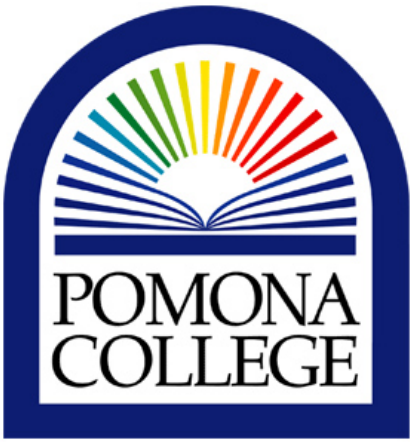


Senior Thesis in Mathematics

Differential Gene Expression Analysis with Microarray and RNA-seq Data

Author:
Jacob Fiksel

Advisor:
Dr. Jo Hardin

Submitted to Pomona College in Partial Fulfillment
of the Degree of Bachelor of Arts

April 3, 2015

Abstract

Over the past 70 years, scientists have focused in on genetic mutations as the cause of many human diseases. However, the human genome consists of 20,000-25,000 protein coding genes [1], making it very difficult to determine the exact mutation that causes each disease. Within the past 30 years the growing use of microarrays has lead to a decrease in the cost of measuring gene expression, both in time and money. More recently, researchers have developed next generation sequencing (NGS) technologies, such as RNA-sequencing (RNA-seq) to obtain a more exact picture of an individual's gene expression levels. Studies which use sequencing data often have a low sample size, which, combined with the high-dimensional nature of genetic data, makes it extremely difficult to make inferences on whether a gene is differentially expressed. To combat this problem, Smyth [2] uses a linear model (LIMMA) to estimate differences in gene expression for each gene amongst different samples measured with microarrays. To improve the estimate of the coefficient variation, Smyth employs an empirical Bayesian hierarchical model to create the moderated t-statistic. More recently, Law *et al.* [3] developed the "variance modeling at the observation level" (VOOM) method to analyze RNA-seq data in the LIMMA pipeline. In a comparison of microarray and RNA-seq data take from the same samples in a study designed to compare kidney and liver cell gene expression, I find that RNA-seq with VOOM is more powerful in its ability to detect differentially expressed genes than microarray with LIMMA.

Contents

Chapter 1

Introduction

The 20,000 protein coding genes that that make up the human genome [1] contain the recipe for the complex network of neurons and tissues that is the human body. Although genetic variation has given rise to life and the human species through evolution, it is also the cause of many deadly diseases. For example, mutations of the BRCA1 and BRCA2 genes are heavily associated with breast and ovarian cancer [4]. Furthermore, Diggs-Andrews *et al.* [5] have hypothesized that neurofibromatosis type 1 presents itself differently based on genetic differences linked to the sex of the patient. In order to give doctors the ability to utilize gene therapy to prevent complications resulting from neurofibromatisis type 1, along with other diseases, it is necessary to identify the exact genes that are over or under-expressed in patients with the disease.

Two commonly used methods to measure genetic expression are mRNA microarrays and RNA-sequencing. Microarrays quantify the expression of a gene by measuring the fluorescence of mRNA labeled with dye that hybridize to the complementary base pairs of known genes. If we are measuring the difference in gene expression between male and female mice with neurofibromatosis type 1, we can obtain microarray data for a male mouse and a female mouse. We then obtain a vector of fluorescence levels for each gene, $y_g^T = (y_{g1}, y_{g2}, y_{g3}, y_{g4})$, where y_{g1} and y_{g2} are the responses from male samples, and y_{g3} and y_{g4} female samples. Using this data, we create a linear model, $y_g^T = \alpha_1 X_1 + \alpha_2 X_2 + \epsilon_i$, where X_1 and X_2 are indicator variables for whether sample y_{gi} came from a male or female, respectively. α_{g1} and α_{g2} are then estimated to create the statistic $\hat{\beta}_g = \hat{\alpha}_{g1} - \hat{\alpha}_{g2}$. We then wish to make an inference on $\hat{\beta}_g$, the estimated difference between male and female

expression.

A common statistical test to determine whether or not $\beta_g = 0$ is a standard t-test. However, microarrays typically have a low number of replicates, which leads to a noisy estimate of the variance of the estimate of β_g, $\hat{\beta}_g$, and thus poor performance from the standard t-test [6].

To work around this problem, Smyth assumes a common distribution on the variance of $\hat{\beta}_g$, σ_g^2, and uses an empirical Bayesian hierarchical model to borrow information across genes to create a moderated t-statistic, $\tilde{t}_g$. Under the null hypothesis, $H_0 : B_g = 0$, $\tilde{t}_g$ is shown to follow a standard t-distribution with added degrees of freedom compared to the standard t-statistic, leading to higher power and a lower false discovery rate.

Despite being commonly used, microarrays force researchers to decide which genes to include on the microarray chip, which does not allow researchers to discover novel potential genes of interest. RNA-sequencing avoids this problem by sequencing small parts of every mRNA in the sample and mapping reads to a reference gene library to quantify expression. However, the discrete nature of RNA-seq data makes it more difficult for researchers to accurately detect differential gene expression. While several statistical methods have been developed to detect differential expression in RNA-seq data, several papers have identified the VOOM as the best method for analyzing RNA-seq data [3] [7] [8] [9].

I plan to compare RNA-seq and microarray data taken from two previous studies that have made their data publicly available. The first dataset is taken from 21 inbred mice, with the dataset containing 10 C57BL/6J (B6) mice and 11 DBA/2J (D2) mice [10]. The second dataset is from human subjects, and allows us to compare samples taken from liver and kidney cells [11]. Neither of these studies used VOOM for the differential expression analysis of the RNA-seq data. Because VOOM utilizes the LIMMA pipeline for the differential expression analysis, I believe that using VOOM and LIMMA allows for a more accurate comparison of the abilities of RNA-seq and microarray data to detect differential expression.

Chapter 2

Microarray Data Normalization

The process of collecting microarray data leads to a high probability of variation in measured fluorescence intensities from different arrays that result from steps such as sample preparation and the hybridization of the samples on the array. This variation is called obscuring variation [12]. When we conduct inference to assess whether genes are differentially expressed, we risk reaching false conclusions unless as much of the obscuring variation is removed as possible. To remove obscuring variation, the microarray data must undergo preprocessing, or normalization. Because Affymetrix GeneChip microarrays are the most commonly used microarray chips [13], I will focus on the standard normalization algorithm for Affymetrix microarray data, the Robust Multi-array Average (RMA) algorithm [14].

2.1 Background removal

The first step in the RMA algorithm, called background removal, is to remove noise in the fluorescence intensities which results from non-specific binding to the probes. The background removal algorithm used in the RMA algorithm only uses perfect match (PM) probes to remove noise, as opposed to PM and mismatch (MM) probes, as background removal via PM-MM attenuates signal and adds bias [14]. To remove background noise, we first assume that the observed probe intensity, PM, is a combination of background noise (BG) and signal (S).

$$PM_{igj} = BG_{igj} + S_{igj} \tag{2.1}$$

for array $i = 1, \ldots, N$, gene $g = 1, \ldots, G$, and probe $j = 1, \ldots, J$. Note that each array has multiple probes per gene, as signal intensity is dependent upon the region of the gene that binds to each probe [15]. The quantity of interest is the signal, given the perfect match probe intensities. Because the signals will eventually be log corrected, the assumption is that the signal $S_{igj} \sim exp(\lambda)$. The assumption for the background noise is that $BG_{igj} \sim \mathcal{N}(\mu, \sigma^2)$. Having observed the perfect match probe intensities, we then want to estimate the background noise and signal, which can do this through calculating $E[S|PM = pm] = E[PM - BG|PM = pm]$. Letting $a = PM_{igj} - \mu - \sigma^2\lambda$, Bolstad [16] finds this to be

$$E[S|PM = pm] = a + b\frac{\phi\left(\frac{a}{\sigma}\right) - \phi\left(\frac{PM_{igj}-a}{\sigma}\right)}{\Phi\left(\frac{a}{\sigma}\right) - \Phi\left(\frac{PM_{igj}-a}{\sigma}\right) - 1} \tag{2.2}$$

Where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf, respectively, of the standard normal distribution function.

Once we have estimates for μ, σ, and λ, using the above equation to estimate $E[S|PM = pm]$ is equivalent to removing the background noise from the perfect match probes.

