Scholars' Mine

Masters Theses   Student Theses and Dissertations

2013

Statistical analysis of microarray data in sleep deprivation

Stephanie Marie Berhorst

Follow this and additional works at: https://scholarsmine.mst.edu/masters_theses

Part of the Statistics and Probability Commons

Department:

Recommended Citation

STATISTICAL ANALYSIS OF MICROARRAY

DATA IN SLEEP DEPRIVATION

by

STEPHANIE MARIE BERHORST

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN APPLIED MATHEMATICS

2013

Approved by

Gayla R. Olbricht, Advisor
Matthew Thimgan
Robert Paige

ABSTRACT

Microarray technology is a useful tool for studying the expression levels of thousands of genes or exons within a single experiment. Data from microarray experiments present many challenges for researchers since costly resources often limit the experimenter to small sample sizes and large amounts of data are generated. The researcher must carefully consider the appropriate statistical analysis to use that aligns with the experimental design employed. In this work, statistical issues are investigated and addressed for a microarray experiment that examines how expression levels change over time as individuals are sleep deprived. Over the course of 48 hours of sleep deprivation, RNA is collected from saliva samples of two study participants. These samples were hybridized to exon microarrays to measure gene and exon expression at three different time points. Five different statistical analyses are conducted to test for expression differences over time. These analyses are carefully scrutinized and a thorough investigation is conducted on the microarray data. The different analyses elicited different findings. Several genes and exons are identified as differentially regulated over time and should be examined closer with regard to their relationship to sleep deprivation.

ACKNOWLEDGMENTS

I would personally like to thank Dr. Gayla Olbricht for giving me the opportunity to work under her supervision for the past few years on the topic of this thesis. The dedication and time she has set aside to assist me in learning and research has been invaluable to nurturing my development and successes along the way. I consider her a mentor and role-model. Without her, I would not have had the opportunity and skills to succeed in this thesis endeavor. I am forever grateful, again thank you.

Another important thanks needs to be extended to Dr. Matt Thimgan, without which this thesis would not be possible either. Thank you for allowing me to use the microarray data for this thesis. It has provided me with excellent opportunities to learn not only about statistics, but also in biological sciences and SAS programming as well.

I would also like to thank Dr. Robert Paige for providing me the opportunity to learn so much in classes that have helped prepare me for working on this thesis. In Statistics for Epidemiological Studies and Regression Analysis I was given tools that can be used not only for work on this thesis, but I fully believe they will extend to future professional projects as well. I certainly appreciate that I what I learned will be something I can use not only in a classroom but in real projects also.

Lastly I would like to thank Research Support Services in IT for all of their help with running SAS on the NIC Cluster as well as general questions with the technology used throughout conducting my thesis.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

1.1 GENERAL OVERVIEW

There are many different reasons why scientists study the human genome. The study of gene expression has commonly focused on detecting significant differences in the expression patterns of genes [1]. These differences are then explored further and studied in conjunction with diseases or inherited traits. Through the use of gene expression studies specific genes have been linked to certain diseases. Also these patterns have greatly aided in predicting the outcomes for these epidemics as well as determining treatments and courses of action. Just as there are many different reasons one might want to study gene expression patterns, there are different tools used to obtain information about the genes. Figure 1.1 shows the different components of a gene.
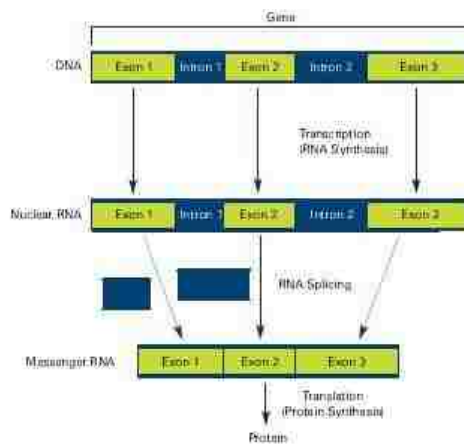
Figure 1.1 Exons Contained Within a Gene

A gene is a section of DNA located on a part of a chromosome containing chemical information necessary to specify certain proteins from RNA [2]. Genes contain relevant information that make each organism's cells and help to pass genetic information and traits to their offspring. Every living being has many different genes, each linked to various biological traits. There are approximately 20,000 genes in the human genome that are responsible for encoding proteins and RNA chains. In the genes of certain organisms that have a membrane bound nucleus, protein sequences are interrupted by stretches of DNA called introns. The introns are taken from the primary transcript during the formation of mRNA and the coding sequence that is left between the introns is called an exon. An exon is a sequence encoded by a gene that remains present within the final mature RNA product of that gene after introns have been removed via RNA splicing. The term exon refers to both the DNA sequence within a gene and to the corresponding sequence in RNA transcripts. Each gene contains many, different numbers of exons, hence there are more exons than there are genes as shown in Figure 1. 1.

