Project Part II

Ana Sofia Almeida | 49292

Métodos Estatísticos em Bioinformática

Mestrado em Bioinformática e Biologia Computacional

Faculdade de Ciências da Universidade de Lisboa

Supervisor Lisete Maria Ribeiro de Sousa

June, 2021

Table of Contents

Introduction

1.1 State of the Art

DNA microarrays have revolutionized molecular biology, and their application in academics, medicine, and the pharmaceutical, biotechnology, agrochemical, and food sectors has exploded in the past years. The number of quantitative data generated by microarrays is one of its most notable features (Bumgarner, 2013).

A microarray is a solid support (such as a membrane or glass microscope slide) on which DNA of known sequence is deposited in a grid-like array (Wildsmith, 2001). Microarrays are most commonly used to measure gene expression throughout RNA extraction from a set of matched samples. Gene expression is context-dependent and can be regulated in a variety of ways such as by region, gene activity, in development, disease states or dynamic response to environmental signals (Pevsner, 2015).

In order to quantify the expression levels of thousands of genes, RNA is often transformed to cDNA, tagged with fluorescence (or radioactivity), and then hybridized on microarrays. The Green label (Cy3) often represents the control cDNA, the Red (Cy5) represents the sample cDNA and the combination of the two is represented by Yellow color and Black representing areas where neither the Control nor Sample cDNA hybridized to the target DNA. Besides of being fast, easy and flexible, microaaray experiments have some drawbacks such as the cost (sometimes an appropriate number of controls or replicates is not affordable), the RNA significance and the quality control, since not much attention is payed to experimental design and there is not enough collaboration with statisticians (Jaksik et al., 2015).

The design of a microarray analysis consists of:

Stage 1: Experimental design

Stage 3: Hybridization to DNA arrays

Stage 2: RNA and probe preparation

Stage 4: Image analysis

Stage 5: Microarray data analysis

Stage 6: Biological confirmation

Stage 7: Microarray databases

1.2 Microarray Data

Samples of mRNA from saphenous vein (vein inside the leg) tissues were obtained from people operated for cardiac from an experiment carried out in a Laboratory of Molecular Cardiology of a Brazilian Institute. These tissues were cultured ex vivo and subjected to two different experimental regimens: arterial and venous. A saphenous vein tissue sample was taken from each patient and subjected to both experimental conditions mentioned. The chip1.txt file comprises the expression levels of genes in tissue subjected to the arterial regime (Art), genes in tissue subjected to the venous regime (Ven), and respective background values (BgArt and BgVen) for the first patient. The sample Art corresponds to the red channel and the sample Ven to the green channel. However, two more patients were operated and two arrays under the same conditions as the first were obtained. The intensities are registered in chip2.txt and chip3.txt.

1.3 Objective

In the scope of the Statistical Methods in Bioinformatics course, the goal of this project is to perform a gene expression analysis derived from microarray data. Therefore, this report is focused on the satage 5 of the microarray design where data normalization methods were performed in order to identify differential expressed genes through the use of genArise and limma packages. All computation analyses were carried out in the R (Team, 2018); the code is provided as an R script.

Microarray Analisys using GenArise

GenArise is an R-based, easy to use tool for working with dual-color microarray data. genArise is a package that includes specific functions for analyzing cDNA microarray data to find genes that are differentially expressed under different growth conditions. Before doing this analysis, genArise performs a number of data modifications to remove low-quality observations and alter measured intensities to make comparisons easier (Ana Patricia Gomez Mayen and L̃ina Riego Ruiz, 2020).

2.1 Normalization Methods

Normalisation is a general term for a collection of methods that are directed at resolving the systematic errors and bias generated by the microarray experimental platform. A variety of measures are frequently performed to guarantee that the data is of high quality and suitable for analysis before using it to answer scientific issues. These measures include: removing flagged features; background subtraction; taking logarithms (Quackenbush, 2002).

2.1.1 Data Cleaning and Transformation

Taking Logarithms (a)

Before proceeding with analysis, it is standard practice to convert DNA microarray data from raw intensities to log intensities. This transition aims to achieve a number of goals: across the intensity range, there should be a fairly uniform distribution of features; at all intensity levels, the variability should be consistent; the distribution of experimental mistakes should resemble that of a normal distribution; the intensities should be distributed in a bell-shaped pattern (Stekel, 2003).

In microarray data analysis, base 2 logarithms are commonly used. The reason for this is that the ratio of raw Cy5 and Cy3 intensities is converted into the difference between the logs of the Cy5 and Cy3 channel intensities (Fig.2.2). As a result, genes that are 2-fold up-regulated have a log ratio of +1, while genes that are 2-fold down-regulated have a log ratio of -1. Genes with a log ratio of 0 are not differentially expressed. The intrinsic symmetry of these log ratios reflects biology and is not reflected in the raw fold difference (Stekel, 2003).

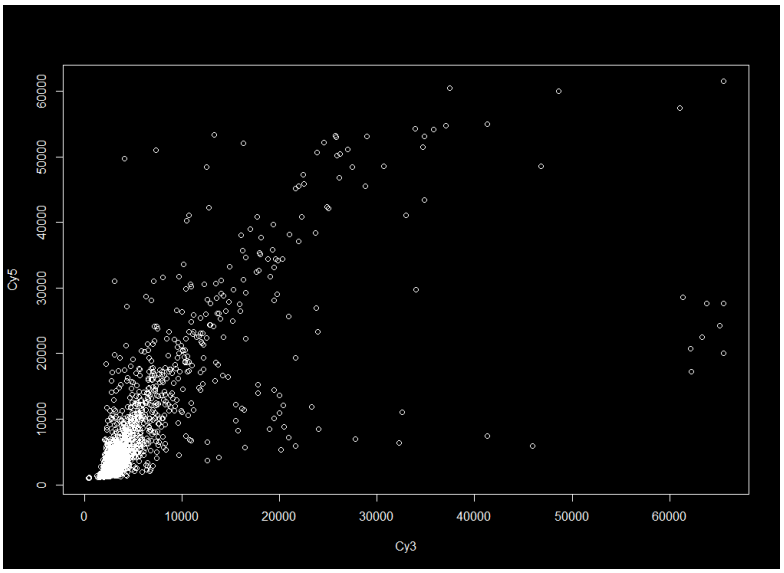


Figure 2.1: Raw intensity scatterplot for chip1; each point on the graph represents a feature in the array, with the x coordinate indicating the Cy3 intensity and the y coordinate indicating the Cy5.