2.2 Across array normalization via quantile normalization

After estimating $E[S|PM = pm]$ to remove background noise from each individual probe, we next consider batch effects for each microarray chip. Because different microarrays may be run on different days or by a different researcher, these events will lead to technical variation for each probe. To normalize the data across arrays and remove batch effects,the RMA algorithm uses quantile normalization, Bolstad *et al.* [17] found RMA to have the greatest effect on removing batch effects when compared to other normalization methods. The goal is for an N-dimensional QQ-plot of probe intensities for N arrays to follow the N-dimensional identity line. The algorithm to perform this normalization is as follows:

1. Arrange the data in a table, where each row represents a probe, and each column represents a different array

2. Identify the smallest measurement for each array and take the average. Replace the original values for the smallest measurements with this average

3. Repeat step 2 for the second smallest measurement on each probe

4. Continue this process until the largest probe measurement on each array has been replaced with the average of the largest probe measurements

Below is an example with a hypothetical experiment in which three arrays measured three probes.

	Array 1	Array 2	Array 3
Probe 1	1	4	2
Probe 2	3	3	6
Probe 3	5	7	5

	Array 1	Array 2	Array 3
Probe 1	**1**	4	**2**
Probe 2	4	**3**	6
Probe 3	5	7	5

	Array 1	Array 2	Array 3
Probe 1	2	**4**	2
Probe 2	**3**	2	6
Probe 3	5	7	**5**

	Array 1	Array 2	Array 3
Probe 1	2	4	2
Probe 2	4	2	**6**
Probe 3	**5**	**7**	4

	Array 1	Array 2	Array 3
Probe 1	2	4	2
Probe 2	4	2	6
Probe 3	6	6	4

The quantile normalization procedure creates arrays with identical measurements throughout. However, each of those measurements is on a different probe.

2.3 Summary and Calculation of Probe Level Intensity via Median Polishing

The final step in the RMA algorithm is to remove experimental effects induced by individual probes and summarize the fluorescence intensities for each gene. After background correcting and normalizing the data, we obtain fluorescence intensities Y_{igj} for each gene on arrays $i = 1, \ldots, N$, genes $g = 1, \ldots, G$, and probes $j = 1, \ldots, J$. To remove individual probe effects and summarize the data, we consider the model

$$Y_{igj} = \mu_{ig} + p_{gj} + \epsilon_{igj} \tag{2.3}$$

where the parameter of interest is μ_{ig}, the average gene intensity for array i and gene g. The probe effect, p_{gj}, can be thought of as the variation due to the specific features of a certain probe, relative to other probes for the same gene. Equation 2.3 is constrained such that $\sum_{j=1}^{J} p_{gj} = 0$. Note that the this step is carried out on the gene level, while the quantile normalization is carried out at the probe level. To estimate μ_{ig}, Irizarry *et al.* [14] first estimate ϵ_{igj} through the median polish procedure [18] and let $\hat{\mu}_{ig} = \frac{1}{J}\sum_{j=1}^{J} Y_{igj} - \hat{\epsilon}_{igj}$. The median polish procedure, along with the summarization step, is as follows:

1. Arrange the background-corrected, normalized, and log_2 transformed fluorescence data in a table, where each row represents an array, and each column represents a different probe

2. Calculate the median of the probe fluorescence intensities for each array and subtract it from each probe in the given array

3. Calculate the median of the fluorescence intensities across the arrays for each probe, and subtract it from the given probe value for each array

4. Repeat the previous two steps until the array and probe-wise medians are sufficiently small

5. Each entry in the new matrix represents the estimated error term, $\hat{\epsilon}_{igj}$. To obtain the corrected fluorescence matrix, subtract this matrix from the original matrix of fluorescence intensities to obtain the expression $Y_{igj} - \hat{\epsilon}_{igj}$

6. Take the average of the probes in each array to obtain the final expression measurement for the gene on each array, $\hat{\mu}_{ig}$

Below is example for a 5×5 fluorescence data matrix, provided by Dan Nettleton [19].

1. Arrange the background-corrected, normalized, and log_2 transformed fluorescence data in a table, where each row represents an array, and each column represents a different probe

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5	Row Medians
Array 1	4	3	6	4	7	4
Array 2	8	1	10	5	11	8
Array 3	6	2	7	8	8	7
Array 4	9	4	12	9	12	9
Array 5	7	5	9	6	10	7

2. Subtract row medians

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
Array 1	0	-1	2	0	3
Array 3	0	-7	2	-3	3
Array 3	-1	-5	0	1	1
Array 4	0	-5	3	0	3
Array 5	0	-2	2	-1	3
Column Medians	0	-5	2	0	3

3. Subtract column medians

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5	Row Medians
Array 1	0	4	0	0	0	0
Array 2	0	-2	0	-3	0	0
Array 3	-1	0	-2	1	-2	-1
Array 4	0	0	1	0	0	0
Array 5	0	3	0	-1	0	0

4. Subtract row medians

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5
Array 1	0	4	0	0	0
Array 2	0	-2	0	-3	0
Array 3	0	1	-1	2	-1
Array 4	0	0	1	0	0
Array 5	0	3	0	-1	0
Column Medians	0	1	0	0	0

5. Subtract column medians

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5	Row Medians
Array 1	0	3	0	0	0	0
Array 2	0	-3	0	-3	0	0
Array 3	-1	0	-2	1	-2	0
Array 4	0	-1	1	0	0	0
Array 5	0	2	0	-1	0	0
Column Medians	0	0	0	0	0	

6. Now that the row and column medians are both 0, the matrix in step 5 represents the estimated residuals, $\hat{\epsilon}_{igj}$ for each observed fluorescence intensity, Y_{igj}. We subtract the residual matrix from our original data, and summarize each row by taking the mean

	Probe 1	Probe 2	Probe 3	Probe 4	Probe 5	Row Means
Array 1	4	0	6	4	7	4.2
Array 2	8	4	10	8	11	8.2
Array 3	6	2	8	6	9	6.2
Array 4	9	5	11	9	12	9.2
Array 5	7	3	9	7	10	7.2

The row means in the example above represent the value μ_{ig}. Because the probe affect, p_{gj}, remains constant across arrays for each probe, we can use the value $\hat{p}_{gj} = Y_{igj} - \hat{\epsilon}_{igj} - \mu_{ig}$. Note in the example above that the probe affinity effects are -.2, -4.2, 1.8, -.2, and 2.8 for probes 1, 2, 3, 4, and 5, respectively.

Chapter 3

Microarray Analysis with LIMMA

3.1 Constructing a Linear Model for Microarray Data Analysis

To make inferences on tens of thousands of genes, Smyth (2004) first creates a linear model for each gene. This statistical technique is called "Linear Modeling for Microarray Data" (LIMMA). Consider an experiment in which 2 microarray samples each are collected from female and male mice. For each gene, the fluorescence data yields a response vector $\mathbf{y}_g = (y_{g_1}, y_{g_2}, y_{g_3}, y_{g_4})^T$, where y_{g_1} and y_{g_2} represent the two male microarray response and y_{g_3} and y_{g_4} represent the two female microarray responses. Using the linear model

$$E \begin{bmatrix} y_{g1} \\ y_{g2} \\ y_{g3} \\ y_{g4} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_{g1} \\ \alpha_{g2} \end{bmatrix} \tag{3.1}$$

we can define a contrast matrix

$$C = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \tag{3.2}$$

to test whether $\beta_g = C^T \begin{bmatrix} \alpha_{g1} & \alpha_{g2} \end{bmatrix}^T = \alpha_{g2} - \alpha_{g1} = 0$.

This example can be generalized to create a suitable linear model for a range of experimental designs. Information is collected on N microarrays,

yielding a response vector $\mathbf{y}_g = (y_{g_1}, \ldots, y_{g_N})^T$ containing normalized fluorescence intensities for the gth gene. To represent the experimental design, one can create a design matrix X of full column rank and a coefficient vector α_g, such that the entries of α_g capture the coefficients of interest for comparing the contribution of each array to $\mathbf{y}_g$. If there are P treatments, one typically creates a $N \times P$ design matrix, where the columns represent each different treatment. The design matrix has a 1 in the nth row and pth column if the pth treatment corresponds to the fluorescence intensities in the nth entry in $\mathbf{y}_g$, and a 0 otherwise, as in equation 3.1. We assume

$$E(y_g) = X\alpha_g \tag{3.3}$$

We also assume

$$var(y_g) = W_g\sigma_g^2 \tag{3.4}$$

where W_g is a non-negative definite weight matrix related to array quality and heteroskedasticity. W_g can be specified by the user, and is especially useful for RNA-seq data, which will be discussed in chapter 4.

In order to make inference on the contrasts between certain coefficients of interest, we define a contrast matrix C such that we can test whether the entries in the vector $\beta_g = C^T\alpha_g$ are equal to 0. A general guideline in creating a contrast matrix is to have each column represent the contrast of interest and each row represent a treatment condition. The ijth entry of C will be 0 if the ith treatment is not of interest in the jth contrast, or will be either 1 or -1, depending on whether the coefficient corresponding to the ith treatment is being added or subtracted in the jth contrast. In equation 3.2, the two rows of C corresponded to α_1 and α_2, while the one column of C corresponded to the contrast $\alpha_2 - \alpha_1$.