Just as there are many different reasons to study gene expression, there are also different methods that can be used to measure gene expression. Microarray analysis is a very popular method in which genes are studied. In a microarray analysis, mRNA is isolated from a tissue sample such as a saliva or blood sample; it is then copied several times through a process called PCR, or polymerase chain reaction [3]. Once a large quantity of mRNA has been made, it is chopped up into millions of short pieces. The short mRNA pieces are then labeled and loaded into a microarray for gene expression quantification as shown in Figure 1.2.
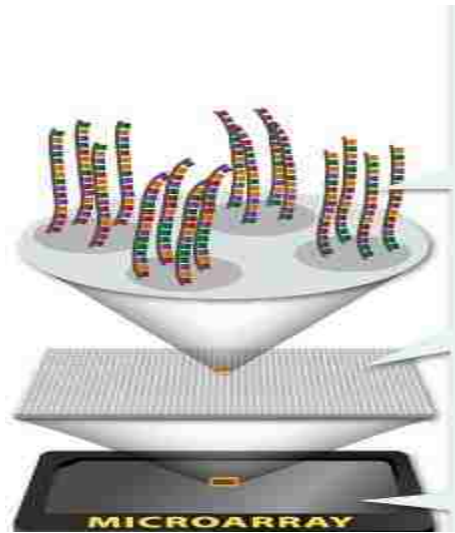
Figure 1.2 Microarray Technologies

After the microarray data are collected, they are generally stored in .CEL files. A statistical software package R Bioconductor [4], more specifically the "Biobase" and "Oligo" libraries can be used to read the gene expression data. The RMA (Robust Multichip Average) method is used to preprocess the data into a usable format. The RMA process consists of three steps; background correction, quantile normalization and lastly summarization [5]. A background correction method adjusts the data for noise and cross hybridization. This step also scales the data appropriately for the analysis. Quantile normalization is the process of reducing unnecessary variation among the biological information. The arrayed data is scaled so that the quantiles of each probeset are equal and have the same distribution across all arrays within a treatment group. Lastly, the data is summarized as a gene expression measure and is sent to the output file i.e. multiple probes from the same gene are summarized so that we have one expression measurement per gene or exon.

1.2 MICROARRAY DATA

Since gene expression measures help to study several biological variants in the

human genome, it makes sense to think that a person's sleeping pattern might affect the

overall gene expression as well. The goals of the following analyses are to determine if

any genes or exons are significantly expressed differently after sleep deprivation and

identify the pattern, if any they follow over time. For the analysis, RNA from two human

subjects was collected at three times: baseline prior to any sleep deprivation, after 24

hours without sleep and finally after 48 hours of being sleep deprived. RNA was

extracted from a saliva sample from each subject and hybridized to an Affymetrix Human

Exon 1.0 ST microarray [6], resulting in expression data from six different microarrays.

Table 1.1 shows the experimental set up where y_{ij} equals the expression level of subject j

at time i for a particular gene or exon.

Table 1.1 Experimental Observations

Microarray Data			
	Baseline	24 Hours	48 Hours
Subject 1	Y_{11}	Y_{21}	Y_{31}
Subject 2	Y_{12}	Y_{22}	Y_{32}

The objective is to detect genes/exons that are differentially expressed between

the different times, giving a measure of which genes/exons may be associated with sleep

