A COMPARISON OF FILTERING AND NORMALIZATION METHODS IN THE STATISTICAL ANALYSIS OF GENE EXPRESSION EXPERIMENTS

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Mackenzie Rosa Marie Speicher

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2020

Fargo, North Dakota

North Dakota State University

Graduate School

Title

A COMPARISON OF FILTERING AND NORMALIZATION METHODS IN THE STATISTICAL ANALYSIS OF GENE EXPRESSION EXPERIMENTS

By

Mackenzie Rosa Marie Speicher

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Dr. Megan Orr

Chair

Dr. Rhonda Magel

Dr. Anne Denton

Approved:

April 6, 2020

Date

Dr. Rhonda Magel

Department Chair

ABSTRACT

Both microarray and RNA-seq technologies are powerful tools which are commonly used in differential expression (DE) analysis. Gene expression levels are compared across treatment groups to determine which genes are differentially expressed. With both technologies, filtering and normalization are important steps in data analysis. In this thesis, real datasets are used to compare current analysis methods of two-color microarray and RNA-seq experiments. A variety of filtering, normalization and statistical approaches are evaluated. The results of this study show that although there is still no widely accepted method for the analysis of these types of experiments, the method chosen can largely impact the number of genes that are declared to be differentially expressed.

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my advisor, Dr. Megan Orr, for her encouragement and guidance throughout the course of this thesis. I am grateful for the amount of time she spent meeting with me, answering my questions and sharing her knowledge on this topic. Additionally, I would like to thank Dr. Rhonda Magel and Dr. Anne Denton for their support and recommendations.

TABLE OF CONTENTS

ABSTRACT iii

ACKNOWLEDGEMENTS iv

LIST OF TABLES vii

LIST OF FIGURES viii

1. INTRODUCTION 1

1.1. Background 1

1.1.1. Differential Expression Analysis 1

1.1.2. Two Color Microarrays 1

1.1.3. RNA-seq 2

1.2. Research Objectives 3

1.3. Organization 3

2. LITERATURE REVIEW 4

2.1. False Discovery Rate 4

2.2. Microarray Analysis 5

2.2.1. Background Correction 5

2.2.2. Transformation 6

2.2.3. Filtering 7

2.2.4. Lowess Normalization 7

2.2.5. Median Normalization 8

2.2.6. Scale Normalization 9

2.2.7. Quantile Normalization 9

2.2.8. Limma 10

2.2.9. SAM-seq 10

2.2.10. Comparison of Methods 11

2.3. RNA-seq Analysis 11
2.3.1. Filtering 11
2.3.2. EdgeR 12
2.3.3. DESeq 13
2.3.4. DESeq2 14
2.3.5. Limma-trend 14
2.3.6. Limma-voom 15
2.3.7. Comparison of Methods 15
3. METHODS 17
3.1. Two-Color Microarray 17
3.2. RNA-seq 18
4. RESULTS AND DISCUSSION 21
4.1. Two-Color Microarray 21
4.2. RNA-seq 24
4.2.1. FI vs FN 24
4.2.2. MI vs MN 28
4.2.3. FN vs MN 31
4.2.4. Interaction 34
4.2.5. Asymmetry in the Distribution of Effect Sizes 37
5. CONCLUSION AND FUTURE RESEARCH 41
5.1. Conclusion 41
5.2. Future Research 41
REFERENCES 42

LIST OF TABLES

Table Page

1. Outcomes for hypothesis testing in gene expression experiments 5

2. Number of genes DDE for different normalization and filtering methods with an alpha value of .085 23

3. Number of genes DDE for the FI vs FN comparison of interest 26

4. Number of genes DDE for the MI vs MN comparison of interest 29

5. Number of genes DDE for the FN vs MN comparison of interest 32

6. Number of genes DDE for the interaction term 35

7. The estimated proportion of equivalently expressed genes, as well as the proportion of upregulated among differentially expressed genes for each comparison of interest 38

LIST OF FIGURES

Figure Page

1. MA plot before lowess normalization. 21

2. MA plot after lowess normalization. 21

3. Venn diagram of the number of DDE genes for different normalization methods applied after filtering with filter method 1. LS stands for loess and scale, LMSQ stands for loess, median, scale and Gquantile. An alpha value of .085 was used to determine significance. 24

4. Venn diagram of DDE genes for FI vs FN comparison. 27

5. Venn diagram of the number of genes DDE for the FI vs FN comparison of interest. Includes various normalization methods of the edgeR packge; TMM, TMMwsp, upperquartile, RLE and none. 27

6. Venn Diagram of DDE genes for MI vs MN. 30

7. Venn diagram of the number of genes DDE for the MI vs MN comparison of interest. This diagram includes the various normalization methods of the edgeR packge; TMM, TMMwsp, upperquartile, RLE and none. 30

8. Venn Diagram of DDE genes for FN vs MN. 33

9. Venn diagram of the number of genes DDE for the FN vs MN comparison of interest. This diagram includes the various normalization methods of the edgeR packge; TMM, TMMwsp, upperquartile, RLE and none. 33

10. Venn diagram of DDE genes for the interaction term. 36

11. Venn diagram of the number of DDE genes for the interaction term. This diagram includes the various normalization methods of the edgeR package; TMM, TMMwsp, upperquartile, RLE and none. 36

12. Histogram showing the difference between the mean log counts per million for female injured vs female naïve. 39

13. Histogram showing the difference between the mean log counts per million for the male injured vs male naïve groups. 39

14. Histogram showing the difference between the mean log counts per million for the male naïve vs female naïve groups. 40

15. Histogram of the mean log counts per million of MI-MN vs the mean log counts per million of FI-FN. 40

1. INTRODUCTION

1.1. Background

1.1.1. Differential Expression Analysis

DNA microarray technologies, developed in the 1990's, are used to provide the gene expression levels of thousands of genes from various organisms (Marioni et al., 2008). By comparing the expression levels of genes in healthy and diseased tissues, these technologies have helped to advance medical and biological research. RNA-seq is a new advancement in gene expression analysis which results in a more direct sequencing of transcripts. A common use of microarray and RNA-seq technology is differential expression (DE) analysis. That is the process of testing for quantitative changes in expression levels between various treatment groups. This is done gene by gene, and the goal is to determine which genes are differentially expressed. RNA-seq and microarray analyses have different processes and advantages that are discussed in greater detail below.

1.1.2. Two Color Microarrays

In two color microarray experiments, DNA sequences are deposited onto a surface that is often referred to as a slide or chip. The sequences are typically arrayed in spots which serve as probes. Messenger RNA is extracted from an organism, dyed with fluorescent dye, and hybridized to the slide. Two color microarrays are hybridized with two samples which are labelled with different colored dyes (Hoen et al., 2004). The most commonly used dyes are Cy3 (green) and Cy5 (red). The expectation is that the mRNA will bind to complementary sequences that are already present on the slide. A laser is used to excite the dye, and the fluorescent intensities are recorded and compared (Kaliyappan et al., 2012). The ratio of the red/green intensity is often used as a relative measure of gene expression.

The linear model for two color microarray experiments is as follows:

$$Y_{ijk} = \mu + \tau_i + \delta_j + s_k + e_{ijk}$$

In this formula, Y_{ijk} represents the normalized signal intensity (NSI) of the experimental unit in the ith treatment group, hybridized with the jth dye, on kth slide. μ represents overall mean NSI, τ_i represents the effect of the ith treatment. δ_j represents the effect of the jth dye (j = 1,2). s_k represents the random effect of the kth slide, and $\boldsymbol{e_{ijk}}$ represents the random error for Y_{ijk}.

Microarray technologies are relatively inexpensive and readily available for use, however, there are several limitations. For example, background hybridization may limit the accuracy of measurements. Additionally, microarrays can only be used on known genomes, and have poor quantification of lowly and highly expressed genes (Zhao et al., 2014). Although they are still being used today, in many cases microarrays are being replaced with newer advancements.

1.1.3. RNA-seq

A new development in gene expression technologies is RNA-seq, which is the direct sequencing of transcripts by high-throughput sequencing technologies. This is done through the use of next generation sequencing. Messenger RNA is extracted from an organism and converted to a library of cDNA fragments. The fragments are then sequenced and mapped to a genome, and the level of expression is recorded (Wang et al., 2009). RNA-seq is advantageous because it offers a wide measurable range of expression levels and does not depend on a known genome (Wang et al., 2009). The data from RNA-seq experiments is in counts, unlike microarray data which is in the form of fluorescent intensities and considered continuous. Therefore, the data analysis must be conducted differently. These technologies are relatively new, so there is still no widely accepted method for data analysis.

Since the data from RNA-seq experiments is in counts, it must follow a discrete distribution. A negative binomial model is most commonly used to model the data, as it allows for over dispersion. This means that the variability of the response is greater than expected based on the mean. The model is as follows:

$$Y_{gi} \sim NB(M_i P_{gi}, \varphi_g)$$

Where M_i is the library size, P_{gi} is the relative abundance of the g^{th} gene from the i^{th} sample and φ_g is the dispersion parameter. This model will be discussed in greater detail in chapter 2.3.

1.2. Research Objectives

1. To use a real data set to compare the various methods of DE analysis in two-color microarray experiments. Lowess, location, scale and quantile normalization methods are compared, as well as different filtering methods. Analysis is conducted in R using the Limma package, SAM-seq and t-tests.

2. To use a real data set to compare the various methods of DE analysis in RNA-seq experiments. Normalization methods implemented by the edgeR, DESeq, DESeq2 and Limma packages are compared, as well as different filtering methods.

