Final Year Project

The Analysis of Gene Expression Data generated by Microarray and RNA-Seq technologies

Author: Amanda Forde 16308191

Supervisor: Dr. Emma Holian

April 2020

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Bachelor of Science (Mathematical Science) is entirely my own work and had not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

ID: *16308191*

Date: *08/04/2020*

Acknowledgements

I would like to extend my sincere gratitude to Dr. Emma Holian for her dedication to assisting me in this project. I am most appreciative of her encouragement and guidance for which without, I would not have been able to complete this report.

I would also like to acknowledge the lecturers and academic staff at the School of Mathematics, Statistics and Applied Maths here in NUI, Galway. I am very grateful for their teaching and continuous support throughout my degree.

Contents

1 Introduction

1.1 Differential Expression Analysis

In a biological system, genetic information has been found to flow from DNA to RNA to proteins. This concept is referred to as the central dogma of molecular biology. Every cell of an organism contains the same genetic information but in spite of this, it is obvious that cells can have very different roles. For example, consider a blood cell and a skin cell in the human body, the tasks carried out by these two cells alone have few similarities. The reason for this is predominantly due to variances in trends of gene activation, which in turn are responsible for the control of the protein construction. This process of manufacturing proteins from the information which is stored in genes is called *gene expression*. At any given time, several inactive genes can exist within a single cell, producing none or very few copies of their respective protein. On the other hand, other genes within the same cell can have considerably high levels of activity and many copies of these particular proteins yielded. Unfortunately, estimation of the abundance of individual proteins has proven to be difficult. Therefore, a quantitative description of the expression of a gene is generally obtained by determining the number of RNA transcripts, which are associated with that gene, present in cell samples.

Comparison of these gene expression profiles across various experimental groups can provide scientists with the ability to gain insight into the properties and functions of genes. Furthermore, changes in gene expression levels between different cells can assist in understanding why cells perform in various ways while different levels of expression in the same cells in contrasting experimental conditions can aid in the determination of differences between these conditions. A simple example is the comparison of expression levels in treated cells to those of untreated cells within subjects. This investigation may be performed in order to evaluate the effect of a particular treatment. In *differential expression analysis*, the primary concern of researchers is the identification of a subset of genes that are differentially expressed between two conditions from a large population of candidate genes. The role of the statistician in this process is to evaluate the magnitude of the difference of mean expression level between two or more conditions for each gene from the available data and assess the statistical significance of this difference. However, these tasks are often most challenging. The datasets acquired tend to be extremely large with tens of thousands of genes but in which there may only be two or three replicates for each condition of interest. In addition, it is most likely that these measurements have been greatly affected by noise. Thus, it is of interest to perform a thorough investigation of the many statistical methods which must be employed in order to overcome issues such as those mentioned, the statistical concepts which have led to their construction and of course, the manner in which these methods are implemented on real-world datasets.

The most popular technologies used in gene expression experiments today are microarrays and the next-generation sequencing technique, RNA-Seq. Even though they are employed for the same purpose of measuring RNA quantities in biological samples, there are significant differences between the two experimental procedures, especially with respect to the form of data they generate. Therefore, it will also be interesting to explore the way in which certain techniques must be altered depending on the source of the data.

1.2 Origin of Data

Before delving into a comprehensive investigation of statistical methodologies, it is important to understand the nature of the raw data obtained from gene expression experiments and appreciate that many pre-processing procedures must take place in order to transform this original data of great technical complexity into an appropriate form for analysis.

Microarrays:

Microarray technologies can be divided into two main categories: cDNA microarrays and oligonucleotide *probe-set* arrays. While sharing many common traits, the two procedures are differentiated by a number of technical details. Since their introduction by Affymetrix, Inc. in 1994, the popularity of high-density

oligonucleotide expression arrays has grown tremendously. In fact, most microarray data being generated today is of this *probe-set* type, also referred to as *single-channel* data. As the data which will be used for illustrative purposes originates from a probe-set array, the methods employed in sourcing this data type will be briefly explored.

A single-channel array is a small glass chip which consists of a series of short, single-stranded DNA sequences. These sequences are known as oligonucleotides and are commonly 25 base pairs in length. In general, each gene is represented by 16-20 pairs of oligonucleotides. These pairs of oligonucleotides are defined as probe sets, as their main purpose is to *probe*, i.e. search for, their specific complementary base sequences. Each probe set consists of a perfect match (PM) probe and a mismatch (MM) probe. PM probes are designed so that they exactly match different small segments of RNA sequences. The partner MM probe of a particular PM probe is created to merely differ from the PM probe by a single substitution of the middle base of the sequence.

An RNA sample is suitably prepared, labeled and subsequently hybridized with the array. Intensity values are obtained for each probe as a result of the production of images from the scanned array. The generated intensities quantify how much hybridization has occurred for each probe. The entire process is based on the fact that if the target sequence of a specific PM probe is present in the sample, the PM probe should hybridize to a stronger degree with this complementary base sequence than its partner MM probe. The PM intensity value contains the true signal as well as stray signal stemming from non-specific binding. The MM probes have been included in order to estimate this extra noise. The intensities from the 16-20 probe pairs representing a specific gene are combined to form a summary measurement for that gene.

RNA-Seq:

Recent years have witnessed the emergence of RNA-Seq, a new technology which can be used to generate gene expression data. Unlike microarrays, which provide continuous numerical measurements to quantify expression levels, RNA-Seq expression profiles are compiled of integer counts. This recorded count data represents the number of sequence reads obtained from a particular sample which have been mapped to each gene.

RNA-Seq is commonly considered to be superior to microarrays for several reasons. Among these include the fact that background signal has been found to be much lower with RNA-Seq, resulting from a lack of issues regarding non-specific hybridization. Furthermore, RNA-Seq is capable of quantifying a wide range of expression levels, in contrast to microarrays which can struggle with the detection of very high or very low expression levels. A brief description of the complex steps of RNA-Seq experimental process is provided as follows.

Prior to sequencing, the RNA must first be extracted from the cell sample. The method proceeds by transforming this population of RNA into numerous cDNA fragments. The collection of these fragments is known as a library. The addition of adapters to each end of the fragments takes place to enable hybridization to a *flowcell*, a small glass slide with microfluidic chambers. Following this step, each individual fragment is clonally amplified to guarantee that the sequencing signal for detection is sufficiently strong. The cDNA library is now of a suitable form to undergo sequencing using the sequencing platform of choice. The depth to which the library is sequenced can depend on several factors. This sequencing process results in a great number of short sequence reads. The genomic origin of each read must then be established in order to determine the RNA transcripts present in the sample at hand. The alignment of reads to the reference genome is referred to as mapping. As the objective of the RNA-Seq experiment is to provide a measurement of gene expression, the number of sequence reads that overlap with known genes are counted once each read has been associated with its locus of origin. Thus, for each sample and each gene, the RNA-Seq process provides a positive integer designed to represent the expression level of the gene in the sample.

1.3 Introduction to `limma` and Linear Models

The `limma` package is a key element of the open-source software designed for bioinformaticians, `Bioconductor`, which can be used through the interface of `R`, (Ritchie et al. (2015), Huber et al. (2015)). It has been developed to provide an accessible and statistically rigorous platform for the analysis of data obtained from gene expression experiments. Originally, `limma` gained popularity for its ability to conduct thorough analyses of microarray data. However, the availability of `limma`'s functions have been recently extended to new technologies such as that of RNA-Seq. Once initial pre-processing and normalisation procedures have taken place, `limma` can supply users with the ability to perform differential expression analyses on data arising from both microarray and RNA-Seq technologies in a very similar manner. For this reason, there will be much focus on the statistical methods used by `limma`.

The combination of many core statistical concepts in a manner which has proven to be most effective can be seen in `limma`. Differential expression is evaluated through the employment of a linear modelling approach in which linear models are fitted to normalised expression measurements of each gene. The adaptability of these models permits the analysis of experiments with complex design as the models can be adjusted to take into account a variety of treatment factors. Another key component of `limma` is its empirical Bayes procedures which borrow information across genes. These methods are included to enable each gene to have its own specific variance estimate and ensure that stable analyses take place even when only a small number of samples are available.

As the use of linear models will be a dominant feature throughout this investigation of differential expression analysis, a brief introduction to the way in which the application of linear models can result in inferences being made regarding gene expression levels is provided as follows. A simple example is considered where the intention is to assess the effect of a proposed treatment, Treatment A say. The question of interest, with respect to differential gene expression, may then be to determine differences in gene expression levels between samples extracted from treated cells and those which have been taken from untreated cells. Following this, the identification of genes which are over-(under-) expressed may most likely be required. In the context of statistical modelling, the measurement of expression level of a particular gene is recognised as the response variable while the characteristic of being either treated or untreated is the predictor categorical variable. If expression measurements are assumed to be normally distributed, a linear regression model for gene g may take the following form:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

for which:

- Y_i specifies the expression level of gene g for sample i,
- β_0 is the mean expression level in untreated samples for this gene,
- β_1 is the mean difference/change in expression level from untreated to treated samples,
- X_i is the treatment status of sample i, a discrete parameter which takes values 0 or 1, with 0 = untreated while 1 = treated,
- ϵ_i represents random noise.

The standard assumption of this particular model is that the error terms, ϵ_i, are independent and identically distributed random variables with $\epsilon_i \sim N(0, \sigma^2)$.

A linear model such as the one above is often used for prediction purposes, in which an estimation for Y_i is desired for a given value of X_i. However, in this application of linear models, it is the values of the coefficients, β_j that are of most importance. It is evident from the equation above that β_1 is the coefficient which captures the magnitude of the difference in expression levels between the two groups of samples, treated and untreated. Therefore, in order to assess if this gene g is differentially expressed, an appropriate hypothesis test is conducted. This hypothesis test takes the form:

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

and the test statistic, namely the t-statistic, defined by

$$t_1 = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)},$$

is used to evaluate how close β_1 is to zero. The computation of this t-statistic occurs using the estimate of β_1, $\hat{\beta}_1$, and its associated standard error, $se(\hat{\beta}_1)$ which have been obtained as a result of fitting the above linear model to the available data. Under the null hypothesis, H_0, the t-statistic follows a t-distribution with $n-p-1$ degrees of freedom, in which n is the number of samples and p is the number of predictor variables in the model. The probability of observing this t-statistic, or indeed a value more extreme than it, under the null hypothesis, can then be obtained. This probability is referred to as a p-value and is often compared with a previously specified significance level, α. If the p-value is less than α, the null hypothesis, $H_0 : \beta_1 = 0$ is rejected and it is concluded that there is sufficient evidence to indicate that the gene g is differentially expressed between the two conditions of interest.

This is indeed a simplified description of the linear modelling procedure as it is extremely rare that raw experimental data would be normally distributed. However, it provides a basic comprehension of the rationality behind the implementation of linear models in differential expression analysis. Furthermore, the parallel nature of performing this same test for many, many genes must be accounted for. This concept of multiple testing and the appropriate adjustments which must be made will be explored in Section 2.

2 Multiple Testing

2.1 The Multiplicity Problem

One of the principal statistical issues that bioinformaticians must regularly face is that of the multiplicity problem. This issue is certainly not unique to gene expression experiments. However, its magnitude is dramatically intensified as each experiment can consist of thousands of genes. In order to identify a small portion of genes which are differentially expressed between experimental conditions from a large pool of candidate genes, thousands of comparison tests must be performed. The result of this is a sharp increase in the likelihood of obtaining false positive results. Therefore, it is clear that when performing differential expression analysis, incorporation of a procedure which appropriately corrects for multiple testing is most important.

First, consider the performance of a hypothesis test for statistical significance of differential expression for one individual gene, similar to that described in Section 1.3. The null hypothesis, H_0 is that the gene shows no difference in expression between the two conditions of interest. A t-statistic is computed and used to produce a p-value, the probability of obtaining the t-statistic under H_0. The gene is recognized as being differentially expressed if this p-value is less than the significance level of this single hypothesis test, α. However, it may be merely random effects that are responsible for this extreme t-statistic and small p-value. If this is the case, an error has been made in identifying the gene as differentially expressed. The null hypothesis, H_0 has been rejected when it is in fact true. This is known as a Type I error and the probability of making this error is exactly α. It follows that the probability of the correct conclusion being deduced for this particular gene is then $1-\alpha$.

For a situation in which many associated hypotheses are tested, the probability that a false positive error, such as the one described above, is committed is witnessed to increase greatly. Consider m associated hypothesis tests. As stated, the probability of not making a Type I error for a single hypothesis test is $1-\alpha$. For m tests, the probability of not making a Type I error reduces to $(1-\alpha)^m$. However, this means that the chance of making at least one false positive error is now represented by the value $1-(1-\alpha)^m$. This number grows at a rapid rate, eventually reaching extremely close to 1, as demonstrated in Figure 1.

In a typical experiment, at least $10,000$ separate hypothesis tests may be executed. Suppose data from a gene expression experiment has been obtained in which exactly 100 genes are known to be differentially

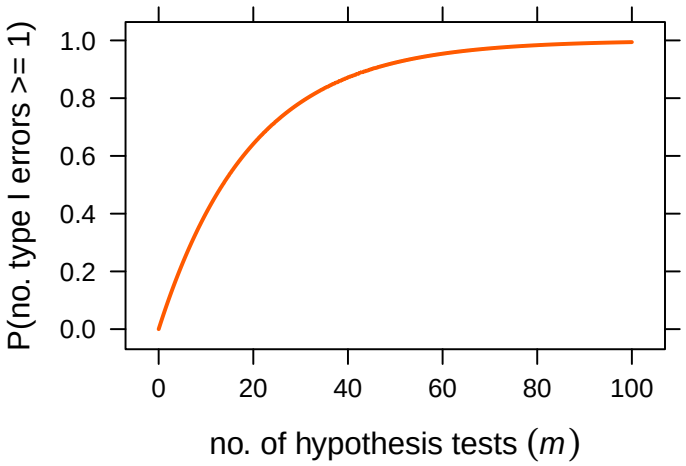

Figure 1: *The number of hypothesis tests, on the x-axis, is plotted against the probability that the number of type I errors committed is greater than or equal to 1, on the y-axis.*

expressed. If a gene-by-gene significance level of $\alpha = 0.05$ is chosen, then approximately 495 out of the ordinary $9,900$ genes would also be expected to be deemed significant merely by chance. Even if the testing procedure accomplishes the task of determining the 100 differentially expressed genes, they will be easily lost in the large pool of 495 false predictions. This results in a false positive rate of $\approx 83.19\%$ in the final set of predicted differentially expressed genes.

It is clear from this simple example that in this process of performing many associated comparison tests in parallel, a hypothesis testing approach which is deemed perfectly valid when applied to a single gene can potentially produce significant errors when analyzing data for thousands of genes. Thus, extreme caution must be taken when attempting to draw conclusions from the results of a testing procedure in which no corrections for multiple testing have been implemented. Recognition of the enormity of this issue has led to the focus of considerable attention to determining methods which can achieve acceptable control of the type I error rate. Commonly used approaches regularly involve adjusting the individual p-values obtained during the hypothesis testing procedure.