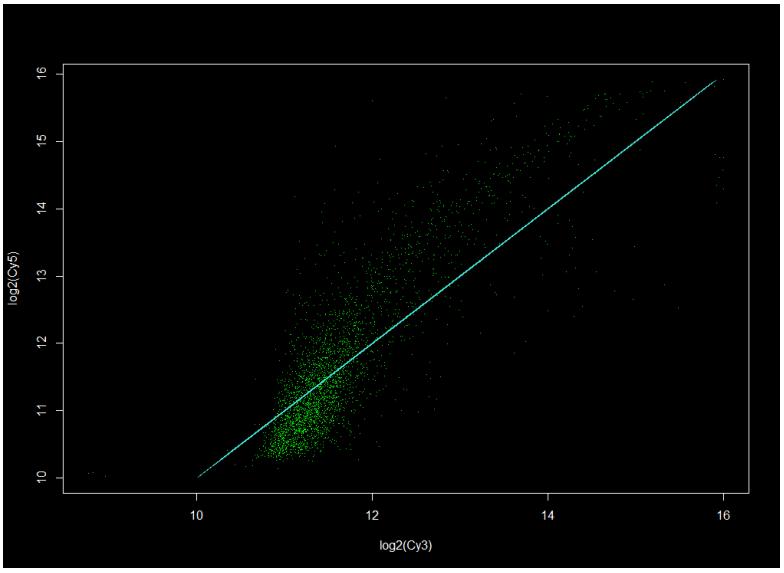
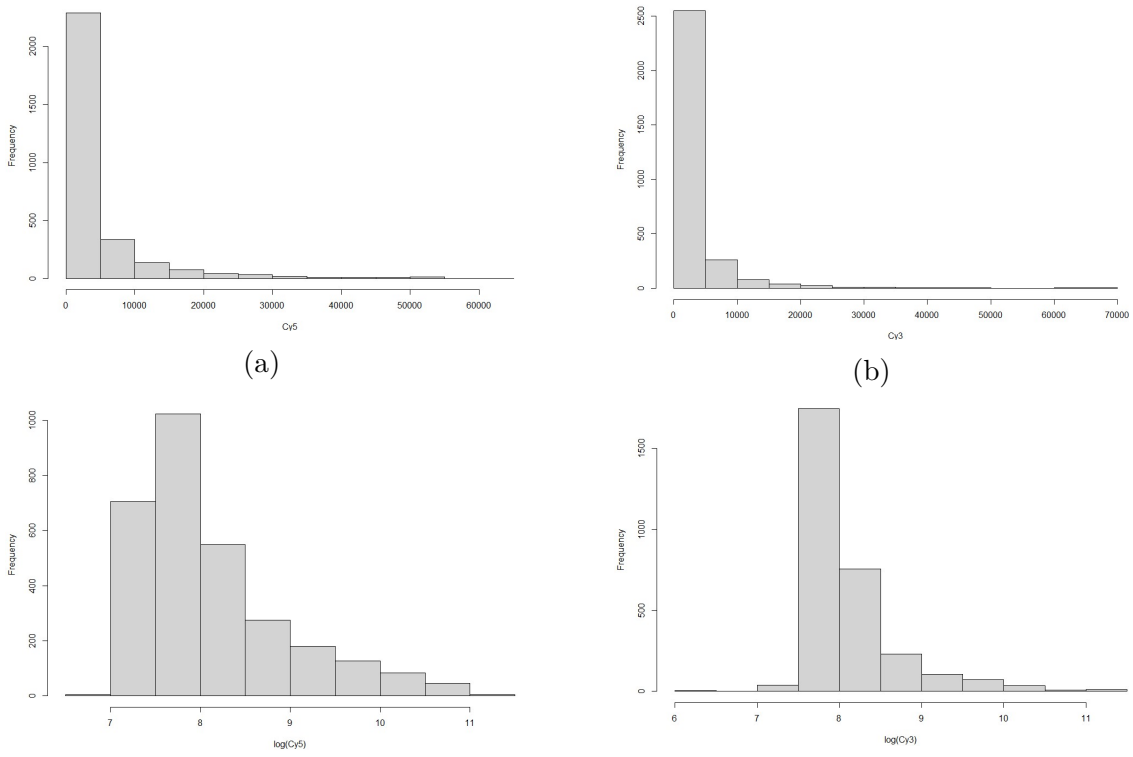


Figure 2.2: Scatterplot of the base 2 log intensities

Figure 2.1 shows that the raw data does not meet the criteria for effective analysis. The majority of the features are in the graph's bottom left corner; variability increases with intensity, and the intensity distribution is not bell-shaped but substantially right-skewed with the majority of features having low intensity and decreasing numbers of features having higher intensity (Fig.2.3a and Fig.2.3b). However, the logged data (Fig.2.2) meets the requirements. The data are evenly distributed across the log intensity scale. With the exception of very high expressed genes, whose intensities are likely to be inaccurate, the variability is essentially constant at all intensities and appears to be regularly distributed; and the intensity distribution (Fig.2.3c and Fig.2.3d) is more bellshaped (although these distributions are also slightly right-skewed).



(a) (b) (c) (d)

Figure 2.3: Histograms of the raw and log Cy3 and Cy5 intensities; (a) The raw intensities for the Cy5 channel. (b) The raw intensities for the Cy3 channel. (c) The log intensities for the Cy5 channel. (d) The log intensities for the Cy3 channel.

Background subtraction (b)

The background signal is assumed to be the result of non-specific hybridization of the labbeled target with the glass, as well as the glass slide's intrinsic fluorescence. When the feature intensity is higher than the background intensity, this approach works effectively (Kooperberg et al., 2002). Through the use of genArise package's function, for each spot in the microarray, the default background correction action is to subtract the Cy3 background intensity (BckgCy3) from the Cy3 foreground intensity, and the Cy5 background intensity (BckgCy5) from the Cy5 foreground intensity. The goal of this function is to extract the value of the real signal of Cy3 and Cy5 from the data file's reported values (Fig.2.4b).

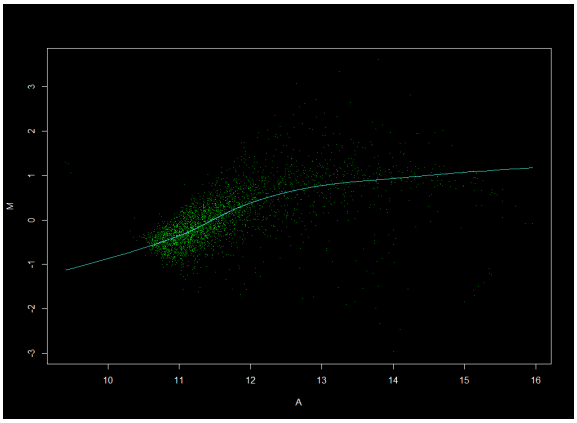
2.1.2 Within-array Normalisation (c)

Microarray is being used to analyse saphenous vein samples subjected to two different conditions and identify genes that are differentially expressed. The two treatments were labeled with two different fluorescent dyes in two different chemical processes, and their intensity was evaluated with two distinct lasers operating at two distinct wavelengths. Furthermore, the array's characteristics are dispersed around the array's surface in various locations. It must be confirmed if the measures of differential gene expression between the two samples reflect real differential gene expression rather than bias and error generated by the experimental procedure such as, for example, the differently integration of Cy3 and Cy5 markers into DNA of varying quantity or the distinct emission response of Cy3 and Cy5 dyes to the excitation laser. According to several studies, log2(ratio) values can have a systematic dependence on intensity, which most usually manifests itself as a divergence from zero for low-intensity locations (Y. H. Yang et al., 2002; I. V. Yang et al., 2002).

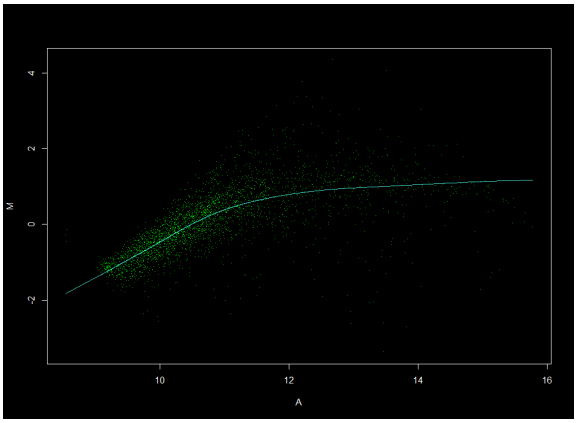
The lowess analysis (Cleveland, 1979) has been proposed as a normalizing method that can remove such intensity-dependent effects in the log2(ratio) data (Y. H. Yang et al., 2002) (I. V. Yang et al., 2002). GenArise uses a local weighted regression (loess) analysis due to location and intensity dependent biases in several trials. Such intensity-dependent effects in log2(ratio) data can be removed with this normaliz-

ation procedure. For this purpose, lowess was applied globally (to the entire microarray data set) instead of locally (to each individual subgrid). In this function, the normalize algorithm will be applied to all observations to get the lowess factor and then fit Cy5 with this factor. Any observation with a ratio (Cy5/Cy3) value of zero will be deleted during the process since they have no change in their expression levels.

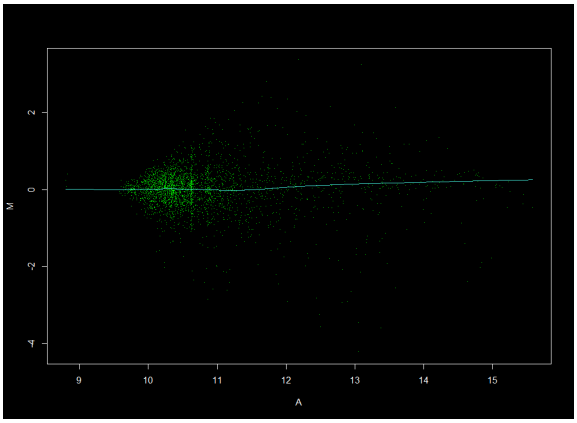
A scatterplot of the log ratio against the average intensity of each feature is a highly effective way of visualizing and normalizing the data. In the microarray literature, such plots are referred to as MA plots (Robinson, McCarthy and Smyth, 2010). Each point represents a feature in these plots, with the x coordinate representing the average value of the Cy3 and Cy5 log intensities and the y coordinate representing the difference between the Cy3 and Cy5 log intensities (i.e., the log ratio). MA plots are geometrically similar to Cy5 vs. Cy3 scatterplots, which are produced by rotating the graph through 45 degrees and then scaling the two axes. If the two channels are responding differently or non-linearly, it is typically clearer. If the two channels behave equally, the data should be symmetrical around a horizontal line through zero; any deviations from this horizontal line show differences in the two channels' reactions. Any linear or non-linear regression that compares the log ratio to the average intensity treats the two channels equally. As a result, such regressions are more reliable and repeatable than regressions of one channel versus the other (Stekel, 2003).