1.3. Organization

The rest of this thesis is organized as follows. In chapter 2, false discovery rate and current filtering/normalization methods for microarray and RNA-seq experiments are reviewed. Chapter 3 describes the methods used for the analysis of real datasets for both two-color microarray and RNA-seq experiments. Chapter 4 presents the results and discussion of the comparison of normalization methods for the two-color microarray and RNA-seq datasets. Lastly, overall conclusions and future research are given in chapter 5.

2. LITERATURE REVIEW

2.1. False Discovery Rate

In both microarray and RNA-seq experiments, we are interested in performing a hypothesis test for each gene in order to test for differential expression. This means that thousands of hypothesis tests are conducted simultaneously. Therefore, it is important to implement some form of multiple testing correction. Controlling FWER (family wise error rate), is a traditional multiple testing strategy. FWER is the probability of one or more false positive, where a false positive is rejecting the null hypothesis when it is actually true. The Bonferroni method is a standard way to control FWER, in which the significance level alpha is divided by the number of tests conducted. This isn't the best method to use in gene expression experiments because of the high dimensionality of the data. The number of genes is often times in the tens of thousands, so the p-value would have to be extremely small in order to reject the null hypothesis. This can result in very low power for detecting differentially expressed genes. For this reason, researchers have proposed controlling FDR (false discovery rate), rather than FWER as the multiple testing correction. The FDR was proposed by Benjamini and Hochberg in 1995. Referring to table 1, FDR can be defined as E(Q) where Q=V/R if R>0, and Q=0 otherwise. In words, FDR is the expected portion of false rejections among all rejections. When compared to FWER, controlling for FDR results in higher power and may also allow for more type 1 errors. However, this is acceptable as long as the number of false rejections is being controlled. For example, if controlling FDR results in a list of DDE (declared to be differentially expressed) genes in which 95% of the genes truly are differentially expressed, this is useful from a scientific standpoint. For this reason, controlling FDR is more appropriate for gene expression experiments. The procedure for controlling FDR is as follows:

- Order the m p-values from smallest to largest, and denote them as $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$
- Find the largest integer k such that

$$p_{(k)} \leq \frac{k\alpha}{m}$$

- Reject the null hypotheses corresponding to the smallest k p-values. This procedure controls the FDR at α.

Table 1

Outcomes for hypothesis testing in gene expression experiments

	Fail to reject H_0	Reject H_0	Total
H_0 true (Gene EE)	U	V	m_0
H_0 false (Gene DE)	T	S	m_1
Total	W	R	m

Note. m is the total number of genes. The values W, R and m are observable quantities. The rest of the values are unobservable quantities.

2.2. Microarray Analysis

2.2.1. Background Correction

Background correction is necessary in two color microarray experiments. The goal is to obtain the target sample intensity without any background noise. Some potential sources of background signal include non-specific binding of sample to the slide, contamination from other particles, or noise from the scanner. There are many different proposed methods of background correction, with the most common being background subtraction. This is where you subtract the

estimated background intensity from the overall signal intensity to obtain the estimated target sample intensity (Ritchie et al., 2007). This process could result in a negative target sample intensity. One solution to correct for this would be to replace any negative values with zero, and add one to all of the values.

Although subtraction is the most common method of background correction, it is not necessarily the best method. Other methods that have been proposed are designed to result in strictly positive target sample intensities. One example is the Edwards method, in which the background is only subtracted from the overall signal intensity if the difference is larger than a certain threshold. If the difference is smaller than that threshold, a monotonic function is used instead of subtraction (Edwards, 2003). Some other methods, such as "half", "minimum", and "normexp" are options in the limma package in R. As previously mentioned, the goal of each of these methods is to correct for background while resulting in positive target sample intensities.

2.2.2. Transformation

Two-color microarray intensities are typically transformed from their original scale to a log base 2 scale. Intensity data are generally skewed right, complicating analysis. Performing a log2 transformation can make the data follow an approximate normal distribution, making analysis more straight forward. This can be done without losing any information. The log2 scale is particularly useful because up and downregulated genes are treated similarly (Quackenbush, 2002). Upregulated genes are genes that are more highly expressed in one condition compared to the second condition, whereas downregulated genes have a lower degree of expression in one condition compared to the second condition. For example, when the fold change (the ratio of R/G) has a value of 2, the log2(ratio) has a value of 1. When the fold change is equal to 0.5, the

log2(ratio) has a value of -1. Clearly, log2 transformation is a convenient tool for the analysis of microarray data.

2.2.3. Filtering

In two-color microarray experiments, multiple hypothesis tests are performed simultaneously. The number of tests being performed is very large, and the proportion of genes that are differentially expressed is typically low. Multiple testing corrections must be implemented, but this results in low power for detecting differentially expressed genes. One way to ameliorate this issue is with filtering. By filtering the data to get rid of unreliable spots, the dimensionality can be reduced, thereby increasing the power for detecting differentially expressed genes.

The appropriate method of filtering can vary from experiment to experiment. One general recommendation is to get rid of spots that are below the detection limit, as well as spots that are above the detection limit (saturated signal). Scanners have a maximum value for which they can detect signal intensities (Quackenbush, 2002). The true signal intensity might actually be much higher than the value that is displayed, therefore these measurements are meaningless. Another filtering method is keeping only spots for which the intensity is significantly different from the corresponding background intensity (Quackenbush, 2002). This can help to increase the reliability of the intensities.

2.2.4. Lowess Normalization

The goal of normalization in gene expression experiments is to eliminate non-biological variation so true biological variation can be observed. One example of non-biological variation that we want to account for is dye-bias. In two color microarray experiments, red dye (Cy5) and green dye (Cy3) are both used on one slide. Differences in the properties of the dyes can result in

a dye effect that needs to be accounted for. Possible reasons for these dye effects may include differences in heat and light sensitivity, efficiency of dye incorporation and scan settings.

Self-self experiments have been conducted in which the same sample is dyed both red and green and hybridized to a slide. In this scenario we would expect the red and green intensities to be the same, since there is no biological variation. However, from experiments like this, we know that red intensities tend to be lower than green intensities. This difference between intensities is not constant across all spots. One proposed method to account for this difference is a type of within-slide normalization called lowess normalization.

Lowess stands for LOcally Weighted polynomial regreSSion. It is used to smooth the MA scatter plot of two-color microarray data. An MA plot is a plot of the intensity log ratio ($M=\log_2 R - \log_2 G$) vs. the mean log intensity ($A=[\log_2 R + \log_2 G]/2$). MA plots are often used in favor of $\log_2 R$ vs. $\log_2 G$ plots because they do a better job of revealing interesting features of the data (Yang, 2002). In lowess normalization, a tricube weight function is used to calculate regression weights for each data point. Neighboring data points within a specific span have weight and are influential on the fitted value (Cleveland, 1979). It is also possible to use robust weights, which are resistant to outliers. A lowess normalized M value for each spot is calculated using the equation $M_j^* = M_j - \widehat{M}_j$. In this equation, $\widehat{M}_j$ is the fitted M-value from LOWESS for the jth spot. This process smooths the MA plot and corrects for red-green differences.

2.2.5. Median Normalization

In addition to lowess, median normalization is another type of location normalization used in two-color microarray experiments (Yang et al., 2002). The purpose is to align the centers for each channel to make them more comparable. One potential choice of median in this process is zero, however the value chosen is not important.

2.2.6. Scale Normalization

Scale normalization is another proposed type of normalization for two-color microarray experiments. It was first suggested by Yang et al. in 2002 and is typically performed after location normalization. It follows the assumption that all log-ratios for a particular channel follow a $N(0, a_j^2\sigma^2)$ distribution. The purpose of scale normalization is to estimate the constant a_j in order to make the variance of each channel comparable. The formula used to estimate $\hat{a}_j$ is $\hat{a}_j = \frac{MAD_j}{C}$. MADj is the median absolute deviation, and is calculated by the formula $MAD_j = median(\left|M_{ij} - median_j(M_{ij})\right|)$. C is the geometric mean, calculated by the formula $C = (\Pi_{j=1}^{J})^{1/J}$. Following scale normalization, each channel should follow a normal distribution with variance σ^2.

2.2.7. Quantile Normalization

Quantile normalization was originally proposed for affymetrix experiments by Bolstad et al. in 2003. The idea was extended to two-color microarray data by Yang and Thorne in 2003. Quantile normalization is a form of between slide normalization in which all quantiles (percentiles) across channels must be equivalent (Yang and Thorne, 2003). This results in each channel sharing a similar distribution. The limma package in Bioconductor allows for different types of quantile normalization. The different options in the package are “quantile”, “Aquantile”, “Tquantile”, “Gquantile”, and “Rquantile”. The “quantile” option results in the intensities following the same distribution across channels. The “Aquantile” option performs quantile normalization on the A-values (average intensities). The “Gquantile” and “Rquantile” options are useful in experiments in which the green channel or red channel is used as a common reference throughout the experiment, respectively. For example, if the “Gquantile” option is used, the distribution of intensities is ensured to be the same for the green channel on all arrays.

2.2.8. Limma

Limma (Linear models for microarray) is a package in R that is used for differential expression analysis of microarray and RNA-seq experiments. Using the Limma package, two-color microarray data can be pre-processed using the background correction and normalization methods discussed above. A linear model is then fit to the data for each gene and an empirical Bayes approach is used to borrow information between genes to better estimate gene-wise variances (Ritchie et al., 2015). When testing for differential expression, a moderated t-statistic or F-statistic is used in which the variances are moderated across genes. That is, information from all genes is taken into account when estimating the variance of each individual gene. This increases the degrees of freedom, making inference more reliable even when sample sizes are small. Limma can be used for a wide variety of experimental designs, making it a common choice for DE analysis.