In this complicated task of differential expression prediction, there are two ways in which false positive rates are commonly understood, the *Family-Wise Error Rate* (FWER) and the *False Discovery Rate* (FDR). The *Family-Wise Error Rate* is the probability that at least one type I error is committed, while the *False Discovery Rate* describes the expected proportion of false positive errors which exist among the total number of rejected hypotheses. Assuming that m hypotheses are being tested, $H^1, H^2, \ldots, H^m$, let R represent the total number of rejected hypotheses and V signify the number of type I errors that have been made. Then,

$$\text{FWER} = P(V \geq 1)$$

$$\text{FDR} = E\left(\frac{V}{R} | R > 0\right) P(R > 0)$$

A procedure is said to control the FWER or FDR at a specific level α, if the FWER or FDR is known to be less than α. Details of methods commonly implemented to control these rates is provided below.

2.2 Family-Wise Error Rate (FWER)

Consider a case where the intention is to carry out m hypothesis tests and m_0 of the null hypotheses are in fact true. Then, the family-wise error rate (FWER) is simply the probability that at least one of the m_0 null hypotheses is rejected as a result of the performed hypothesis testing procedure.

One of the most simplistic approaches to ensuring that this FWER remains at a specified acceptable value, α, is the Bonferroni correction, (Bland and Altman (1995)). When controlling the FWER for a total number of m associated comparison tests, the task is to determine the significance level required for each individual hypothesis test in order to ensure that the global Type I error is less than or equal to α. It is perhaps intuitive that in order to meet this requirement, the Type I error of each individual test must be reduced to a number much smaller than α, depending on the number of statistical tests being carried out. First, p-values are generated for each hypothesis test corresponding to each gene. Let p_i represent the p-value for gene i. The Bonferroni correction method proceeds by implementing a Type I error of $\frac{\alpha}{m}$ for each individual hypothesis test. All hypotheses for which $p_i < \frac{\alpha}{m}$ are subsequently rejected. Equivalently, each p_i may be corrected through multiplication by m to form a new value, $\tilde{p_i} = p_i m$ and $\tilde{p_i}$ is compared with α. In a similar fashion, if $\tilde{p_i} < \alpha$, the null hypothesis for gene i, H_0^i is rejected. Thus, the Bonferroni correction procedure guarantees weak control of the experiment-level Type I error, the FWER at level α.

However, it is known that a fundamental issue of many approaches which focus on controlling the FWER, regardless of number of hypotheses being tested, is that of over-conservativeness. It can be deduced that after implementation of the Bonferroni correction method, if gene i is found to be significant, then it is truly differentially expressed. However, due to the stringency of the procedure for a large number of m tests, a gene which is expressed to a substantially different extent between experimental groups could still be deemed insignificant. This strictness results in a considerable loss of available power to detect expression changes for particular genes, leading to many missed findings. Thus, in the context of gene expression analysis, methods which aim to control the FWER are generally deemed unsuitable. This is primarily due to the fact that as the number of genes increases, the corrected probability of a Type I error at gene-level becomes very small. At this level, it is very difficult to succeed in rejecting the null hypotheses which are actually false. Attempts to counteract this conservativeness include raising the FWER to a figure greater than 0.05. However, increasing this probability of at least one false positive cannot control the amount of false positives allowed.

2.3 False Discovery Rate (FDR)

Essentially, the aim of overcoming the multiple testing issue is to find an approach which not only locates as many genes as possible which are expressed differently in contrasting experimental conditions but also ensures that a reasonably small proportion of false positives have been incurred. In order to achieve this, it has been recognized that emphasis on the proportion of false positives among the rejected hypothesis tests is more appropriate. The false discovery rate (FDR) was introduced by Benjamini and Hochberg (1995) and originally defined it as the expectation of this proportion. More formally, an FDR approach is designed to control the expected proportion of false positives among the set of rejected hypotheses, R at some prescribed value δ, in which V is the number of falsely rejected genes:

$$\text{FDR} = E\left(\frac{V}{R}|R > 0\right) P(R > 0).$$

FDR approaches to multiple testing have gained great popularity in data analysis which attempts to recognise differentially expressed genes. This is predominantly due to the less stringent nature of the false discovery rate, which allows it to be more powerful than other correction methods. Consider an experiment which has produced data for a total of $10,000$ genes, with 100 of these genes being diagnosed as differentially expressed on account of the enforcement of a hypothesis testing procedure. The null hypothesis, H_0^i was rejected for each of these 100 genes. As an example, an FDR of 50 would lead to 50 truly differentially expressed genes among this group of 100 candidate genes. A set of predictions such as this would in fact be considered suitable for a follow-up verification experimental method such as polymerase chain reaction (PCR).

Benjamini and Hochberg Method:

As stated above, Benjamini and Hochberg (1995) are responsible for the development of the original procedure of controlling the FDR. The testing method may be described as follows.

The p-values, $p_1, p_2, \ldots, p_m$ are generated gene-by-gene from a total of m hypothesis tests, under m individual null hypotheses. These m p-values are assumed to be independent. Suppose that m_0 of these null hypotheses are actually true. It is worth noting that at this stage of the process, which hypotheses are true and which are false remains unknown. The unadjusted p-values computed in this first step are now re-ordered from smallest to largest. The ith smallest of all p-values is denoted as $p_{(i)}$ with corresponding null hypothesis $H_0^{(j)}$, so that $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(g)}$. With δ representing the chosen false discovery rate, q_i is defined as:

$$q_i = \frac{i}{m}\delta, \quad i = 1, \ldots, m.$$

Each p-value, $p_{(i)}$ is compared with its complementary q_i and the hypothesis test with the highest rank, i.e. largest value i, for which the inequality $p_{(i)} \leq q_i$ holds is sought out. If the highest ranking test which satifies this condition is denoted as $H_0^{(k)}$, then all tests of rank $1, 2, \ldots, k$ are declared to be significant as for all $i = 1, 2, \ldots, k$, $p_{(i)} \leq q_i$. The null hypothesis $H_0^{(1)}, ..., H_0^{(k)}$ are rejected while the null hypotheses $H_0^{(k+1)}, ..., H_0^{(m)}$ are accepted.

In fact, it can be proven that this procedure precisely controls the FDR to the level $\frac{m_0}{m}\delta$, which is at the very most δ. In general, all adjustment procedures tend to result in the reduction of statistical power. However, the power losses are drastically smaller for the BH method than the Bonferrroni procedure detailed above.

The construction of this method to control the false discovery rate assumes that the m hypothesis tests are all independent. With gene expression data, the idea that the numerous performed hypothesis tests would be truly independent is most unlikely. That said, there is little proof as to the exact form which a dependence structure may take between the expression levels of various genes. It was proven by Benjamini and Yekutieli (2001) that the BH procedure does in fact control the FDR at level less than or equal to $\frac{m_0}{m}\delta$ for the case in which there is a positive regression dependence between the different hypothesis tests. The BH method was also extended by Benjamini and Yekutieli (2001) to take into an account an arbitrary dependence structure in which q_i in the above explanation is redefined as:

$$q_i = \frac{\delta i}{g \sum_{j=1}^{g} \frac{1}{j}}.$$

A negative characteristic of this BY method is that it is generally more conservative than the BH procedure. For this reason, the BH procedure is still often preferred. In fact, the default method used by the `limma` software package for adjusting p-values to account for multiple testing is the BH method while options to use the Bonferroni correction and BY methods are also provided.

3 Analysis of Microarray Data

3.1 Normalisation of Microarray Data

Throughout the microarray experimental process, which has been briefly detailed in Section 1.2, the introduction of many sources of variation can occur as there are numerous steps required which are not exactly reproducible. Technical error can be introduced at any technical stage of the experiment such as sample preparation and even, array manufacturing. As the intent is to detect the biological differences in gene expression, comparing data from arrays in which variation of non-biological origin hasn't been minimized

can lead to extremely misleading results. Therefore, implementation of normalisation procedures which aim to remove technical variation are essential prior to analysis.

Many methods for extracting gene expression measurements from the raw intensity data obtained from high-density oligonucleotide arrays which incorporate normalisation steps have been proposed over the last two decades. One method which attempts to perform this task efficiently is that suggested by Irizarry (2003), known as the robust multi-array average (RMA) method. Let PM_{ijn} denote the intensity obtained for an individual PM probe. The values n, i and j represent the specific gene, RNA sample and probe set, respectively. This probe set number identifies the physical position in the gene of the oligonucleotide in question. The method may be separated into three individual steps.

First, background correction of each array must be performed in order to adjust for the effects of optical noise and hybridization that is not associated with specific binding between probes and their target sequence. This task is completed with the assistance of a transformation, $B(\cdot)$. A model of the form $PM_{ijn} = bg_{ijn} + s_{ijn}$ is imposed on each probe, in which bg_{ijn} symbolizes the background signal in the sample i and s_{ijn}, the true signal. It is assumed that each array i has a mean background signal of $E(bg_{ijn}) = \beta_i$. Then, in order to remove this background effect, an appropriate transformation is specified by $B(PM_{ijn}) \equiv E(s_{ijn}|PM_{ijn})$. With $B(PM_{ijn}) > 0$, $B(\cdot)$ is defined as the closed-form transformation obtained when assuming s_{ijn} to be exponential and bg_{ijn} normal.

Following this, technical errors are removed using quantile normalisation. The aim of this approach is to ensure that the probe intensity distribution is the same for all arrays so that data can be compared across samples. Probe level data from all arrays are mapped by this normalisation so that an I-dimensional quantile-quantile (q-q) plot follows the I-dimensional identity line, in which I is the total number of arrays.

Finally, summarization takes place. This involves the combination of multiple intensities into a single number which can be understood as a measurement of gene expression. Let Y denote the background-adjusted, normalized and log-transformed PM intensities. This Y is assumed to follow the linear model:

$$Y_{ijn} = \mu_{in} + \alpha_{jn} + \epsilon_{ijn}$$

in which α_j specifies a probe affinity effect and ϵ_{ij} represents the standard error terms, independently and identically distributed with a mean of zero. Model parameters are estimated using a robust procedure in order to safeguard against outliers. Log scale expression levels for array i are provided by the estimate of μ_i. This is referred to as the robust multi-array average (RMA).

3.2 The `leukemiasEset` Dataset

The dataset chosen to gain a comprehension of the fundamental concepts of differential gene expression analysis was obtained from the `Bioconductor` package, `leukemiasEset`. The data is presented within an object of the `ExpressionSet` class, a form appropriate for insertion into the `limma` analysis pipeline. Gene expression data from a total of 60 samples of individual patients is contained in this object. 12 samples are derived from each of four different leukemia types: Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Lymphocytic Leukemia (CLL) and Chronic Myeloid Leukemia (CML). In addition, control samples were included from 12 healthy non-leukemia individuals (NoL). These 60 samples are a small subset of the samples collected by the Microarray Innovations in Leukemia (MILE) study (Kohlmann et al. (2008), Haferlach et al. (2010)).

The gene expressions were measured by high-density oligonucleotide arrays, namely Affymetrix Human Genome U133 Plus 2.0, in which mononuclear cells from bone marrow tissues of each patient were hybridized. The arrays were subjected to the RMA normalisation procedure, outlined in Section 3.1 above. The description of this procedure provides an understanding of the way in which the single log-intensity values for each gene in each sample have been produced. The total number of $20,172$ genes have been identified with their corresponding Gene Ensembl IDs. These expression values are contained within a large matrix in which rows represent different genes and the columns relate to the various samples. The expression values of the first 4 genes which have been generated for the first 4 `ALL` samples are shown directly below.

```
                ALL_0151 ALL_0153 ALL_0154 ALL_0157
ENSG00000000003 3.386743 3.687029 3.360517 3.459388
ENSG00000000005 3.539030 3.836208 3.246327 3.063286
ENSG00000000419 9.822758 7.969170 9.457491 9.591018
ENSG00000000457 4.747283 4.866344 4.981642 5.982854
```

Preliminary exploration of the data may include the production of visual summaries of the expression values for certain genes using a boxplot. The plots underneath, in Figure 2, depict how the expression measurements for two genes vary across the 5 different conditions of this dataset.

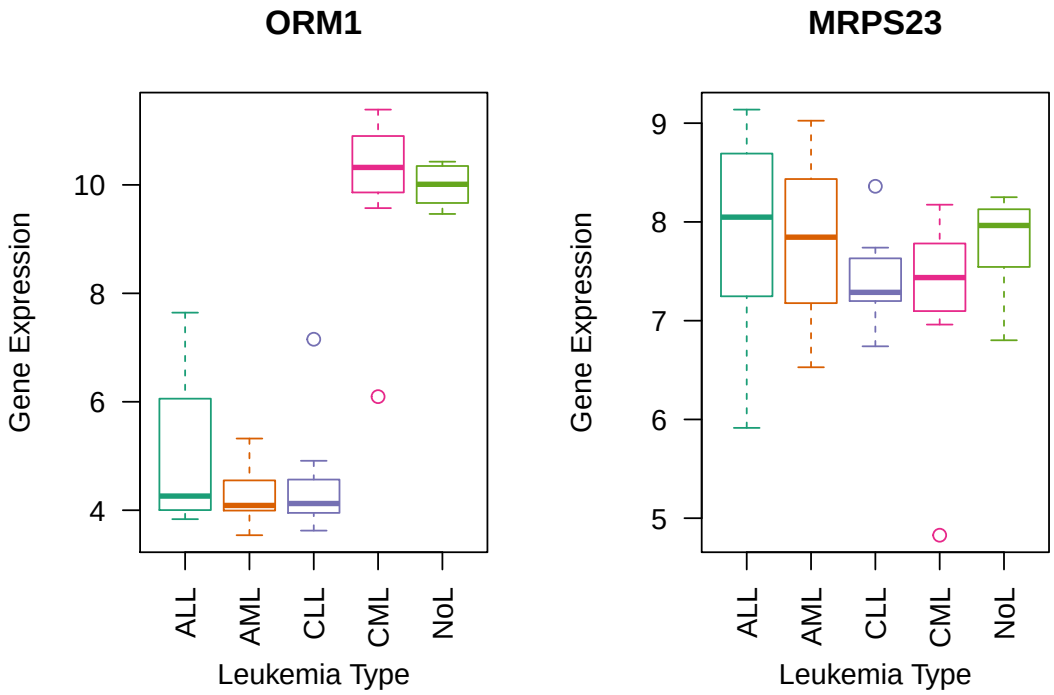

Figure 2: *Boxplots showing gene expression levels across five different leukemia types for ORM1 gene, on the left, and MRPS23 gene, on the right.*

In the first boxplot, it is clear that there seems to be a substantial difference in expression levels for the ORM1 gene, identified by Ensembl ID: ENSG00000229314, between `NoL` samples and samples from `AML`, `ALL` and `CLL` patients. This indicates that this gene may be differentially expressed with respect to a contrast between groups `NoL` and `AML`, for example. The second boxplot shows little change in the means across the various types of leukemia. It would thus be expected that the MRPS23 gene, with Ensembl ID: ENSG00000181610, would not be differentially expressed when two particular conditions of this group are compared.
Certainly, constructing boxplots and relying on judgments made by eye is an inadequate method in the assessment of which genes are differentially expressed. However, this visualisation technique can provide an insight into what may be expected in further analysis.

3.3 Application of Linear Models with `limma`

As mentioned previously in Section 1.2, the `limma` package utilizes a linear modelling approach in order to perform analyses on microarray experiments and have success in the identification of genes that are differentially expressed. For data which has been obtained from a single-channel oligonucleotide experiment

such as that provided in the `leukemiasEset` package, the linear modelling process is most similar to the common analysis of variance or multiple regression procedures. However, in this instance, a linear model is fitted to each individual gene. It is generally the case that `limma` is presented with a large number of genes, possibly tens of thousands, in which it must apply this linear modelling technique. For example, information relating to $20,172$ different genes is witnessed in this Leukemia dataset.