deprivation. The initial experiment examined 22,011 genes and 1,411,399 exons, however some of the biological units did not have reliable evidence supporting the transcription or mapping of the genomic sequence. Possible evidence levels included for the probe sets are core, extended and full. The most reliable evidence level obtained is classified as "core". This classification refers to probe sets fully supported by RefSeq [7] and mRNA GenBank [8] or can be assigned reliably to a particular transcript cluster. Core was the only evidence level classification that was used for our analysis. After discarding the genes and exons whose evidence level was identified as extended or full we were left with 17,874 genes and 284,805 exons. A common theme throughout this study is testing the difference between the three time points, baseline, 24 and 48 hours. Several different tests were conducted to measure the difference between the three comparisons. This study deals with a few statistical issues as well. First of all there is a small sample size, and a lot of data being tested. Also there is a question of dependency in the biological units being compared. After both of these issues are combined into analyzing one set of data, tradeoffs are made in each test that is conducted. If one assumes that the data are dependent over time, fewer degrees of freedom are available with an already small sample. On the other hand, if the data is assumed to be independent over time, the variance will likely be inflated. Both of these options result in lowering the power of the tests conducted, so choosing an appropriate analysis is important. The study below analyzes the biological data at both the gene and exon level. As mentioned previously, there can be many exons per gene, so the exon level analysis is considerably lengthier. However the exon level data is specifically of interest because it is possible that in the saliva samples we may see cases where exons within the same gene show different

expression patterns. Also saliva degrades RNA, so specific exons may contain the most information. It is important to test at the exon level since the gene level analysis results may miss important differences, for example in a single exon when the probes are summarized across the whole gene.

2. STATISTICAL ANALYSIS - DEFINITIONS

2.1 T-TEST

The t-test is a statistical test that evaluates whether or not a population mean is equal to some value [9]. A t-test can be performed for one sample, two samples, and paired observations. In all three of these various t-tests, the true population mean and standard deviation are unknown. This is why a statistical test is conducted to test whether the true population mean is equal to some value. A one-sample t-test will test whether the population from which a sample was drawn has a mean value equal to a specific number thought to be the true mean. An independent two-sample t-test compares the means of two normally distributed populations. The two random samples taken from populations are assumed to be independent, meaning the first sample should not influence the second. Generally the test being conducted for an independent t-test is whether there is a difference between the two population means. The paired t-test on the other hand looks at the mean of the differences of two matched pairs or samples. In the paired t-test the population mean that is being tested is the mean difference or the average difference between the two sets of observations. Generally the hypothesis being tested for the paired t-test is that the mean difference is equal to zero. It is logical to assume that due to sampling error the sample mean will not be exactly equal to some theoretical value or the difference between two sample means will not be exactly zero even if this is true in the population. This is often due to random sampling error. The t-test tells us whether or not the true population mean differs from the value in question due to random sampling error

or due to true population differences. The main difference between the independent *t*-test and the paired *t*-test is that the paired test deals with how the data are collected on the statistical units in question. If the data measurements are taken on the same statistical units or we expect the data to have some correlation, a paired *t*-test is appropriate. If the difference being measured between the observations should not have any correlation or dependencies an independent *t*-test is the better analysis.

2.1.1. Independent Two Sample T-Test: General Definition. An independent two sample *t*-test is a statistical test that measures the population mean difference between two independent populations [9]. In laymen's terminology, this means that two different populations are being observed or analyzed, and the measured response in one group should not influence the response in the other group in any way.

2.1.1.1 Assumptions. When performing an independent t-test, certain criteria should be met. The measured observations for each group should follow a normal distribution. This can be tested by looking at a normal probability plot and/or plotting a frequency histogram for the response variables. The normal probability plot that is observed should resemble an approximately straight line and the frequency histogram should be unimodal and fairly symmetric. Next, the measured observations must be obtained from different/unrelated statistical units. This should be the case for both the statistical units that are being compared to one another in different groups as well as the units in the same population group. Lastly, the variances for the two groups are assumed to be equal to one another.

2.1.1.2 Hypothesis tests. The hypotheses being tested for the independent t-test are given below.

Null Hypothesis: The population means from the two unrelated groups are equal:

H_o: μ_i - $\mu_{i'}$ = 0.

Alternative Hypothesis: The population means from the two groups are not equal:

H_a: μ_i - $\mu_{i'}$ ≠ 0.

Here we define y_{ij} to be the response value for the jth individual (j=1,…,n) under

treatment i (i=1,2).

2.1.1.3 Test statistic. The *t*-statistic for the independent *t*-test is computed by:

$$T = \frac{\overline{y_1} - \overline{y_2}}{s_p\sqrt{\frac{1}{n} + \frac{1}{n}}}$$. Where we assume an equal sample size for the respective groups equal to

n and $\overline{y_1}$ and $\overline{y_2}$ are the sample means for the two independent groups respectively. s_p is

the pooled sample standard deviation for the data and s^2_p is the pooled sample variance

for all of the data calculated by $s^2_p = \frac{(n-1)s_1^2 + (n-1)s_2^2}{n+n-2}$. The pooled standard deviation

is simply the square root of the pooled variance. The significant level of a hypothesis test

represents the probability of committing a Type I error or a false positive result.