2.2.9. SAM-seq

Significant analysis of microarrays (SAM) is a method for analyzing microarray data that was proposed in 2001 (Tusher et al., 2001). First, a test statistic is calculated for each of the m genes. These statistics are then ordered from smallest to largest. This process is repeated for all B possible permutations of the data. For each gene, the expected relative difference is found by averaging the statistics across all B permutations. Genes are declared to be differentially expressed (DDE) when the absolute value of the difference between the statistic and the expected relative difference is greater than some value delta. Delta is chosen by the user based on the desired cutoff for controlling the FDR. The SAM method has since been extended to be used for RNA-seq data in addition to microarray data.

2.2.10. Comparison of Methods

Research has been done to determine appropriate methods for the analysis of two-color microarray experiments. In 2004, Smyth simulated data in order to compare various methods of DE analysis. He found that the moderated t-statistic implemented by the limma package outperformed the simple two-sample t test (Smyth, 2004). Similarly, a study conducted in 2010 used simulated data to compare methods of analysis (Jeanmougin et al., 2010). It was discovered that limma was superior to the t-test because it consistently had higher power and lower FDR. SAM had weak performance compared to limma. Although SAM effectively controlled FDR, it had low power when the sample size was small. Based on this research, there is some consensus that limma performs reasonably well when it comes to DE analysis. This is particularly true when sample sizes are small. As sample sizes increase, other methods may also perform adequately.

2.3. RNA-seq Analysis

2.3.1. Filtering

As previously mentioned, filtering is important in the data analysis of gene expression experiments. It can help to reduce the dimensionality of the data, thereby increasing the power to detect differentially expressed genes. Various methods of filtering RNA-seq data sets have been proposed. The DESeq2 package automatically filters the data prior to analysis. It removes genes with mean normalized count below a certain threshold. The default chooses a threshold that maximizes the number of genes that are DDE (Love et al., 2014). The edgeR package filters with the function filterByExpr, which utilizes the method proposed by Chen and Smyth (2016). Genes must have a certain degree of expression before they are likely to be translated into a protein. Therefore, genes that are minimally expressed across treatments can be removed, as they are

unlikely to display significant differential expression. Chen and Smyth recommend that genes with a count of less than 10-15 across all treatments be removed before subsequent analysis. However, it is best to filter based on counts per million in order to account for differences in library size. Another consideration is whether filtering should be done before or after normalization. Some research has indicated that filtering after normalization yields more accurate results (Lin et al., 2016). The parameters in certain normalization methods are estimated based on read count. If filtering is done prior to this estimation, the results can be changed. Research is still being done to determine which filtering methods are best, and whether filtering should be performed before or after normalization.

2.3.2. EdgeR

EdgeR is a bioconductor package that was initially developed for SAGE (serial analysis of gene expression). However, it is used to analyze count data, and therefore has applications to RNA-seq data analysis (Robinson et al., 2009). The edgeR package assumes that data are modeled with a negative binomial distribution, which is appropriate for experiments with biological replication. This is because biological replication tends to result in overdispersion, meaning that the variability of the response is greater than expected based on the mean. The negative binomial distribution accounts for this overdispersion. The model is as follows:

$$Y_{gi} \sim NB(M_i P_{gi}, \varphi_g)$$

Where M_i is the library size, P_{gi} is the relative abundance of the g^{th} gene from the i^{th} sample and φ_g is the dispersion parameter. This reduces to a Poisson distribution when the dispersion parameter is equal to zero (Robinson et al., 2009). Dispersions are estimated using a conditional

maximum likelihood and are shrunken to a common value using an empirical Bayes approach. An exact test similar to Fisher's exact test is used to determine differential expression.

One normalization method that is used by the edgeR package is the Trimmed mean of M-values (TMM) method, proposed by Robinson and Oshlack in 2010. This method accounts for the library size variation between samples. Both M-values and A-values are trimmed, and then a weighted average is found using precision weights (Robinson and Oshlack, 2010). This has shown to be an effective normalization method for RNA-seq experiments. Other options in the edgeR package include TMMwsp, RLE, and UQ. TMMwsp (TMM with singleton pairing) is a modification of the TMM method, which is useful for data that contain a large number of zeros. In the RLE (relative log expression) method, a scaling factor is calculated by taking the median ratio of each sample to the median library. In the upper quartile (UQ) method, reads with no counts are removed, and the 75th percentile of remaining counts are used to calculate the scale factors.

2.3.3. DESeq

DESeq was originally proposed by Anders and Huber in 2010. It is another method for testing for differentially expressed genes between treatment groups and is available for use in the R bioconductor package. Similar to edgeR, it assumes that the data are modeled by a negative binomial distribution in order to account for the overdispersion. However, the method by which the normalization factors are estimated differs from edgeR. DESeq uses the median of ratios method, which utilizes the following formula:

$$\hat{s}_j = median_i \frac{k_{ij}}{(\Pi_{v=1}^{m} k_{iv})^{1/m}}$$

In the above formula, k_{ij} stands for the read count from the i^{th} gene in the j^{th} treatment group. The read count for each gene is divided by the geometric mean of all read counts from

that gene. Once that ratio is found, the median is calculated and used as the normalization factor. Expression strength and variance are also estimated. Data from genes with similar expression strength are pooled to obtain a better estimate of the variance. When testing for differential expression, DESeq uses an exact test that is similar to the method used in edgeR.

2.3.4. DESeq2

DESeq2 was proposed as an improvement from DESeq in 2014. It uses the same median of ratios normalization method as DESeq. One key difference is that DESeq2 is meant to test for the strength of, rather than just the presence of, differential expression. This is done by using shrinkage estimation for dispersion and fold changes. Accurate dispersion estimation is very important, however it can be difficult due to the small number of replicates used in gene expression experiments. When estimating dispersions, the initial estimates are calculated, as well as the sample means of normalized counts. A smooth curve is fit, which provides an estimate of expected dispersion values (Love et al., 2014). The gene-wise dispersion estimates are then shrunk toward the predicted values, giving a final dispersion estimate. DESeq2 also shrinks log fold changes toward zero, with shrinkage being stronger for genes with lower read counts. This is because ratios can be noisier for weakly expressed genes. When testing for differential expression, DESeq2 uses a Wald test, which uses the LFC estimates. This differs from the exact tests which are used by DESeq and edgeR.

2.3.5. Limma-trend

The limma package (linear models for microarray) was originally proposed for analyzing microarray data. However, its methods have been extended, and limma can now be used for the analysis of RNA-seq data. This is done by converting count data into log counts per million and proceeding with analysis as if it were microarray data. With the limma-trend method, the mean-

variance relationship is estimated using an empirical Bayes approach (Ritchie et al., 2015). A trendline is fit through a scatterplot of the square root of the standard deviation vs. average logCPM. This is used to estimate the prior variance for each gene. Prior degrees of freedom are also calculated, and a moderated t-test is used to determine differential expression.

2.3.6. Limma-voom

Like limma-trend, limma voom converts count data into logCPMs and proceeds with analysis as if it were microarray data. However, unlike limma-trend, precision weights are used to model the mean-variance relationship. Voom models the mean-variance trend of the logCPM values at the individual observation level, rather than at the gene level. This can be beneficial, because count sizes may vary significantly for the same gene across samples (Law et al., 2014). The mean-variance trend is first estimated at the gene level, and then is interpolated to predict variances at the individual observation level. The inverse squared predicted standard deviation is used as the precision weight for each observation (Law et al., 2014). Limma voom and limma-trend have both shown to effectively control for type one error rate. Voom is more powerful when library sizes vary significantly between samples. LogCPM values are normalized for sequencing depth, but other normalization methods can also be performed. The bioconductor package recommends using scale normalization, as well as the TMM normalization method implemented by edgeR (Bioconductor, also Law et al).

2.3.7. Comparison of Methods

Research has been done to determine which methods are best for differential expression analysis in RNA-seq experiments. Performances are evaluated based on factors such as power for determining DE genes, as well as false discovery rate. One study conducted in 2014 used a simulated dataset to compare the methods of edgeR, voom, DESeq and DESeq2. They found that

all methods performed adequately, but edgeR performed best (Williams et al., 2014). Another study conducted in 2017 compared the methods of limma, edgeR and DESeq2. A dataset was used in which 1001 genes were known to be DE. They determined that DESeq2 was best at discovering DE genes, but edgeR and limma both performed adequately (Sahraeian et al., 2017). A third study used "gold standard" analysis, in which all experimental units were first used to determine DE genes. They then resampled using smaller samples of various sizes to compare various methods for determining differential expression. The researchers found that edgeR, limma, DESeq and DESeq2 all performed excellently (Schurch et al., 2016). There is some consenses that edgeR and DESeq2 perform reasonably well. However, based on current research, there is still no "best" method for the analysis of RNA-seq experiments. Many factors can affect which method is most appropriate, such as the design of the experiment, the number of replicates per treatment group, the amount of biological variation present, etc. When determining sample size, it is best to use as many biological replicates as possible. Gene expression experiments usually do not have a large number of replicates, due to the high expense and the lack of resources and funding. Schurch et al. recommend having at least 6 replicates per treatment group, whereas Williams et al. recommend using a sample size of at least 3 per treatment group.

