0.04	105	108	103	96.7%	2/5
0.05	120	116	114	96.6%	6/2

Table 3 and 4 show the DEGs uniquely identified after each normalization method were applied and the difference in p-value for each gene is compared between the methods. When the robust spline method was applied, the limmaGUI identified more DEGs than the print tip loess method. Therefore it is not surprising that more genes were identified uniquely when the robust spline normalization method was used. There were 6 genes in the robust spline method that did not overlap with the print tip method. As can be seen in table 3 the p-value is indicated after both methods were used. Two genes were identified uniquely after the print tip normalization method was applied. For both these genes the p-values are very close to the cut off value after both normalizations, so the effect for those genes were marginal.

Table 4 shows seven genes that were identified uniquely after the robust spline normalization method was applied. The first 5 genes in the table all have a p-value close to the cut off value 0.05 where one gene actually seems to be affected by the normalization methods. The p-value for this gene increased with almost 0.01 when the print tip method was applied.

Table 3 - DEGs that were identified uniquely by the print tip method and p-value in the robust spline DEGs.

Gene number	ID	Name	p-value robust spline	p-value print tip	p-value difference
1	fb58a04	11-I7	0.05366	0.04940	0.00426
2	fb85d02	18-F4	0.05382	0.04720	0.00662

Table 4 - DEGs that were identified uniquely by the robust spline method and p-value in the print tip loess DEGs.

Gene number	ID	Name	p-value print tip	p-value robust spline	p-value difference
1	fb92g06	20-D11	0.05061	0.04197	0.00864
2	fb41c04	7-F7	0.05087	0.04684	0.00403
3	fb94d01	20-J2	0.05263	0.04728	0.00535
4	fb37c04	6-F7	0.05263	0.04889	0.00374
5	fb26b03	3-I6	0.05263	0.04889	0.00198
6	fb57g04	11-H7	0.05948	0.04969	0.00979

6 Conclusion

In this thesis two normalization methods were compared to see what effect they would have on the identification of DEGs. The majority of DEGs identified after each normalization method was applied are overlapping between the two methods. The overlap is not appreciably affected by different p-value cut offs but remains in the range 96%-97%. There was one cut off value 0.02 that did differ slightly from the other with percentage 93.3%. This is though not large enough difference to consider it as a major effect. Thus the main conclusion is that the choice of normalization method does not have a major effect on the resulting list of DEGs.

Boes and Neuhäuser (2005) compared the following seven normalization methods: scaling, invariant set, quantile, robust quantile, qspline, cyclic loess, and contrast on two datasets. Their conclusion was that quantile normalization performed the best in reducing obscuring variation but more investigation on more datasets is necessary to find out if this is true for a wide range of data. Even though Boes and Neuhäuser had a different viewpoint in their comparison, there are still some similarities between the present study and their study. They emphasized the lack of publicly available datasets appropriate for comparing normalization methods. I agree with this, especially in this thesis since the most beneficial test in finding out which normalization method performs better in finding DEGs is to have a dataset that has known DEGs. To the best of my knowledge no such public datasets are available.

Bolstad et al. (2003) conducted a similar experiment with the following five normalization methods: scaling, invariant set, quantile, cyclic loess, and contrast on two datasets. They came to the same conclusion as Boes and Neuhäuser that the quantile normalization method performs favorably but more datasets are needed to test different conditions. They also concluded that the three complete data methods performed comparably at reducing variability across arrays. The two normalization methods compared in this thesis were both complete normalization methods as well which means that they make use of data from all arrays in an experiment to form the normalizing relation. The same conclusion was found in this thesis work, which is that the two normalization methods have a similar effect on identifying DEGs.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. 2002. *Molecular Biology of the cell.* 4th edition, Garland Science, New York, pp. 375-376.

Duggan, D. J., Bittner, M., Chen, Y., Meltzner, P. and Trent, J. M. 1999. Expression profiling using cDNA microarrays. *Nature Genetics,* 21, pp. 10–14.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U. and Speed, T. P. 2003. Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. *Biostatistics*, 4, pp. 249-264.

Kerr, M. K, and Churchill, G. A. 2001. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl. Acad. Sci*, 98, pp. 8961-8965.

Lockhart, D. J. and Winzeler, E. A. 2000. Genomics, gene expression and DNA arrays. *Nature - Macmillan Magazines,* 405, pp. 827-836.

Mehta, G. *International Council for Science* [online]. Available from: http://www.doylefoundation.org/icsu/glossary.htm [Accessed 06.03.01]

Reiner, A., Yekutieli, D. and Benjamini, Y. 2003. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics,* 1(3), pp. 368-375.

Sebastiani, P., Gaussoni, E., Kohane, I. S. and Ramoni, M. 2003. Statistical challenges in functional genomics. *Statistical science,* 18(1), pp. 33–70.

Smyth, G. K. and Speed, T. 2003. Normalization of cDNA Microarray Data. *Methods,* 31, pp. 265-273.

Smyth, G. K. 2004. Linear models and empirical bayes for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology,* 3(1), Article 3.