Consider the assumption that for a set of n samples, expression measurements for a large quantity of genes have been obtained and combined in a response vector, $y_g^T = (y_{g1}, \cdots, y_{gn})$. For single-channel data, y_g^T represents the log-intensities which have been yielded for a gene g across all n microarrays. These expression measurements are assumed to be of a suitable form for analysis, i.e. appropriate normalisation has taken place resulting in $y_{gi} \sim N(\mu_{gi}, \sigma_g^2)$. As has been previously noted, in the leukemia dataset at hand, a single expression value has been produced for each of the $20,172$ genes for each of the 60 arrays.

A linear model is fitted to each gene using the `limma` function `lmFit()`. The intention is that this linear model would completely model the systematic part of the available data. Thus, the linear model may be given in a general form as:

$$E(y_g) = X\alpha_g,$$

where X is the *design* matrix and α_g is the coefficient vector.

The construction of the design matrix which specifies the linear model is as follows. With log-intensity data of this type, design matrices can be created exactly as in classical linear model practice with univariate data. The overall experimental layout is essentially based on a small number of biological factors, such as the type of leukemia a patient has been diagnosed with. Therefore, it is these factors which are recognised as being the independent variables that influence the response variable, a measurement of gene expression, denoted y_{gi}. These predictors are of a categorical nature. The design matrix can be easily used to illustrate the manner in which the researcher intends to code this biological information. Each row of the matrix usually corresponds to an experimental array and each column is used as a descriptor of the origin of the samples which have been applied to the arrays.

An appropriate design matrix can be created in `limma` using `model.matrix()`. In relation to the example discussed throughout, the design matrix is formed with a total of 60 rows and 5 columns, a column for each type of leukemia presented in the dataset. The first six rows of this design matrix are of the form:

```
  NoL ALL AML CLL CML
1   0   1   0   0   0
2   0   1   0   0   0
3   0   1   0   0   0
4   0   1   0   0   0
5   0   1   0   0   0
6   0   1   0   0   0
```

The value 1 which is present in the `ALL` column of these rows indicate that the first six samples have been obtained from patients suffering with Acute Lymphoblastic Leukemia. The remaining rows of the design matrix are of a similar nature. Take for example a row representing an array to which cells from healthy bone marrow tissue have been applied, a 1 will be visible in the `NoL` column of this row.

It is known that there are several ways in which the analysis of a microarray experiment can be approached and questions of interest investigated with the assistance of `limma`. However, one of the simplest strategies is to construct a straightforward design matrix as has been done above and to then proceed by extracting specific contrasts of interest from the fit. This is the primary reason that specification of the model in the manner above does not include an intercept coefficient or a reference level. Therefore, all possible comparisons may now take place between these initial coefficients, identified by the vector α_g for gene g, in order to answer questions that may be proposed. In order to accomplish this, an additional matrix, the *contrast* matrix, must be established. This *contrast* matrix C permits the extraction of constants of biological interest β_g from the coefficient vector α_g as $\beta_g = C^T\alpha_g$.

For illustrative purposes, the intention will be to determine the genes which are differentially expressed between the two groups, `NoL` and `AML`. In order to investigate this appropriately, the extraction of the difference between the coefficients specifically related to these two groups as a contrast must occur. The suitable contrast matrix for this question of interest is assembled in `limma` using `makeContrasts()` and is found to be:

```
      Contrasts
Levels AML-NoL
   NoL      -1
   ALL       0
   AML       1
   CLL       0
   CML       0
```

The variance associated with the response vector for gene g, y_g may be given as:

$$var(y_g) = W_g \sigma_g^2,$$

with W_g, a known non-negative definite weight matrix. Therefore, the linear model for gene g is understood to have residual variance σ_g^2, which has sample value s_g^2 and degrees of freedom d_g. Indeed, estimation of the variability of each gene in the data set is the key focus of this first step of fitting a separate linear model to each gene. It is worth noting that this process allows for the accommodation of different variances for each gene. It is important that the systematic part must be modelled in the correct manner so that random variation, which plays an important role in hypothesis testing, can be distinguished.

The result of the application of linear models is the generation of coefficient estimators $\hat{\alpha}_g$ of α_g. It is discovered that in the leukemia dataset, for gene g, $\hat{\alpha}_g$ is merely found to be a vector where the values are equivalent to the mean log-intensity of the gene across samples for each leukemia type. This is due to the fact that a "group-means" parametrization has been chosen. In addition, the production of the estimators s_g^2 of σ_g^2 and the estimated covariance matrices $var(\hat{\alpha}_g) = V_g s_g^2$ where V_g is a positive definite matrix, independent of s_g^2, takes place. Following this, the contrast estimators $\hat{\beta}_g = C^T \hat{\alpha}_g$ together with the estimated covariance matrix $var(\hat{\beta}_g) = C^T V_g C s_g^2$ can be attained in using $\beta_g = C^T \alpha_g$. The `limma` function `contrasts.fit()` computes $\hat{\beta}_{gj}$ and associated standard errors.

When the desire is to obtain a number of genes that are considered to be differentially expressed between two experimental conditions and if the contrast matrix has been set up in the correct manner, then what is essentially of interest is the evaluation of whether the contrast values β_{gj} are equal to zero. Therefore, in the usual context of hypothesis testing, the null hypotheses $H_0 : \beta_{gj} = 0$ are to be tested.

In the current situation where a single contrast between two groups, `NoL` and `AML`, is of interest, β_{gj} can be denoted simply as β_g as it is a single value rather than a vector. The null hypothesis is established as $H_0 : \beta_g = 0$ along with the alternative hypothesis of $H_A : \beta_g \neq 0$. It is rather easy to see that testing this null hypothesis is analogous to testing if $\alpha_{g3} - \alpha_{g1} = 0$ where α_{g1} is considered to be the average expression of gene g in all `NoL` samples and α_{g3}, the average expression in all `AML` samples. Rejecting the null hypothesis in this instance provides evidence for differential expression between the two conditions being compared.

However, in order to test the hypothesis $H_0 : \beta_{gj} = 0$ which requires the computation of a test statistic, assumptions regarding the underlying distributions of the contrast estimators $\hat{\beta}_g$ must be made. These estimators $\hat{\beta}_g$ are assumed to be normally distributed with mean β_g and covariance matrix $C^T V_g C \sigma_g^2$. In addition to this, the residual variances s_g^2 are understood to approximately follow a scaled chi-square distribution. The unscaled standard deviations are determined as the square roots of the diagonal elements of $C^T V_g C$. Allowing v_{gj} to represent the jth diagonal element of $C^T V_g C$, i.e. $(C^T V_g C)_{jj}$, these distributional assumptions may be summarized by

$$\hat{\beta}_{gj} | \beta_{gj}, \sigma_g^2 \sim N(\beta_{gj}, v_{gj} \sigma_g^2)$$

and

$$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi^2_{d_g}$$

in which the residual degrees of freedom for the linear model for gene g is given by d_g.

At this point, under the assumptions which have been outlined, the ordinary t statistic for the jth contrast for gene g can be derived using the contrast estimators $\hat{\beta}_g$ and its variance estimators:

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{v_{gj}}}.$$

This t-statistic is known to follow an approximate t-distribution on d_g degrees of freedom.

Consider a particular gene g from our data, ORM1, which has corresponding Ensembl ID: ENSG00000229314. Upon the specification of both the design and contrast matrices which have been detailed and the subsequent application of `lmFit()` to the normalized data stored in the `ExpressionSet` object, it was found that this specific contrast estimator $\hat{\beta}_{ORM1}$ was computed to be -5.7265 with standard deviation 0.9994, together with an unscaled standard deviation of 0.4082. The t-statistic for ORM1 which has $d_g = 55$ degrees of freedom was determined as:

$$t_g = \frac{\hat{\beta}_g}{s_g\sqrt{v_g}} = \frac{-5.7265}{(0.9994)(0.4082)} \approx -14.0351$$

In this case, ignoring any multiple testing corrections, at a significance level of $\alpha = 0.05$, the null hypothesis is rejected if $|\ t_g| > 2.004045$.. Thus, it is clear to see that the obtained t-statistic provides a very strong indication of differential expression for this chosen gene between the two conditions.

As is to be expected, each gene tested generally possesses a different test statistic. However, this test statistic is very much dependent on the variance of the distribution. For each gene, variability of expression values can be substantially different which indicates that the probability of detection of differential expression varies from gene to gene. For example, low variability can result in the rejection of the null hypothesis even if the mean difference between conditions is of a relatively small value. Therefore, the computation of ordinary t-statistics for each gene, such as in the illustration above, is deemed as an unsuitable approach in correctly identifying differentially expressed genes. The empirical Bayes method employed by `limma` is a technique used to counteract this issue associated with potentially very different variabilities in gene expression for certain genes through *variance shrinkage*. The approach is based on the borrowing of information across the large number of available genes which ultimately aids in more correct inferences being made in relation to each gene. In fact, the empirical Bayes method is considered to be one of the most influential statistical methods that have been developed to date in the analysis of genomics data.

3.4 Empirical Bayes Method in `limma`

Moderated t-statistic:

The current empirical Bayes method used by `limma` was proposed by Smyth (2004). This approach was designed to cope with the burden of instability which commonly presented itself in the variance estimation of gene expression measurements attained from microarray experiments. The aim was to achieve this by developing an improved version of the test statistic in the context of hypothesis testing. An empirical Bayes procedure is used to borrow information across the large number of genes. The process basically combines gene-specific variability with global variability in order to create a new form of the ordinary t-statistic. This new t-statistic is known as a *moderated* t-statistic, in which the posterior residual standard deviation has replaced the original standard deviation.

The construction of this technique by Smyth (2004) may be summarized in the following steps:

i) The model is built in a linear modelling framework, such as that described above in Section 2.3, in which data has been made available from samples which naturally fall into a number of different experimental groups.

ii) The marginal distributions of the observed statistics assist in the estimation of the hyperparameters, denoted d_0 and s_0^2, of the hierarchical Bayesian model, from the data.
iii) Upon implementation of the hierarchical model, a *moderated* t-statistic is obtained for each gene and in addition, a posterior odds statistic for each gene can be reformulated in terms of this *moderated* t-statistic using the calculated posterior residual standard deviations.

The gene-wise residual sample variances are witnessed to "shrink" towards a common value. With a small number of arrays, as is most often the case, the emergence of more stable inferences is ensured from this shrinkage. It is less likely that t-statistics of a large magnitiude, which have arisen due to the under-estimation of sample variances, will prevail under this approach. The reason for this is that small sample variances tend to be offset heavily by the posterior variance s_g^2.

The *moderated* t-statistic, $\tilde{t}_{gj}$ has a comparable interpretation to the ordinary t-statistic and may be used for inference about the contrast of interest, β_{gj} in a similar fashion as before. Additionally, a corresponding p-value can be easily obtained from the moderated t-statistic. The moderated t-statistic follows a t-distribution, under the null hypothesis $H_0 : \beta_{gj} = 0$, but one in which the degrees of freedom has been increased from d_g to $d_g + d_0$. The augmented degrees of freedom of $\tilde{t}_{gj}$ are a reflection of the extra information which has been borrowed across the ensemble of genes. A detailed explanation of the construction of the *moderated* t-statistic is given in the following section.

Hierarchical Model:

The formation of the hierarchical model begins with the specification of prior distributions for the unknown parameters, coefficients β_{gj} and variances σ_g^2. These prior distributions are intended to decribe how the parameters may vary across genes. This model proposed by Smyth (2004) is based on the assumption that the estimators $\hat{\beta}_g$ and s_g^2 from different genes are independent. Even though this is an unrealistic assumption in the context of microarray experiments, it has been shown that the derived methodology makes qualitative sense when there is evidence of dependence among genes.

The unknown variances σ_g^2 are supplied with a prior distribution of the form:

$$\frac{1}{\sigma_g^2} \sim \frac{1}{d_0 s_0^2} \chi^2_{d_0}$$

in which s_0^2 is a prior estimator with d_0 degrees of freedom.

Consider a given contrast of interest, β_{gj}. This is assumed to be non-zero with a known probability, $P(\beta_{gj} \neq 0) = \rho_j$. Thus, ρ_j is regarded as the expected proportion of genes which are truly differentially expressed between the two treatment conditions which j represents. For this fraction of the gene population, then the coefficient β_{gj} is assigned the prior distribution:

$$\beta_{gj} | \sigma_g^2, \beta_{gj} \neq 0 \sim N(0, v_{0j}\sigma_g^2)$$

This equates to a prior observation of zero with unscaled variance v_{0j} and provides a description of the expected distribution of the log-expression change that occurs from experimental condition A to experimental condition B for genes which are considered to be differentially expressed.

Besides the introduction of a mixing proportion ρ_j, the prior distibution defined is recognised as a standard conjugate prior for the normal model of $\hat{\beta}_{gj}$. Under this model, the posterior mean of σ_g^{-2} given s_g^2 is computed to be $\tilde{s}_g^{-2}$ in which

$$\tilde{s}_g^{-2} = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}.$$

Thus, this posterior value of the variance is seen to be equivalent to an adjusted version of the observed variance, one which has been squeezed towards the prior value. The degree to which the observed variances

are shrunk is very much dependent on the relative sizes of the observed and prior degrees of freedom. As a consequence of these results, the *moderated* t-statistic is defined as

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}.$$

It is merely a substitution of variances which has transformed the classical t-statistic into the *moderated* version.

It is noted that this distributional result assumes d_0 and s_0^2 to be given values. In practice, the hyperparameters, d_0, s_0 and v_{0j}, which have been introduced can be estimated in a stable fashion from the data.

The hierarchical model is built in `limma` using the function `eBayes()`, which evaluates *moderated* t-statistics for each gene for each contrast. In the context of the example relating to the leukemia dataset, d_0 and s_0 have been estimated as 2.5554 and 0.0935, respectively. Returning to gene ORM1, its moderated t-statistic is computed to be:

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g \sqrt{v_g}} = \frac{-5.7265}{(0.9791)(0.4082)} = -14.3263.$$

Implementation of the empirical Bayes method has reduced the original standard deviation of 0.9994 to the value of the posterior residual standard deviation of 0.9791. This in turn has resulted in an increase in the test statistic from -14.0351 to -14.3263. Under the null hypothesis, $\beta_{gj} = 0$, this moderated t-statistic of -14.3263 has an associated p-value of $1.2353063 \times 10^{-20}$ with the augmented posterior degrees of freedom, 57.5554.

Posterior odds and the B-statistic:

In this Bayesian framework, given the unconditional distributions of both the *moderated* t-statistic, $\tilde{t}_{gj}$ and the residual sample variance s_g^2, the posterior odds that a gene g is differentially expressed can be easily computed. The odds that $\beta_{gj} \neq 0$ for the gene g is given by:

$$O_{gj} = \frac{\rho_j}{1-\rho_j} \left(\frac{v_{gj}}{v_{gj}+v_{0j}} \right)^{1/2} \left(\frac{\tilde{t}_{gj}^2 + d_0 + d_g}{\tilde{t}_{gj}^2 \frac{v_{gj}}{v_{gj}+v_{0j}} + d_0 + d_g} \right)^{(1+d_0+d_g)/2}$$

.