3. METHODS

3.1. Two-Color Microarray

Data was used from a study that involved 28 privately owned dogs of ages ranging from 1 to 8 years old. Eight of the dogs served as healthy controls, and twenty of the dogs were diagnosed with atopic dermatitis. Blood samples were collected, and microarray sequencing was performed. (Majewska et al., 2016). A two-color reference design was used for this experiment, with the common reference sample consisting of a pool of RNA from 13 healthy dogs. The common reference sample was dyed with cy3, and the samples from the 28 dogs in the study were dyed with cy5. Twenty-eight two-color microarrays were performed. Prior to analysis, lowess normalization was performed and the data was filtered. Genes without expression were removed. For the twenty slides from the investigative samples, the combined median signal intensity (across all channels) was calculated. Similarly, for the eight slides with samples from healthy dogs, the combined median signal intensity (across all channels) was calculated. Genes were removed if both of these medians were less than 100. The limma package in R was used to test for differential expression. The linear model for this experiment is as follows:

$$Y_{ijk} = \mu + \tau_i + \delta_j + s_k + e_{ijk}$$

In this formula, Y_{ijk} represents the normalized signal intensity (NSI) of the dog in the i^{th} treatment group, hybridized with the j^{th} dye on k^{th} slide. μ represents overall mean NSI, τ_i represents the effect of the i^{th} treatment (i = 1,2), δ_j represents the effect of the j^{th} dye (j = 1,2), s_k represents the random effect of the k^{th} slide (k = 1,2, … , 28) and e_{ijk} represents the random error for Y_{ijk}.

We would like to test the following set of hypotheses for each gene:

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

Where $\mu_i = \mu + \tau_i$ from the model above. μ_i represents the population mean expression value from the i^{th} treatment group.

Data was accessed on the Gene Expression Omnibus data repository under the number GSE76119 (Edgar et al., 2002). Analysis was performed in R using the limma package. The data was analyzed multiple times with different normalization and filtering methods, and the number of DDE genes were compared. An FDR of 0.085 was used, as that was the significance level chosen by the researchers in the original paper. Various combinations of the following normalization methods were used: lowess, scale, median and Gquantile. The different filtering methods include the method discussed above (Method 1), and an additional filtering method in which genes with median signal intensity less than or equal to the background intensity were removed (Method 2). This method was chosen because it was implemented in a similar experiment in which dogs were under observation (Thomson et al., 2005). The analysis was repeated with filtering performed both before and after normalization, and comparisons were made. In addition to the limma method, the analysis was also repeated with a standard t-test and SAM-seq.

3.2. RNA-seq

RNA-seq data was used from a study that compared sex differences in rats after nerve injury (Stephens et al., 2017). Data was accessed on the Gene Expression Omnibus data repository under the number GSE100122 (Edgar et al., 2002). A total of eight rats were used in the experiment. There were four female rats and four male rats, and the rats of each sex were

randomly assigned to two treatment groups: injured and naive. The rats in the injured group received CCI surgery to the sciatic nerve, while the naive group served as a control. All surgeries were conducted by the same individual as to avoid variations in technique. All animals were euthanized 14 days post-op, and RNA was extracted for next generation sequencing. This is a two-factor factorial experiment, with the two factors being gender and status. Data was analyzed in R using various bioconductor packages, including edgeR, DESeq, DESeq2 and Limma. There were four comparisons of interest: female injured (FI) vs female naïve (FN), male injured (MI) vs. male naïve (MN), female naïve vs. male naive, and the interaction term. The interaction corresponds to the difference between female and male rats in terms of their response to injury. The analysis was repeated with a variety of different filtering methods. For filtering method A, the filterbyexpr function was used in R. This method was chosen because it is commonly used in RNA-seq experiments. As previously discussed, this function implements the filtering method recommended by Chen in 2016. The number of genes was reduced from 32,623 to 16,988 after filtering. For filtering method B, genes that did not have counts per million (CPM) greater than or equal to 2 in at least 2 samples were removed. This method was chosen because it was implemented in an experiment with a similar design that was conducted in mice (Li et al., 2019). The number of genes remaining after filtering was 13,583. The analysis was repeated with filtering performed both before and after normalization, and comparisons were made. An FDR of 0.05 was used, as that was the significance level chosen by the researchers in the original paper.

Different methods will be compared by looking at the number of genes that are DDE in each case. This will give an idea as to which techniques are more conservative/liberal. Additionally, the lists of DDE genes resulting from the application of each method will be

compared. Venn diagrams were created to visually assess the overlap in DDE genes among the various methods.

4. RESULTS AND DISCUSSION

4.1. Two-Color Microarray

From figures 1 and 2, we can see MA plots before and after lowess normalization. Before lowess normalization, we can see that the vast majority of the points lie above the M=0 line. This indicates that in most cases, the red signal intensities were higher than the green. Following lowess, there is random scatter around the M=0 line. This demonstrates that lowess normalization corrected for the red-green differences in signal intensities.

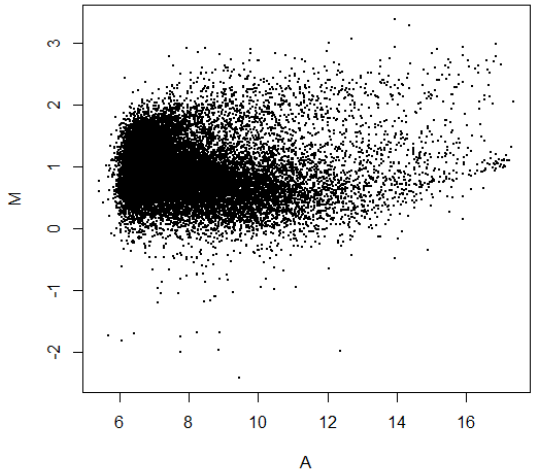

Figure 1. MA plot before lowess normalization.

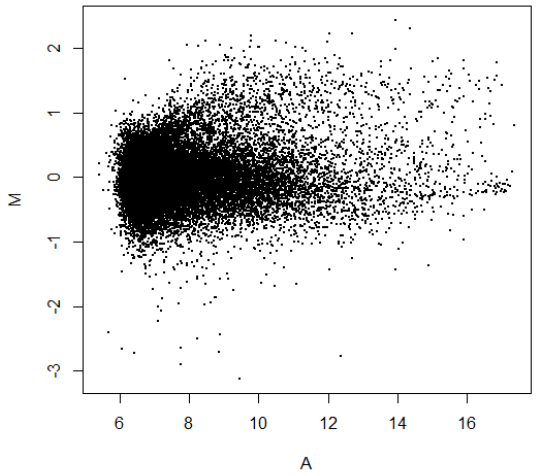

Figure 2. MA plot after lowess normalization.

The data analyzed in this study is from a real experiment, and therefore the number of genes that are truly differentially expressed is not known. For this reason, it can't be determined which normalization method performed best. However, we can see that the method of normalization used did impact the number of genes that were declared differentially expressed. Table 2 shows the number of DDE genes for the various filtering and normalization methods. When no filtering or normalization was performed, an extremely large number of genes were DDE. This could be partially due to the fact that no lowess normalization was performed. Without lowess normalization, there is no correction for the red-green differences. The genes may be DDE because of non-biological variation rather than true biological differences.

Referring to Figure 3, we can see that when the same filtering method was used, there was overlap in the DDE genes among the different methods. The results were more conservative when more normalization methods were used. For example, when no lowess normalization was performed, 9,178 genes were DDE. When only lowess and scale normalization were performed, this number was reduced to 110 genes. When lowess, median, scale and Gquantile normalization were all performed, 59 genes were DDE. For comparison, following lowess normalization with the limma package, SAM-seq was used to test for differential expression and 338 genes were DDE. The Venn diagram shows that these same 338 genes were DDE in at least one other method, indicating that there is some consistency between SAM-seq and limma. A t-test was also performed gene by gene, which result in 0 DDE genes.

Using the q-value function in R, an estimate of the proportion of equivalently expressed genes was obtained. This was determined to be approximately 0.897, meaning that roughly 10% of the genes remaining after filtering were expected to be differentially expressed. With the variety of methods used for analysis, most resulted in 0 DDE genes. This is because there is not a

very high degree of differential expression present in this dataset. Even genes that truly are differentially expressed may have small differences that are difficult to be noticed in analysis. Additionally, there is relatively low power for detecting differentially expressed genes due to the large number of hypothesis tests being conducted. If this process was repeated for a dataset with a larger degree of differential expression, more meaningful inferences could be made.

Although the "ground truth" is not known, there were three methods of analysis that yielded potentially trustworthy results; SAM-seq, limma with loess and scale normalization, and limma with loess, median, scale and Gquantile normalization. These techniques resulted in 338, 110 and 59 DDE genes, respectively. We believe that there is some amount of differential expression present in this dataset, so any methods that resulted in 0 DDE genes are likely too conservative and are not scientifically useful. Additionally, methods that resulted in a very large number of DDE genes may be unreliable, because there are likely a lot of false discoveries. Therefore, filtering method 1 along with any of the three approaches mentioned above may all be appropriate methods of analysis for this particular dataset.