(a)



(b)



(c)

Figure 2.4: Plots of log ratio as a function of average intensity (MA) and normalisation; (a) MA plot of raw data. (b) MA plot with background subtraction. (c) MA plot after global normalization.

The data points of patient 1, have been fitted with a line, revealing a clear pattern in the Cy5 and Cy3 responses(Fig.2.4a). The Cy3 channel responds more strongly at low intensity levels, while the Cy5 channel responds more strongly at high in-

tensities. Assuming that most genes are not differentially expressed, this line shows an artifact rather than differential expression. Even after performing a background subtraction (Fig.2.4b) the results observed are very similar. The log ratios can be linearly normalised by subtracting the fitted value on the line from each log ratio. The line fit does not appear to be an exact fit to the data's center (Fig.2.4b). The data looks to flatten out at high intensities, implying that a non-linear fit would provide more accurate results. By subtracting the fitted value on the line from the log ratio of each feature, the data has been normalized to the line seen in Fig.2.4b. This line is changed to a horizontal line through zero. The highest-intensity points are located above the line (Fig.2.4c).

Selecting Differentially Expressed Genes (d)

The Z score transformation approach for normalizing data is a well-known statistical method that has been utilized in microarray studies to compare changes in gene expression between experimental and control groups (Vawter et al., 2001; Virtaneva et al., 2001). The purpose of genArise is to identify which genes have strong evidence of differential expression. This function calculates an intensity-dependent Z-score to identify differentially expressed genes. A sliding window algorithm is used to calculate the mean and standard deviation within a window surrounding each data point, as well as establish a Z-score, which represents the number of standard deviations a data point is from the mean.

$$z_i = (R_i - \text{mean(R)}) \ / \ \text{sd(R)}$$

where z_i is the z-score for each element, R_i is the log-ratio for each element, and sd(R) is the standard deviation of the log-ratio.

If a gene is not differentially expressed, its expected value in both experimental conditions of the sample is the same. The Z-score in this situation will be standardized, i.e. will have mean zero and unit standard deviation. There are 2994 genes in patient 1 data. If none of theses genes are differentially expressed, the Z-scores should have a mean of zero and a variance of one. In fact, there appear to be numerous genes for which the mean expression levels in arterial (Art) and venous regimen (Ven) do

not differ, as the standard deviation is very close to 1 (sd = 0.99). Furthermore, because the mean Z-score is negative (-0.00057), it appears that the dominant pattern is for genes to be expressed at a higher level in the venous regimen compared to the arterial regimen. The conventional threshold for statistical significance is a p-value smaller than 0.05, which corresponds to the Z-score being greater than 2 in magnitude (Cheadle et al., 2003). Many genes fit this requirement, nonetheless this does not imply that they are differently expressed.

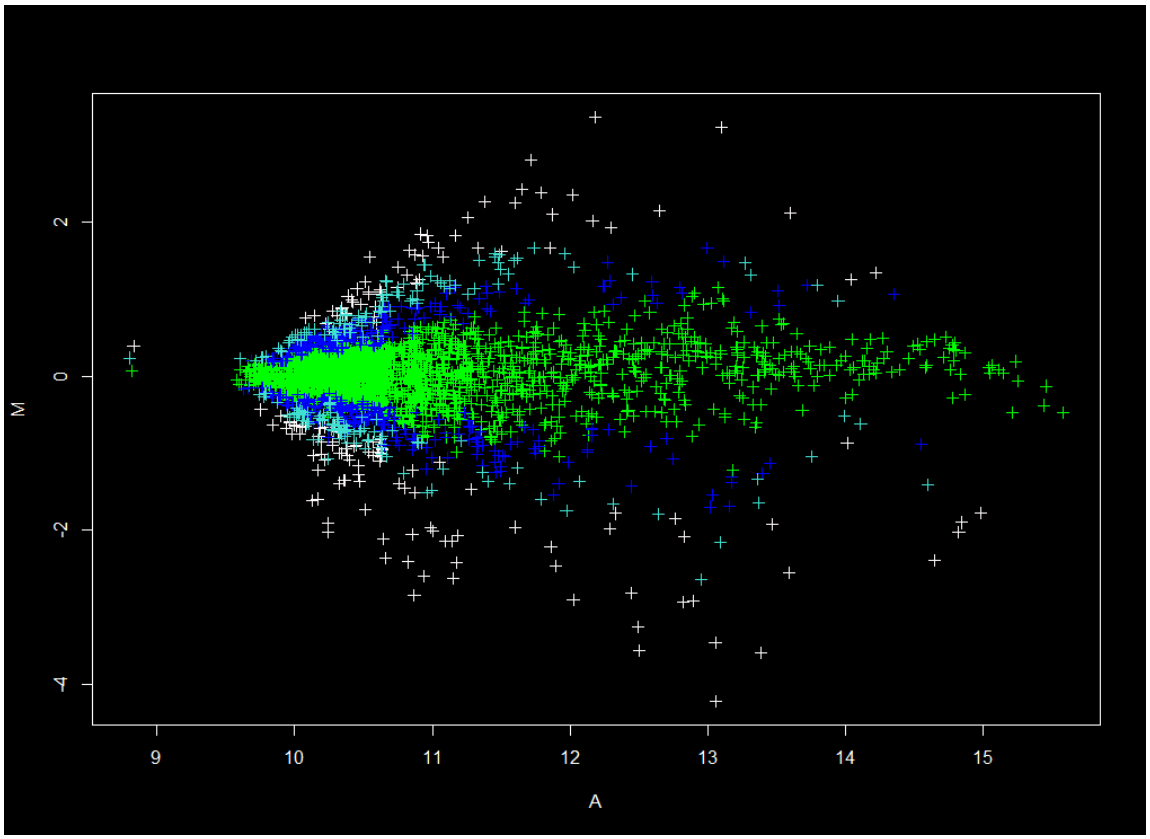


Figure 2.5: Data after Z-score Analysis

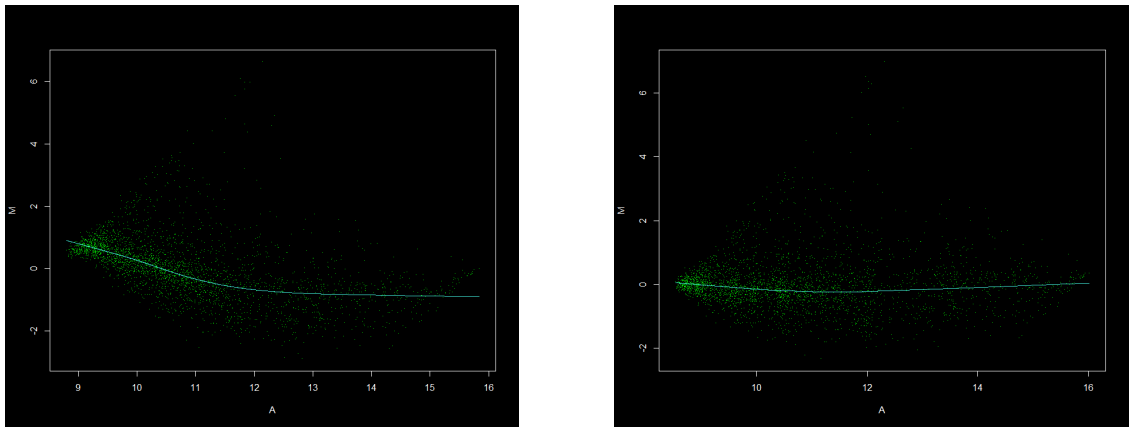
The fact that Z scores are proportional to the intensity of the initial hybridization signal makes it easier to analyze Z score values directly in visualization analysis. The Z score's value is directly proportional to the differential hybridization values, i.e. larger positive Z scores represent the most highly expressed genes, whereas lower negative Z scores represent the least expressed genes. As a result, Z scores provide a convenient and straightforward way to visualize and understand large amounts of data in their natural biological context (Cheadle et al., 2003).