Smyth, G. K. and Wettenhall, J. M. 2004. LimmaGUI: A graphical user interface for linear modeling of microarray data. *Bioinformatics*, 20(18), pp. 3705-6

Smyth, G. K. 2005. Limma documentation: Normalize Single Microarray Using Shrunk Robust Splines [online]. Available from: http://finzi.psych.upenn.edu/R/library/limma/html/normalizeRobustSpline.html [Accessed 06.05.31]

Smyth, G. 2006. *The limma Package* [online]. Available from: http://bioinf.wehi.edu.au/limma [Accessed 06.08.24]

U.S. Dept. of Energy. 1997. *Human Genome Program Report* [online]. Available from: http://www.ornl.gov/sci/techresources/Human_Genome/publicat/97pr/09gloss.html [Acc essed 06.03.10]

Ward, V. L. director. *Access Excellence Resource Center* [online]. Available from: http://www.accessexcellence.org/RC/VL/GG/central.html [Accessed 06.03.09]

Wikipedia Foundation. 2005. *DNA microarrays* [online]. Available from: http://en.wikipedia.org/wiki/Image:Microarray-schema.gif [Accessed 06.03.01]

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

Yang, Y. H. and Dudoit, S. 2006. *Introduction to the Bioconductor marray package: Input component* [online]. Available from: http://inn.weizmann.ac.il/bioconductor/packages/bioc/1.8/vignettes/marray/inst/doc/marrayInput.pdf [Accessed 06.05.24]

Appendix A

Table 1- Differentially expressed genes from the print tip loess method. Gene ID is a unique name for the gene. Here a gene with a p-value below 0.05 is identified as differentially expressed. The colored rows represent other cut off values. Red is for cut off value 0.01, blue is for 0.02, purple for 0.03 and green for 0.04.

Nr.	ID	Name	p-value
1	fb94h06	20-L12	0.001993
2	fb40h07	7-D14	0.001993
3	fc22a09	27-E17	0.001993
4	fb85f09	18-G18	0.001993
5	fc10h09	24-H18	0.001993
6	fb85a01	18-E1	0.001993
7	fb85d05	18-F10	0.001993
8	fb87d12	18-N24	0.001993
9	fb85e07	18-G13	0.001993
10	fb37b09	6-E18	0.002087
11	fb26b10	3-I20	0.002087
12	fb24g06	3-D11	0.002087
13	fc18d12	26-F24	0.002122
14	fb37e11	6-G21	0.002133
15	fb50g12	9-L23	0.002133
16	fb32f06	5-C12	0.002133
17	fb23d08	2-N16	0.002144
18	fb36g12	6-D23	0.002899
19	fb38a01	6-I1	0.004123
20	fb22a12	2-I23	0.004123
21	fb84a05	18-A9	0.004123
22	fc24c10	27-N19	0.004123
23	fb51f10	9-O20	0.004221
24	fb32g01	5-D1	0.004252
25	fc13b07	25-A14	0.004425
26	fb54e03	10-K5	0.004425
27	fb50b07	9-I14	0.004425
28	fb39c03	6-N5	0.004425
29	fb87f03	18-O6	0.004461
30	fb48b12	9-A24	0.004662
31	fb26g09	3-L17	0.004911
32	fb58g10	11-L19	0.005042
33	fb56f07	11-C14	0.005043
34	fb17b10	1-E20	0.005410
35	fb87c12	18-N23	0.006479
36	fb85d06	18-F12	0.006484
37	fb53c04	10-F7	0.006729
38	fc22f05	27-G10	0.007306

39	fb26f09	3-K18	0.007678
40	fb86b05	18-I10	0.009366
Cut off value 0.01			
41	fb66d09	13-J18	0.010810
42	fb32a09	5-A17	0.010810
43	fb55f03	10-O6	0.011605
44	fb97g03	21-H5	0.011721
45	fb67f06	13-O12	0.012188
46	fb25c05	3-F9	0.012188
47	fb85d01	18-F2	0.012425
48	fc10b01	24-E2	0.014576
49	fb93d12	20-F24	0.015110
50	fb52g05	10-D9	0.015110
51	fc07e05	23-K9	0.015125
52	fb42a12	7-I23	0.015125
53	fb87g09	18-P17	0.015417
54	fb87b04	18-M8	0.015417
55	fc20d08	26-N16	0.015417
56	fb63g03	12-P5	0.015445
57	fb86h09	18-L18	0.016106
58	fb54a01	10-I1	0.016590
59	fc24a05	27-M9	0.017552
60	fb42g07	7-L13	0.017552
61	fb23b08	2-M16	0.017552
62	fb20d05	2-B10	0.018905
63	fb49b04	9-E8	0.018924
64	fc14f05	25-G10	0.019207
65	fb99e11	21-O21	0.019837
Cut off value 0.02			
66	fb55c06	10-N11	0.020521
67	fb52d01	10-B2	0.020521
68	fb84a07	18-A13	0.021228
69	fc13c06	25-B11	0.021228
70	fb66f05	13-K10	0.022659
71	fb37d02	6-F4	0.023410
72	fb55b06	10-M12	0.023680
73	fc08d12	23-N24	0.023680
74	fc06b10	23-E20	0.023680
75	fb50e05	9-K9	0.023680
76	fc20c01	26-N1	0.023680
77	fb96f01	21-C2	0.023680
78	fb57b04	11-E8	0.023680
79	fb54f05	10-K10	0.023680
80	fb34b09	5-I18	0.024097
81	fb19f06	1-O12	0.025587
82	fb33f11	5-G22	0.026087