Table 2

Number of genes DDE for different normalization and filtering methods with an alpha value of .085

Normalization methods used	No Filter (45,210 genes)	Filter Method 1 (20852 genes remain)	Filter Method 1 after normalization (20852 genes remain)	Filter Method 2 (40,244 genes remain)	Filter Method 2 after normalization (40,244 genes remain)
None	30588	9178	9178	28196	28196
Loess	0	0	0	0	0
Loess and Scale	0	110	0	0	0
Median and Scale	0	0	0	0	0
Loess,Median,Scale,Gquantile	0	59	0	0	0

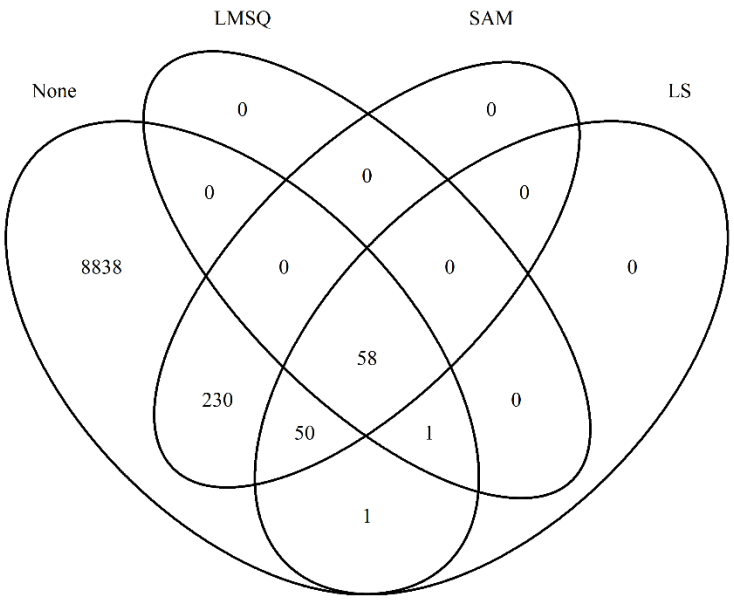

Figure 3. Venn diagram of the number of DDE genes for different normalization methods applied after filtering with filter method 1. LS stands for loess and scale, LMSQ stands for loess, median, scale and Gquantile. An alpha value of .085 was used to determine significance.

4.2. RNA-seq

As before, the data analyzed in this study is from a real experiment, and therefore the number of genes that are truly differentially expressed is not known. However, we can see that the method of normalization used did impact the number of genes that were declared differentially expressed. Results for each comparison of interest can be found below.

4.2.1. FI vs FN

For the comparison of the female injured versus female naive groups, approximately 2000 genes were declared differentially expressed (Table 3). As previously discussed, three different methods of filtering were performed; no filtering, method A, and method B. With these three methods, there were 32623, 16988, and 13583 genes remaining, respectively. Across all

normalization methods, no filtering resulted in the smallest number of genes that were DDE. This is as expected, because fewer genes remaining means fewer hypothesis tests are conducted. This results in higher power for detecting differential expression. When the same filtering method was used after normalization, fewer genes were DDE. DESeq2 resulted in the highest number of DDE genes, regardless of which filtering method was used. This could be due to asymmetric distribution of effect sizes, which will be discussed in greater detail later on. From figure 4, we can see that most DDE genes are shared between all methods. Specifically, there are 1267 genes that were DDE regardless of which method was used. Therefore, we can be reasonably confident that these 1267 genes truly are differentially expressed. DESeq2 differed the most, as there were 503 genes that were not DDE with any other method. From figure 5, we can see that there was a lot of overlap between the various edgeR normalization methods. Although there were differences among the various methods of analysis, it is reassuring to see that there is some overlap in the lists of DDE genes. This demonstrates that it may be useful for researchers to use multiple methods of analysis and look for consistency among the results.

Table 3

Number of genes DDE for the FI vs FN comparison of interest

FI vs FN	No Filtering (32,623 genes remain)	Filter method A (16,988 genes remain)	Filter method B (13,583 genes remain)	Filter method A after normalization (16,988 genes remain)
DESeq2	2619	2624	2752	-
edgeR w/TMM	1522	2107	2114	1970
edgeR w/TMMwsp	1510	2109	2116	1926
edgeR w/RLE	1545	2098	2099	1966
edgeR w/upperquartile	1547	2113	2131	2009
edgeR w/none	1602	2231	2240	2107
Limma trend	633	1917	2111	1891
Limma voom	1187	1598	1767	1552
DESeq	1511	1547	1628	1700

Note. Various normalization and filtering methods were applied, and differential expression was determined at the 0.05 significance level.

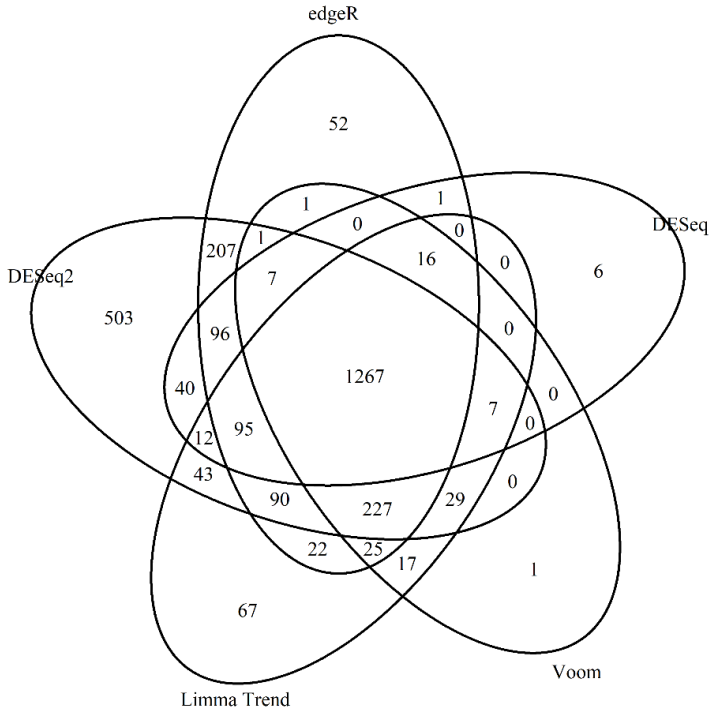

Figure 4. Venn diagram of DDE genes for FI vs FN comparison.

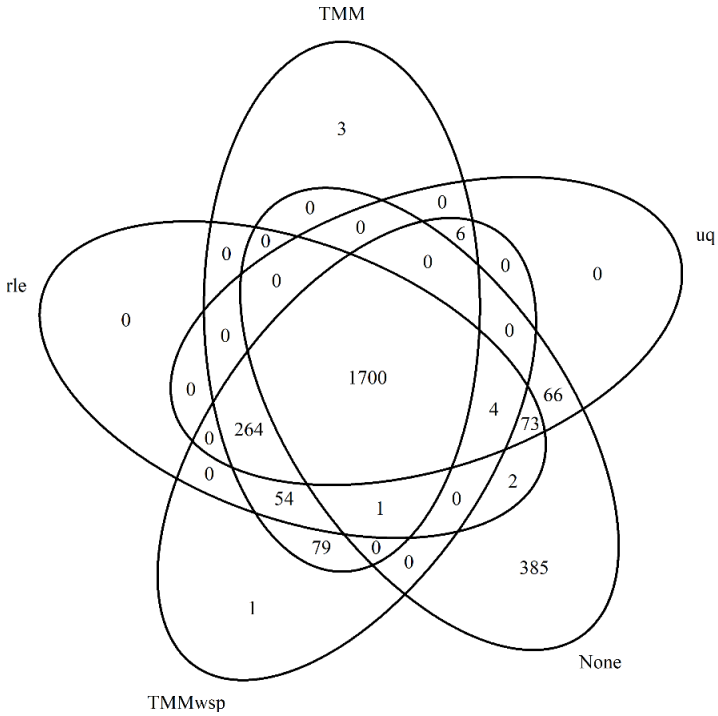

Figure 5. Venn diagram of the number of genes DDE for the FI vs FN comparison of interest. Includes various normalization methods of the edgeR packge; TMM, TMMwsp, upperquartile, RLE and none.

4.2.2. MI vs MN

From table 4, we can see that for the comparison of the male injured versus male naive groups, approximately 6000 genes were declared differentially expressed. Unlike the other comparisons of interest, voom seemed to result in the highest number of DDE genes when compared with the other methods. No filtering resulted in the fewest number of DDE genes, as expected. However, filtering method A resulted in more DDE genes than filtering method B. This is unexpected because there are more genes remaining after filtering with method A. Therefore, we would expect fewer genes to be declared differentially expressed. Upon further investigation, it was determined that 503 of the genes that were filtered out with method B were DDE in method A. This could be one potential explanation for why more genes were DDE with filtering method A. From figure 6, we can see that the majority of DDE genes are shared between all methods. Specifically, there are 5264 genes that were DDE regardless of which method was used. Therefore, we can be reasonably confident that these 5264 genes truly are differentially expressed. Voom differed the most, as there were 204 genes that were not DDE with any other method. From figure 7 we can see that there was a lot of overlap between the various edgeR normalization methods.

Table 4

Number of genes DDE for the MI vs MN comparison of interest

MI vs MN	No Filtering (32,623 genes)	Filter method A (16,988 genes)	Filter method B (13,583 genes)	Filter method A after normalization (16,988 genes)
DESeq2	6181	6230	6170	-
edgeR w/TMM	5210	6066	5707	5894
edgeR w/TMMwsp	5255	6061	5706	5845
edgeR w/RLE	5183	6088	5695	5858
edgeR w/upperquartile	5271	6078	5634	5908
edgeR w/none	5135	6020	5608	5815
Limma trend	4631	6661	6374	6703
Limma voom	6291	6933	6545	6983
DESeq	5380	5441	5345	5758

Note. Various normalization and filtering methods were applied, and differential expression was determined at the 0.05 significance level.