Zscore.plot function allows for the visualisation of a MA plot where each feature is color coded depending on weather they are less than 1 standard deviation from the mean (green), between 1 and 1.5 standard deviations (blue), between 1.5 and 2 standard deviations (cyan), more than 1.5 standard deviations (yellow), or more than 2 standard deviations from the mean (white) (Fig.2.5). It was possible to identify a total of 156 differentially expressed genes under both treatment conditions (Art and Ven), with 96 being down regulated (z-score $<$ -2) and 60 being up regulated (z-score $>$ 2).

2.1.3 Between-array Normalisation (e)

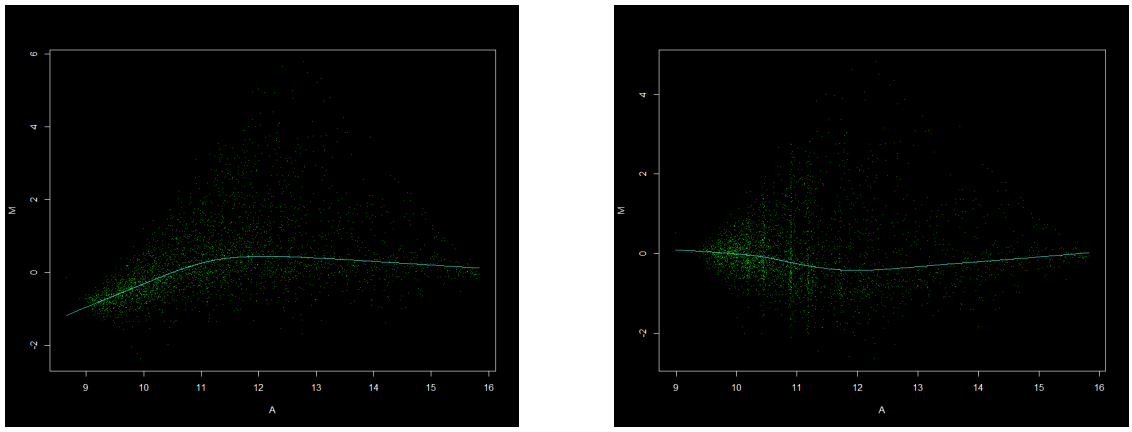
For microarrays, the Z score transformation methodology corrects data internally within a single hybridization, and individual gene hybridization values are presented as a unit of standard deviation from the normalized mean of zero. Because the correction is done before the sample-to-sample comparison, it is comparison-independent. Comparisons across samples or across experiments are then performed on equivalently transformed data. Gene expression data from several microarray studies may be compared across experiments using this method (Cheadle et al., 2003). However, before it could be applied it was necessary to perform normalization methods in both patient 2 and patient 3 arrays taking in count the process previously done for patient 1 array. Therefore, background subtraction followed by global normalization was carried out (Fig.2.6 and Fig.2.7). Similarly to patient 1 array (Fig.2.4b), in patient 3 array the Cy3 channel appears to responds more strongly at low intensity levels while the Cy5 channel responds more strongly at high intensities. On the other hand, in patient 2 array the opposite is verified. Nonetheless, the data looks

to flatten out at high intensities for both patient 2 and patient 3 arrays, suggesting once again that a non-linear fit would provide more accurate results. The data of both arrays was then normalized by subtracting the fitted value on the line from the log ratio of each feature. The highest-intensity points are the ones found above the horizontal line (Fig.2.6b and Fig.2.7b).



(a) (b)

Figure 2.6: Patient 2 plots of log ratio as a function of average intensity (MA) and normalisation; (a) MA plot of raw patient 2 data (background-subtracted) (b) MA plot of patient 2 data after global normalization.



(a) (b)

Figure 2.7: Patient 3 array normalization; (a) MA plot of raw patient 3 data (background-background-subtracted) (b) MA plot of patient 3 data after global normalization.

Visualising the Data: Box Plots (f)

The box plot is a method that allows for the visualization of several distributions simultaneously. It is a great way to compare genes log intensities or log ratios across multiple microarrays. A box plot depicts a distribution as a central box with whiskers on either side. The mean of the distribution is shown by the line through the center of the box. The standard deviation of the distribution is shown by the box. The horizontal lines that surround the box represent the distribution's extreme values. There are three common ways to normalize data so that arrays can be compared fairly. They all start with the premise that differences in distributions between arrays are caused by experimental circumstances and not by biological variability (Stekel, 2003).

Centering is one data normalization method in which the data is centered to make sure that all of the distributions' means and standard deviations are the same. For each measurement on the array the mean measurement is subtracted and divide by the standard deviation. The mean of the measurements on each array will be zero after centering, and the standard deviation will be (Stekel, 2003).

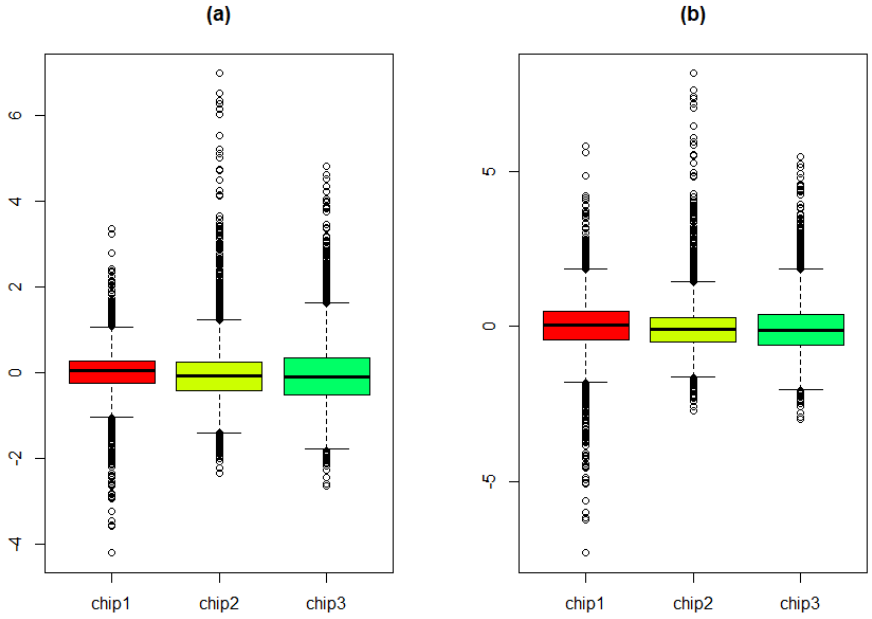


Figure 2.8: Boxplots; (a) Boxplot of raw log ratios of 3 patients arrays (background subtration). (b) Boxplot of centered data of 3 patients arrays.

Differential expression analysis (g)

For each gene in the patients data there is a measurement of expression for each treatment condition. These were normalized, logged and combined into a log ratio followed by a Z-score transformation for each patient, which describes numerically the extent to which the gene is differentially expressed, and whether it is up-regulated or down-regulated. The goal is to be identify the genes that are consistently differentially regulated across all patients, with a view to asserting what genes are differentially expressed in each treatment condition. Therefore, z-scores higher than threshold of -2 and 2 were analysed for all pairwise patient comparisons. A total of 7 genes were found to be differential expressed when comparing patient 1 and 2 (Fig.2.9), with the highest number being 12 genes when comparing patient 2 and 3 (Fig. 2.11). Regarding patient 1 and 3, a total of 9 genes were found (Fig.2.10).