83	fb61c07	12-F13	0.026539
84	fb97b10	21-E20	0.028330
85	fb58h09	11-L18	0.028330
86	fb93h07	20-H14	0.028330
87	fc22d09	27-F18	0.028330
88	fb97b02	21-E4	0.029746
89	fb52g12	10-D23	0.029746
Cut off value 0.03			
90	fb52e01	10-C1	0.030888
91	fb24d04	3-B8	0.031599
92	fb27e07	3-O13	0.031599
93	fb20g12	2-D23	0.031599
94	fb92c06	20-B11	0.031599
95	fb17a09	1-E17	0.031599
96	fb42e06	7-K11	0.031599
97	fc16d04	25-N8	0.032637
98	fc18d08	26-F16	0.032637
99	fc18d02	26-F4	0.032637
100	fc18c02	26-F3	0.032727
101	fc17d01	26-B2	0.033026
102	fc24e05	27-O9	0.033323
103	fc05d06	23-B12	0.036937
104	fb37d04	6-F8	0.036937
105	fc23h02	27-L4	0.036937
106	fb48a11	9-A21	0.037145
107	fc11b06	24-I12	0.037182
108	fb97e03	21-G5	0.038337
Cut off value 0.04			
109	fb49b11	9-E22	0.040099
110	fc11c12	24-J23	0.040099
111	fb61e03	12-G5	0.045374
112	fb85d02	18-F4	0.047198
113	fc21b02	27-A4	0.047348
114	fb43f06	7-O12	0.048073
115	fb96f10	21-C20	0.048196
116	fb58a04	11-I7	0.049398

Table 2 - Differentially expressed genes from the robust spline method. For explanation, see table 1.

Nr.	ID	Name	p-value
1	fc22a09	27-E17	0.002469

2	fb94h06	20-L12	0.002469
3	fb37b09	6-E18	0.002469
4	fb85f09	18-G18	0.002469
5	fb40h07	7-D14	0.002469
6	fc10h09	24-H18	0.002469
7	fb85e07	18-G13	0.002469
8	fb85d05	18-F10	0.002469
9	fb87d12	18-N24	0.002469
10	fb37e11	6-G21	0.002469
11	fb26b10	3-I20	0.002469
12	fc18d12	26-F24	0.002643
13	fb32f06	5-C12	0.002643
14	fb23d08	2-N16	0.002998
15	fb85a01	18-E1	0.002998
16	fb24g06	3-D11	0.003431
17	fb39c03	6-N5	0.003431
18	fb50g12	9-L23	0.003431
19	fc13b07	25-A14	0.003449
20	fb22a12	2-I23	0.003449
21	fb50b07	9-I14	0.003492
22	fc24c10	27-N19	0.003585
23	fb38a01	6-I1	0.003663
24	fb36g12	6-D23	0.003706
25	fb51f10	9-O20	0.004651
26	fb84a05	18-A9	0.004651
27	fb32g01	5-D1	0.004880
28	fb58g10	11-L19	0.004880
29	fb48b12	9-A24	0.005102
30	fb56f07	11-C14	0.005102
31	fb17b10	1-E20	0.005157
32	fb87f03	18-O6	0.005621
33	fb54e03	10-K5	0.005859
34	fb26g09	3-L17	0.006103
35	fb87c12	18-N23	0.006412
36	fb85d06	18-F12	0.006521
37	fb53c04	10-F7	0.007389
38	fb26f09	3-K18	0.008587
39	fb66d09	13-J18	0.009360
	Cut off value 0.01		
40	fc22f05	27-G10	0.010163
41	fb55f03	10-O6	0.010246
42	fb86b05	18-I10	0.010246
43	fc07e05	23-K9	0.011374
44	fb52g05	10-D9	0.013117
45	fb25c05	3-F9	0.013117
46	fb85d01	18-F2	0.013117

47	fb97g03	21-H5	0.013156
48	fb67f06	13-O12	0.013380
49	fb32a09	5-A17	0.014694
50	fb42a12	7-I23	0.014780
51	fb63g03	12-P5	0.016605
52	fb93d12	20-F24	0.017644
53	fb42g07	7-L13	0.017788
54	fc14f05	25-G10	0.017788
55	fb49b04	9-E8	0.017788
56	fc20d08	26-N16	0.017788
57	fb23b08	2-M16	0.017788
58	fb87b04	18-M8	0.017788
59	fb86h09	18-L18	0.017788
60	fc13c06	25-B11	0.018013
61	fb50e05	9-K9	0.018013
62	fc10b01	24-E2	0.018013
63	fb99e11	21-O21	0.018401
64	fb20d05	2-B10	0.018401
65	fb87g09	18-P17	0.018401
66	fb55c06	10-N11	0.018401
67	fb52d01	10-B2	0.018401
68	fb96f01	21-C2	0.019454
69	fb34b09	5-I18	0.019694
70	fb61c07	12-F13	0.019694
	Cut off value 0.02		
71	fc06b10	23-E20	0.020456
72	fb37d02	6-F4	0.021629
73	fc20c01	26-N1	0.021999
74	fb54a01	10-I1	0.022374
75	fb17a09	1-E17	0.022374
76	fb66f05	13-K10	0.022374
77	fb97b10	21-E20	0.022374
78	fb19f06	1-O12	0.022451
79	fc08d12	23-N24	0.022592
80	fb33f11	5-G22	0.023080
81	fb55b06	10-M12	0.023080
82	fc22d09	27-F18	0.024448
83	fb57b04	11-E8	0.026257
84	fb42e06	7-K11	0.026649
85	fb54f05	10-K10	0.027269
86	fb93h07	20-H14	0.029169
87	fb58h09	11-L18	0.029211
	Cut off value 0.03		
88	fc18c02	26-F3	0.030739
89	fb27e07	3-O13	0.031518
90	fb97b02	21-E4	0.031518