Following this, consider the B-statistic, $B_{gj} = \log O_{gj}$. This is defined as the log-odds that gene g is differentially expressed with respect to a given contrast, β_{gj}. Suppose a B-statistic of 1.25 is obtained for gene g. The interpretation of this result is that the odds that gene g is differentially expressed is $e^{1.25} \approx 3.49$, i.e. about three and a half to one. Equivalently, it may be stated that gene g is differentially expressed with probability $\frac{3.5}{1+3.5} \approx 0.78$. A calculated B-statistic of zero equates to a 50-50 chance of differential expression for the corresponding gene.

`limma` computes a B-statistic for each individual gene with the expectation that 1% of all genes are differentially expressed. Thus, in order to calculate this statistic, ρ_j is assigned the value of 0.01. Recall the gene ORM1 and the value obtained for its *moderated* t-statistic of -14.3263. Combining this with the estimation of 10.0658 for the prior unscaled variance v_{0j}, a value for the odds statistic O_g of this gene g may be computed as follows:

$$O_g = \frac{0.01}{1-0.01} \left(\frac{0.4082}{0.4082+10.0658} \right)^{1/2} \left(\frac{(-14.3263)^2 + 2.5554 + 55}{(-14.3263)^2 \left(\frac{0.4082}{0.4082+10.0658} \right) + 2.5554 + 55} \right)^{(1+2.5554+55)/2} \approx 5.036 \times 10^{15}.$$

Its corresponding B-statistic is given as $B_g = \log(5.036 \times 10^{15}) \approx 36.1554$. The fact that such a large number has been obtained strongly suggests that this gene is differentially expressed.

The main purpose of the B-statistic is to rank genes in order of evidence for differential expression. It is often the case that only a limited number of genes can be followed up for further study and thus, a gene

ranking procedure is most important. Notice that if the values d_g and v_{gj} are assumed to be constant across the set of genes, B_{gj} is a monotonically increasing function of the absolute value of the *moderated* t-statistic, $|\tilde{t}_{gj}|$. This would indicate an equivalence between the B-statistic and the *moderated* t-statistic in terms of ranking genes. In fact, it is often found that even when d_g and v_g do vary, which would be the case if quality weights were introduced or the dataset contained missing values, both statistics normally rank genes in the same order.

However, even though the B-statistic has the ability to be interpreted in a very intuitive manner, it is the moderated t-statistic that is often preferred. This preference is primarily based on the B-statistic's dependence on extra hyperparameters, including a prior estimate for the proportion of differentially expressed genes, ρ_j.

3.5 Discussion and Visualisation of Results

As stated, `limma` assesses for differential expression using the *moderated* t-statistic. Computation of moderated t-statistics for each gene enables the production of corresponding p-values. As expected, a method to correct for multiple testing is incorporated into the `limma` pipeline which adjusts these p-values. The default method for this adjustment is the BH method discussed in Section 2.3. This procedure controls the false discovery rate (FDR) at a level less than the threshold value of 0.05 by selecting all genes with adjusted p-values below 0.05 as differentially expressed.

The `limma` function `topTable()` provides a summary of the results of this process. The output contains a list of ordered genes accompanied by their estimated statistics, in relation to the specific contrast of interest. The genes are ranked based on their individual *moderated* t-statistics, in a decreasing order. The top six differentially expressed genes in which the contrast in question, `AML-NoL`, has been specified, together with their summary statistics, are shown in the table below:

```
           logFC  AveExpr         t      P.Value    adj.P.Val        B
OLFM4  -6.944266 6.547371 -14.74016 3.386891e-21 6.229650e-17 37.37184
PADI4  -4.128657 6.669315 -14.42716 8.993595e-21 6.229650e-17 36.45408
CHIT1  -4.642819 5.957630 -14.36226 1.102905e-20 6.229650e-17 36.26209
ORM1   -5.726453 6.749225 -14.32627 1.235306e-20 6.229650e-17 36.15537
CAMP   -6.818293 9.188380 -13.41648 2.291015e-19 9.242872e-16 33.39741
CHI3L1 -5.873373 7.642469 -13.03379 8.076630e-19 2.715363e-15 32.20243
```

Average expression values, $\log_2$-fold changes and B-statistics are displayed as well as moderated t-statistics, p-values and adjusted p-values. The `AveExpr` column provides the average $\log_2$ expression level for each gene across all samples obtained in the experiment. In the case of the leukemia dataset, as the expression values were previously log-transformed, these values merely equate to the mean of each row in the matrix of expression measurements, `exprs(leukemiasEset)`. The `logFC` column provides the contrast value. In general, this is representative of a $\log_2$-fold change between two experimental conditions. Here, the `logFC` quantity for a gene g is the mean expression value of gene g of the control condition, `NoL`, across all 12 samples subtracted from its mean expression value in all samples sourced from an `AML` patient.

Upon examination of the table outputted, it can be seen that the top 4 most differentially expressed genes have all been assigned the same adjusted p-value. It is of interest to investigate why this has occurred. Glancing at the explanation of the BH method, detailed in Section 2.3, may result in a deduction that for a gene in position i, after re-ordering due to p-value size has taken place, then the adjusted p-value for this gene is given by $p_i^* = \frac{n}{i} p_{(i)}$. It is recalled that the BH method, with a desired false discovery rate threshold of q, works by obtaining the largest i such that the i-th smallest p-value is found to be less than $\frac{i}{n}q$. Then, the adjusted p-values that are viewed here, which have been obtained using the `p.adjust()` function, are simply the smallest value of q for each test, by definitioin of `p.adjust()`. This may not necessarily be equal to $\frac{n}{i} p_{(i)}$ for test i. A geometric interpretation of this is given as follows. Consider plotting i on the x-axis and the corresponding p-value, $p_{(i)}$ on the y-axis. A plot which includes the 4 smallest p-values can be seen in

Figure 3. A red line is drawn through the largest point with slope $\frac{p_4^*}{m}$ in which $p_4^* = 6.229650 \times 10^{-17}$. This adjusted p-value is noted to be equivalent to $\frac{n}{i}p_{(i)} = \left(\frac{20172}{4}\right) 1.235306 \times 10^{-20}$ in which 1.235306×10^{-20} is the original p-value for this gene in position 4. However, the other three p-values are all found to lie above this line. This indicates that if these adjusted p-values were computed in the same way as above, they would in fact be greater than $p_4^* = 6.229650 \times 10^{-17}$. By definition of the BH method, this would be incorrect and thus, the adjusted p-values p_3^*, p_2^* and p_1^* must be assigned the smallest value of q which in this case is equivalent to $p_4^* = 6.229650 \times 10^{-17}$. Basically, it may be deduced that witnessing many adjusted p-values of the same value merely highlights the manner in which the `p.adjust()` function based on the BH method ensures that ranking genes according to their new adjusted p-values will not alter the order in which genes have been ranked with respect to their original p-value.

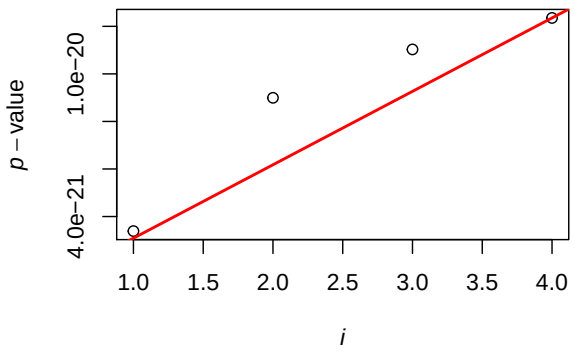

Figure 3: *Position when all p-values have been ordered from smallest to largest, on the x-axis, is plotted against p-value, on the y-axis, for the four smallest p-values.*

Additionally, the outcome of each hypothesis test may be summarized simply using `decideTests()`.

```
       AML-NoL
Down      1518
NotSig   16881
Up        1773
```

From this, it is easily deduced that upon implementation of the `limma` analysis procedure, $3,291$ out of the total $20,172$ genes show evidence for differential expression. The expression levels of $1,518$ genes have been found to be downregulated, while $1,773$ are upregulated. The expression of a gene is defined as being upregulated, with respect to two conditions A and B, if there is evidence of a significant increase in the expression measurement from A to B.

It was also noted that if the Bonferroni method had been used in order to correct for multiple testing in place of the BH method, only a mere 463 genes would have been regarded as differentially expressed. This is a clear demonstration of the issue of over-conservativeness that is often associated with the Bonferroni method. It would thus be a major concern here that many truly differentially expressed genes have not been identified. In a similar fashion, the BY method was implemented, resulting in $1,283$ significant genes. Again, it is indicated by this figure that the BY method is a more stringent approach to overcoming the multiplicity problem than the BH procedure.

In order to visualise these results, suitable plots have been produced, as can be seen in Figure 4. A volcano plot can be witnessed on the left. For each gene, the volcano plot construction involves plotting its log2 fold change against a measure of statistical significance, such as -log(p-value) or the B-statistic. In the volcano

plot below, the posterior log-odds of differential expression, i.e. B-statistic, has been chosen. Downregulated genes are located towards the left of the plot while upregulated genes lie to the right. The two most significant genes from above, OLFM4 and PADI4, have been highlighted on the plot with the last 6 digits of their respective Ensembl Gene IDs. Their position in the volcano plot in the top left corner emphasizes the size of their B-statistic values accompanied by large negative change in their expression measurements from `NoL` to `AML` samples. In fact, these two genes seem to be joined by many more differentially expressed genes which possess similar traits.

The plot on the left is an MA plot which was originally created for visualising genomic data obtained from microarray experiments. The MA plot illustrates the relationship between the change in expression (M) from one condition of interest to the other and the average expression level (A) across all samples for each of the $20,172$ genes. Genes deemed significant as a result of the previously detailed analysis are clearly pinpointed with upregulated genes coloured in red and downregulated genes coloured in green.

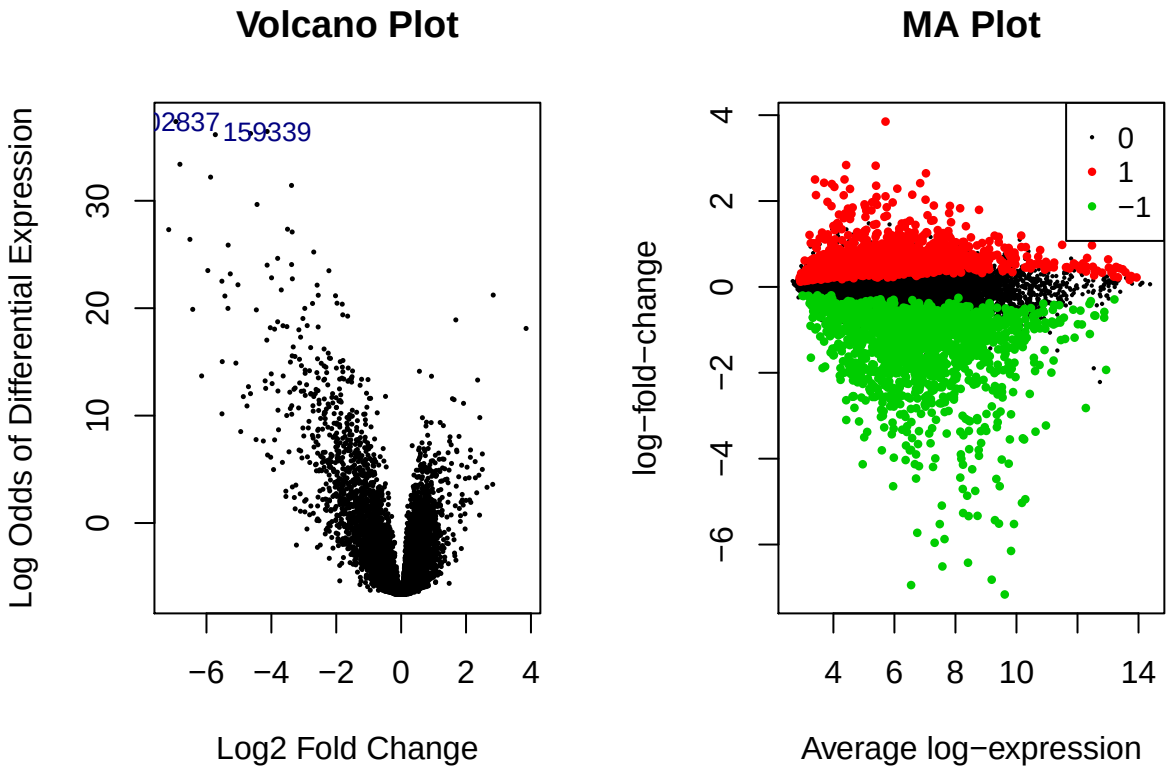

Figure 4: *Plots of the results of the analysis performed on the leukemia dataset with a volcano plot, on the left, and an MA plot, on the right.*

3.6 Comparing Results of Different Approaches

In addition to this analysis, it was of interest to investigate the effect that merely inserting the 12 `AML` samples together with the 12 `NoL` samples, in order to determine differentially expressed genes between these two conditions, into the `limma` pipeline would have on the analysis. Above, all 60 samples were included and the wish to specifically examine the `AML-NoL` contrast was subsequently stated. However, this other method would only allow these 24 samples to be introduced to the `limma` process. Therefore, the differences between the results obtained when the `limma` process has knowledge of information regarding the other 36 samples

compared with those acquired when these other leukemia samples are discarded prior to analysis are to be examined. The analysis in which all samples are included is to be referred to as the *Full Dataset* approach while the second method which only uses the samples related to the two conditions of interest, `AML` and `NoL` is to be known as the *Subset* approach.

The `limma` linear modelling analysis procedure was implemented in the same manner as above on this dataset with a reduced number of samples. It is important to note that it is the exact same question that is being asked in both cases. An illustration of the comparison of the two sets of yielded results is provided in Figure 5. It can be witnessed from this diagram that as a consequence of the *Subset* approach, $2,653$ out of the total $20,170$ genes were determined as being statistically significant. This number is much less than $3,291$, the number of genes found to be differentially expressed in the previous analysis which incorporates the entire dataset with 60 samples. It is of most interest to examine the genes which have been regarded as differentially expressed as a result of just one approach. The existence of 260 genes which have been deemed significant in the *Subset* method only is clearly visible from the diagram. It can also be seen that there are a total of 898 genes which solely show evidence for differential expression in the *Full Dataset* approach.