	Id	chip1	chip2
1	Id2065	2.142968	7.411959
2	Id2096	-2.047047	5.953666
3	Id2293	2.100720	3.791967
4	Id2360	-4.263769	2.230356
5	Id2615	2.020467	2.730120
6	Id346	-3.024238	2.303104
7	Id371	2.684353	2.415394

Figure 2.9: Differently expressed genes in common between patient 1 and 2

	Id	chip1	chip3
1	Id1333	-2.089688	2.237557
2	Id2091	-3.602278	2.899724
3	Id2092	-2.319838	5.451917
4	Id2354	-3.101514	2.332039
5	Id2359	-4.534983	2.134339
6	Id241	2.137866	2.854713
7	Id253	-5.032710	-2.070046
8	Id2657	-2.368212	3.189057
9	Id80	-3.704293	2.830578

Figure 2.10: Differently expressed genes in common between patient 1 and 3

	Id	chip2	chip3
1	Id1286	3.039321	2.393107
2	Id1319	4.956010	2.198549
3	Id1861	2.591825	2.080910
4	Id2001	-2.284987	2.003673
5	Id2073	2.304369	4.525736
6	Id2188	3.736945	2.036081
7	Id2319	3.093042	2.755878
8	Id2630	-2.007369	2.323451
9	Id2864	4.068148	3.492663
10	Id2878	5.244739	3.087481
11	Id437	2.807645	-2.577587
12	Id663	2.086078	2.653135

Figure 2.11: Differently expressed genes in common between patient 2 and 3

Differential expression analysis using Limma

Limma is a package for analyzing gene expression data generated by microarray or RNA-seq technology (Ritchie et al., 2015). The use of linear models to assess differential expression in multifactor design experiments is a basic competence. Limma allows the comparison of many RNA targets at the same time.

3.1 Empirical Bayesian Method

Identifying genes which are differentially expressed across specified conditions in designed microarray experiments represents a massive multiple testing problem in which one or more tests are conducted for each of tens of thousands of genes. The problem is difficulted by the fact that measured expression levels are frequently non-normally distributed, with non-identical and dependent distributions between genes. Allowance needs to be made in the analysis of microarray experiments for the amount of multiple testing, throught the control of the familywise error rate or the false discovery rate, even though this reduces the power available to detect changes in expression for individual genes (Ge, Dudoit and Speed, 2003). However, the parallel structure of microarray inference enables some compensating possibilities for borrowing information from the ensemble of genes, which can help infer about each gene separately. Empirical Bayes method is one strategy that can be used to accomplish this where data from all the genes in a replicate set of experiments are combined into estimates of parameters of a prior distribution (Lönnstedt, 2001). These parameter estimates are then combined at the gene level with means and standard deviations to form a statistic B which is a Bayes log posterior odds. B can then be used to determine if differential expression has occurred. Limma makes use of this method to moderate the standard errors of the estimated log-fold changes.

This results in more stable inferences and improved power, especially for experiments with small numbers of arrays.

Let N be the number of genes in a microarray experiment, n the number of replicates for each gene, and M_{ij} =log R_{ij} - log G_{ij} , i = 1,...,N and j = 1,...,n, the log ratios of the green and red intensities for each gene. Mij is regarded as random variables from a normal distribution with mean μ_i and variance θ_i^2, so that, independently and identically:

$$M_{ij}|\mu_i, \sigma_i \sim N(\mu_i, \sigma_i)$$

Let I_i be the indicator for wheter a gene is differentially expressed or not:

$$I_i = \begin{cases} 0, & \mu_i = 0 \\ 1, & \mu_i \neq 0 \end{cases}$$

Most genes have $\mu_i = 0$, but a small proportion p of genes have some $\mu_i = 0$, indicated by $I_i = 1$ as opposed to $I_i = 0$. The parameters (μ_i, θ_i^2) are treated

The authors propose a target measure - a logarithm of posterior odds for being differentially expressed for each gene i:

$$B_i = \ln \frac{P(I_i=1|\mathbf{M}_i)}{P(I_i=0|\mathbf{M}_i)} = \ln \frac{p\, f_{I_i=1}(\mathbf{M}_i)}{(1-p)\, f_{I_i=0}(\mathbf{M}_i)};$$

where p is the proportion of differentially expressed genes; M_i = (M_{i1},....,M_{iJ} and gene i is differentially expressed if $B_i > 0$.

3.2 Differential expression Analysis (h)

First a linear model was applied to the previously normalized data of the 3 patients, which fully models the systematic part of the data. Limma topTable() function summarizes the linear model's results, performs hypothesis tests, and adjusts the p-values for multiple testing. Results include (log2) fold changes, standard errors, t-statistics and p-values. topTable() displays a number of summary statistics for the top genes and the specified contrast. The value of the contrast is shown in the

logFC column. This usually refers to a log2-fold difference between two or more experimental conditions, although it can also refer to a log2-expression level. The AveExpr column shows the gene's average log2-expression level across all arrays and channels in the experiment. Column t is the moderated t-statistic. Column P.Value is the associated p-value and adj. P.Value is the p-value adjusted for multiple testing. Benjamini and Hochberg's (BH) method, one popular form of adjustment, was applied in order to control the false discovery rate [1]. If the goal is to control or estimate the false discovery rate, the adjusted values are commonly referred to as q-values. The meaning of "BH" q-values is as follows: if all genes with a q-value less than a certain threshold, such as 0.05, are selected as differentially expressed, the estimated proportion of false discoveries in the selected group is kept below the threshold, in this case 5%.

The log-odds that the gene is differently expressed are represented by the B-statistic (lods or B). A B-statistic of zero corresponds to a 50-50 chance that the gene is differentially expressed. The B-statistic is automatically adjusted for multiple testing by assuming that 1% of the genes, or some other percentage specified by the user in the call to eBayes(), are expected to be differentially expressed. In this report proportion of 5% and 10% were tested. Benjamini and Hochberg's control of the false discovery rate assumes independence between genes. The B-statistic probabilities are based on the same assumptions but they also need a prior estimation for the proportion of probes that are differentially expressed.

A volcano plot typically displays some measure of effect on the x-axis, such as the log fold change, and the statistical significance on the y-axis, in this case the log odds of differential expression. The basic use of volcano plots is to survey genes that could be selected by one differential expression criterion but not the other. However, the genes identified as differentially expressed should respect the selection according to both the statistical method and the log fold change. Statistical speaking, many times, the differences between expression level in both methods are significant because the statistic method is able to recognize differences that the fold change is not because the fold change is not robust.

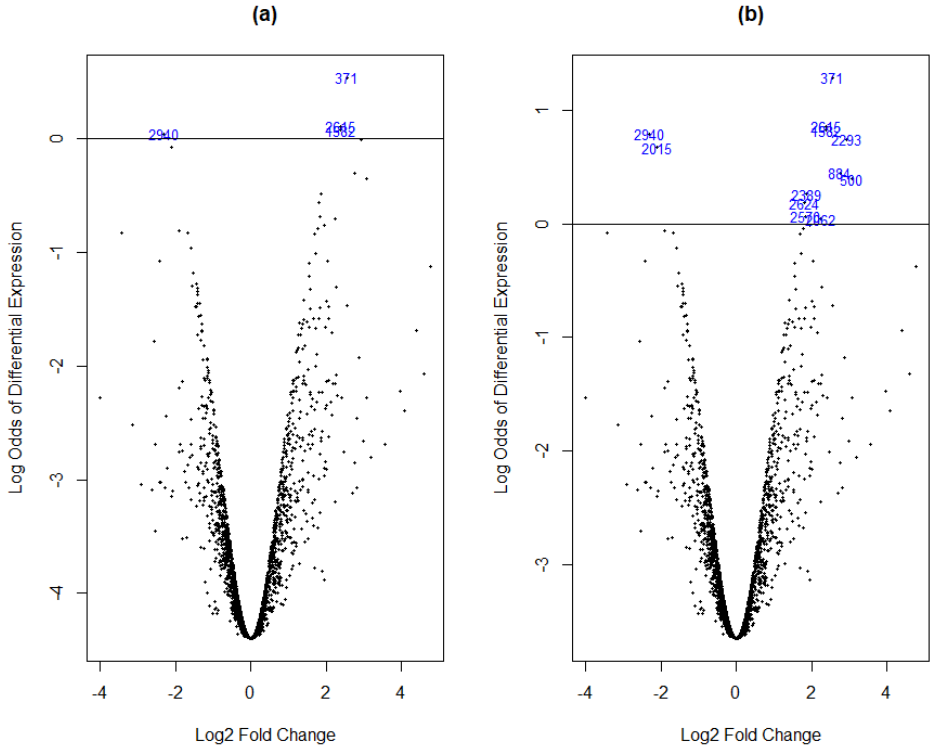


Figure 3.1: Volcano plots comparison between patient 1 and 2; (a) Volcano plot with proportion = 0.05. (b) Volcano plot with proportion = 0.1.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
Id371	2.549874	2.549874	5.956589	0.003043528	0.9852702	0.53766831
Id2615	2.375294	2.375294	5.207697	0.005136814	0.9852702	0.11003277
Id1582	2.380985	2.380985	5.152864	0.005349649	0.9852702	0.07564226
Id2940	-2.322399	-2.322399	-5.100596	0.005562451	0.9852702	0.04244208