91	fc24a05	27-M9	0.031518
92	fb52g12	10-D23	0.032544
93	fc17d01	26-B2	0.034093
94	fc16d04	25-N8	0.034093
95	fb84a07	18-A13	0.034100
96	fb52e01	10-C1	0.035253
97	fb92c06	20-B11	0.035312
98	fc05d06	23-B12	0.035342
99	fb24d04	3-B8	0.036098
100	fc18d02	26-F4	0.036098
101	fc18d08	26-F16	0.037961
102	fc21b02	27-A4	0.038236
103	fb49b11	9-E22	0.038453
104	fc11b06	24-I12	0.038764
105	fb37d04	6-F8	0.039616
Cut off value 0.04			
106	fc11c12	24-J23	0.040819
107	fc24e05	27-O9	0.041972
108	fb92g06	20-D11	0.041972
109	fc23h02	27-L4	0.043683
110	fb41c04	7-F7	0.046839
111	fb43f06	7-O12	0.046839
112	fb61e03	12-G5	0.046839
113	fb20g12	2-D23	0.046839
114	fb96f10	21-C20	0.047275
115	fb94d01	20-J2	0.047275
116	fb37c04	6-F7	0.048887
117	fb48a11	9-A21	0.048887
118	fb26b03	3-I6	0.048887
119	fb97e03	21-G5	0.049043
120	fb57g04	11-H7	0.049685

Appendix B

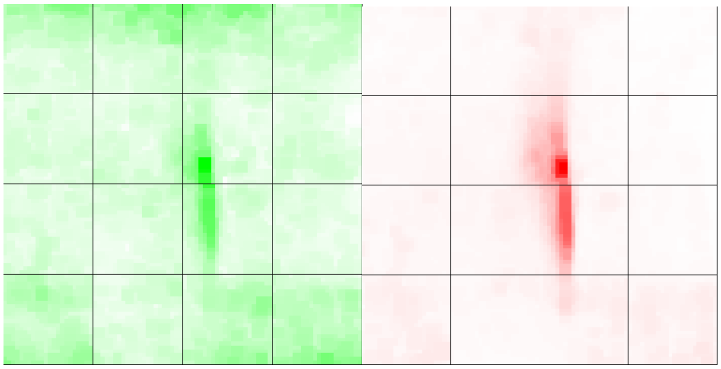


Figure 1 – Red and green background images for replicate slide 1.

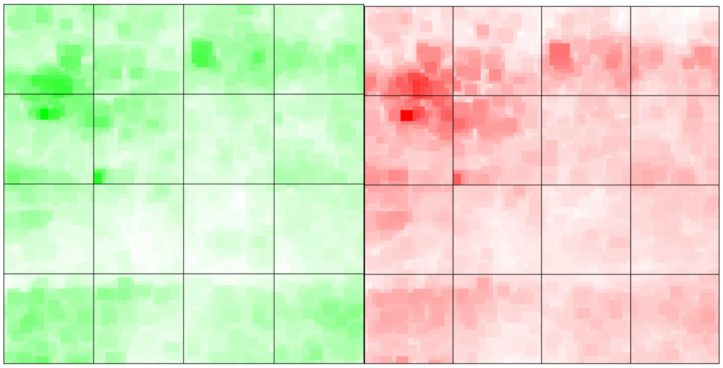


Figure 2 – Red and green background images for replicate slide 2.

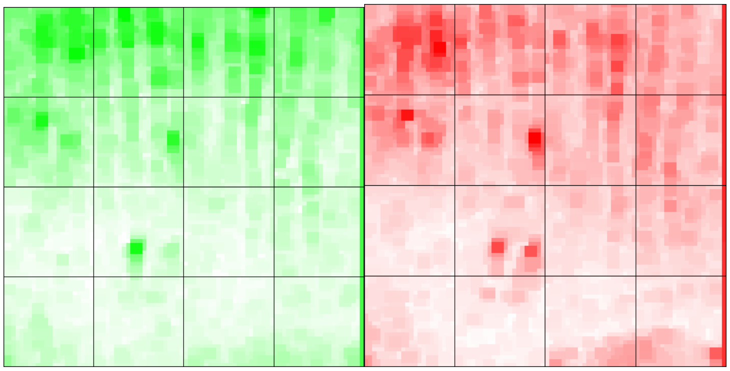


Figure 3 – Red and green background images for replicate slide 3.

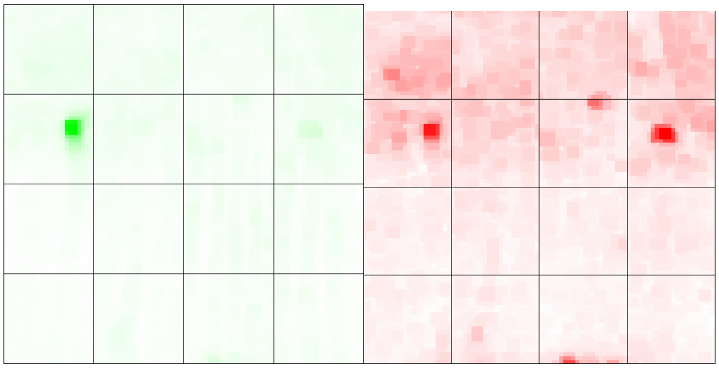


Figure 4 – Red and green background images for replicate slide 4.