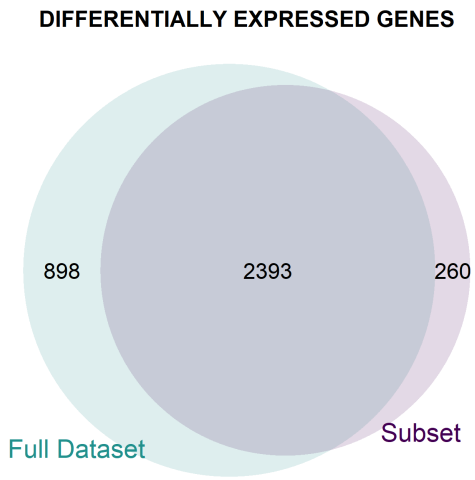

Figure 5: *Venn diagram depicting differences in the sets of differentially expressed genes obtained as a result of implementing the two outlined approaches.*

The reasons why certain genes may be present in the list of significant genes as a result of one method and not the other are investigated. Firstly, a gene which was discovered to be differentially expressed when all 60 samples were included in the analysis but failed to show evidence for differential expression in the *Subset* approach is considered.
In the *Full Dataset* approach, the moderated t-statistic of this gene, namely the ZNF747 gene, with Ensembl ID: ENSG00000169955, was calculated as:

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g\sqrt{v_g}} = \frac{0.4105}{(0.3548)(0.4082)} = 2.8335$$

while in the second method, it was computed to be:

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g\sqrt{v_g}} = \frac{0.4105}{(0.3778)(0.4082)} = 2.6609.$$

These two numbers are reasonably close in value with the only difference stemming from an increase in $\tilde{s_g}$ in the second method. With similar t-statistics, it may be questioned why the results are different. However, it must be remembered that the degrees of freedom are completely different when p-values are being calculated. With 57.55538 degrees of freedom, the first *moderated* t-statistic has corresponding p-value of 0.0063379 while

the second *moderated* t-statistic, with the much smaller degrees of freedom of 24.289799 has a p-value of 0.0136. After adjustment for multiple testing, these p-values are increased to 0.0426 and 0.0778 respectively. It may be concluded that the reason for this gene's inclusion in the first set of differentially expressed genes is predominantly as a result of the larger degrees of freedom which has generated a very small p-value. This stems from the form that the t_{n-p-1}-distribution takes when analyzing samples which have a smaller value for $n-p-1$ representing the added uncertainty associated with this analysis involving small sample sizes. The adjusted p-values are much closer in value but as the chosen FDR is $\delta = 0.05$, $H_0 : \beta_g = 0$ is rejected in the second case.

A similar investigation may be performed for the S100A13 gene, with Ensembl ID: ENSG00000189171, which was identified as significant under the second approach but not the first. In this instance, the moderated t-statistic of this gene was calculated as:

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g\sqrt{v_g}} = \frac{0.6432}{(0.645)(0.4082)} = 2.4426$$

while in the second, it was computed to be:

$$\tilde{t}_g = \frac{\hat{\beta}_g}{\tilde{s}_g\sqrt{v_g}} = \frac{0.6432}{(0.2958)(0.4082)} = 5.3253.$$

Following this, the corresponding p-values are discovered to be 0.0177 and 0.000018, respectively, with the degrees of freedom the same as above. The adjusted p-values were subsequently computed to be 0.0835 and 0.00092. After scrutinizing these produced values, it is thus inferred that the reason that this gene is included in the second set of differentially expressed genes is based on the large decrease in its variability estimate when the `CLL`, `CML` and `ALL` samples have been excluded.

From these observations, it is deduced that there exist both advantages and disadvantages to the two approaches outlined above. The main factors which have influenced dissimilarities between results are sample size and changes in variability estimates. The larger the sample size, the greater the degrees of freedom associated with each hypothesis test. As noted, for similar estimates of both variability and fold changes, a much smaller p-value is obtained for the case in which the degrees of freedom are increased. Therefore, it is to be expected that using information from all 60 samples would result in a larger set of significant genes in which a false discovery rate of $\delta = 0.05$ is still maintained. This may be seen as a positive characteristic of the *Full Dataset* approach as a larger set of differentially expressed genes can often prove useful in downstream analyses such as gene set testing. However, performing the analysis with the entire dataset permits the introduction of additional sources of variability. This can greatly increase estimates of variability, as demonstrated above. It is understood from this brief inspection that when performing differential expression analysis with respect to a specific contrast of interest, it is crucial that one comprehends both the pros and cons of different approaches that may be taken as well as the way in which their method of choice executes the required tasks.

4 Analysis of RNA-Seq Data

4.1 Statistical Methodologies and RNA-Seq Data

Over the last two decades, many different statistical methods have been introduced for the specific purpose of application in differential gene expression analysis of RNA-Seq data. These methods have been designed to assist in overcoming the challenges that integer count data present in this setting. Differential gene expression tools that have been proven to perform very well are `edgeR` (Robinson, McCarthy, and Smyth (2009)), `DESeq/DESeq2` (Anders and Huber (2010), Love, Huber, and Anders (2014)) and `limma-voom` (Law et al. (2014)), which can all be used in `R`. These various procedures all use a wide range of statistical techniques in order to ensure detection of many truly differentially expressed genes even when data from a

limited set of samples is available. The strategies applied in each case have notable similarities. Regression-based models are used to estimate changes in gene expression levels for each gene while a hypothesis test is carried out to determine if this change in expression between two conditions of interest should be considered significant. However, the tools mentioned can be clearly separated into two distinct groups:

1) modelling of the count data directly takes place in order to make inferences: `edgeR` and `DESeq/DESeq2`
2) normal-based methodologies are implemented upon transformation of the counts: `limma-voom`

Due to the fact that there has been much focus on the use of normal linear models for microarray data with `limma`, it is of interest to investigate how a similar approach may be taken with data generated by RNA-Seq technologies. Therefore, the latter of these two methods will be explored in much detail. This will include examining how unequal variabilities across different count sizes is managed. That said, the use of count-based modelling techniques along with their limitations will be first briefly discussed.

Use of count distributions and their limitations:

RNA-Seq data takes the form of integer counts which represent the number of sequence reads mapped to a specific gene. Because of this, modelling this data appropriately using count distributions has been an area of much research. The Poisson distribution is a popular distribution used to model count variables which possesses the convenient characteristic of its mean and variance being equivalent. However, it would be most naive to assume that this distribution is suitable for modelling purposes here. Biological replicates have been shown to introduce greater variability, resulting in a distribution which takes a similar shape to that of an overdispersed Poisson rather than the basic Poisson distribution. This overdispersion can be expressed using a negative binomial distribution, a distribution considered to be similar to the Poisson distribution but in which the variance is permitted to have a greater magnitude than the mean. Both `edgeR` and `DESeq/DESeq2` procedures incorporate the fitting of these negative binomial models to the observed count data so that variability of both technical and biological origin can be accounted for.

Using a negative binomial model necessitates the estimation of two parameters, both the mean and the variance, from the dataset. The availability of merely a few replicates for most RNA-Seq experiments makes accurate estimation of these parameters difficult. Both `DESeq/DESeq2` and `edgeR` have aimed to improve the reliability of its inferences through information borrowing across genes. With `DESeq`, it is presumed that all genes share the same parameter of dispersion for which estimation can occur with the use of a conditional maximum likelihood method involving the entire dataset. Clearly, this can often be quite a restrictive approach. An empirical Bayes method based on weighted conditional likelihood was introduced by Robinson and Smyth (2007) which was designed to allow gene-wise dispersion estimates. These dispersion estimates for each individual gene are shrunk towards the common dispersion, moderating the degree of overdispersion across genes.

It is known that the use of count distributions, such as the negative binomial, can have many limitations in comparison to normal distributions, as a result of the mathematical intractability of this type of probability distribution. Hypothesis testing with these distributions is often based on tests such as the Wald test and likelihood ratio test which are only asymptotically valid. A non-asymptotic exact test was introduced by Robinson and Smyth (2007). However, this test assumes that the negative binomial dispersion parameters are equivalent for the two groups of interest. In fact, many of the negative binomial methods are inclined to regard the estimated dispersions as known parameters, failing to take into account uncertainty in estimation.

It was noted by Soneson and Delorenzi (2013) that the presence of an outlier in the dataset could result in a considerable increase in the type I error rate of methods which are based on the use of negative binomial modelling. Additionally, for samples sizes which were large in number, `DESeq2` was seen to act most conservatively. On the other hand, `edgeR` was regularly overly liberal as it deemed many genes as significant, including those that were truly differentially expressed as well as a great amount which weren't differentially expressed in reality. It was concluded by Soneson and Delorenzi (2013) that these various observations were to be explained by the different ways in which dispersion parameters were estimated.

Prior to the introduction of RNA-Seq technologies, a large body of statistical tools had been devised in order to perform detailed analyses on microarray data. However, many of these rely on the data being normally distributed such as incorporating quality weighting into the analyses, conduction of gene set testing and use of random effects. Unfortunately, these methodologies are inaccessible to RNA-Seq data when it is modelled using count distributions. The prospect of the extension of RNA-Seq analyses to include the techniques mentioned above motivated the establishment of the `limma-voom` procedure by Law et al. (2014).

4.2 A Local RNA-Seq Mouse Dataset

In order to investigate the steps which must be taken in the analysis of RNA-Seq data, a local RNA-Seq dataset has been obtained from research conducted to support studies in breast cancer in conjunction with the National Breast Cancer Research Institute, NBCRI. The samples used in this experiment have been extracted from murine cells. The dataset has been found to be quite limited, comprising of a mere total of 4 samples and $1,908$ genes. It is also unfortunate that many of these genes have very low counts across all samples as will be seen. Nevertheless, the dataset will prove useful for illustrative purposes. Two replicates of each condition are provided. Count data has originated from secreted extracellular vesicles and from cells. Thus, the question of interest which may be proposed here is to simply identify which genes are differentially expressed between these two groups, given the data at hand.

The integer count data which has been generated for the first 4 genes is displayed directly below. Genes have been annotated in the form `GeneIDX` while the four samples may be identified as `RNA4T1A`, `RNA4T1B`, `EV4T1A` and `EV4T1B`. It should be clear that the first two samples are those derived from the cell while the second two correspond to the secreted extracellular vesicles. The sample information available for this dataset merely specified the samples' origin.

```
        RNA4T1A RNA4T1B EV4T1A EV4T1B
GeneID1      52      53      0      1
GeneID2       8       8      1      0
GeneID3   33560   28196   3215   3823
GeneID4      53      51      2      1
```

Appropriate filtration and normalisation procedures are to be performed first on the data using `edgeR`. These are necessary so that `limma-voom` can be implemented and normal linear models can be applied to this RNA-Seq dataset.

4.3 Filtration and Investigation of Sample Distributions

In general, the first step in the analysis of RNA-Seq data is the removal of genes which have very low counts across all samples. As a result of small count values, these genes are incapable of showing much evidence for differential expression and thus, are likely to negatively affect certain statistical approximations that are required to be made in the analysis process. Furthermore, they add to the multiple testing burden by decreasing the power to extract genes which are truly differentially expressed between conditions. There are many ways in which filtration of lowly expressed genes may occur. With RNA-Seq data, it is often the case that significant differences can exist in the depth to which samples have been sequenced. This gives reason as to why merely comparing counts in order to remove these genes could lead to great errors. Therefore, a common approach is to convert all counts to counts-per-million values and subsequently, define a specific threshold.

Suppose an experiment has been performed in which a total of n RNA samples have been obtained. The samples are sequenced, generating RNA-Seq libraries for each individual sample. These libraries contain recordings of the amount of reads which have been mapped to each gene for their corresponding sample. Thus, the total number of mapped reads for a particular sample is defined as its library size. This data may

be represented by a matrix of counts, denoted by r_{gi}, for samples $i = 1, \ldots, n$ and genes $g = 1, \ldots, G$, and the library size for sample i is given as:

$$R_i = \sum_{g=1}^{G} r_{gi}.$$

Variance in library sizes indicates that the read count obtained for a given gene is not only dependent on its expression level but is also proportional to the sequencing depth of the library. Counts-per-million (*cpm*) are obtained by dividing each individual read count by its associated library size, in millions. Equivalently,

$$\text{cpm}_{gi} = \frac{r_{gi}}{R_i} \times 10^6.$$

This provides a basic measure which can be compared across all samples, even if library sizes are drastically different from one another. A basic interpretation of the *cpm* value of a particular gene, cpm_{gi}, is that it exhibits the number of mapped reads that would be obtained for that gene if the library size was reduced to 1 million.

There is an enormous difference between library sizes for the two groups of this dataset. The first two samples, namely `RNA4T1A` and `RNA4T1B` have a total number of mapped reads of $951,042$ and $932,341$, respectively. In great contrast, the library sizes of the other two samples are greatly reduced with a count sum of $72,604$ for `EV4T1A` and $84,002$ for `EV4T1B`. The barplot below, in Figure 6, displays these values.

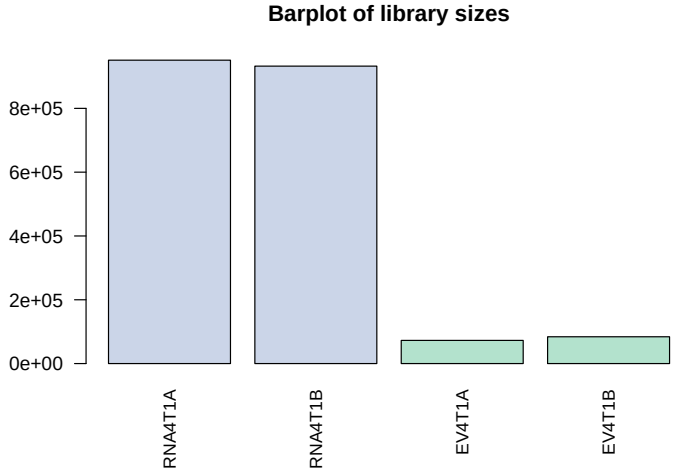

Figure 6: *Barplot of the library sizes of each of the four samples.*

`edgeR`'s `cpm()` function is implemented to generate counts-per-million values. Take for instance, the count related to the first gene, `GENEID1` in the first sample, `RNA4T1A` of 52. Its corresponding counts-per-million was obtained to be:

$$\text{cpm}_{11} = \frac{r_{11}}{R_1} \times 10^6 = \frac{52}{951042} \times 10^6 \approx 58.8828.$$

Following this conversion for each observation, a *cpm* threshold must be chosen which describes each gene to be either expressed or unexpressed. In general, a good threshold is considered to be one which is equivalent to a read count of about 10 to 15. In this RNA-Seq dataset at hand, it was seen that library sizes cover a vast range from approximately $70,000$ to $1,000,000$. It was decided that an appropriate means of filtration would be to first compute the median library size. This was estimated to be $\approx 508,200$, allowing a rough *cpm* threshold of $\frac{12.5}{508200} \times 10^6 \approx 25$ to be defined. As there are 2 replicates for each group, the action of retaining genes which exceed this threshold in at least 2 samples is favoured. A similar approach to filtration of lowly expressed genes is taken by the `filterByExpr()` function found in the `edgeR` package.

This filtration process resulted in the removal of a total of 834 genes from the dataset, leaving a remainder of 264 which will be kept for further analysis. This set of 264 genes are all considered to be *expressed* in at least

one group. In general, it would be expected that a much larger proportion of genes would be kept. However, retaining more genes which have low count figures from our dataset by reducing the threshold could have a severely negative impact in downstream analyses such as modelling the mean-variance relationship.

Moreover, the voom-plot displayed on the left in Figure 9 provides a visual check that correct filtration has taken place. If filtering of lowly-expressed genes is insufficient, a drop in variance levels can be observed at the lower end of the expression scale due to very small counts. If this is observed, the filtering step should be returned to and the expression threshold applied to the dataset increased. The fact that there is no evidence of a decrease in variance levels to the left of the voom-plot due to very low counts supports the threshold that has been chosen here for filtration purposes.

The effect of filtration can be visually represented through the production of a density plot. Currently, data is in the form of counts-per-million, which is not normally distributed. It is difficult to conduct a proper examination of the distribution of the data in this form, and thus, conversion to the log-scale is necessary. The density plot, visible in Figure 7 includes merely the data of the retained genes for each of the four samples.