For the remainder of the derivations in this paper, I will assume a fitting of the linear model by least squares, although the original LIMMA paper does not make this assumption. After obtaining estimates for α_g and σ_g^2, $\hat{\alpha}_g$ and s_g^2 respectively, and the covariance matrices

$$Cov(\hat{\alpha}_g) = V_g\sigma_g^2 \tag{3.5}$$

where $V_g = (X^TX)^{-1}$ [20] is a positive definite matrix, we can derive the covariance matrices for the contrast estimators, $\hat{\beta}_g = C^T\hat{\alpha}_g$:

$$
\begin{aligned}
Cov\left(\hat{\beta}_g\right) &= Cov\left(C^T\alpha_g\right) && (3.6)\\
&= E[(C^T\alpha_g - E[C^T\alpha_g])(C^T\alpha_g - E[C^T\alpha_g])^T] && (3.7)\\
&= E[C^T(\alpha_g - E[\alpha_g])(C^T(\alpha_g - E[\alpha_g]))^T] && (3.8)\\
&= C^T E[(\alpha_g - E[\alpha_g])^2]C && (3.9)\\
&= C^T Cov\left(\hat{\alpha}_g\right)C && (3.10)\\
&= C^T V_g C\sigma_g^2 && (3.11)
\end{aligned}
$$

Because we want to test whether $\beta_{gj} = 0$ for genes $g = 1, \ldots, G$, and contrasts $j = 1, \ldots, J$ we need to derive the distributions for $\hat{\beta}_{gj}$ and s_g^2. Note that $Var(\hat{\beta}_{gj}) = \nu_{gj}\sigma_g^2$, where ν_{gj} is the jth diagonal element of the matrix $C^T V_g C$. Given the response vector, $\mathbf{y}_g$, the distributional assumption for $\hat{\beta}_{gj}$ can be summarized by

$$\hat{\beta}_{gj}|\beta_{gj}, \sigma_g^2 \sim N\left(\beta_{gj}, \nu_{gj}\sigma_g^2\right) \tag{3.12}$$

which agrees with [20]. To derive the distribution for s_g^2, we first let $S_g^2 = \sum_{i=1}^{N}(Y_{gi} - X_{i1}\hat{\alpha}_{g1} - \cdots - X_{ip}\hat{\alpha}_{gp})^2$, where p is the number of experimental conditions. Also, note that the quantity $s_g^2 = \frac{S_g^2}{n-p}$ is an unbiased estimator of σ_g^2 [20]. For the purposes of the following calculations, I will assume that the responses $\mathbf{y}_g$ are normal, although Smyth does not make this assumption. Given this assumption, it follows that $S_g^2/\sigma_g^2 \sim \chi_{d_g}^2$ [20], and we use the fact $s_g^2 = \frac{\sigma_g^2}{n-p}\frac{S_g^2}{\sigma_g^2}$ to derive the conditional distribution of s_g^2 given σ_g^2 and the residual degrees of freedom, $d_g = N - p$ [20]:

$$s_g^2|\sigma_g^2 \sim \frac{\sigma_g^2}{d_g}\chi_{d_g} \tag{3.13}$$

These assumptions allow us to make inference on whether $\beta_{gj} = 0$ using the ordinary t-statistic

$$t_{gj} = \frac{\hat{\beta}_{gj}}{s_g\sqrt{\nu_{gj}}} \tag{3.14}$$

which follows a standard t-distribution on d_g degrees of freedom.

3.2 Using a Hierarchical Model to Improve the Estimate for σ_g^2

In conducting ordinary t-tests, microarray data presents the problem of a low sample size, which leads to a poor estimate of the true variation, σ_g^2. A large estimate of the variation will lead failing to identify a truly differentially expressed gene as differentially expressed, giving the test lower power, while a small estimation of the variation leads to classifying non-differentially expressed genes as significant, giving the test a higher false positive rate. However, microarrays present the opportunity to gather information from all genes together in order to create a pooled estimation of the variance. While using a pooled variation estimate rather than s_g^2 does not allow for the estimate of the variance for the t-statistic to take the behavior of individual genes into account, it does suggest the idea of using Bayesian statistics to add prior information to each of our gene variation estimates.

Smyth (2004) uses a hierarchical empirical Bayesian model, in which the parameters for the prior on σ_g^2 are estimated from the data. Because

$$s_g^2|\sigma_g^2 \sim Gamma\left(\frac{d_g}{2}, \frac{2\sigma_g^2}{d_g}\right) \tag{3.15}$$

we want a conjugate prior distribution for $\frac{1}{\sigma_g^2}$. Note that we put a prior on $\frac{1}{\sigma_g^2}$, rather than σ_g^2, due to mathematical ease. The most convenient distribution is the Gamma with estimated hyperparameters d_0 and s_0^2:

$$\frac{1}{\sigma_g^2} \sim Gamma\left(\frac{d_0}{2}, \frac{2}{d_0 s_0^2}\right) \tag{3.16}$$

To estimate the prior degrees of freedom and sample variance, d_0 and s_0^2, let $z_g = logs_g^2$. Smyth shows that the z_g follow a Fisher's z-distribution [2]. The theoretical mean and variance of the z-distribution are then matched to the empirical mean and variance of the z_g, which leads to the estimate $s_0^2 = \frac{1}{G}\sum_{g=1}^{G} s_g^2$ [21]. The derivation of the estimate of d_0 is outside the scope of this paper, but more details can be found in [2].

To make the estimate of σ_g^2 less susceptible to extreme values, we will instead use the inverse of the expected value for the posterior for $\frac{1}{\sigma_g^2}$ given s_g^2. To find the posterior distribution, we use the following calculations:

$$\xi\left(\frac{1}{\sigma_g^2}|s_g^2\right) \propto f\left(s_g^2|\sigma_g^2\right)\xi\left(\frac{1}{\sigma^2}\right) \tag{3.17}$$

$$\propto \left(\frac{d_g}{\sigma^2}\right)^{d_g/2} e^{-\frac{d_g s_g^2}{2\sigma_g^2}} \left(\frac{1}{\sigma^2}\right)^{-\frac{d_0}{2}-1} e^{-\frac{d_0 s_0^2}{2\sigma^2}} \tag{3.18}$$

$$\propto \left(\frac{1}{\sigma^2}\right)^{-\frac{d_0+d_g}{2}-1} e^{-\frac{d_0 s_0^2 + d_g s_g^2}{2\sigma^2}} \tag{3.19}$$

which shows that

$$\xi\left(\frac{1}{\sigma^2}|s_g^2\right) \sim Gamma\left(\frac{d_0+d_g}{2}, \frac{2}{d_g s_g^2 + d_0 s_0^2}\right) \tag{3.20}$$

and

$$\left(E\left[\frac{1}{\sigma^2}|s_g^2\right]\right)^{-1} = \frac{d_g s_g^2 + d_0 s_0^2}{d_0 + d_g} \tag{3.21}$$

The moderated estimate of the variation, denoted $\tilde{s}_g^2$, gives a weighted average of our estimate for the prior variation, s_0^2, which is found using all of the data, and the individual gene wise variation, s_g^2. Because we have added information from the pooled gene wise variance, we now replace s_g^2 with $\tilde{s}_g^2$ in the t-statistic. This new statistic, $\tilde{t}_{gj}$, is called the moderated t-statistic and is defined as:

$$\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g\sqrt{\nu_{gj}}} \tag{3.22}$$

3.3 The Distribution of the Moderated t-statistic

Now that we have precisely defined the moderated t-statistic, it is of interest to find its distribution in order to make use of the statistic in inference. Specifically, we want to find the distribution of $\tilde{t}_{gj}$ when our null hypothesis, $\beta_{gj} = 0$, is true. To do this, we will derive the joint distribution of $\tilde{t}_{gj}$, and s_g^2, and show that the two statistics are independent with the following marginal distributions:

$$s_g^2 \sim s_0^2 F_{d_g, d_0} \tag{3.23}$$

and

$$\tilde{t}_{gj}|\beta_{gj} = 0 \sim t_{d_0+d_g} \tag{3.24}$$

To begin this proof, let $f_{\tilde{t}_{gj},S_g^2|\beta_{gj}}\left(\tilde{t}_{gj}, s_g^2|\beta_{gj}=0\right)$ and $F_{\tilde{t}_{gj},S_g^2|\beta_{gj}}\left(\tilde{t}_{gj}, s_g^2|\beta_{gj}=0\right)$ denote the p.d.f. and c.d.f., respectively, of the joint distribution of $\tilde{t}_{gj}$ and s_g^2 given $\beta_{gj}=0$. Recall that

$$f_{\tilde{t}_{gj},S_g^2|\beta_{gj}}\left(\tilde{t}_{gj}, s_g^2|\beta_{gj}=0\right) = F'_{\tilde{t}_{gj},S_g^2|\beta_{gj}}\left(\tilde{t}_{gj}, s_g^2|\beta_{gj}=0\right) \tag{3.25}$$