Figure 3.2: Toptable of patient 1 and 2 with proportion = 0.05

	logFC	AveExpr	t	P.Value	adj.P.Val	B
Id371	2.549874	2.549874	5.956589	0.003043528	0.9852702	1.28488272
Id2615	2.375294	2.375294	5.207697	0.005136814	0.9852702	0.85724717
Id1582	2.380985	2.380985	5.152864	0.005349649	0.9852702	0.82285666
Id2940	-2.322399	-2.322399	-5.100596	0.005562451	0.9852702	0.78965648
Id2293	2.946343	2.946343	5.027128	0.005878992	0.9852702	0.74229039
Id2015	-2.095665	-2.095665	-4.922322	0.006368777	0.9852702	0.67327756
Id884	2.772793	2.772793	4.593628	0.008255229	0.9852702	0.44541588
Id500	3.066451	3.066451	4.520711	0.008760152	0.9852702	0.39243275
Id2389	1.876364	1.876364	4.346789	0.010120793	0.9852702	0.26234662
Id2624	1.799007	1.799007	4.246914	0.011015664	0.9852702	0.18522864
Id2570	1.826850	1.826850	4.097988	0.012530885	0.9852702	0.06687831
Id2062	2.248776	2.248776	4.068955	0.012854350	0.9852702	0.04333079

Figure 3.3: Toptable of patient 1 and 2 with proportion = 0.1

In order to identify genes differentially expressed the data of the three patients were all pairwise compared. Genes that are highly dysregulated are farther to the left and right sides, while highly significant changes appear higher on the plot. As previously mentioned, for all pairwise comparisons was tested a proportion of differentially expressed genes of 5% and 10%. It can be verified that for all cases (Fig.3.1, Fig.3.4, Fig.3.7 and Fig.3.10) more differentially expressed genes are found when the proportion is set to 10% when comparing to the 5%. In order to gain insights on the B-statistic values, the maximum values were calculated for all pairwise comparisons and it was verified that in all cases, these values varied across all comparisons but none of them was much higher than 1, with the lowest being 0.5 and the highest 1.99. If B-statistic is zero, the probability of gene being differentially expressed is the same as not being differentially expressed. If the goal is to determine the genes that are differentially expressed it means that the B-statistic must be higher than

zero to considered the gene to be differentially expressed, therefore the cuttof value established was zero.

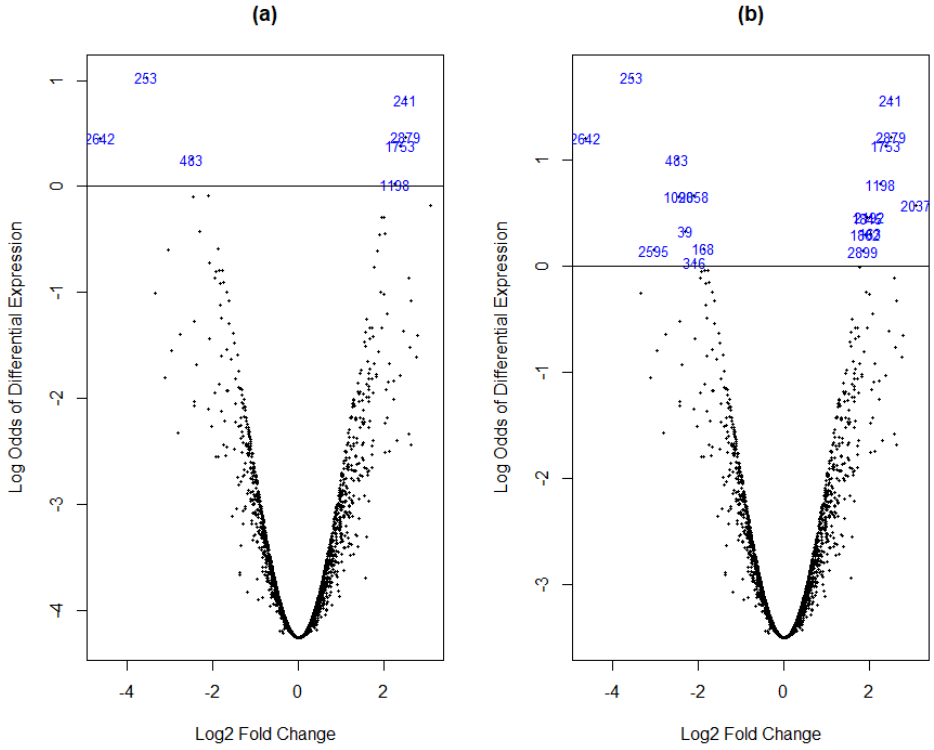


Figure 3.4: Volcano plots comparison between patient 1 and 3; (a) Volcano plot with proportion = 0.05. (b) Volcano plot with proportion = 0.1.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
Id253	-3.551378	-3.551378	-4.801920	0.001288479	0.9733623	1.0267855
Id241	2.496289	2.496289	4.610558	0.001655025	0.9733623	0.8168104
Id2879	2.486903	2.486903	4.305189	0.002495447	0.9733623	0.4676777
Id2642	-4.621645	-4.621645	-4.293712	0.002534942	0.9733623	0.4542193
Id1753	2.370985	2.370985	4.236508	0.002742095	0.9733623	0.3867822
Id483	-2.490510	-2.490510	-4.128733	0.003183636	0.9733623	0.2581043
Id1198	2.247126	2.247126	3.935748	0.004177028	0.9733623	0.0224719

Figure 3.5: Toptable of patient 1 and 3 with proportion = 0.05

	logFC	AveExpr	t	P.Value	adj.P.Val	B
Id253	-3.551378	-3.551378	-4.801920	0.001288479	0.9733623	1.77399985
Id241	2.496289	2.496289	4.610558	0.001655025	0.9733623	1.56402477
Id2879	2.486903	2.486903	4.305189	0.002495447	0.9733623	1.21489212
Id2642	-4.621645	-4.621645	-4.293712	0.002534942	0.9733623	1.20143366
Id1753	2.370985	2.370985	4.236508	0.002742095	0.9733623	1.13399655
Id483	-2.490510	-2.490510	-4.128733	0.003183636	0.9733623	1.00531874
Id1198	2.247126	2.247126	3.935748	0.004177028	0.9733623	0.76968630
Id2058	-2.112519	-2.112519	-3.846555	0.004744232	0.9733623	0.65856795
Id1098	-2.462692	-2.462692	-3.839713	0.004791026	0.9733623	0.64998669
Id2037	3.073798	3.073798	3.777184	0.005242205	0.9733623	0.57120116
Id2192	1.981831	1.981831	3.689978	0.005948683	0.9733623	0.46022946
Id1845	1.940954	1.940954	3.683796	0.006002479	0.9733623	0.45231483
Id39	-2.319309	-2.319309	-3.584595	0.006940268	0.9733623	0.32447734
Id163	2.006378	2.006378	3.571030	0.007080149	0.9733623	0.30687466
Id1862	1.882961	1.882961	3.561632	0.007178807	0.9733623	0.29466331
Id168	-1.884085	-1.884085	-3.461206	0.008328908	0.9733623	0.16333343
Id2595	-3.034144	-3.034144	-3.450502	0.008462539	0.9733623	0.14924566
Id2899	1.832135	1.832135	3.440583	0.008588378	0.9733623	0.13617797
Id346	-2.085929	-2.085929	-3.359445	0.009695118	0.9733623	0.02875116

Figure 3.6: Toptable of patient 1 and 3 with proportion = 0.1

When comparing patient 1 and 2 a total of 4 and 12 genes were found to be differentially expressed when the proportion of 5% and 10% was established, respectively (Fig.3.2 and Fig.3.3). On the other hand, when comparing patient 1 and 3, a total of 7 and 19 genes were found to be differentially expressed when the proportion of 5% and 10% was established, respectively (Fig.3.5 and Fig.3.6). When comparing patient 2 and 3, a total of 6 and 18 genes were found to be differentially expressed when the proportion of 5% and 10% was established, respectively (Fig.3.8 and Fig.3.9). Lastly, when comparing all three patients a total of 3 and 11 genes were found to be differentially expressed when the proportion of 5% and 10% was established, respectively (Fig.3.11 and Fig.3.12).