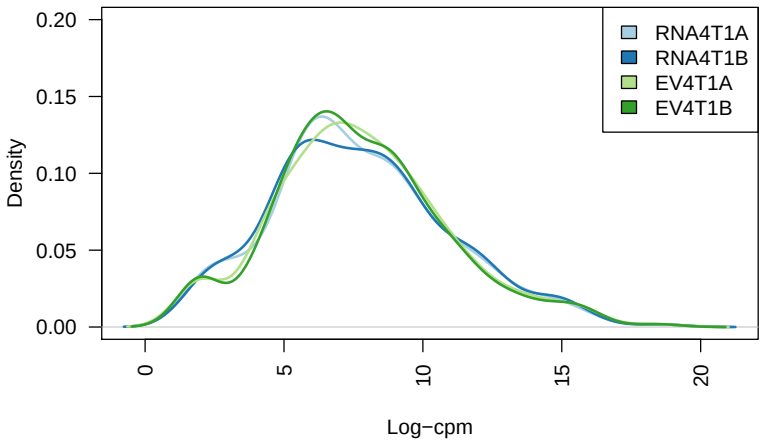

Figure 7: *Density plot of count-per-million values on the log-scale for each sample after filtration.*

A plot which displays the expression distribution for each sample, such as the one produced, is helpful in finding samples which may be dissimilar to others. The presence of a sample lying more to the extreme left or right of the plot would strongly emphasize that further investigation of this particular sample should take place before proceeding. With respect to this dataset, the distributions of the counts-per-million values on the log-scale are viewed to be quite similar throughout all samples. They are by no means identical in nature but great differences are not visible.

It is important to note that the data is still considered to be *unnormalised*. Even though it is not the case here, large differences between median values for different samples is often witnessed. It would therefore be expected that if a differential expression analysis was performed on two samples for which this was an issue, almost all genes would be down-regulated in one sample while up-regulated in the other. This problem is overcome by effectively increasing or decreasing the expression of all genes in each sample to a level in which proper comparisons can take place using the normalisation procedures detailed in the next section.

4.4 Normalisation of Read Counts

One of the most fundamental steps in attempting to gain information from large quantities of RNA-seq data regarding differential expression, is the processing procedures employed to make the data suitable for

implementation of statistical techniques. An example of one such method is the filtration of lowly expressed genes detailed previously. However, similar to microarray data, another essential task is that of normalisation. This permits precise parameter estimation through ensuring that all samples of the experiment possess similar distributional patterns in their respective expression values. Just as with microarray data, factors besides those of biological origin, such as technical errors arising from sample preparation and the sequencing process have the potential to cause variation between samples. Furthermore, the number of mapped reads for a particular gene and sample, r_{gi}, tends to be influenced by the composition of the population of RNA from which the sample originates. For instance, consider an individual condition which has a large set of highly expressed genes that are distinct to that condition. As a consequence of this, it is likely that there will only be a small number of reads mapped to the other genes. Thus, if this is not taken into account, the existence of bias will be clear in the analysis results in which there may be a large number of upregulated genes when this condition is compared to another.

The step of normalisation is employed to eliminate systematic technical effects, minimising the possibility for skewed analyses. In Section 3.1, the RAM normalisation procedure, which incorporates quantile normalisation to counteract the influence of technical errors on data obtained from microarrays, is discussed. However, there exist many obvious differences between the manner in which RNA-Seq data is generated and the production of microarray data. For this reason, the normalisation methods used when dealing with microarrays are not directly transferable.

A normalisation procedure created specifically for RNA-Seq data, namely the trimmed mean of M-values normalisation method (TMM), was presented by Robinson and Oshlack (2010). This empirical method was designed with the aim of eradicating composition biases across samples. Suitable scaling factors are evaluated with the assistance of the raw count data. These normalisation factors are then available for incorporation into the statistical model which has been chosen for the identification of differentially expressed genes. A description of the TMM technique is provided as follows.

It was suggested by Robinson and Oshlack (2010) that robust estimation of ratios of RNA production between samples could be obtained by using the means of log-expression ratios which are both weighted and trimmed. A mean which is considered to be *trimmed* is equivalent to the average value obtained following the elimination of a specified upper and lower percentage of the data. The log-fold change in expression level for gene g in sample i, relative to a particular reference sample i', is defined as:

$$M_{gi}^{i'} = \log_2 \left(\frac{r_{gi}/R_i}{r_{gi'}/R_{i'}} \right)$$

with its absolute expression level given by:

$$A_g = \frac{1}{2} \log_2 (r_{gi}/R_i \cdot r_{gi'}/R_{i'})$$

for $r_{gi}, r_{gi'} \neq 0$.

In order to determine appropriate scaling factors, the TMM method first trims both the log-fold changes, M values, and the absolute expression levels, A values. The default trimming percentages of 30% for M values and 5% for A values are generally used. After the data has been trimmed, a weighted mean of the M values is acquired. Weighting occurs in accordance with inverse variances, with the aim of accounting for the greater variance that is shown in log-fold changes of genes with lower counts. This weighted mean is therefore intended to robustly represent the M values which were originally observed. Then, the normalisation factor for sample i, in which sample i' is used as a reference, is explicitly computed to be:

$$\log_2 \left(TMM_i^{(i')} \right) = \frac{\sum_{g \in G^*} w_{gi}^{i'} M_{gi}^{i'}}{\sum_{g \in G^*} w_{gi}^{i'}}$$

where

$$M_{gi}^{i'} = \frac{\log_2 (r_{gi}/R_i)}{\log_2 (r_{gi'}/R_{i'})} \text{ and } w_{gi}^{i'} = \frac{R_i - r_{gi}}{R_i r_{gi}} + \frac{R_{i'} - r_{gi'}}{R_{i'} r_{gi'}}$$

and $r_{gi}, r_{gi'} > 0$.

Instances in which $r_{gi} = 0$ or $r_{gi'} = 0$ are discarded prior to the above calculation of the scaling factor. G^* is used to symbolize the collection of genes remaining after the trimming process. For the case in which data associated with several samples has been obtained, one sample is designated as the reference sample. Scaling factors are subsequently calculated for every other sample with respect to this chosen sample. It is clear from the formula given above that the normalisation factor of the sample which is being used as a reference is equivalent to 1.

The `calcNormFactors()` functon in `edgeR` executes the TMM normalisation procedure, resulting in the generation of scaling factors. For symmetrical purposes, the factors are altered so that they multiply to unity across all samples. Modification of observed library sizes occurs through multiplication by their corresponding scaling factor in order to obtain effective library sizes. It is noted that unlike many other normalisation strategies, TMM does not make any adjustments to the raw data. Instead, the outcome of its implementation is that it is these effective library sizes that will be used in downstream analyses in place of the original library sizes. As stated, TMM normalisation scales relative to one specific sample, known as the reference sample. This sample is determined by the `calcNormFactors()` function as the sample which possesses the library whose upper quartile is nearest to the mean upper quartile.

The effect of TMM-normalisation on this local dataset is considerably mild as anticipated from inspection of the density plot in Figure 7. This is evident from the magnitude of the scaling factors, which are all relatively close to 1. The scaling factors were determined by the `calcNormFactors()` function to be:

RNA4T1A	RNA4T1B	EV4T1A	EV4T1B
0.9520907	0.9701609	1.0833825	0.9993004

A normalisation factor less than 1 indicates that the counts-per-million values will be scaled upwards in that sample. In the case of `RNA4T1A`, the scaling factor is noted to be approximately 0.95. This transforms the current library size of $949,118$ into an effective library size of $949118 \times 0.95 \approx 903646$. Conversely, a factor greater than 1 leads to downscaling the observed counts. The scaling factor computed for `EV4T1A` of roughly 1.08 increases its library size from $72,452$ to $72452 \times 1.08 \approx 78248$.

4.5 Linear Modelling and Count Data: `limma-voom`

After these primary actions have been taken, attention is turned to the identification of which statistical methodologies can be applied to the dataset in order to successfully detect genes that are truly differentially expressed. As discussed previously, there has been much debate about what methods are most appropriate when analysing count data from RNA-Seq experiments. The possibility for the transformation of count data into normally distributed data with greater tractability by means of *precision weights* was introduced by Law et al. (2014). This approach was created in order to gain access to the wide variety of techniques which have been established for microarray data in the `limma` software package, such as quality weighting and gene set testing. The method constructed by Law et al. (2014) was named *voom* and is implemented in `R` using the `voom()` function from `limma`.

The various stages of the method may be summarized as follows:

1. Firstly, each individual observed read count is transformed to a log-counts-per-million (log-*cpm*) value, using the effective library sizes obtained through TMM normalisation.
2. Next, a linear model is fitted to these log-*cpm* values, in which the specified experimental design is taken into consideration.
3. Application of this linear model supplies residual standard deviations for each gene. A mean-variance trend is then fitted to these standard deviations as a function of average log-count across all genes and samples.
4. This obtained mean-variance relationship is subsequently used to estimate the standard deviation of each log-*cpm* value as a function of its associated fitted count value. For each individual observation, the squared inverse of the estimated standard deviation becomes its own unique precision weight.

5. The log-*cpm* values together with their associated precision weights are ready to be analysed for evidence of differential gene expression with the `limma` package and its functions.

4.5.1 Transformation: log-*cpm*s

Production of the boxplots above supports the requirement of transformation of the raw counts to log-counts-per-million following normalisation in order to make this dataset available for `limma`'s linear modelling analysis procedure, as detailed in Section 3. The distribution of the log-*cpm*s for each sample shown in Figure 7 indicate the possibility for these values to be treated in a similar manner to the log-intensities generated by microarray experiments. For a given gene g, the difference in its log-*cpm* value between two conditions of interest can be understood to be equivalent to a log-fold change in expression level. However, a crucial difference between these two forms of measurements is that it simply cannot be assumed that log-*cpm*s have constant variances. Many statistical analyses operate by assuming that all variables of the dataset possess the same variance. If this assumption that the data is homoscedastic in nature is violated, then fitting linear models directly to the data and carrying out statistical tests can lead to greatly inaccurate results. Thus, it would be incorrect to believe that the log-*cpm* values can be directly entered into the `limma` pipeline as the log-intensities were.

First, a formal definition of the manner in which log-*cpm* quantities are obtained from the raw count data is provided as follows. Using the notation previously introduced, the corresponding log-*cpm* for a particular count of mapped reads to gene g in sample i may take the form:

$$y_{gi} = \log_2\left(\frac{r_{gi}+0.5}{R_i+1}\times 10^6\right).$$

Notice that y_{gi} is not simply equal to $\log_2(\text{cpm}_{gi})$, with $\text{cpm}_{gi} = \frac{r_{gi}}{R_i}$. A small offset of 0.5 is added to the count number, r_{gi} in order to ensure that the log of zero will not be taken. This inclusion of an offset guarantees that no log-*cpm* values will be lost and results in a reduction of variability for those genes which still have relatively low counts. Furthermore, the library size R_i endures an offset of 1 which is designed to ensure that the fraction, $\frac{r_{gi}+0.5}{R_i+1}$ is strictly between 0 and 1. With `limma`'s `voom()` function, original count values are converted to log-*cpm*s using this exact format but in which R_i is the *effective* library size.

Illustration of this calculation is provided with the aid of this local dataset. Take for instance the first gene, denoted by `GeneID1`. The number of mapped reads to this gene in the first sample, namely `RNA4T1A`, was observed to be $r_{11} = 52$. After normalisation using `calcNormfactors` function was carried out, the effective library size was noted to be $R_1 = 903,646$. Computation of this particular log-*cpm* value may now be performed as follows:

$$y_{11} = \log_2\left(\frac{r_{11}+0.5}{R_1+1}\times 10^6\right) = \log_2\left(\frac{52+0.5}{903646+1}\times 10^6\right) \approx 5.8604.$$

4.5.2 Variance of log-*cpm*s

As mentioned, it cannot be assumed that the log-*cpm* values have non-constant variances. In fact, a common characteristic of probability distributions for count data is that of heteroskedasticity. Take for example the Poisson distribution, a distribution frequently used to describe count data, in which the variance increases with the mean. A visual demonstration of the heteroscedasticity concept is obtained by plotting the mean against the standard deviation across the four samples for the log-*cpm* values of each gene, as can be seen in Figure 8. Existence of variability is anticipated throughout the plot. However, it is most apparent from the left-hand side of the plot that the variance seems to be much greater for lower log-*cpm* values, specifically those less than 5. This is a clear indication of the dependence of the variance on the mean. It is of interest to investigate further what exact form the mean-variance trend of read counts on the log scale should take.

An approximate quadratic mean-variance relationship for count data obtained from RNA-Seq experiments was postulated by McCarthy, Chen, and Smyth (2012). It was proposed that considering the variance

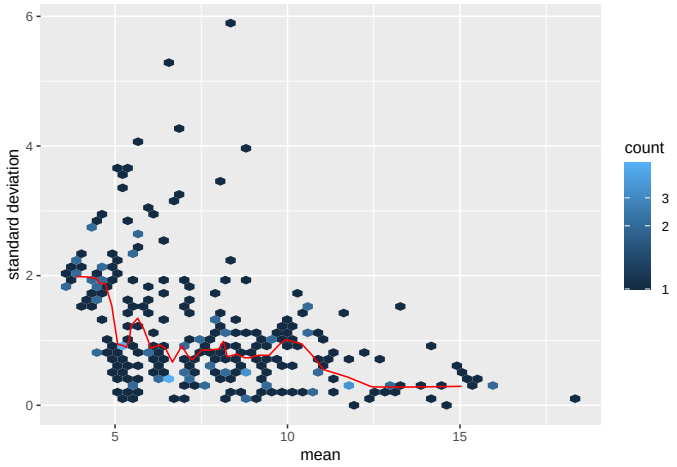

Figure 8: *Plot of mean, on the x-axis, against standard deviation, on the y-axis, for each gene showing the mean-variance trend of the data.*

formula of a mixture distribution, this relationship should take the form of

$$\text{var}(r_{gi}) = \mu_{gi} + \phi_g \mu_{gi}^2$$

in which $\mu_{gi} = E(r_{gi})$ where r_{gi} is the number of sequence reads mapped to gene g for sample i and $\sqrt{\phi_g}$ is said to represent the biological coefficient of variantion, CV_{bio}. However, CV_{bio} can also incorporate variation resulting from technical errors such as sample preparation as well as the true biological variation that exists between samples. With knowledge that the coefficient of variation is simply the ratio of standard deviaton to the mean, division by μ_{gi}^2 in the above formula permits the construction of the following:

$$\text{CV}^2(r_{gi}) = \frac{1}{\mu_{gi}} + \phi_g = \text{CV}_{tech}^2 + \text{CV}_{bio}^2.$$

It is evident that $\text{CV}_{tech}^2 = \frac{1}{\mu_{gi}}$, a term arising from variation introduced by the sequencing process, is a decreasing function of the expected count size, while CV_{bio}^2 remains at a constant level.