Because of this relation, we can find $F_{\tilde{t}_{gj},S_g^2|\beta_{gj}}\left(\tilde{t}_{gj}, s_g^2|\beta_{gj}=0\right)$ to derive $f_{\tilde{t}_{gj},S_g^2|\beta_{gj}}\left(\tilde{t}_{gj}, s_g^2|\beta_{gj}=0\right)$. We will use the definition of $\tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g\sqrt{\nu_{gj}}}$ to derive the c.d.f.

$$\begin{aligned}
F_{\tilde{t}_{gj},S_g^2|\beta_{gj}}\left(\tilde{t}_{gj}, s_g^2|\beta_{gj}=0\right) &= Pr(\frac{\hat{\beta}_{gj}}{\tilde{s}_g\sqrt{\nu_{gj}}} \leq \tilde{t}_{gj}, S_g^2 \leq s_g^2|\beta_{gj}=0) && (3.26)\\
&= Pr(\hat{\beta}_{gj} \leq \tilde{t}_{gj}\tilde{s}_g\sqrt{\nu_{gj}}, S_g^2 \leq s_g^2|\beta_{gj}=0) && (3.27)\\
&= F_{\hat{\beta}_{gj},s_g^2|\beta_{gj}}(\tilde{t}_{gj}\tilde{s}_g\sqrt{\nu_{gj}}, s_g^2|\beta_{gj}=0) && (3.28)
\end{aligned}$$

Finally, noting that $\tilde{t}\tilde{s}\sqrt{\nu_{gj}} = \hat{\beta}$, we take the derivative of $F_{\hat{\beta}_{gj},s_g^2|\beta_{gj}}$ with respect to $\tilde{t}_{gj}$ and s_g^2 to find

$$f_{\tilde{t}_{gj},S_g^2|\beta_{gj}}\left(\tilde{t}, s_g^2|\beta=0\right) = f_{\hat{\beta}_{gj},s_g^2|\beta_{gj}}\left(\hat{\beta}, s_g^2|\beta=0\right)\tilde{s}\sqrt{\nu_{gj}} \tag{3.29}$$

Under the assumption that that $\hat{\beta}_{gj}$ and s_g^2 are independent we can find $f_{\hat{\beta}_{gj},s_g^2|\beta_{gj}}\left(\hat{\beta}_{gj}, s_g^2|\beta_{gj}=0\right)$. While the assumption that that $\hat{\beta}_{gj}$ and s_g^2 are independent is not necessarily true, the moderated t-statistic still outper-

forms other methods of inference using the independence assumption.

$$f_{\hat{\beta}_{gj},s_g^2|\beta_{gj}}\left(\hat{\beta}_{gj},s_g^2|\beta_{gj}=0\right)=\int_{-\infty}^{\infty}f_{\hat{\beta}_{gj},s_g^2,\sigma_g^{-2}|\beta_{gj}}\left(\hat{\beta}_{gj}=x,s_g^2,\sigma_g^{-2}|\beta_{gj}=0\right)d(\sigma_g^{-2}) \tag{3.30}$$

$$=\int_{-\infty}^{\infty}f_{\hat{\beta}_{gj},s_g^2|\sigma_g^{-2},\beta_{gj}}\left(\hat{\beta}_{gj}=x,s_g^2|\sigma_g^{-2},\beta_{gj}=0\right)f_{\sigma_g^{-2}}(\sigma_g^{-2})d(\sigma_g^{-2}) \tag{3.31}$$

$$\begin{aligned}&=\int_{0}^{\infty}f_{\hat{\beta}_{gj}|\sigma_g^{-2},\beta_{gj}}\left(\hat{\beta}_{gj}=x|\sigma_g^{-2},\beta_{gj}=0\right)\\&\times f_{s_g^2|\sigma_g^{-2},\beta_{gj}}\left(s_g^2|\sigma_g^{-2},\beta_{gj}=0\right)f_{\sigma_g^{-2}}(\sigma_g^{-2})d(\sigma_g^{-2})\end{aligned} \tag{3.32}$$

$$\begin{aligned}&=\int_{0}^{\infty}\frac{1}{\left(2\pi\nu_{gj}\sigma_g^2\right)^{\frac{1}{2}}}e^{-\frac{\hat{\beta}_{gj}^2}{2\nu_{gj}\sigma_g^2}}\\&\times\left(\frac{d_g}{2\sigma_g^2}\right)^{\frac{d_g}{2}}\frac{s^{2(d_g/2-1)}}{\Gamma\left(d_g/2\right)}e^{-\frac{d_gs_g^2}{2\sigma_g^2}}\\&\times\left(\frac{d_0s_0^2}{2}\right)^{d_0/2}\frac{\sigma_g^{-2(d_g/2-1)}}{\Gamma\left(d_0/2\right)}e^{-\sigma_g^{-2}\frac{d_0s_0^2}{2}}d_g\sigma_g^{-2}\end{aligned} \tag{3.33}$$

$$\begin{aligned}&=\frac{\left(\frac{d_0s_0}{2}\right)^{d_0/2}\left(d_g/2\right)^{d_g/2}s^{2(d_g/2-1)}}{(2\pi\nu_{gj})^{1/2}\Gamma\left(d_o/2\right)\Gamma\left(d_g/2\right)}\\&\times\int_{0}^{\infty}\sigma_g^{-2(1/2+d_0/2+d_g/2-1)}e^{-\sigma_g^{-2}\left(\frac{\hat{\beta}_{gj}^2}{2\nu_{gj}}+\frac{d_gs_g^2}{2}+\frac{d_0s_0^2}{2}\right)}d(\sigma_g^{-2})\end{aligned} \tag{3.34}$$

Where the functions in equation 3.33 come from the assumptions made in equations 3.12, 3.13, and 3.16. Note that the integral in equation 3.34 can be written as

$$\frac{\Gamma(\alpha)}{\beta^\alpha}=\int_{0}^{\infty}x^{\alpha-1}e^{-\beta x}dx \tag{3.35}$$

with $x = \sigma_g^{-2}$, $\alpha = 1/2 + d_0/2 + d_g/2$, and $\beta = \frac{\hat{\beta}_{gj}^2}{2\nu_{gj}} + \frac{d_g s_g^2}{2} + \frac{d_0 s_0^2}{2}$. Thus,

$$\int_0^\infty \sigma_g^{-2(1/2+d_0/2+d_g/2-1)} e^{-\sigma_g^{-2}\left(\frac{\hat{\beta}_{gj}^2}{2\nu_{gj}}+\frac{d_g s_g^2}{2}+\frac{d_0 s_0^2}{2}\right)} d(\sigma_g^{-2}) \tag{3.36}$$

$$= \Gamma\left(1/2 + d_0/2 + d_g/2\right) \left(\frac{\frac{\hat{\beta}_{gj}^2}{\nu_{gj}} + d_g s_g^2 + d_0 s_0^2}{2} \right)^{-(1+d_0+d_g)/2} \tag{3.37}$$

Looking at the constant $\frac{\left(\frac{d_0 s_0}{2}\right)^{d_0/2} (d_g/2)^{d d_g/2} s^{2(d_g/2-1)}}{(2\pi\nu_{gj})^{1/2}\Gamma(d_o/2)\Gamma(d_g/2)}$, we note that

$$(2\pi\nu_{gj})^{1/2} = (2\nu_{gj})^{1/2}(\pi)^{1/2} = (2\nu_{gj})^{1/2}\Gamma(1/2) \tag{3.38}$$

To derive that:

$$f\left(\hat{\beta}_{gj}, s_g^2 | \beta_{gj} = 0\right) = \frac{\left(\frac{1}{2\nu_{gj}}\right)^{1/2} \left(\frac{d_0 s_0}{2}\right)^{d_0/2} (d_g/2)^{d_g/2} s^{2(d_g/2-1)}}{\Gamma\left(1/2\right)\Gamma\left(d_o/2\right)\Gamma\left(d_g/2\right)}$$
$$\times \Gamma\left(1/2 + d_0/2 + d_g/2\right) \left(\frac{\frac{\hat{\beta}_{gj}^2}{\nu_{gj}} + d_g s_g^2 + d_0 s_0^2}{2} \right)^{-(1+d_0+d)/2} \tag{3.39}$$

$$= \frac{\left(\frac{1}{2\nu_{gj}}\right)^{1/2} \left(\frac{d_0 s_0}{2}\right)^{d_0/2} (d_g/2)^{d_g/2} s^{2(d_g/2-1)}}{D\left(1/2, d_0/2, d_g/2\right)} \left(\frac{\frac{\hat{\beta}_{gj}^2}{\nu_{gj}} + d_g s_g^2 + d_0 s_0^2}{2} \right)^{-(1+d_0+d_g)/2} \tag{3.40}$$

where $D(\cdot)$ is the Dirichlet function.