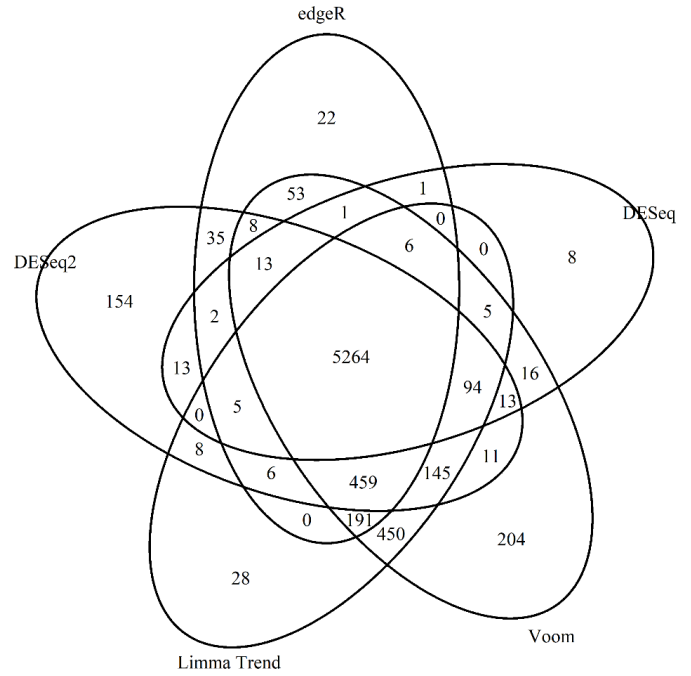

Figure 6. Venn Diagram of DDE genes for MI vs MN.

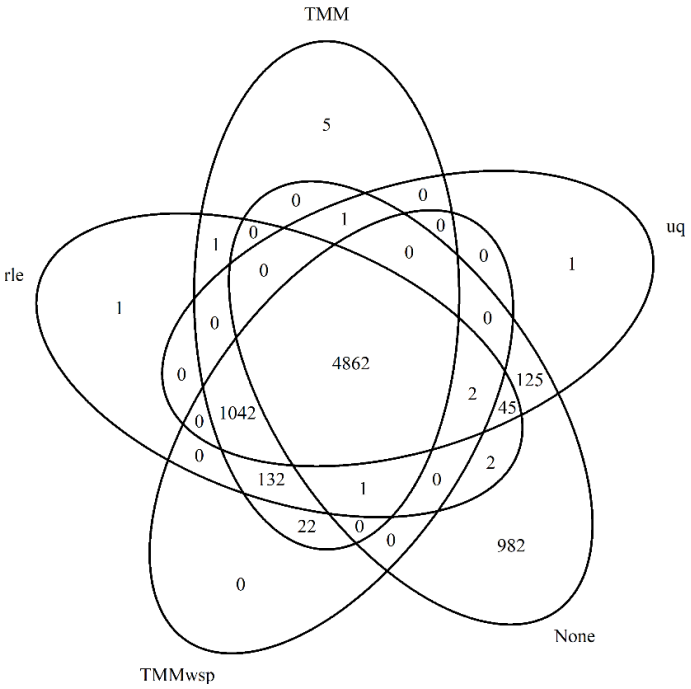

Figure 7. Venn diagram of the number of genes DDE for the MI vs MN comparison of interest. This diagram includes the various normalization methods of the edgeR packge; TMM, TMMwsp, upperquartile, RLE and none.

4.2.3. FN vs MN

For the comparison of the female naive versus male naive groups, approximately 2000 genes were declared differentially expressed (table 5). As expected, the fewest number of genes were DDE when no filtering was performed, followed by method A. Filtering method B resulted in the largest number of DDE genes. Again, fewer genes were found DDE when filtering was performed after normalization. Of all the methods, DESeq2 resulted in the highest number of DDE genes. From figure 8, we can see that there was a lot of overlap in the DDE genes from the various methods. Specifically, there are 1249 genes that were DDE regardless of which method was used. Therefore, we can be reasonably confident that these 1249 genes truly are differentially expressed. DESeq2 differed the most, as there were 566 genes that were not DDE with any other method. Referring to figure 9, we can see that the majority of DDE genes were shared between all edgeR normalization methods.

Table 5

Number of genes DDE for the FN vs MN comparison of interest

FN vs MN	No Filtering (32,623 genes remain)	Filter method A (16,988 genes remain)	Filter method B (13,583 genes remain)	Filter method A after normalization (16,988 genes remain)
DESeq2	2653	2680	2813	-
edgeR w/TMM	1405	2023	2089	1902
edgeR w/TMMwsp	1398	2024	2089	1885
edgeR w/RLE	1453	2060	2104	1943
edgeR w/upperquartile	1473	2076	2092	1974
edgeR w/none	1467	2089	2106	1959
Limma trend	97	1654	1995	1620
Limma voom	574	1860	2147	1820
DESeq	1449	1534	1698	1669

Note. Various normalization and filtering methods were applied, and differential expression was determined at the 0.05 significance level.

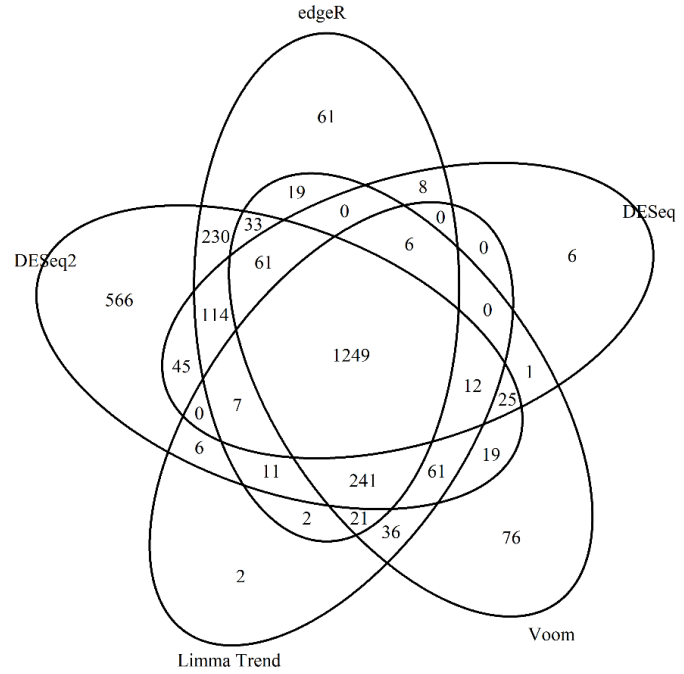

Figure 8. Venn Diagram of DDE genes for FN vs MN.

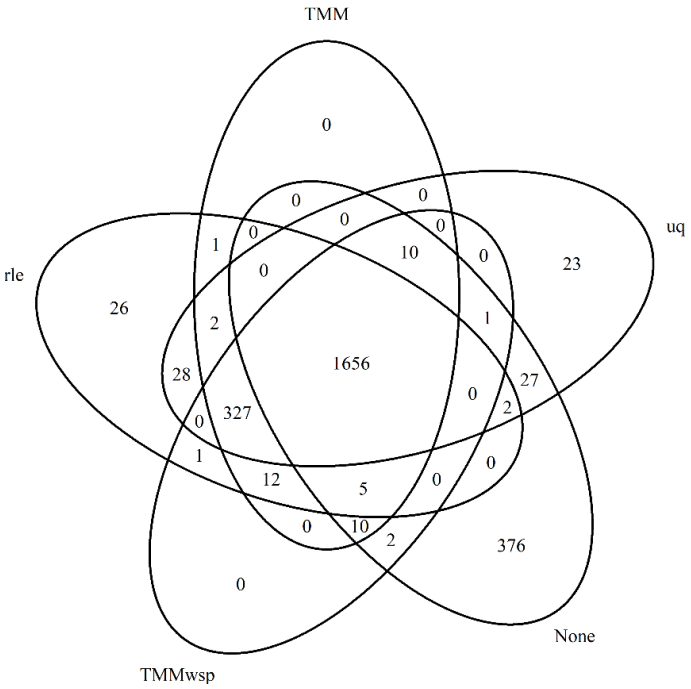

Figure 9. Venn diagram of the number of genes DDE for the FN vs MN comparison of interest. This diagram includes the various normalization methods of the edgeR packge; TMM, TMMwsp, upperquartile, RLE and none.

4.2.4. Interaction

The researchers were interested in comparing the differential expression between male and female rats in terms of their response to injury. This is equivalent to testing the interaction term. Table 6 shows the number of DDE genes for this comparison for various filtering and normalization methods. There is a lot more variation in the results than there was for the previous comparisons of interest. This is likely due to the fact that this is a more complex comparison. DESeq2 by far resulted in the most DDE genes when compared to the other methods. Figure 10 shows that there were 821 genes unique to DESeq2 that were not significant in any other method. Specifically, there are 101 genes that were DDE regardless of which method was used. Therefore, we can be reasonably confident that these 101 genes truly are differentially expressed. For the most part, more genes were DDE when there were fewer genes remaining after filtering. One exception to this is when edgeR was used, more genes were DDE when no filtering was performed. One potential explanation for this is that the quasi-likelihood test was used instead of the exact test that was used for the other comparisons of interest. The quasi-likelihood test is more appropriate in this case, because there are two factors involved. From figure 11 we can see that there was a lot of overlap between the various edgeR normalization methods.

Table 6

Number of genes DDE for the interaction term

Interaction	No Filtering (32,623 genes remain)	Filter method A (16,988 genes remain)	Filter method B (13,583 genes remain)	Filter method A after normalization (16,988 genes remain)
DESeq2	1573	1588	1659	-
edgeR w/TMM	962	669	906	663
edgeR w/TMMwsp	1037	665	906	642
edgeR w/RLE	1024	662	878	662
edgeR w/upperquartile	1040	726	1035	730
edgeR w/none	1063	813	1078	822
Limma trend	117	280	550	272
Limma voom	0	104	463	106
DESeq	557	683	801	732

Note. Various normalization and filtering methods were applied, and differential expression was determined at the 0.05 significance level.