It was concluded by Law et al. (2014) that through the performance of a simple linearization calculation, the standard deviation of log-*cpm* values are in fact approximately equivalent to $\text{CV}(r_{gi})$, the coefficient of variation of count size. From above,

$$y_{gi} \approx \log_2\left(\frac{r_{gi}}{R_i} \times 10^6\right) = \log_2(r_{gi}) - log_2(R_i) + 6\log_2(10).$$

Treating R_i, the library size, as a constant, it is easily seen that $\text{var}(y_{gi}) \approx \text{var}(\log_2(r_{gi})$. Taylor's theorem indicates that $\log_2(r_{gi}) \approx \mu_{gi} + \frac{r_{gi} - \mu_{gi}}{\mu_{gi}}$, in which μ_{gi} is of a relatively large value. Thus,

$$\text{var}(y_{gi}) \approx \text{var}\left(\frac{r_{gi}}{\mu_{gi}}\right) = \frac{1}{\mu_{gi}^2}\text{var}(r_{gi}) = \frac{1}{\mu_{gi}} + \phi_g.$$

It is therefore to be expected that the standard deviation of log-*cpm*s is a steadily decreasing function of the mean count size, for counts which are of small to medium magnitude. For greater observed count sizes, demonstration of asymptotic behaviour at a certain level, which is dependent on biological variability, should be likely.

The establishment of the *voom* method by Law et al. (2014) originated from the idea that precise modelling of mean-variance relationships is fundamental in the design of methodologies with great statistical power. It

was thus proposed that correct modelling the mean-variance trend of log-cpm values at individual observation level should supersede the aim of exact specification of the probability distribution. In order to achieve this, a comprehension of the way in which the variability depends on count size, which has been discussed above, was deemed crucial.

4.5.3 `voom`: Variance Modelling at Observational Level

A detailed description of the *voom* procedure is as follows. Firstly, for each gene, a linear model is fitted to its normalised log-*cpm* values, y_{gi} by ordinary least squares. This linear model is of the form $E(y_{gi}) = \mu_{gi} = x_i^T \beta_g$. Experimental design is incorporated into the model as the vector of covariates, given by x_i, can be used to describe the manner in which treatment factors are assigned to the various samples. β_g represents a vector containing unknown coefficients which represent $\log_2$-fold changes between the conditions of interest.

The application of these gene-wise linear models result in the production of regression coefficient estimates ($\hat{\beta}_g$), together with fitted log-*cpm* values, $\hat{\mu}_{gi} = x_i^T \hat{\beta}_g$ and residual standard deviations, s_g. These fitted log-*cpm* values can be transformed into predicted counts by:

$$\hat{\lambda}_{gi} = \hat{\mu}_{gi} + \log_2(R_i + 1) - \log_2(10^6).$$

In addition, average log-*cpm* values, $\bar{y}_g$ are calculated for each gene and their conversion into an average number of log-counts can take place:

$$\tilde{r}_g = \bar{y}_g + log_2(\tilde{R}) - log_2(10^6).$$

The geometric mean of library sizes added to 1 is denoted by $\tilde{R}$.

The method proceeds by fitting a LOWESS (locally weighted scatterplot smoothing) curve to the square-root residual standard deviations, $\sqrt{s_g}$ as a function of the mean log-count, $\tilde{r}_g$ for each gene, in order to acquire a statistically robust mean-variance trend. It is because of their approximate symmetric distribution that square-root standard deviations are used. By interpolating this LOWESS curve between ordered $\tilde{r}_g$ values, the piecewise linear function, lo() may be specified. This function, lo() is used to predict the standard deviations of each indiviudal observation with the result of $\text{lo}(\hat{\lambda}_{gi})$ taken as equating to the square-root of the predicted standard deviation of y_{gi}. The *voom* precision weight for a specific observation is thus defined as its inverse predicted variance, $w_{gi} = \text{lo}(\hat{\lambda}_{gi})^{-4}$. Inclusion of weights into the linear modelling process effectively eradicates the existence of a mean-variance relationship in the log-*cpm* values.

Log-*cpm* values y_{gi} accompanied by their respective weights w_{gi} may now be satisfactorily inputted into the `limma`'s linear modelling pipeline. In fact, many of the functions in `limma` can successfully facilitate quantitative weights. Hence, this method proposed by Law et al. (2014) has opened up the possibility of performing thorough microarray-like analyses on RNA-Seq data, in which ignorance of the mean-variance relationship of log-*cpm* values at observational level is not required.

The effect of the application of the `voom()` function to the dataset is highlighted in Figure 9 below. To the left is a depiction of the original mean-variance relationship of the log-*cpm* values. This plot, known as the `voom`-plot, presents a steadily decreasing trend resulting from variation which has arisen from both technical and biological sources. As a consequence of fitting linear models to log-*cpm* values, residual variances are obtained. It is the quarter-root of these variances that can be seen to be plotted against the average log-*cpm* value for each gene. The second plot, generated using the `plotSA` function in `limma` is added to demonstrate the removal of the dependence of the variance on the mean. Similarly, square-root residual standard deviations are plotted against their corresponding mean log-*cpm* values. The residual standard deviations in this instance are obtained as a result of fitting linear models with `lmFit` and implementing `eBayes`, in which the *voom* precision weights are now taken into account. The horizontal blue line highlights the average of all square-root residual standard deviations.

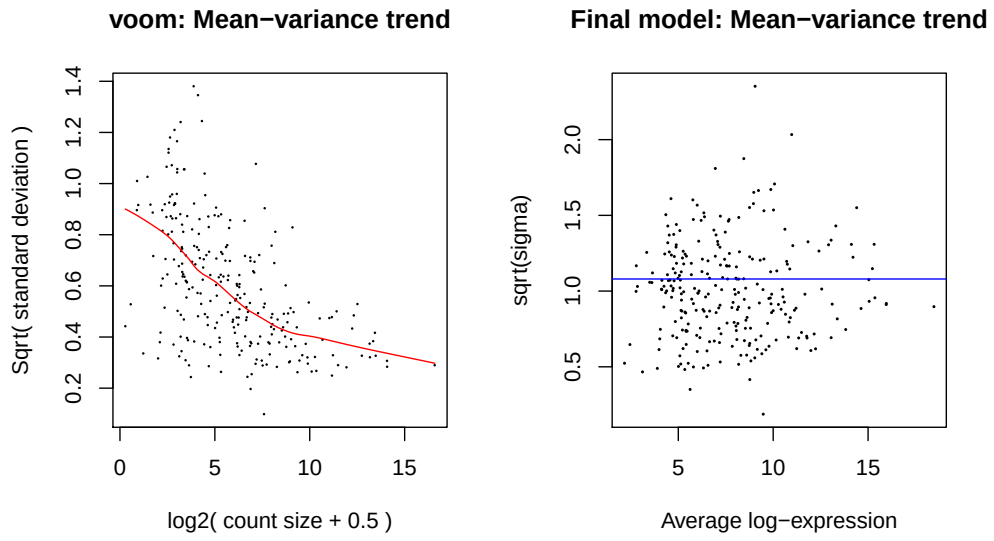

Figure 9: *For each gene, the mean, on the x-axis, is plotted against the square-root standard deviation, on the y-axis, before (left) and after (right) the voom function is applied.*

4.6 Results of the Analysis of RNA-Seq Data

As previously stated, the log-*cpm* values accompanied by their *voom* precision weights are entered into the same linear modelling pipeline that has been described in significant detail for microarray data in Section 3. Upon completion of this process, a total of 164 genes were considered to show evidence of differential expression, 74 downregulated and 90 upregulated, in which the contrast of interest was specified as `Cell-SecretedEV`.

The summary statistics for the top 4 differentially expressed genes, ordered according to their p-values, is included below. The `logFC` for a particular gene may be interpreted as the average of its log-*cpm* values in the two `SecretedEV` samples subtracted from the average of the two `RNACell` samples, while its `AveExpr` is intuitively its mean log-*cpm* value across all samples. The moderated t-statistics and their corresponding p-values have been computed as described in Section 3. The p-values have been adjusted using the BH method which corrects for multiple testing by controlling the false discovery rate. It can be seen from this table that if ranking was to occur with respect to their B-statistics, the genes would be ordered in a different manner. This attribute is due to the introduction of weights which has resulted in the unscaled standard deviations, v_{gj} no longer being constant across the genes. Thus, it now cannot be deduced that the B-statistic, B_{gj} is a monotonically increasing function of the absolute value of the moderated t-statistic, $|\tilde{t}_{gj}|$.

```
              logFC  AveExpr         t      P.Value    adj.P.Val        B
GeneID713  -6.912208 8.765715 -21.53126 4.472199e-13 8.748242e-11 19.87161
GeneID62  -10.708523 8.039360 -20.98119 6.627456e-13 8.748242e-11 18.15124
GeneID195 -11.887197 5.092344 -17.33674 1.177454e-11 1.036160e-09 14.99547
GeneID258  -6.012078 8.061566 -15.94870 4.082689e-11 2.694575e-09 15.69905
```

A volcano plot has been produced to highlight the genes that have been found to be statistically significant as a result of the performed analysis. The top 4 differentially expressed genes from the table above are named in this plot, which can be viewed in Figure 10. The large $\log_2$-fold changes of these genes are evident, especially `GeneID62` and `GeneID195` which have `logFC` values of -10.71 and -11.89, respectively. As with all volcano plots, downregulated genes, coloured in red, lie to the extreme left of the plot while upregulated genes, coloured in green, can be seen on the right.

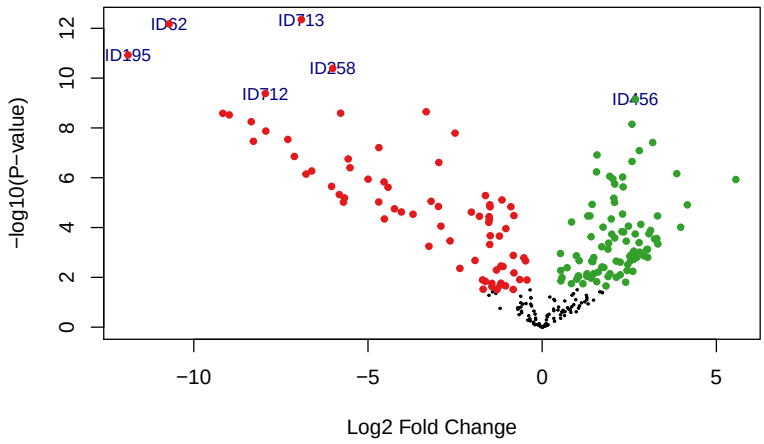

Figure 10: *Volcano plot depicting the results of the differential expression analysis performed on this local mouse dataset.*

5 Hierarchical Clustering

5.1 Introduction to Hierarchical Clustering

Hierarchical clustering is a method which has been used since the very beginning of microarray data analysis. Being an easily visualised and intuitive technique, it has proven to be most useful in performing tasks such as assessing similarity of gene expressions and ensuring that basic global patterns of the dataset in hand are as expected. It is generally assumed that similar patterns of expression are shown by technical and biological replicates while pattern of two different experimental conditions would be significantly more varied. Thus, the principal aim of hierarchical clustering is to provide definitive clustering which characterizes a set of genes or samples with respect to a given distance metric. It is referred to as an unsupervised method due to the fact that the number of clusters are not set in advance and there is no prior knowledge of experimental design.

In the following, the task of clustering a set of samples will be considered. However, it is important to note that the clustering of genes into small sets is also a topic of significant interest. Clustering of genes is regularly used to assist in the identification of genes belonging to the same molecular pathway. Before hierarchical clustering is performed, two important decisions are required:

- How should one compute the degree of similarity between pairs of samples?
- How should this concept of similarity be implemented in order to obtain clusters of samples?

A large variety of cluster analysis methodology exists, with the degree of sample similarity in many of these methods based on a specific definition of *distance* between a pair of samples. A popular approach for the evaluation of this concept of distance between microarray/RNA-Seq samples is that of the Pearson correlation coefficient, r, as discussed in Ewens and Grant (2005). This correlation coefficient is a measure of the strength of the linear relationship between two variables. It is defined as the covariance of two variables divided by the product of standard deviations. Pairwise correlations are computed in `R` using the `cor()` function. Given an estimate of the correlation, r, between expression levels of two genes, a distance between these genes may be consequently defined in various ways. If the sign of r is deemed not to be of importance, a possible distance measure is d = 1 - $\mid r \mid$. This is exactly 0 when $r = \pm 1$, indicating that genes which are judged to be perfectly correlated are taken as having no distance between them. The Euclidean distance

is also commonly used in order to measure the distance between two vectors of read counts. However, for RNA-Seq data, large differences in the library sizes of two samples can greatly affect the Euclidean distance.

With this concrete definition for distance between samples, a choice must now be made as to how the inter-cluster distance is to be judged. As above, there are multiple options for linkage functions, with the most popular being:

1. *complete-linkage*: distance between two clusters of samples is defined as the largest distance between any two neighbours.
2. *average-linkage*: understood to have a very similar meaning to complete-linkage but now, the average distance between the individual components of the two clusters is considered.

Two samples are regarded as being neighbours if they belong to different clusters. It has been established that complete linkage often outperforms average linkage.

The result of this clustering technique is a complete tree with individual samples as leaves and the root as the convergence point of all branches. Unlike real trees, these hierarchical trees are drawn with a root on top and branches developing underneath. This diagram produced by hierarchical clustering is known as a dendrogram. It clearly illustrates a hierarchy of categories based on their degree of similarity. Pairwise comparisons of the individual samples allow for the grouping together of similar samples into clusters, resulting in the creation of the dendrogram. These comparisons can be made using a hierarchical clustering strategy which is either agglomerative or divisive in combination with specific distance and linkage functions. The method of tree construction which commences with individual genes and works up towards the root is agglomerative, more commonly known as the *bottom-up* approach. At the start, n clusters are present, each consisting of a single sample. In the case where the clustering algorithm is being applied to the set of genes, then the n clusters would each hold an individual gene. At each time point of the algorithm, the distances from each cluster to every other cluster are computed based on a specification of inter-cluster distance. Then, this hierarchical agglomerative clustering method repeatedly merges the two nearest clusters into a new single super-cluster until construction of the entire tree has taken place.

By default, the `hclust` function in `R` uses the complete linkage method for hierarchical clustering clusters genes or samples using the agglomerative approach previously detailed. The dendrogram obtained upon use of this function for the 60 samples present in the `leukemiasEset` dataset is included in Figure 11. In this instance, the correlation coefficient, r has been chosen as a suitable distance measure. The samples have been labelled according to their respective leukemia types. It is interesting to see that division into 5 clear separate groups has not occurred. However, many similarly coloured samples seem to appear close to one another. Additionally, it is witnessed that all `CLL` samples are grouped alone in a distinctly defined cluster.

Unfortunately, this technique of hierarchical clustering also has its limitations, with one of the most prominent being that it is not an iterative procedure. Thus, great sensitivity to noise, outliers and even, the different cluster distance measures exists. It is also noted that the process performs significantly better if implemented on a limited dataset, with less than approximately 300 genes or samples.

5.2 Heat Maps

Irrespective of which platform, microarrays or RNA-Seq, has been used to generate gene expression data, one of the most popular forms of visualisation of this data is through the production of a heat map. This heat map construction often relies on the clustering techniques detailed above as two separate dendrograms, one for samples and one for genes, are included in the image. The heat map exhibits the data in grid format, in which the different colours of the components of the grid are used to display the degree of change in expression for each gene across the range of samples. The genes are represented by the rows and the columns of the grid are used to represent the samples. This graphical representation of the data can make the identification of groups of genes and samples with common features much easier.