Going back to the joint probability of $\tilde{t}_{gj}$ and s_g^2, and letting $B(\cdot,\cdot)$ denote

the Beta function, we find

$$f_{\tilde{t}_{gj},S_g^2}\left(\tilde{t}_{gj},s_g^2|\beta_{gj}=0\right)=\tilde{s_g}\sqrt{\nu_{gj}}f_{\hat{\beta}_{gj},s_g^2}\left(\hat{\beta}_{gj},s_g^2|\beta_{gj}=0\right) \tag{3.41}$$

$$=\tilde{s_g}\sqrt{\nu_{gj}}\frac{\left(\frac{1}{2\nu_{gj}}\right)^{1/2}\left(\frac{d_0s_0}{2}\right)^{d_0/2}(d_g/2)^{d_g/2}s_g^{2(d_g/2-1)}}{D\left(1/2,d_0/2,d_g/2\right)}$$

$$\times\left(\frac{\frac{\hat{\beta}_{gj}^2}{\nu_{gj}}+d_gs_g^2+d_0s_0^2}{2}\right)^{-(1+d_0+d_g)/2} \tag{3.42}$$

$$=\left(\frac{d_0s_0^2+d_gs_g^2}{d_0+d_g}\right)^{1/2}(d_0s_0)^{d_0/2}(d_g)^{d_g/2}s_g^{2(d_g/2-1)}$$

$$\times\frac{\Gamma\left(1/2+d_0/2+d_g/2\right)}{\Gamma(1/2)\Gamma(d_0/2)\Gamma(d_g/2)}\frac{\Gamma(d_0/2+d_g/2)}{\Gamma(d_0/2+d_g/2)}$$

$$\times\left(\frac{\hat{\beta}_{gj}^2}{\nu_{gj}}+d_gs_g^2+d_0s_0^2\right)^{-(1+d_0+d_g)/2} \tag{3.43}$$

$$=\left(\frac{d_0s_0^2+d_gs_g^2}{d_0+d_g}\right)^{1/2}\frac{(d_0s_0)^{d_0/2}(d_g)^{d_g/2}s_g^{2(d_g/2-1)}}{B(d_g/2,d_0/2)B(1/2,d_0/2+d_g/2)}$$

$$\times(d_0s_0^2+d_gs_g^2)^{-\frac{1+d_0+d_g}{2}}\left(1+\frac{\hat{\beta}_{gj}^2}{d_gs_g^2+d_0s_0^2\nu_{gj}}\right)^{-(1+d_0+d_g)/2} \tag{3.44}$$

$$=\frac{(d_0s_0^2)^{\frac{d_0}{2}}d_g^{\frac{d_0}{2}}s_g^{2\frac{d_0}{2}-1}}{B(\frac{d_g}{2},\frac{d_0}{2})(d_0s_0^2+d_gs_g^2)^{\frac{d_0}{2}+\frac{d_g}{2}}}$$

$$\times\frac{(d_0+d_g)^{-\frac{1}{2}}}{B(\frac{1}{2},\frac{d_0}{2}+\frac{d_g}{2})}\left(1+\frac{\tilde{t}_{gj}^2}{d_0+d_g}\right)^{-\frac{1+d_0+d_g}{2}} \tag{3.45}$$

The derivation above shows that conditional joint p.d.f of $\tilde{t}_{gj}$ and s_g^2 is the product of a standard t-distribution and F-distribution. Thus, $\tilde{t}_{gj}$ and s_g^2 are independent with

$$s_g^2\sim s_0^2F_{d_g,d_0} \tag{3.46}$$

and

$$\tilde{t}_{gj}|\beta_{gj}=0\sim t_{d_0+d_g} \tag{3.47}$$

Result 3.47 show that $\tilde{s}_g^2$ not only produced a weighted average of the pooled and individual residual variances, but also leads to a t-statistic which follows a t-distribution with increased degrees of freedom. The increase in degrees of freedom, from d_g to $d_0 + d_g$, reflects the added information from the Bayesian hierarchical model, and results an increase in power [22]. In addition, Smyth finds that the moderated t-statistic has a much lower false discovery rate (FDR) than the standard t-statistic in a range of simulated conditions [2].

Chapter 4

RNA-seq Analysis with VOOM

4.1 Introduction to RNA-seq

While microarrays are currently the most popular sequencing technology, they limit researchers' abilities to accurately measure end expression in several ways [23] [24]. First, microarrays rely on complementary hybridization, forcing researchers to use a limited number of genes whose sequence is already known. Second, probes measuring genes that are naturally low in transcription abundance will pick up more background noise. Finally, relative to next-generation sequencing (NGS) technologies, there is a high amount of technical variation, which requires the use of imperfect preprocessing methods, such as RMA.

Because of these issues, NGS technologies that directly measure gene expression, such as RNA-seq [25] have grown in popularity. RNA-seq technology sequences short RNA reads by adding fluorescent nucleotides to the complimentary DNA (cDNA) strand. The fluorescent bases allow the exact sequence to be captured and then mapped to a reference genome. If the species being investigated does not have a reference genome, technologies such as [26] exist for *de novo* assembly of full-length transcripts.The number of reads mapped to a gene on the reference genome is known as that gene's "count". A large reason that researchers are turning to RNA-seq is that it does not limit the number of genes that can be sequenced.

Despite the fact that RNA-seq does not rely on prior knowledge of the genome of interest, as researchers can create a *de novo* reference genome, the process is still dependent upon conditions imposed by researchers, namely

in determining sequencing depth. As sequencing depth increases, the number of differentially expressed genes detected also increases [27], most likely due to the increased ability to detect differences in genes with low levels of expression. This is especially of note, as the higher costs of RNA-seq experiments compared to microarray experiments may prevent researchers to use a sufficiently high sequencing depth [27]. With a lower sequencing depth, sampling error can lead to deflated and/or inflated read counts [27], as opposed to higher sequencing depth, where the read counts more accurately reflect true expression levels. Perkins *et al.* find that some genes are only detected as differentially expressed at lower sequencing depth due to these misleading read counts [27]. This result indicates that lower sequencing depth can lead to increased false positive results, despite finding less differentially expressed genes than experiments with higher sequencing depth. As I will briefly discuss in section 5.2, the use of different bioimformatics tools to map the short reads onto the genome of interest can lead to dramatically different results. The interplay between biology, computer science, and statistics in this process underscores the need for researchers in all of those fields to work together in order to obtain accurate and interpretable results.

4.2 Modeling RNA-seq Data

Despite the benefits that RNA-seq provides over microarray technology, the count-based nature of the data makes it difficult to develop statistical tests for differential expression [3]. Law *et al.* argue that methods which use distributions made to model count data, such as the negative binomial (NB) distribution, rely on knowing the true variation of the data [3]. More importantly, most statistical tools developed for inference, such as the standard and moderated t-test, rely on normally distributed data, or data which can be suitably transformed to be then considered approximately normal.

Relying on the theory that statistically powerful hypothesis tests strongly rely on correctly modeling the relationship between the mean and variance [28], Law *et al.* focus on estimating the relationship between the number of counts and the standard deviation through "variance modeling at the observational level" (VOOM) [3], and then using a weighted LIMMA analysis. Because gene counts are normalized by the number of mapped reads (in millions) in each sample to get counts per million (cpm) values, genes with different counts in sample 1 and sample 2 can have the same measured re-

sponse. For example, if a gene gets 200 counts for a sample with 2 million reads, while getting 100 counts in a different sample with 1 million reads, the cpm measurement will be 1 for both samples. However, Law *et al.* observe that observations with small log-count values have much higher variation than observations with large log-count values [3]. Thus, observations with the same log-cpm values may have different count values, resulting in highly unequal variances, which violates a main assumption of LIMMA.

Once this relationship is established non-parametrically, the inverse of the estimated variance for each observation can be used as that observation's weight in the LIMMA pipeline, where the log-cpm values can be used instead of fluorescence intensities for the observed values $\mathbf{y}_g$. This method does not rely on distributional assumptions, such as the NB exact test [29], which requires that the variation of the data is known, leading to poor performance when the variance is estimated [3]. This is especially true when small sample sizes, sometimes as low as three [3], lead to especially imprecise estimates of the variance. An additional advantage of VOOM is that it allows RNA-seq data to be analyzed by a method that is easy to implement and is designed for experiments with small sample sizes that are concerned with relative gene expression .