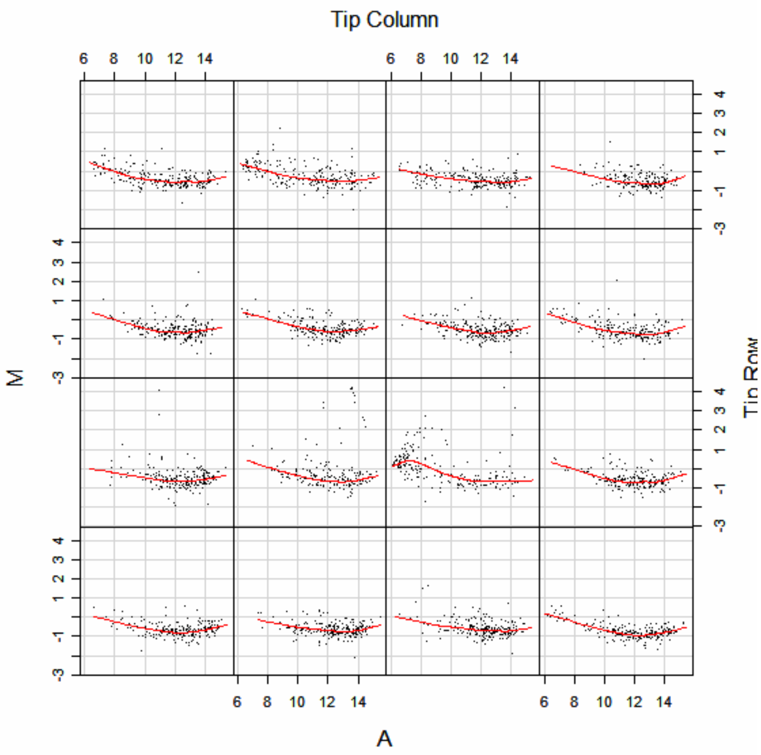


Figure 5 – A print tip group MA-plot for replicate slide 1.

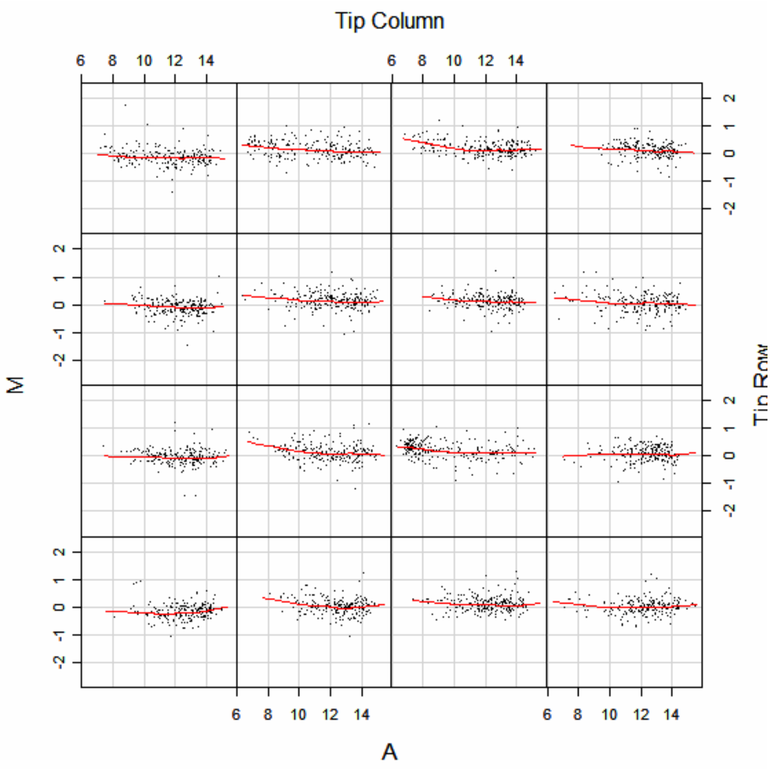


Figure 6 – A print tip group MA-plot for replicate slide 2.

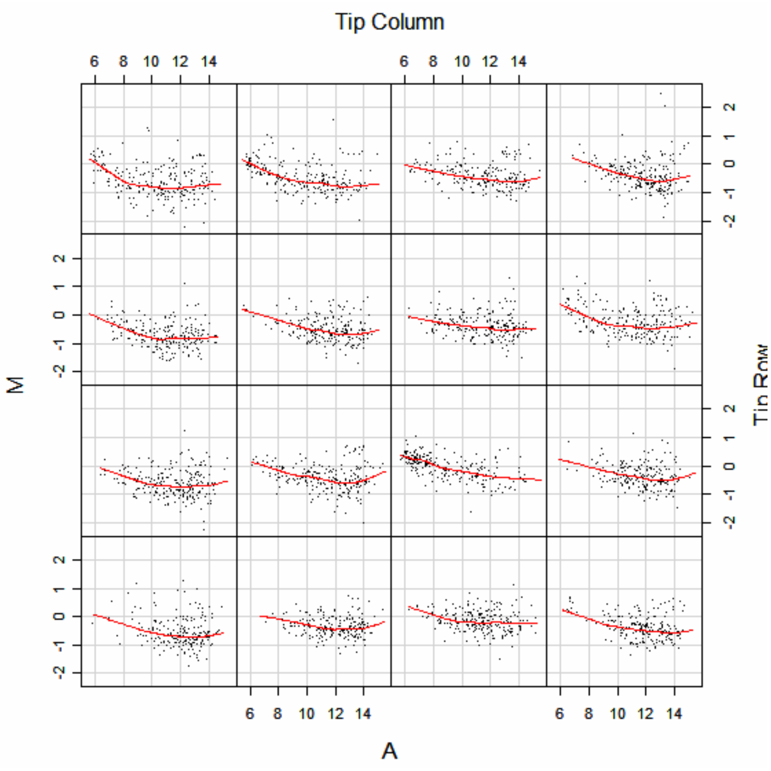


Figure 7 – A print tip group MA-plot for replicate slide 3.