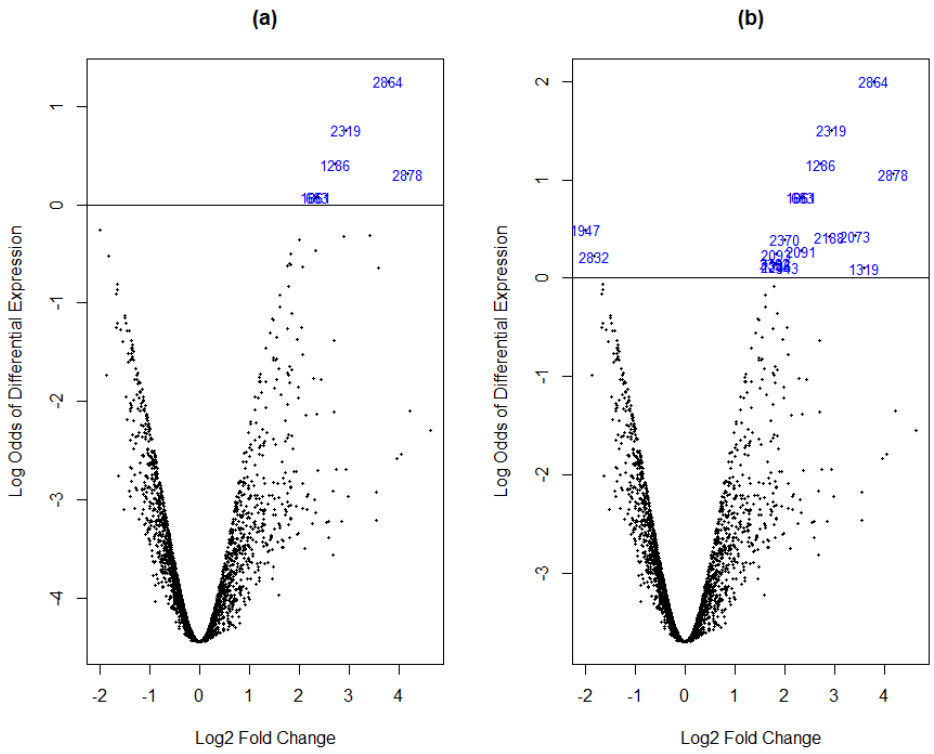


Figure 3.7: Volcano plots comparison between patient 2 and 3; (a) Volcano plot with proportion = 0.05. (b) Volcano plot with proportion = 0.1.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
Id2864	3.780405	3.780405	8.978736	0.001089894	0.8953005	1.25197344
Id2319	2.924460	2.924460	7.245268	0.002366697	0.8953005	0.75699592
Id1286	2.716214	2.716214	6.349901	0.003779688	0.8953005	0.41263309
Id2878	4.166110	4.166110	6.126040	0.004287051	0.8953005	0.31442807
Id663	2.369606	2.369606	5.638101	0.005721450	0.8953005	0.08099626
Id1861	2.336368	2.336368	5.622936	0.005774904	0.8953005	0.07328686

Figure 3.8: Toptable of patient 2 and 3 with proportion = 0.05.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
Id2864	3.780405	3.780405	8.978736	0.001089894	0.8953005	1.9991878
Id2319	2.924460	2.924460	7.245268	0.002366697	0.8953005	1.5042103
Id1286	2.716214	2.716214	6.349901	0.003779688	0.8953005	1.1598475
Id2878	4.166110	4.166110	6.126040	0.004287051	0.8953005	1.0616425
Id663	2.369606	2.369606	5.638101	0.005721450	0.8953005	0.8282107
Id1861	2.336368	2.336368	5.622936	0.005774904	0.8953005	0.8205013
Id1947	-1.981414	-1.981414	-5.023795	0.008485403	0.8953005	0.4917687
Id2073	3.415052	3.415052	4.924019	0.009078872	0.8953005	0.4321547
Id2188	2.886513	2.886513	4.905053	0.009197399	0.8953005	0.4206562
Id2370	2.005455	2.005455	4.858052	0.009499447	0.8953005	0.3919275
Id2091	2.315560	2.315560	4.673486	0.010810314	0.8953005	0.2758332
Id2093	1.824710	1.824710	4.628265	0.011164793	0.8953005	0.2465730
Id2832	-1.818165	-1.818165	-4.588249	0.011490439	0.8953005	0.2204072
Id2392	1.811556	1.811556	4.484352	0.012391962	0.8953005	0.1512544
Id1386	1.824872	1.824872	4.459617	0.012619229	0.8953005	0.1345289
Id224	1.752294	1.752294	4.443825	0.012767005	0.8953005	0.1237968
Id343	2.066578	2.066578	4.430929	0.012889250	0.8953005	0.1150024
Id1319	3.577279	3.577279	4.412942	0.013062166	0.8953005	0.1026887

Figure 3.9: Toptable of patient 1 and 2 with proportion = 0.1.

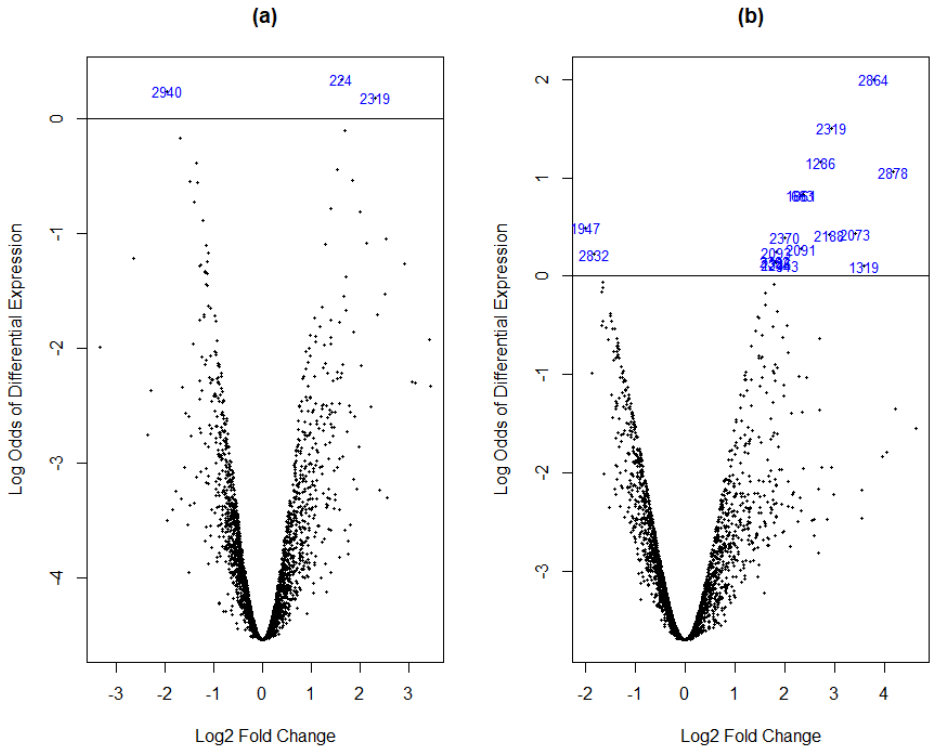


Figure 3.10: Volcano plots comparison between all 3 patients; (a) Volcano plot with proportion = 0.05. (b) Volcano plot with proportion = 0.1.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
Id224	1.628801	1.628801	4.772446	0.003512559	0.8928125	0.3450071
Id2940	-1.947708	-1.947708	-4.662050	0.003920404	0.8928125	0.2415555
Id2319	2.301616	2.301616	4.602037	0.004164589	0.8928125	0.1844333

Figure 3.11: Toptable of all patients with proportion = 0.05

	logFC	AveExpr	t	P.Value	adj.P.Val	B
Id224	1.628801	1.628801	4.772446	0.003512559	0.8928125	1.09222151
Id2940	-1.947708	-1.947708	-4.662050	0.003920404	0.8928125	0.98876992
Id2319	2.301616	2.301616	4.602037	0.004164589	0.8928125	0.93164765
Id2370	1.687747	1.687747	4.314142	0.005604495	0.8928125	0.64882548
Id478	-1.685016	-1.685016	-4.246400	0.006020588	0.8928125	0.58013491
Id1493	-1.355634	-1.355634	-4.042095	0.007502954	0.8928125	0.36797157
Id2915	1.522461	1.522461	3.984888	0.007988916	0.8928125	0.30721339
Id2615	1.844578	1.844578	3.894591	0.008829756	0.8928125	0.21010872
Id1565	-1.483357	-1.483357	-3.886100	0.008913813	0.8928125	0.20090153
Id104	-1.322279	-1.322279	-3.879013	0.008984649	0.8928125	0.19320785
Id2013	-1.405041	-1.405041	-3.723920	0.010704836	0.8928125	0.02255758

Figure 3.12: Toptable of all patients with proportion = 0.1.