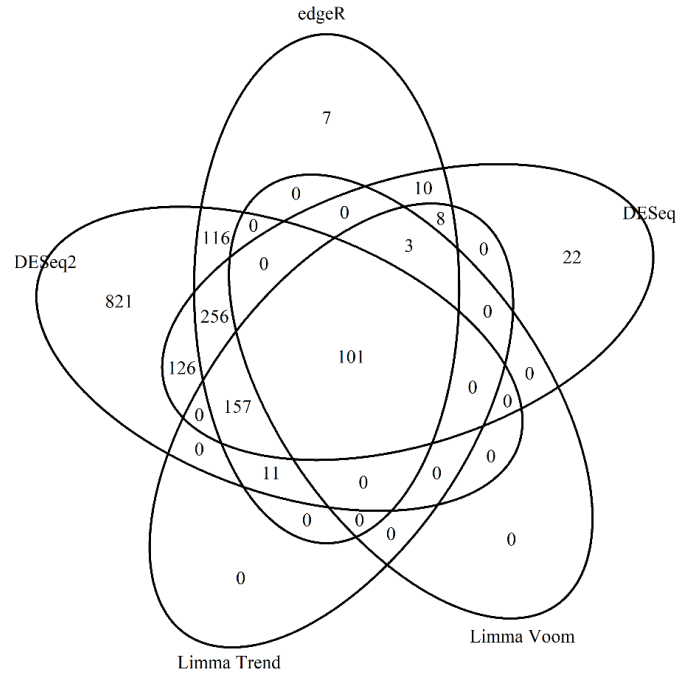

Figure 10. Venn diagram of DDE genes for the interaction term.

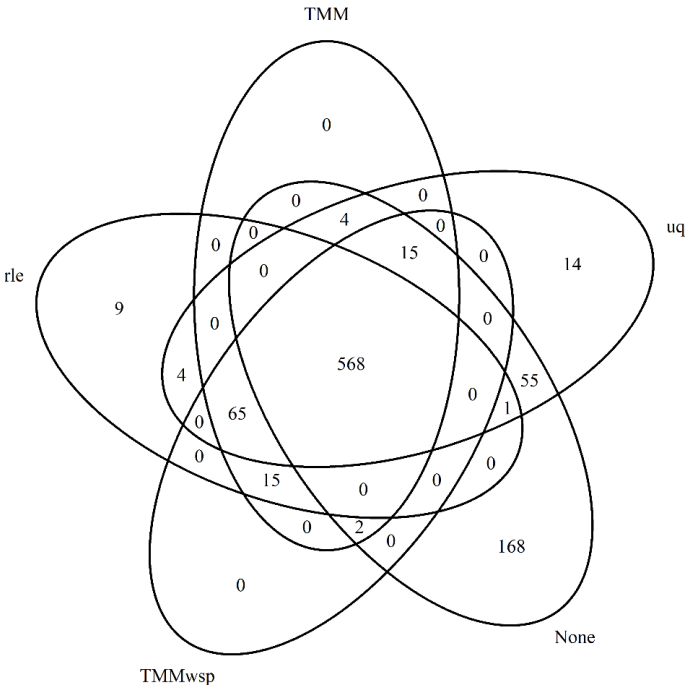

Figure 11. Venn diagram of the number of DDE genes for the interaction term. This diagram includes the various normalization methods of the edgeR package; TMM, TMMwsp, upperquartile, RLE and none.

4.2.5. Asymmetry in the Distribution of Effect Sizes

Using the q-value function in R, the proportion of equivalently expressed genes was estimated for each comparison of interest. As seen in table 7, these proportions are estimated to be approximately 0.3 for the MI vs MN comparison, and approximately 0.5 for all other comparisons of interest. These proportions were used to estimate the proportion of upregulated genes among all DE genes. This was estimated under the assumption that observed effect sizes of EE genes have a symmetric distribution around zero, meaning that half should be positive, and half should be negative. Using the formulas

$$\hat{m}_0 = m\hat{\pi}_0$$

$$UR = P - \frac{\hat{m}_0}{2}$$

$$DR = N - \frac{\hat{m}_0}{2}$$

$$proportion\ of\ upregulated\ genes\ among\ DE\ genes = \frac{UR}{UR + DR}$$

Where $\hat{m}_0$ is the estimated number of EE genes, UR is the estimated number of upregulated genes, DR is the estimated number of downregulated genes, P is the number of genes with a positive observed effect size, and N is the number of genes with a negative observed effect size.

Figures 12-15 show histograms of the difference in average logCPM for all comparisons of interest. Although these histograms do not clearly demonstrate asymmetry, there are more negative than positive differences in all cases. For example, for the FI vs. FN comparison, there are 10,240 positive differences and 6,748 negative differences, indicating asymmetry. Additionally, the proportions of upregulated genes among DE genes was found to be 0.700 for FI vs. FN, 0.302 for MI vs. MN, 0.334 for FN vs. MN and 0.657 for the interaction term. Since these values differ from 0.5, the distribution of effect sizes is asymmetric for all comparisons of

interest. The results of a previous study indicate that when the distribution of effect sizes is asymmetric, the FDR is elevated when DESeq2 is used. When the FDR was supposed to be controlled at 5%, in many cases the actual FDR was elevated above 20% (Kotoka and Orr, 2017). This is a potential explanation as to why DESeq2 generally had a much higher number of DDE genes when compared to other methods.

Table 7

The estimated proportion of equivalently expressed genes, as well as the proportion of upregulated among differentially expressed genes for each comparison of interest

Comparison	Estimated proportion of EE genes	Proportion of upregulated among DE genes
FI vs FN	0.487	0.700
MI vs MN	0.374	0.302
FN vs MN	0.498	0.334
Interaction	0.494	0.657

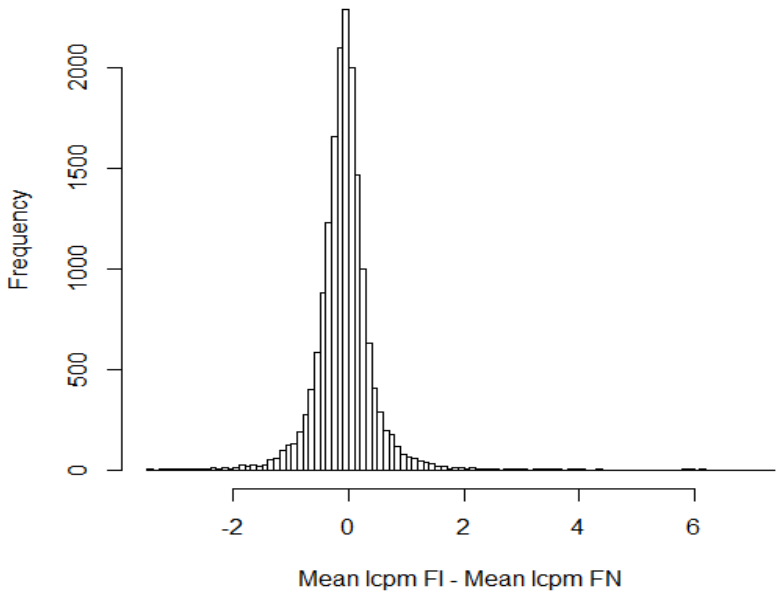

Figure 12. Histogram showing the difference between the mean log counts per million for female injured vs female naïve.

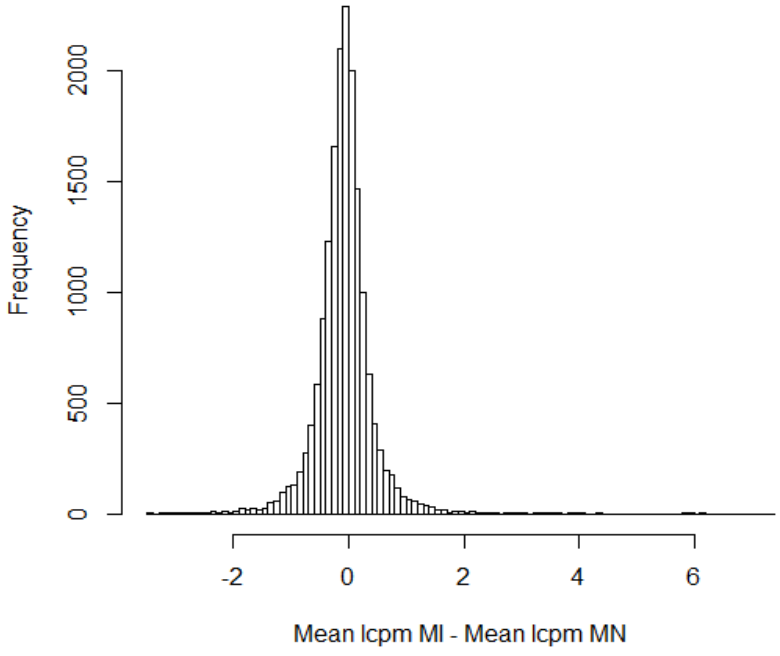

Figure 13. Histogram showing the difference between the mean log counts per million for the male injured vs male naïve groups.