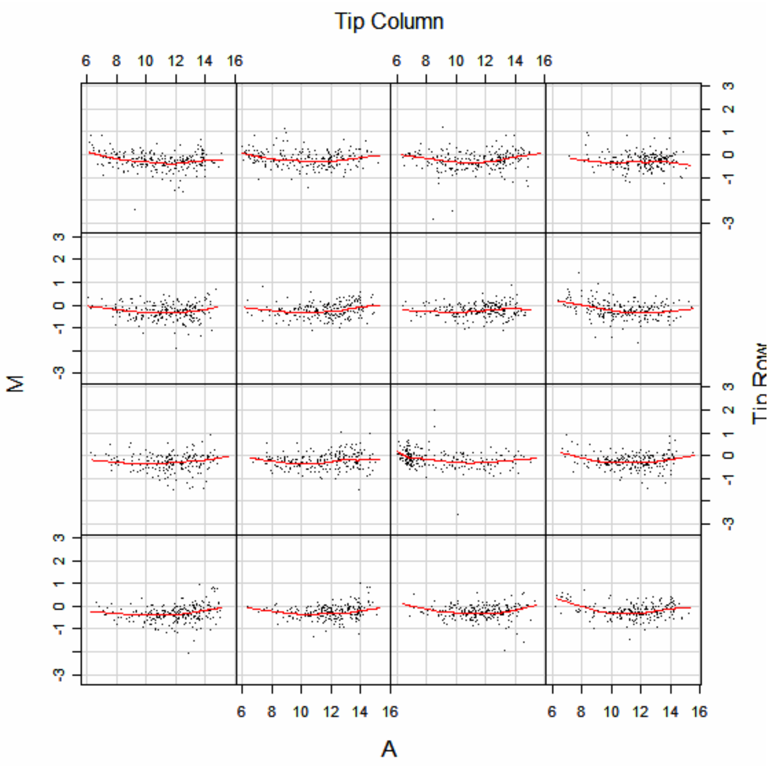


Figure 8 – A print tip group MA-plot for replicate slide 4

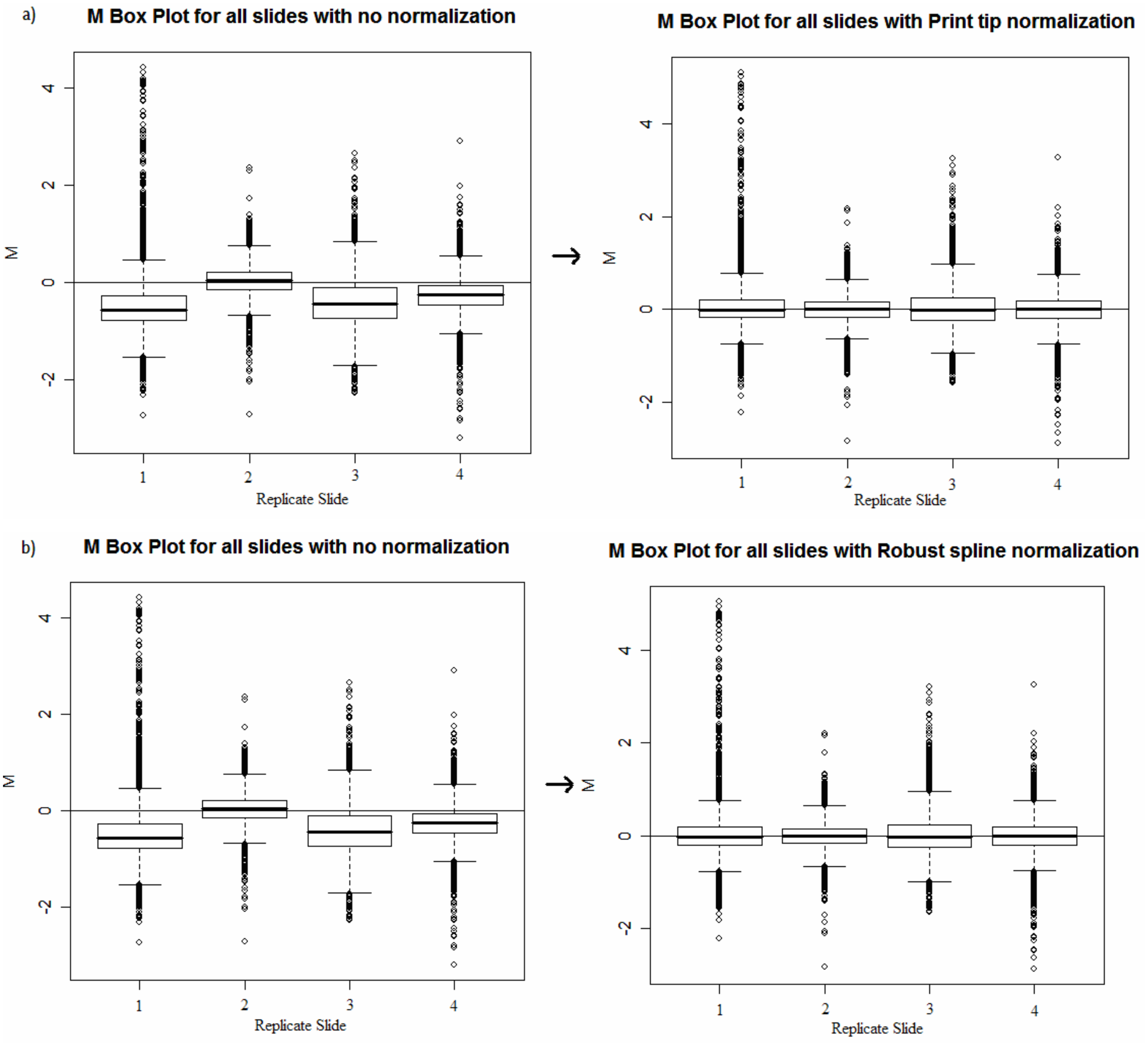


Figure 9 – Boxplots of all replicate plates side by side. Figure a) shows boxplots of the un-normalized swirl data and then the corresponding boxplots of the print tip normalized data. The figure shows how the print tip loess method equalizes some of the scale differences. Figure b) shows the same but with boxplots of the robust spline method instead of the print tip method. Figure b) also shows how the scale differences are equalized with the robust spline method. However, there are some scale differences which can be equalized with scale normalization.