4.3 Differential Expression Using RNA-seq Data in the LIMMA Pipeline

After obtaining raw RNA-seq count data, the data is first normalized using trimmed means of M-values (TMM) normalization [30]. As mentioned in section 4.2, the counts are further normalized by dividing each count by the corresponding library size (in millions) and then taking the log_2 of this value. VOOM seeks to estimate the variation of individual observations. However, obtaining a sample variance requires at least two observations from the same sample, and RNA-seq only allows one observations of the number of counts for each gene in a sample. To solve this problem, Law *et al.* estimate the mean-variance relationship on a gene level, rather than on an individual sample level. This relationship is then used to predict sample variances for individual observations, based on the observed count value. The VOOM method is as follows:

1. Create a design matrix, as in section 3.1, and fit a linear model using

ordinary least squares to the individual log-cpm values ($\mathbf{y}_{gi}$) for genes $g = 1, \ldots, G$ obtained via RNA-seq. In matrix terms, the model is:

$$E[\mathbf{y}_g] = X\alpha_g \tag{4.1}$$

where X is the design matrix

2. For each gene, calculate the residual standard deviations, s_g, coefficient estimates $\hat{\alpha}_g$, and fitted values for each observation, $\hat{\mu}_{gi} = x_i\hat{\alpha}_g$, where x_i represents the ith row, or ith sample, in the design matrix X

3. Compute the mean log-cpm for each gene, $\bar{y}_g$. Note that if the design matrix X is created as in section 3.1, the $\hat{\mu}_{gi}$ will be the average log-cpm for each experimental condition, while $\bar{y}_g$ is the average log-cpm across all conditions. Because $\bar{y}_g$ is an average of log-counts normalized by the number of mapped reads per sample (library size), we can convert $\bar{y}_g$ to an average log-count value, $\tilde{r}_g$, by multiplying by the geometric mean of the total number of mapped reads per sample. Mathematically:

$$\tilde{r}_g = \bar{y}_g + log_2\left(\frac{\tilde{R}+1}{10^6}\right) \tag{4.2}$$

where $\tilde{R}$ is the geometric mean of the library sizes

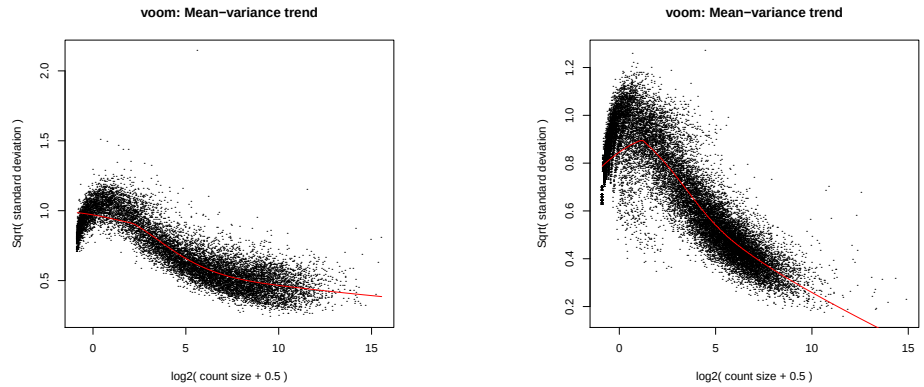
4. Having obtained the average log-count and an estimate of the residual standard deviation for each gene, we can now estimate the relationship between the two statistics. To do this, we fit a LOWESS curve [31] to $\sqrt{s_g}$ as a function of $\tilde{r}_g$ (Figure 4.1). The LOWESS curve provides a smooth trend between mean log-count and standard variation, and is statistically robust [32]

5. Convert fitted log-cpm values, $\hat{\mu}_{gi}$, to fitted count values, $\hat{\lambda}_{gi}$ by replacing $\bar{y}_g$ with $\hat{\mu}_{gi}$ in the equation in step 3

6. Find the VOOM precision weights, w_{gi} for the log-cpm value y_{gi} by taking the inverse of the squared predicted standard deviation for $\hat{\lambda}_{gi}$ using the fitted LOWESS curve [31], lo():

$$w_{gi} = lo(\hat{\lambda}_{gi})^{-4} \tag{4.3}$$

The w_{gi} are then used as the ith diagonal entry in the definite weight matrix W_g specified in equation 3.2



(a) (b)

Figure 4.1: Gene wise variances versus gene wise means. A LOWESS trend, represented by the red line, is fitted to the data. **(a)** Bottomly Mice Data. **(b)** Marioni Human Data.

4.4 Performance of VOOM Relative to Other Methods of Analyzing RNA-seq Data

Law *et al.* performed several simulations of RNA-seq data to compare VOOM to other methods designed specifically for RNA-seq analysis, such as edgeR [33], DESeq [34], and PoissonSeq [35]. The comparisons measured ability to control the type I error rate, power to detect true differential expression, and the number of false discoveries. Six samples of 10,000 genes each were simulated, with the distribution of each sample modeled on an observed distribution for a real RNA-seq dataset. Three samples were considered as condition 1, while the other three samples were considered as condition 2.

To compare the methods on ability to control the type-I error rate, th simulations varied the number of total counts for each sample. In addition, the gene counts were created such that there was no true differential expression between condition 1 and condition 2, which implies that the distribution of P values should be uniform if the method controls the type-I error rate correctly. Using a P value cutoff of .01, VOOM is the only method where the percentage of genes with $P < .01 \approx .01$, whereas other methods are too liberal or conservative.

Law *et al.* then compare the methods on their power to detect true differ-

ential expression by simulating the six samples such that in each condition, 100 random selected genes were significantly upregulated. When the sample sizes are all equal, VOOM finds the third most differentially expressed genes out of all methods using $FDR < .01$, only behind edgeR and PoissonSeq. However, VOOM has a lower false discovery rate (FDR) than the other two most powerful methods. When sample sizes vary, VOOM finds the second most differentially expressed genes out of all methods using $FDR < .01$, only behind edgeR, while still maintaining a lower FDR than edgeR.

Finally, Law *et al.* compare the number of false discoveries versus the number of genes detected as differentially expressed for all of the methods. For all values of number of genes detected as differentially expressed, VOOM has the lowest number of false discoveries.

Several recent papers have compared RNA-seq analysis methods to VOOM, and all find that VOOM performs extremely well relative to other methods. Soneson and Delorenzi [9] perform simulations with a range of samples sizes and number of truly differentially expressed genes. They find that VOOM performs extremely well in almost all conditions, and was computationally fast. Rapaport *et al.* use a spike-in data set to evaluate the RNA-seq analysis methods and find that VOOM generally outperforms methods designed specifically for RNA-seq data. Finally, Seyednasrollah *et al.* [7] recommend using VOOM over other methods due to performance, ease of use, and computational speed.

Chapter 5

Microarray and RNA-seq Data Comparison

5.1 Sources of Data and Pre-processing Methods

To compare the abilities of microarray and RNA-seq to identify differentially expressed genes, I used VOOM and LIMMA to analyze the Bottomly mouse data set [10] and the Marioni human data set [11]. The Bottomly RNA-seq data was downloaded from ReCount [36], while the Marioni RNA-seq data was downloaded directly from www.genome.org. Both of the microarray data sets were publicly available on the GEO database (Accession numbers GSE26024 and GSE11045 for the Bottomly and Marioni data, respectively). To obtain microarray data, the Bottomly study used the Affymetrix MOE 430 2.0 array, while the Marioni study used the Affymetrix HG-U133 Plus 2.0 array. Illumina sequencing was used to obtain RNA-seq data for both studies. The Bottomly data set contains 10 RNA-seq samples for C57BL/6J (B6) mice, 11 RNA-seq samples for DBA/2J (D2) mice, and 10 microarray samples for each strain. The Marioni data set contains 7 RNA-seq samples for both liver and kidney cell samples, and 3 microarray samples for both cell types.

I used the RMA algorithm [14] described in Chapter 2 to normalize the microarray data for both studies. To filter the RNA-seq data, I used the protocol described in the respective papers. For the Bottomly data, I kept genes which had at least one read for the 10 B6 samples and at least one

read for the 11 D2 samples. The Bottomly data originally had mapped reads for 36,536 genes, and had 12,839 genes (35 %) available for analysis after filtering. Marioni used a less conservative filtering method, keeping samples which had at least one read across all samples and cell types. The Marioni data originally had mapped reads for 32,000 genes, and had 22,925 (72 %) for analysis after filtering. After filtering, I used the TMM normalization method briefly mentioned in section 4.3.