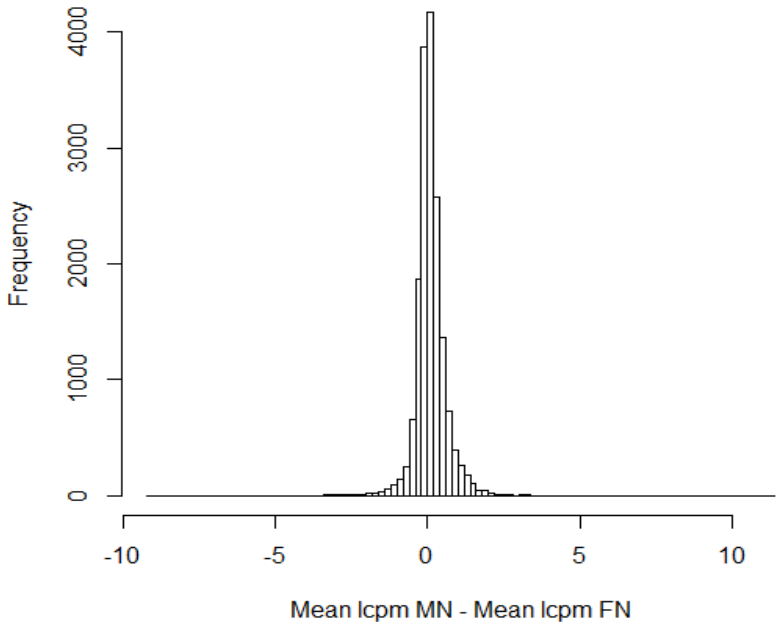

Figure 14. Histogram showing the difference between the mean log counts per million for the male naïve vs female naïve groups.

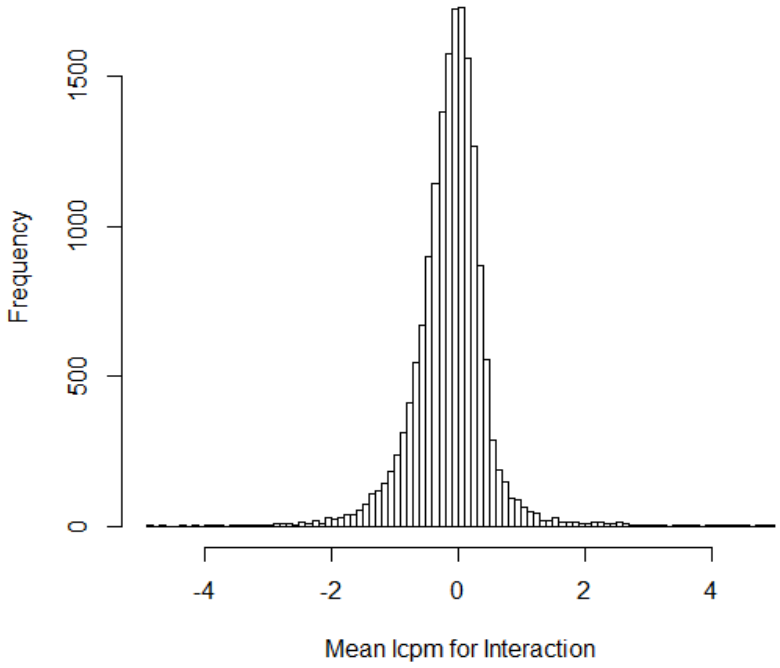

Figure 15. Histogram of the mean log counts per million of MI-MN vs the mean log counts per million of FI-FN.

5. CONCLUSION AND FUTURE RESEARCH

5.1. Conclusion

There is not a widely accepted method for the analysis of microarray and RNA-seq experiments, and it is clear to see why. The small number of replicates in these experiments, along with the very large number of genes make these datasets difficult to analyze. It is considered common practice for the data to be filtered and normalized, but the most appropriate method for doing so may vary from experiment to experiment. Factors such as the complexity of the experimental design, the organism on which the experiment is performed, as well as the number of replicates should all be considered before choosing a normalization method. The results of this study show that which normalization and filtering methods are used can largely impact the number of genes that are declared differentially expressed in gene expression experiments. A wise approach may be to use multiple methods for analysis and look for consistency among the results.

5.2. Future Research

In continuing research on this topic, I would consider using simulated data to compare normalization methods. That way the actual performance of the methods could be determined and compared. The true power for detecting differentially expressed genes, as well as the type one error rate could be observed. I would also consider repeating the two-color microarray experiment on a dataset that has a larger degree of differential expression. It would be interesting to see how the normalization methods differ when there is a larger expected proportion of differentially expressed genes between treatment groups.

REFERENCES

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biology*, 11:R106

Benjamini Y and Hochberg Y. (1995). Controlling the False Discovery Rate: A Practical and Power Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B, 57, 289-300

Bolstad, B. M., Irizarry R. A., Astrand, M., and Speed, T. P. (2003), A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 19, 185-193.

Chen Y, Lun ATL, and Smyth, GK (2016). From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi likelihood pipeline. F1000Research 5, 1438. http://f1000research.com/articles/5-1438

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. JASA 74 829-836.

Edgar R, Domrachev M, Lash AE.
Gene Expression Omnibus: NCBI gene expression and hybridization array data Repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10

Edwards, D. E. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies Bioinformatics 19, 825-833.

Hoen, P. A. C. t. (2004). Intensity-based analysis of two-colour microarrays enables efficient and flexible hybridization designs. *Nucleic Acids Research*, *32*(4). doi: 10.1093/nar/gnh038

Jeanmougin, M., Reynies, A. D., Marisa, L., Paccard, C., Nuel, G., & Guedj, M. (2010). Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLoS ONE*, *5*(9). doi: 10.1371/journal.pone.0012336

Kaliyappan, K., Palanisamy, M., Govindarajan, R., & Duraiyan, J. (2012). Microarray and its applications. Journal of Pharmacy and Bioallied Sciences, 4(6), 310. doi: 10.4103/0975 7406.100283

Kotoka, E., & Orr, M. (2017). Modifying SAM-seq to account for asymmetry in the distribution of effect sizes when identifying differentially expressed genes. *Statistical Applications in Genetics and Molecular Biology*, *16*(5-6). doi: 10.1515/sagmb-2016-0037

Law, CW, Chen, Y, Shi, W, Smyth, GK (2014). Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology 15, R29. http://genomebiology.com/ 2014/15/2/R29

Li, Q., Cheng, Z., Zhou, L., Darmanis, S., Neff, N., Okamoto, J., Gulati, G., Bennett, M., Sun, L., Clarke, L., Marschallinger, J., Yu, G., Quake, S., Wyss-Coray, T. and Barres, B., 2019. Developmental Heterogeneity of Microglia and Brain Myeloid Cells Revealed by Deep Single-Cell RNA Sequencing. *Neuron*, 101(2), pp.207-223.e10.

Lin, Y., Golovnina, K., Chen, Z. *et al.* Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual *Drosophila melanogaster* . *BMC Genomics* 17, 28 (2016) doi:10.1186/s12864-015-2353-z

Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15:550. https://doi.org/10.1186/s13059-014-0550-8

Majewska, A., Gajewska, M., Dembele, K., Maciejewski, H., Prostek, A., & Jank, M. (2016). Lymphocytic, cytokine and transcriptomic profiles in peripheral blood of dogs with atopic dermatitis. *BMC Veterinary Research*, *12*(1). doi: 10.1186/s12917-016-0805-6

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., & Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, *18*(9), 1509–1517. doi: 10.1101/gr.079558.108

Ritchie, ME, Phipson, B, Wu, D, Hu, Y, Law, CW, Shi, W, and Smyth, GK (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 43, e47. http://nar.oxfordjournals.org/content/43/7/e47

Ritchie, M. E., Silver, J., Oshlack, A., Silver, J., Holmes, M., Diyagama, D., Holloway, A., and Smyth, G. K. (2007). A comparison of background correction methods for two-colour microarrays. Bioinformatics 23, 2700-2707. http://bioinformatics.oxfordjournals.org/content/ 23/20/2700

Robinson, M. D., Mccarthy, D. J., & Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. doi: 10.1093/bioinformatics/btp616

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology, 11(3). doi: 10.1186/gb-2010-11 3-r25

Sahraeian, S.M.E., Mohiyuddin, M., Sebra, R. *et al.* Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nat Commun* **8,** 59 (2017). https://doi.org/10.1038/s41467-017-00050-4

Schurch, N.J. et al. (2016) How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA, 22, 839-851.

Smyth, G. K., and Speed, T. P. (2003). Normalization of cDNA microarray data. Methods 31, 265-273.

Smyth, G. K. (2004). Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statistical Applications in Genetics and Molecular Biology*, *3*(1), 1–25. doi: 10.2202/1544-6115.1027

Stephens, K. E., Zhou, W., Ji, Z., He, S., Ji, H., Guan, Y., & Taverna, S. D. (2017). Sex differences in gene regulation in the dorsal root ganglion after nerve injury. doi: 10.1101/152652

Quackenbush J: Microarray data normalization and transformation. Nature Genetics 2002, Suppl 32:496-501

Thomson, S. A. M., Kennerly, E., Olby, N., Mickelson, J. R., Hoffmann, D. E., Dickinson, P. J., … Breen, M. (2005). Microarray Analysis of Differentially Expressed Genes of Primary Tumors in the Canine Central Nervous System. *Veterinary Pathology*, *42*(5), 550–558. doi: 10.1354/vp.42-5-550

Tusher, V.G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences. 98, 5116-5121.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-seq: A revolutionary tool for transcriptomics. Nature Reviews.Genetics, 10(1), 57-63. doi:http://dx.doi.org.ezproxy.lib.ndsu.nodak.edu/10.1038/nrg2484

Williams, A.G. et al. (2014) RNA-seq Data: Challenges in and recommendations for experimental design and analysis. Current Protocols in Human Genetics, Unit 11.3.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cDNA microarray data. In Microarrays: Optical Technologies and Informatics, M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds), Proceedings of SPIE, Volume 4266, pp. 141-152.

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

Yang, Y. H., and Thorne, N. P. (2003). Normalization for two-color cDNA microarray data. In: D. R. Goldstein (ed.), Science and Statistics: A Festschrift for Terry Speed, IMS Lecture Notes - Monograph Series, Volume 40, pp. 403-418.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE*, *9*(1). doi: 10.1371/journal.pone.0078644