Appendix C

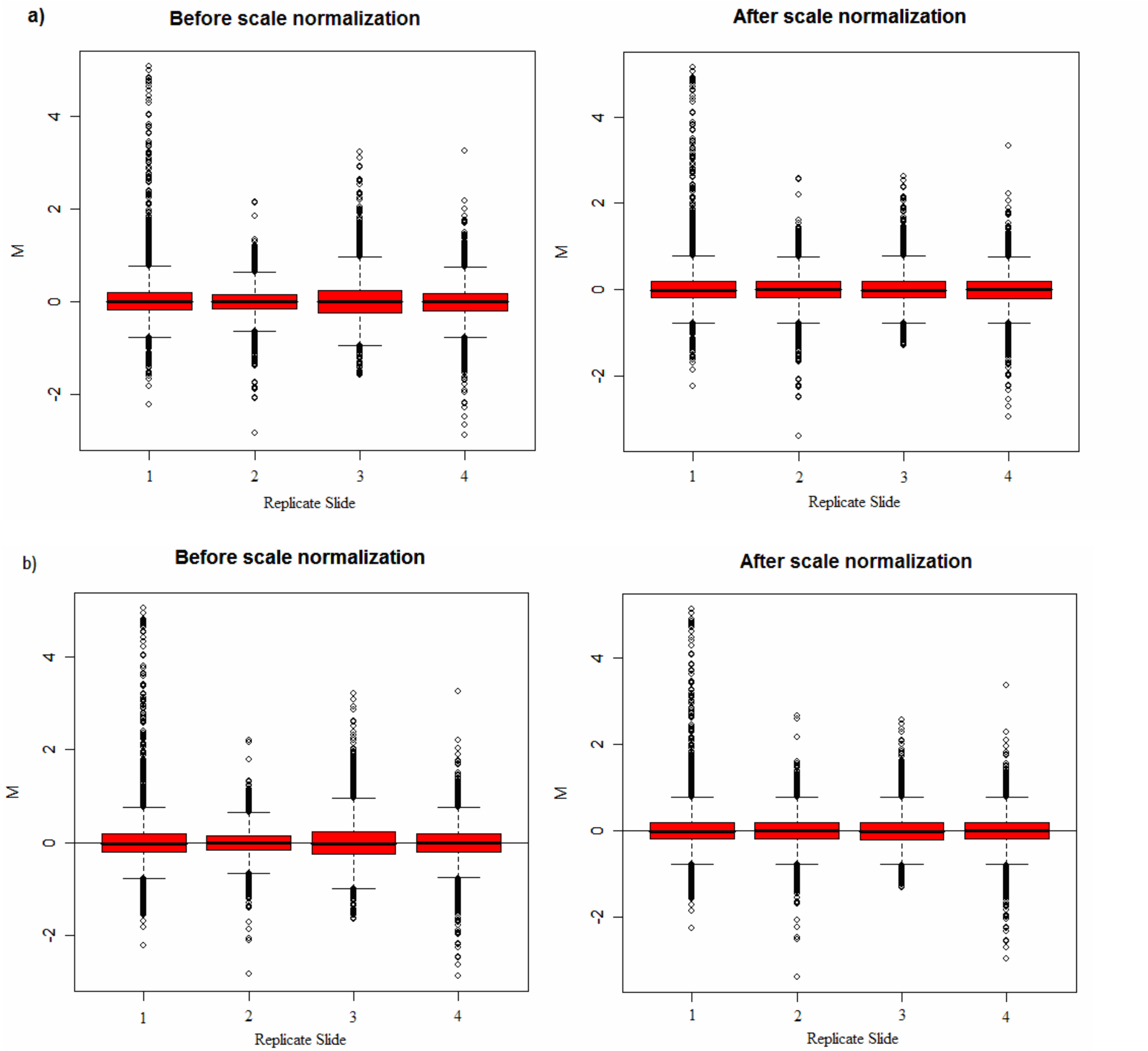


Figure 1 – Boxplot of each replicate plate side by side. Figure a) shows a boxplot of the print tip normalized data before scale normalization and then after scale normalization. The figure shows how the boxes of the replicate slides are more equal after the scale normalization. Figure b) shows

the same but a boxplot with the robust spline method instead of the print tip method. The figure also shows how the boxes are more equal between replicate slides after the scale normalization.