5.2 Results

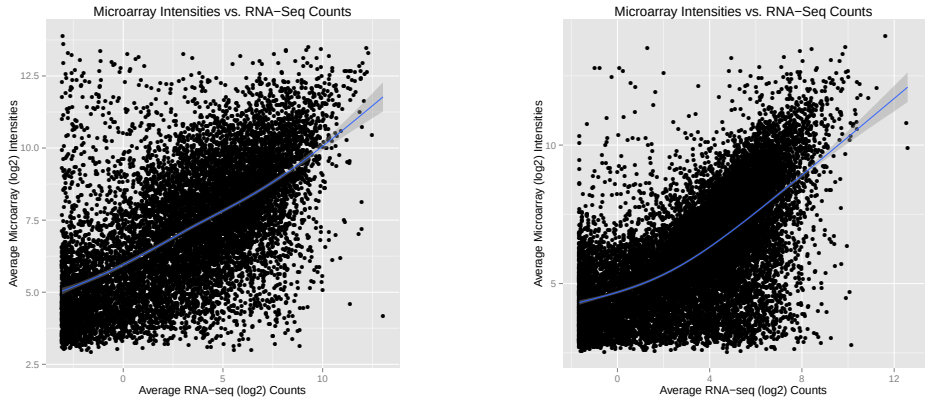
Before applying differential expression analysis, I first compared the average log_2 RNA-seq counts to the average log_2 microarray fluorescence intensities (Fig. 5.1) for genes used in the analysis of both platforms. While both data sets showed positive correlation between RNA-seq counts and microarray intensities, the differences between the data sets become apparent through this comparison. Most notably, a larger percentage of genes with low microarray fluorescence intensity have higher RNA-seq counts in the human data. However, a higher percentage of genes with a low average RNA-seq count have higher microarray fluorescence intensity in the mice data, compared to the human data. Despite this, the human data still has a much higher percentage of differentially expressed genes when using the microarray technology, although this is most likely due to greater difference in gene expression for liver and kidney cells than for different breeds of mice.

VOOM and LIMMA were applied to RNA-seq and microarray data, respectively, to obtain p-values for each gene. The p-values were then adjusted using the Benjamini-Hochburg adjustment [37], and a false discovery rate (FDR) of .01 was used to determine differential expression. The Bottomly data set had 1,426 genes (11%) differentially expressed using microarray data, and 607 genes (5%) differentially expressed using RNA-seq data. The Marioni data set had 10,694 differentially expressed genes (47%) using RNA-seq data, and 8,095 differentially expressed genes (35%) using microarray data. Using the Ensembl annotation, I compared the genes found to be differentially expressed by VOOM and LIMMA. Figure 5.2 shows the overlap between the genes called differentially expressed for both technologies. The Marioni data set finds 43 % of all differentially expressed genes to be called differentiaIly expressed by both technologies, while the Marioni data set only finds 13 % of all differentially expressed genes to be called differentiaIly expressed by both

technologies (Fig 5.2). In addition, the Bottomly data (Fig 5.2.a) has a much larger number of differentially expressed genes found uniquely by LIMMA, as compared to VOOM, while the opposite held in the Marioni data (Fig. 5.2.b). However, it is of note that the results differ greatly from the original findings in the Bottomly study, which found 1,727 genes to be differentially expressed in the RNA-seq data using edgeR [33], and 1,652 genes to be differentially expressed in the Affymetrix microarray data using LIMMA. In addition, the Bottomly paper reported different number of genes before and after filtering compared to the data that I obtained from the ReCount database. Although ReCount database and the Bottomly paper both use Bowtie [38] to map the short reads onto a reference genome, I believe a difference in the filtering process between Bottomly and the creators of the ReCount database can account for the data discrepancy .

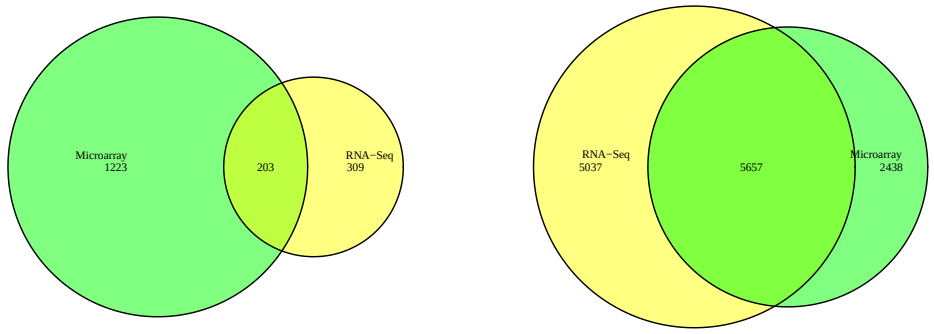
To see if there were any trends amongst genes found to be uniquely expressed by one of the technologies, I again compared the average log_2 expression of microarray and RNA-seq data for both datasets, except this time genes were filtered to be only those that were found to be differentially expressed in at least one of the technologies (Fig. 5.3). In the Bottomly data set (Fig. 5.3.b), many of the genes found to be differentially expressed only in the microarray data (green dots) were in the region of genes that had low RNA-seq counts, but higher microarray fluorescence intensities. In the Marioni human data, the trend for genes found to be differentially expressed only by microarray technology is somewhat close to the trend for genes found to be differentially expressed by both technologies (red dots). However, the genes found to be differentially expressed only by RNA-seq technology (blue dots) tend to have lower microarray fluorescence intensity for a given RNA-seq count.

Finally, I compared the log_2 fold change from RNA-seq and microarray data (Fig 5.4). Blue dots indicate genes that were not found to be differentially expressed in any technology. The genes found uniquely differentially expressed by one of the technologies (purple dots for RNA-seq and green dots for microarray) fall on the x and y-axis for both data sets, which is to be expected. Although the relatively low number of differentially expressed genes in the Bottomly data set does not allow us to see trends as clearly as in the Marioni data, the higher number of genes found to be differentially expressed by both technologies (red dots) in the Marioni data set demonstrates the high concordance in fold change between microarray and RNA-seq data.



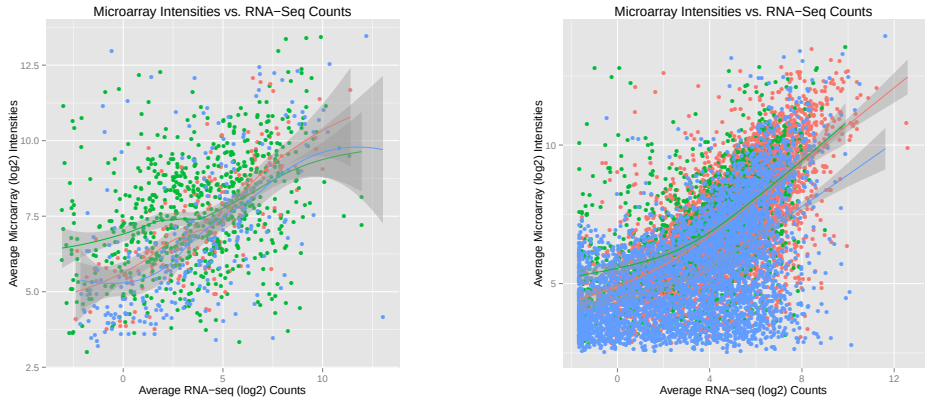
(a) (b)

Figure 5.1: Comparison of the the average log_2 RNA-seq counts to the average log_2 microarray fluorescence intensities for genes that appeared in both the RNA-seq and microarray datasets and for which Ensembl annotation was available. **(a)** Bottomly Mice Data. **(b)** Marioni Human Data.



(a) (b)

Figure 5.2: Overlap between genes found to be differentially expressed in RNA-seq and microarray technologies, based on Ensemble annotations. The number of genes found to be differentially expressed by both technologies appears in the overlapping region of each venn diagram. **(a)** Bottomly Mice Data. **(b)** Marioni Human Data.



(a) (b)

Figure 5.3: Comparison of the the average log_2 RNA-seq counts to the average log_2 microarray fluorescence intensities for genes found to be differentially expressed in at least one of the samples. Blue indicates that the gene was differentially expressed only in RNA-seq data, green indicates that the gene was differentially expressed in only microarray data, and red indicates that the gene was differentially expressed in both technologies. **(a)** Bottomly Mice Data. **(b)** Marioni Human Data.

(a) (b)

Figure 5.4: Comparison of the the log_2 fold change in RNA-seq vs microarray data. Blue indicates that the gene was not found to be differentially expressed in any technology, purple indicates that the gene was differentially expressed only in RNA-seq data, green indicates that the gene was differentially expressed in only microarray data, and red indicates that the gene was differentially expressed in both technologies. **(a)** Bottomly Mice Data. **(b)** Marioni Human Data.