The `gplots` package in `R` provides access to the creation of heat maps using the `heatmap.2` function. This function was employed to assemble the heat map included below in Figure 12. For demonstration purposes,

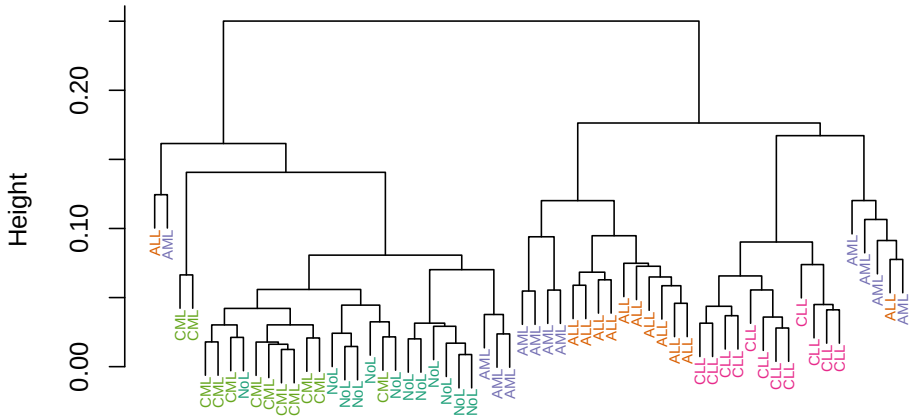

Figure 11: *Dendrogram obtained when samples from the leukemia dataset have been clustered according to correlation coefficients and complete-linkage is used to define inter-cluster distance.*

the log-intensity data from two conditions, `AML` and `NoL` for the 100 most variable genes across these 24 samples from the `luekemiasEset` dataset has been used. Euclidean distances have been computed by `heatmap.2` in order to manufacture the two dendrograms seen on the outer edges of the diagram. In this plot, a bright orange colour symbolizes genes showing evidence for upregulation while darker green and blue shades suggest the presence of downregulated genes.

Further inspection of the sample dendrogram suggests that making a cut at a certain height shows the possibility of defining three separate clusters of samples. All the `NoL` samples have been clearly clustered together on the right. This is supported by obvious dissimilarities in the colour palettes of the two halves of the heat map. Furthermore, it is of interest to note the behaviour of the group of four `AML` samples which have been regarded as being closer in distance to the cluster of `NoL` samples than to the other eight `AML` samples. This characteristic is also portrayed well in the heat map as there is an obvious lack of green in the corresponding columns of these `AML` samples. Overall, many downregulated genes, which are depicted by the green and blue colours, are visible on the left third of the heat map. This region corresponds to samples obtained from `AML` patients. This observation is accompanied by a strong orange signal being broadcast from the right of the heat map, indicating numerous upregulated genes in all 12 `NoL` samples.

In a similar manner, a heatmap has been generated for the local RNA-Seq mouse dataset. A more simplistic heatmap can thus be viewed in Figure 13, which contains only 4 samples. The log counts of the 264 genes which were retained after filtration have been employed to produce this image. The sample dendrogram on top has clearly divided the samples into two groups as expected. The `RNACell` samples are positioned to the left while the `SecretedEV` are on the right. Clear similarities can be seen in the colour profiles of the two samples in each group. In this heatmap, the colour red is representative of upregulated genes while blue emphasizes downregulated genes. In several instances, a band of red can be seen on one half of the plot with shades of blue visible in the section of the other half corresponding to this band. Therefore, upon inspection of this heatmap, it is evident that there exist great differences in gene expression patterns between the two groups of the dataset.

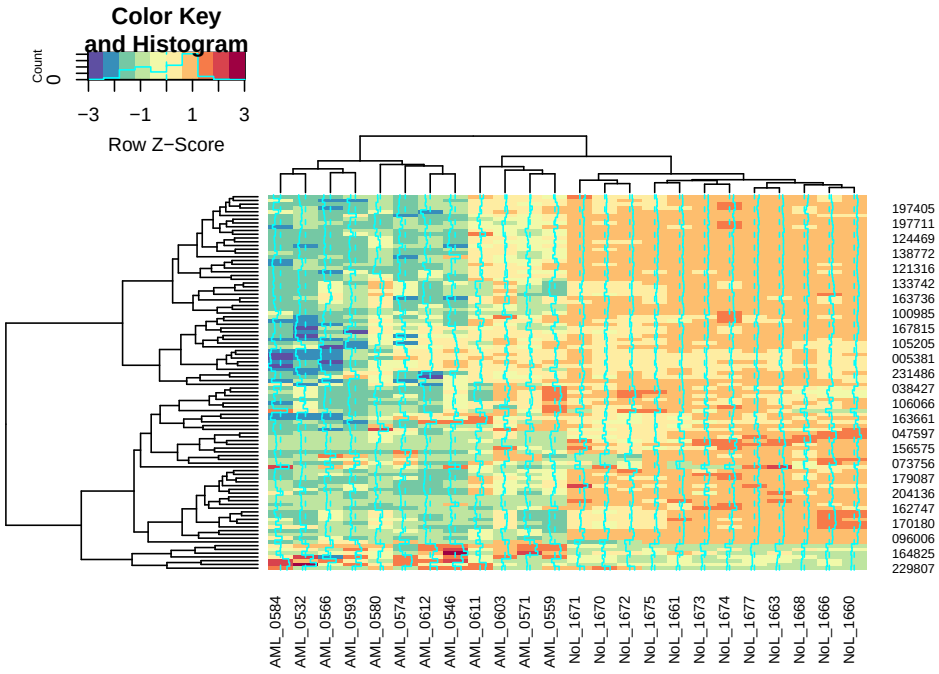

Figure 12: *Heatmap produced using the 12 NoL and 12 AML samples and the 100 most variable genes across these samples from the leukemia dataset.*

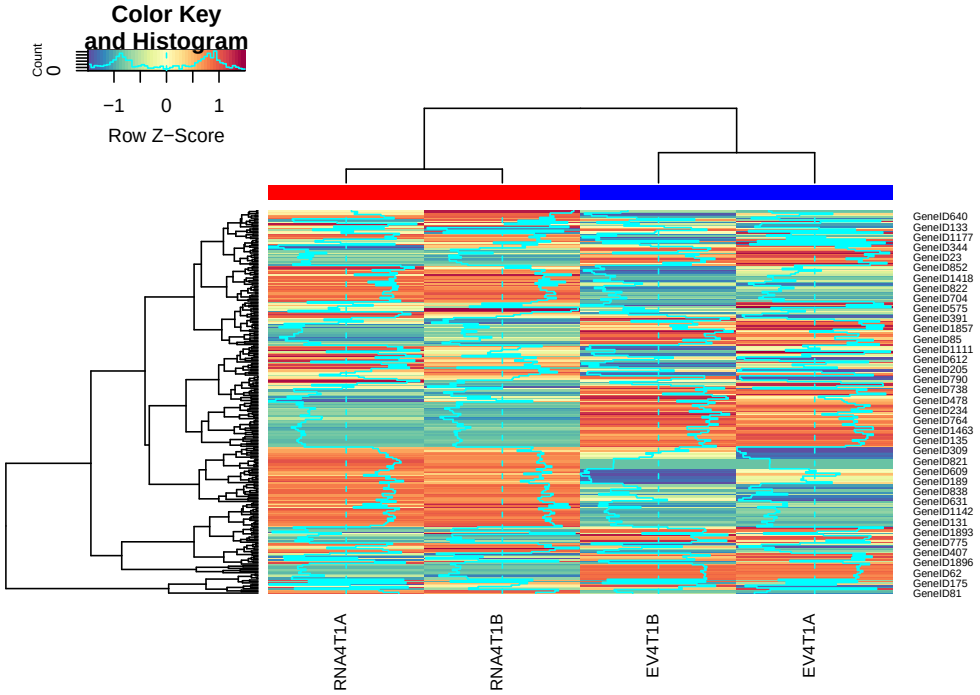

Figure 13: *Heatmap produced using the genes which have been retained after filtration and all four samples in the local mouse RNA-Seq dataset.*

6 Conclusion

This project has predominantly focused on investigating the multiple statistical methods commonly used in the analysis of gene expression data. Not only has a comprehensive understanding of the theoretical concepts behind the construction of these techniques been gained but their implementation on real world datasets, such as the `leukemiasEset` dataset and the local mouse RNA-Seq dataset, has also been explored. Specific questions which arose through preliminary research of this subject, as discussed in Section 1, have been answered in an in-depth manner. As demonstrated, completeness of the project was achieved by using `R` and its packages, in particular `limma` and `edgeR`, for both investigative and illustrative purposes.

The issue of conducting multiple associated hypothesis tests was detailed in Section 2. Certain techniques commonly used to overcome this problem were explored and their individual advantages and disadvantages identified. It was seen that the BH method, which was implemented in both analyses, continues to be the multiple testing correction method of choice in many cases. If more time had been available, it would have been interesting to investigate more modern methods which attempt to control false discoveries, such as those described by Korthauer et al. (2019), and to see how they could be incorporated into differential gene expression analyses of real world data.

Throughout the project, the many challenges, in addition to multiple testing, which arise when analyzing this form of data were discovered. A key issue is related to variance estimation as only a very small number of samples are often available. It was seen that empirical Bayes procedures, such as that applied in `limma`, can be used to assist in making more stable inferences. It was most interesting to witness the important role of hierarchical Bayesian models in this context and to discover new concepts such as the *moderated* t-statistic. As this data can be hugely affected by noise, the necessity of implementation of various normalisation procedures was also seen. Two of these techniques were looked at in detail, namely RMA normalisation and TMM normalisation. It was observed that even though these methods have been designed for essentially the same purpose of normalisation, there exist very few similarities between them as in general, each method has been created with application to a particular data type in mind.

Another main component of this project was the identification of the differences which exist between data sourced from microarray and RNA-Seq technologies and the effect that this has on the way in which the analysis is performed. On one hand, the data is in the form of numerical continuous measurements while on the other, it is count integer data which is generated. A fascinating aspect of the project was seeing how the methods which had been originally used for microarrays could in fact be made available for this count data. Understanding the possibility for this required an investigation of many complex concepts, in particular the modelling of the mean-variance relationship. The advantages of using normal-based approaches made accessible by determining mean-variance trends over modelling the data directly using count distributions were also studied. This was perhaps one of the most interesting parts of the project as exploration of it exposed a multitude of new concepts. An intriguing topic for further study would be the application of this approach of mean-variance modelling to other areas in which the aim is to gain valuable information from tricky count data. In addition, it would be interesting to analyze and compare the results obtained when count distributions, such as the negative binomial distribution which is used by `edgeR` and `DESeq2`, are used to model the data from the RNA-Seq dataset.

Engaging with the two datasets, especially the local dataset, provided the opportunity to add another dimension to the project in which the learned statistical techniques could be fully appreciated. For example, when performing the analysis on the mouse dataset, certain judgements such as those regarding appropriate filtration procedures were required to be made. Thus, this emphasized the importance of having a solid foundation in the reasoning behind the usage of methods in order to successfully complete accurate analyses. Furthermore, knowledge of how various functions operate provided the ability to quickly detect errors, preventing the generation of incorrect results.

The final section provides a brief introduction to the topic of hierarchical clustering and details the outcome of its application to the datasets at hand. The importance of techniques such as this, particularly those which incorporate visual images, is highlighted as they can greatly assist in the formation of expectations regarding analysis results. Indeed, the exploration of other forms of principal component analyses which can complement this study is another possible direction in which an extension of this project could take.

7 Bibliography

Anders, Simon, and Wolfgang Huber. 2010. "Differential Expression Analysis for Sequence Count Data." *Genome Biology* 11 (10): r106.

Benjamini, Yaov, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing Under Dependency." *The Annals of Statistics* 29 (4): 1165–88.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)* 57 (1): 289–300.

Bland, J. M., and D. G. Altman. 1995. "Multiple Significance Tests: The Bonferroni Method." *BMJ (Clinical Research Ed.)* 310 (6973): 170.

Ewens, W. J., and Gregory R. Grant. 2005. *Statisical Methods in Bioinformatics: An Introduction.* New York, NY: Springer.

Haferlach, Torsten, Alexander Kohlmann, Lothar Wieczorek, Giuseppe Basso, Geertruy Te Kronnie, Marie-Christine B'en'e, John De Vos, and et al. 2010. "Clinical Utility of Microarray-Based Gene Expression Profiling in the Diagnosis and Subclassification of Leukemia: Report from the International Microarray Innovations in Leukemia Study Group." *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 28 (15): 2529–37.

Huber, W., V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21.

Irizarry, R. A. 2003. "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics* 4 (2): 249–64.

Kohlmann, Alexander, Thomas J. Kipps, Laura Z. Rassenti, James R. Downing, Sheila A. Shurtleff and Ken I. Mills, Amanda F. Gilkes, and et al. 2008. "An International Standardization Programme Towards the Application of Gene Expression Profiling in Routine Leukaemia Diagnostics: The Microarray Innovations in Leukemia Study Prephase." *British Journal of Haematology* 142 (5): 802–7.

Korthauer, Keegan, Patrick K. Kimes, Claire Duvallet, Alejandro Reyes, Ayshwarya Subramanian, Mingxiang Teng, Chinmay Shukla, and et al. 2019. "A Practical Guide to Methods Controlling False Discoveries in Computational Biology." *Genome Biology* 20 (118).

Law, Charity W., Monther Alhamdoosh, Shian Su, Gordon K. Smyth, and Matthew E. Ritchie. 2016. "RNA-Seq Analysis Is Easy as 1-2-3 with Limma, Glimma and edgeR." *F1000Research* 5.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. "Voom: Precision Weights Unlock Linear Model Analysis Tools for Rna-Seq Read Counts." *Genome Biology* 15 (2): r29.

Love, Michael I., Huber Wolfgang, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for Rna-Seq Data with Deseq2." *Genome Biology* 15 (550).

McCarthy, Davis J., Chen Yunshun, and Gordon K. Smyth. 2012. "Differential Expression Analysis of Multifactor Rna-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): 4288–97.

McLachlan, Geoffrey J., Kim-Anh Do, and Christophe Ambroise. 2004. *Analyzing Microarray Gene Expression Data.* New Jersey: John Wilwy & Sons.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.

Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. "limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies." *Nucleic Acids Research* 43 (7): e47.

Robinson, Mark D., McCarthy Davis J., and Gordon K. Smyth. 2009. "EdgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.

Robinson, Mark D., and Alicia Oshlack. 2010. "A Scaling Normalization Method for Differential Expression Analysis of Rna-Seq Data." *Genome Biology* 11.

Robinson, Mark D., and Gordon K. Smyth. 2007. "Moderated Statistical Tests for Assessing Differences in Tag Abundance." *Bioinformatics* 23 (21): 2881–7.

Smyth, Gordon K. 2004. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology* 3 (1).

Smyth, Gordon K., Matthew Ritchie, Natalie Thorne, James Wettenhall, Wei Shi, and Yifang Hu. 2019. *Limma: Linear Models for Microarray and Rna-Seq Data User's Guide*. Melbourne, Australia: Bioinformatics Division, The Walter; Eliza Hall Institute of Medical Research.

Soneson, Charlotte, and Mauro Delorenzi. 2013. "A Comparison of Methods for Differential Expression Analysis of Rna-Seq Data." *BMC Bioinformatics* 14 (91).

Sorin, Drăghici. 2012. *Statistics and Data Analysis for Microarrays: Using R and Bioconductor*. London: Chapman; Hall.