Conclusions

4.1 Results Comparison (i)

When it comes to the outcomes, using fixed tresholds in the Z-score method did not allow the detection of common genes that were differentially expressed across all three patients. However, by modifing eBayes function parameters and setting the expected proportions to 5% and 10% it was possible to identify 3 and 11 genes, respectively, that were commonly differentially expressed between all patients. Furthermore, when doing pairwise comparisons between the three patients were identified between 4 and 19 common genes.

The intersection analysis of differentially expressed genes found in the pairwise comparisons of the two approaches revealed that the genes found did not overlap. This highlights the differences of both statistical methods. With this in mind, deciding which approach performs better in the data given is not an easy choice and not having additional information on the data complicates this task. Therefor, one strategy that can be considered is to use all genes identified as differentially expressed in both approaches for further analysis. Notwithstanding, it was verified that patient 2 and patient 3 data presented similar values of z-scores which was subsequently reflected in the number of common genes with differential expression in these two arrays, in contrast to patient 1, which had always less genes in common with patient 2 and 3. In the case of B-statistic this was not the case with the total number of genes found for all pairwise comparisons being similar.

As previously mention in introduction, gene expression is context-dependent and can be regulated in a variety of ways which makes it difficult to establish patterns across different samples and thus explaining the differences observed across the patients.

References

Ana Patricia Gomez Mayen, G̃ustavo Corral Guille <gcorral@ifc.unam.mx> and G̃erardo Coello Coutino <gcoello@ifc.unam.mx> L̃ina Riego Ruiz (2020). *genArise: Microarray Analysis tool*. R package version 1.66.0. URL: http://www.ifc.unam.mx/genarise (cit. on p. 3).

Bumgarner, Roger (2013). 'Overview of DNA Microarrays: Types, Applications, and Their Future'. en. In: *Current Protocols in Molecular Biology* 101.1. _eprint: https://currentprotocols.onlinelibrary.wiley.com/doi/pdf/10.1002/0471142727.mb2201s101, pp. 22.1.1–22.1.11. ISSN: 1934-3647. DOI: 10.1002/0471142727.mb2201s101. URL: https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/0471142727.mb2201s101 (visited on 12th June 2021) (cit. on p. 1).

Cheadle, Chris et al. (May 2003). 'Analysis of Microarray Data Using Z Score Transformation'. en. In: *The Journal of Molecular Diagnostics* 5.2, pp. 73–81. ISSN: 1525-1578. DOI: 10.1016/S1525-1578(10)60455-2. URL: https://www.sciencedirect.com/science/article/pii/S1525157810604552 (visited on 12th June 2021) (cit. on pp. 10, 11).

Cleveland, William S. (Dec. 1979). 'Robust Locally Weighted Regression and Smoothing Scatterplots'. In: *Journal of the American Statistical Association* 74.368. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1979. pp. 829–836. ISSN: 0162-1459. DOI: 10.1080/01621459.1979.10481038. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481038 (visited on 12th June 2021) (cit. on p. 6).

Ge, Youngchao, Sandrine Dudoit and Terence P. Speed (June 2003). en. In: *Test* 1, pp. 1–77. (Visited on 13th June 2021) (cit. on p. 16).

Jaksik, Roman et al. (Sept. 2015). 'Microarray experiments and factors which affect their reliability'. In: *Biology Direct* 10. ISSN: 1745-6150. DOI: 10.1186/s13062-015-0077-2. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4559324/ (visited on 12th June 2021) (cit. on p. 1).

Kooperberg, Charles et al. (Jan. 2002). 'Improved Background Correction for Spotted DNA Microarrays'. In: *Journal of Computational Biology* 9.1. Publisher: Mary Ann Liebert, Inc., publishers, pp. 55–66. DOI: 10.1089/10665270252833190. URL: https://www.liebertpub.com/doi/abs/10.1089/10665270252833190 (visited on 12th June 2021) (cit. on p. 6).

Lönnstedt, Ingrid (2001). 'Replicated microarray data'. en. In: *undefined*. URL: /paper/Replicated-microarray-data-L%C3%B6nnstedt/056e7774cd48d90dad92fbac784baa1837459 (visited on 13th June 2021) (cit. on p. 16).

Pevsner, Jonathan (2015). *Bioinformatics and functional genomics*. Third edition. Chichester, West Sussex, UK ; Hoboken, New Jersey: John Wiley and Sons, Inc. ISBN: 978-1-118-58176-6 978-1-118-58169-8 (cit. on p. 1).

Quackenbush, John (Dec. 2002). 'Microarray data normalization and transformation'. en. In: *Nature Genetics* 32.S4, pp. 496–501. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng1032. URL: http://www.nature.com/articles/ng1032z (visited on 12th June 2021) (cit. on p. 3).

Ritchie, Matthew E. et al. (Apr. 2015). 'limma powers differential expression analyses for RNA-sequencing and microarray studies'. eng. In: *Nucleic Acids Research* 43.7, e47. ISSN: 1362-4962. DOI: 10.1093/nar/gkv007 (cit. on p. 16).

Robinson, Mark D., Davis J. McCarthy and Gordon K. Smyth (Jan. 2010). 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data'. eng. In: *Bioinformatics (Oxford, England)* 26.1, pp. 139–140. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btp616 (cit. on p. 7).

Stekel, Dov (2003). *Microarray Bioinformatics*. Cambridge: Cambridge University Press. ISBN: 978-0-521-52587-9. DOI: 10.1017/CBO9780511615535. URL: https://www.cambridge.org/core/books/microarray-bioinformatics/F6E7A3C5C7E57C8DC8086A43 (visited on 12th June 2021) (cit. on pp. 3, 4, 7, 13).

Team, R Core (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/ (cit. on p. 2).

Vawter, Marquis P et al. (July 2001). 'Application of cDNA microarrays to examine gene expression differences in schizophrenia'. en. In: *Brain Research Bulletin*. Neuropathology of Severe Mental Illness: Studies from the Stanley Foundation Neuropathology Consortium 55.5, pp. 641–650. ISSN: 0361-9230. DOI: 10.1016/S0361-9230(01)00522-6. URL: https://www.sciencedirect.com/science/article/pii/S0361923001005226 (visited on 12th June 2021) (cit. on p. 9).

Virtaneva, K. et al. (Jan. 2001). 'Expression profiling reveals fundamental biological differences in acute myeloid leukemia with isolated trisomy 8 and normal cytogenetics'. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 98.3, pp. 1124–1129. ISSN: 0027-8424. DOI: 10.1073/pnas.98.3.1124 (cit. on p. 9).

Wildsmith, S E (Feb. 2001). 'Microarrays under the microscope'. In: *Molecular Pathology* 54.1, pp. 8–16. ISSN: 13668714. DOI: 10.1136/mp.54.1.8. URL: https://mp.bmj.com/lookup/doi/10.1136/mp.54.1.8 (visited on 12th June 2021) (cit. on p. 1).

Yang, Ivana V. et al. (Oct. 2002). 'Within the fold: assessing differential expression measures and reproducibility in microarray assays'. In: *Genome Biology* 3.11, research0062.1. ISSN: 1474-760X. DOI: 10.1186/gb-2002-3-11-research0062. URL: https://doi.org/10.1186/gb-2002-3-11-research0062 (visited on 12th June 2021) (cit. on p. 6).

Yang, Yee Hwa et al. (Feb. 2002). 'Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation'. In: *Nucleic Acids Research* 30.4, e15–e15. ISSN: 0305-1048. DOI: 10.1093/nar/30.4.e15. URL: https://doi.org/10.1093/nar/30.4.e15 (visited on 12th June 2021) (cit. on p. 6).