5.3 Discussion

Despite the difference in results obtained in my study and in the Bottomly paper, the ability of RNA-seq to detect a greater number of differentially expressed genes observed with the Marioni data are comparable with the results obtained in the original study, along with results obtained in other studies [23] [27]. Zhao *et al.* observed that RNA-seq is more sensitive in detecting changes genes with relatively low expression levels [23].

Figure 5.2b shows that RNA-seq with VOOM finds over twice as many unique differentially expressed genes than microarray with LIMMA. Looking at the expression levels of the differentially expressed genes in both technology, figure 5.3b shows that many genes that were detected as uniquely differentially expressed by RNA-seq technology and VOOM had low measures of fluorescence intensity, but had a range of log-counts with RNA-seq. Furthermore, figures 5.1 and 5.4 demonstrate the concordance between microarray and RNA-seq technology, which should give researchers confidence in the

ability of RNA-seq to accurately measure gene level expression. In addition, the ability of RNA-seq to identify novel transcripts when researchers create *de novo* reference genomes should be seen as a huge advantage over microarrays. This advantage will likely lead to more researchers choosing RNA-seq over microarrays, and thus will have to decide on a statistical method for detecting differential expression. I believe the results discussed above and the results of other studies discussed in section 4.4 is a convincing argument for using VOOM to identify differentially expressed genes using filtered and normalized RNA-seq data.

Chapter 6

Conclusion

In this paper I have examined the RMA normalization method for microarray data, the empirical Bayesian hierarchical model used in LIMMA, and how mean-variance modeling can be used to analyze RNA-seq data in the LIMMA pipeline. The moderated t-statistic, using $\tilde{s}_g^2$ instead of s_g^2 in the denominator, still follows a t-distribution with added degrees of freedom in comparison to the standard t-statistic. The moderated t-test has added power over the standard t-test, while also having a lower FDR, giving reason as to why LIMMA has become the standard for analyzing microarray data. When the mean-variance trend is correctly specified for RNA-seq data using VOOM, analyzing RNA-seq data with LIMMA outperforms methods created specifically for RNA-seq count data. An item of interest would be to explore the research of Sartor *et al.* [39] in creating a mean-variance relationship for microarray data within the empirical Bayesian hierarchical model used in LIMMA.

Given the results with the Marioni data set in section 5.2, I believe RNA-seq has the potential to replace microarrays as the most common method for measuring relative gene expression changes. Although I was not able to find the FDR for either technology, RNA-seq with VOOM is much more powerful for detecting differential expression than microarray with LIMMA, especially for genes with relatively low expression levels.

Although I briefly mentioned TMM normalization for RNA-seq data, I did not compare this method to other normalization algorithms for RNA-seq data. As RNA-seq grows in popularity, researchers will have to decide on a standard method for normalization, or else results will continue to vary between studies with the same samples of interest [13]. However, RNA-seq

is a relatively new technology, and as the scientific community continues to focus their research on RNA-seq, so will the statistical community.

Bibliography

[1] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 2004.

[2] Gordon K. Smyth. Limma: Linear models for microarray data. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, 2005.

[3] Charity W Law, Yunshun Chen, Wei Shei, and Gordon K Smyth. Voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biology*, 2014.

[4] Mary-Claire King, Joan H. Marks, and Jessica B. Mandell. Breast and ovarian cancer risks due to inherited mutations in brca1 and brca2. *Science*, 2003.

[5] Kelly A. Diggs-Andrews, Jacquelyn A. Brown, Scott M. Gianino, Joshua B. Rubin, David F. Wozniak, and David H. Gutman. Sex is a major determinant of neuronal dysfunction in neurofibromatosis type 1. *Annals of Neurology*, 2014.

[6] Caimiao Wei, Jiangning Li, and Roger E. Bumgarner. Sample size for detecting differentially expressed genes in microarray experiments. *BMC Genomics*, 2004.

[7] Fatemah Seyednasrollah, Asta Laiho, and Laura L. Elo. Comparison of software packages for detecting differential expression in rna-seq studies. *Briefing*, 2013.

[8] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E. Mason, Nicholas D. Socci, and Doron Be-

tel. Comprehensive evaluation of differential gene expression analysis methods for rna-seq data. *Genome Biology*, 2013.

[9] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of rna-seq data. *BMC Bioinformatics*, 2013.

[10] Daniel Bottomly, Nicole A.R. Walter, Jessica Ezzell Hunter, Priscila Darakijian, Sunita Kawane, Kari J. Buck, Robert P. Searles, Michael Mooney, Shannon K. McWeeney, and Robert Hitzemann. Evaluating gene expression in c57bl/6j and dba/2j mouse striatum using rna-seq and microarrays. *PLOS ONE*, 2011.

[11] John C. Marioni, Christopher E. Mason, Shrikant M. Mane, Matthew Stephens, and Yoav Gilad. Rna-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 2008.

[12] Alexander John Hartemink. *Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks*. PhD thesis, Massachusetts Institute of Technology, 2001.

[13] David B. Allison, Xiangqin Cui, Grier P. Page, and Mahyar Sabripour. Microarray data analysis: From disarray to consolidation and consensus. *Nature Reviews Genetics*, 2006.

[14] Rafael A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer-Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 2003.

[15] Cheng-Chung Chou, Chun-Houh Chen, Te-Tsui Lee, and Konan Peck. Optimization of probe length and the number of probes per gene for optimal microarray analysis of gene expression. *Nucleic Acids Research*, 2004.

[16] Benjamin M. Bolstad. *Low-level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD thesis, University of California, Berkeley, 2004.

[17] Benjamin M. Bolstad, Rafael A. Irizarry, Magnus Astrand, and Terence P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 2003.

[18] John W. Tukey. *Exploratory Data Analysis*. Pearson-Wesley, 1977.

[19] Dan Nettleton. Normalization and construction of expression measures for affymetrix genegene data. Presentation Given for Iowa State Statistics 416, January 2009.

[20] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*. Fourth. Pearson, 2011.

[21] Belinda Phipson, Stanley Lee, Ian J. Majewski, Warren S. Alexander, and Gordon K. Smyth. Empirical bayes in the presence of exceptional cases, with application to microarray data. Technical report, Bioinformatics Division, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia, 2013.

[22] Xiangqin Cui, J.T. Gene Hwang, Jing Qui, Natalie J. Blades, and Gary A. Churchill. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, 2005.

[23] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of rna-seq and microarray in transcriptome profiling of activated t cells. *PLOS ONE*, 2014.

[24] Francesa Finotello and Barbara Di Camillo. Measuring differential gene expression with rna-seq: Challenges and strategies for data analysis. *Briefings in Functional Genomics*, 2014.

[25] Ugrappa Nagalakshmi, Zhong Wang, Karl Waem, Chong Shou, Dabashsish Raha, Mark Gerstein, and Michael Snyder. The transcriptional landscape of the yeast genome defined by rna sequencing. *Science*, 2008.

[26] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson andIdo Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen andEvan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv

Regev. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature Biotechnology*, 2010.

[27] James R Perkins, Ana Antunes-Martins, Margarita Calvo, John Grist, Werner Rust, Ramona Schmid, Tobias Hildebrandt, Matthias Kohl, Christine Orengo, Stephen B McMahon, and David LH Bennett. A comparison of rna-seq and exon arrays for whole genome transcription profiling of the l5 spinal nerve transection model of neuropathic pain in the rat. *Molecular Pain*, 2014.

[28] Peter McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

[29] Mark D. Robinson and Gordon K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 2007.

[30] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Bi*, 2010.

[31] William S. Cleveland. Lowess: A program for smoothing scatterplots by robust locally weighted regression. *American Statistician*, 1981.

[32] Alicia Oshlack, Dianne Emslie, Lynn M. Corcoran, and Gordon K Smyth. Normalization of boutique two-color microarrays with a high proportion of differentially expressed probes. *Genome Biology*, 2007.

[33] Mark D. Robinson, David J. McCarthy, and Gordon. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 2010.

[34] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 2010.

[35] Jun Li, Daniela M. Witten, Iain M. Johnstone, and Robert Tibshirani. Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 2012.

[36] Alyssa C Frazee, Ben Langmead, and Jeffrey T Leek. Recount: A multi-experiment resource of analysis-ready rna-seq gene count datasets. *BMC Bioinformatics*, 2011.

[37] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, 1995.

[38] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biology*, 2009.

[39] Maureen A. Sartor, Craig R. Tomlinson, Scott C. Wesselkamper, Siva Sivaganesan, George D Leikauf, and Mario Medvedovic. Intensity-based hierarchical bayes method improves testing for differentially expressed genes in microarray experiments. *BMC Bioinformatics*, 2006.