Appendix D

Table 1 – The overlapping DEGs between the robust spline method and the print tip method.

Nr.	ID	Name	p-value
1	fb94h06	20-L12	0.001993
2	fb40h07	7-D14	0.001993
3	fc22a09	27-E17	0.001993
4	fb85f09	18-G18	0.001993
5	fc10h09	24-H18	0.001993
6	fb85a01	18-E1	0.001993
7	fb85d05	18-F10	0.001993
8	fb87d12	18-N24	0.001993
9	fb85e07	18-G13	0.001993
10	fb37b09	6-E18	0.002087
11	fb26b10	3-I20	0.002087
12	fb24g06	3-D11	0.002087
13	fc18d12	26-F24	0.002122
14	fb37e11	6-G21	0.002133
15	fb50g12	9-L23	0.002133
16	fb32f06	5-C12	0.002133
17	fb23d08	2-N16	0.002144
18	fb36g12	6-D23	0.002899
19	fb38a01	6-I1	0.004123
20	fb22a12	2-I23	0.004123
21	fb84a05	18-A9	0.004123
22	fc24c10	27-N19	0.004123
23	fb51f10	9-O20	0.004221
24	fb32g01	5-D1	0.004252
25	fc13b07	25-A14	0.004425
26	fb54e03	10-K5	0.004425
27	fb50b07	9-I14	0.004425
28	fb39c03	6-N5	0.004425
29	fb87f03	18-O6	0.004461
30	fb48b12	9-A24	0.004662
31	fb26g09	3-L17	0.004911
32	fb58g10	11-L19	0.005042
33	fb56f07	11-C14	0.005043
34	fb17b10	1-E20	0.005410
35	fb87c12	18-N23	0.006479
36	fb85d06	18-F12	0.006484
37	fb53c04	10-F7	0.006729
38	fc22f05	27-G10	0.007306
39	fb26f09	3-K18	0.007678
40	fb86b05	18-I10	0.009366

41	fb66d09	13-J18	0.010810
42	fb32a09	5-A17	0.010810
43	fb55f03	10-O6	0.011605
44	fb97g03	21-H5	0.011721
45	fb67f06	13-O12	0.012188
46	fb25c05	3-F9	0.012188
47	fb85d01	18-F2	0.012425
48	fc10b01	24-E2	0.014576
49	fb93d12	20-F24	0.015110
50	fb52g05	10-D9	0.015110
51	fc07e05	23-K9	0.015125
52	fb42a12	7-I23	0.015125
53	fb87g09	18-P17	0.015417
54	fb87b04	18-M8	0.015417
55	fc20d08	26-N16	0.015417
56	fb63g03	12-P5	0.015445
57	fb86h09	18-L18	0.016106
58	fb54a01	10-I1	0.016590
59	fc24a05	27-M9	0.017552
60	fb42g07	7-L13	0.017552
61	fb23b08	2-M16	0.017552
62	fb20d05	2-B10	0.018905
63	fb49b04	9-E8	0.018924
64	fc14f05	25-G10	0.019207
65	fb99e11	21-O21	0.019837
66	fb55c06	10-N11	0.020521
67	fb52d01	10-B2	0.020521
68	fb84a07	18-A13	0.021228
69	fc13c06	25-B11	0.021228
70	fb66f05	13-K10	0.022659
71	fb37d02	6-F4	0.023410
72	fb55b06	10-M12	0.023680
73	fc08d12	23-N24	0.023680
74	fc06b10	23-E20	0.023680
75	fb50e05	9-K9	0.023680
76	fc20c01	26-N1	0.023680
77	fb96f01	21-C2	0.023680
78	fb57b04	11-E8	0.023680
79	fb54f05	10-K10	0.023680
80	fb34b09	5-I18	0.024097
81	fb19f06	1-O12	0.025587
82	fb33f11	5-G22	0.026087
83	fb61c07	12-F13	0.026539
84	fb97b10	21-E20	0.028330
85	fb58h09	11-L18	0.028330
86	fb93h07	20-H14	0.028330

87	fc22d09	27-F18	0.028330
88	fb97b02	21-E4	0.029746
89	fb52g12	10-D23	0.029746
90	fb52e01	10-C1	0.030888
91	fb24d04	3-B8	0.031599
92	fb27e07	3-O13	0.031599
93	fb20g12	2-D23	0.031599
94	fb92c06	20-B11	0.031599
95	fb17a09	1-E17	0.031599
96	fb42e06	7-K11	0.031599
97	fc16d04	25-N8	0.032637
98	fc18d08	26-F16	0.032637
99	fc18d02	26-F4	0.032637
100	fc18c02	26-F3	0.032727
101	fc17d01	26-B2	0.033026
102	fc24e05	27-O9	0.033323
103	fc05d06	23-B12	0.036937
104	fb37d04	6-F8	0.036937
105	fc23h02	27-L4	0.036937
106	fb48a11	9-A21	0.037145
107	fc11b06	24-I12	0.037182
108	fb97e03	21-G5	0.038337
109	fb49b11	9-E22	0.040099
110	fc11c12	24-J23	0.040099
111	fb61e03	12-G5	0.045374
112	fc21b02	27-A4	0.047348
113	fb43f06	7-O12	0.048073
114	fb96f10	21-C20	0.048196