Commonly, studies set the significance level alpha (α) to be 0.05 or 0.1 and obtain the t

critical value by $t_{\alpha/2,2(n-1)}$. We compare the *t*-statistic to the *t*-critical value and reject the

null hypothesis if $|T| > t_{(\alpha/2,2(n-1))}$. A p value is obtained after calculating the *t*-statistic. The

p value represents the probability of obtaining a test statistic at least as extreme as what

we did, if the null hypothesis is really true. We can think of the p value as the probability

that the difference between the two sample means happened due to pure chance and not

because there is a true difference. For example, a p-value of .05 suggests that there is a

5% probability that the difference between the means obtained from the sample was indeed random or by occurred chance. In other words, assuming that in the true population there was no difference in the means whatsoever, and this experiment was repeated one time after another, approximately one in every 20 replications might elicit a false positive claim that there is sample evidence to suggest a difference between population means. The smaller the p value the less likely we are to commit a Type I error. This is why we report p values rather than solely rejection conclusions.

2.1.2. Independent Two Sample T-Test: Microarray Data. For this analysis, gene expression differences are examined between three time intervals. The expression measures are taken from the beginning of the trial (baseline), after 24 hours of sleep deprivation, and finally after each subject had gone 48 hours without sleep. An independent t-test is performed for each gene/exon three times. The first test compared the mean expression of the two subjects at baseline to that at 24 hours, the second test at baseline and 48 hours and the last at 24 and 48 hours. Although there is reason to believe the two samples are not independent since the data were collected on the same person, first an independent sample t-test is conducted to examine whether or not the genes and exons are correlated over time. It is logical to assume that some of the genes will be correlated while others may not. These results are compared to those of a paired t-test to examine this correlation.

2.1.2.1 Assumptions. For our microarray data, we assume that the biological expressions of the two populations follow a normal distribution. Since the sample size is only two, we are not able to draw conclusions based on the appropriate tests for normality. We would always be able to draw a straight line between only two different

observations. However previous studies with larger sample sizes report that the gene expressions follow the normal distribution [10]. Furthermore, the t-test is robust against departures from normality. Our study examines expression measures for two individuals over a 48 hour time period of sleep deprivation. The data was analyzed looking at differences from the same individual amongst the times measured; however it is not unreasonable to assume that some genes and exons wouldn't be related over time. Also one subject's expression measure should not influence the others, thus the subjects are independent of one another. Finally, we assume that the variance is similar at the three different time points.

2.1.2.2 Hypothesis tests. The specific hypotheses being tested for each gene or exon are given below. The expression measure is represented by yij for the jth subject (j=1, 2) at time i (i=1, 2, 3) for an individual gene or exon. The mean for each treatment time is calculated and compared to one another.

Null Hypothesis:

1.) There is no difference in mean expression measure between the baseline and 24 hours

of sleep deprivation: $\mu_1 - \mu_2 = 0$

2.) There is no difference in mean expression measure between the baseline and 48 hours

of sleep deprivation: $\mu_1 - \mu_3 = 0$

3.) There is no difference in mean expression measure between 24 and 48 hours of sleep

deprivation: $\mu_2 - \mu_3 = 0$

Alternative Hypothesis:

1.) There is a difference between the two means at baseline and 24 hours: $\mu_1 - \mu_2 \neq 0$

2.) There is a difference between the two means at baseline and 48 hours: $\mu_1 - \mu_3 \neq 0$

3.) There is a difference between the two means at 24 and 48 hours: $\mu_2 - \mu_3 \neq 0$

2.1.2.3 Test statistic. A t-statistic is calculated multiple (17,874 times for the gene level analysis and 284,805 times for the exon level analysis) times. The two sample treatment means are subtracted from one another and compared to the pooled standard deviation or $s_p = \sqrt{\dfrac{s_1^2 + s_2^2}{2}}$. In our study the sample size is two for all three treatments. This gives us two degrees of freedom. After the t-statistics are all computed, they are compared to the t-critical value $T \sim t(\alpha/2,2)$, or $|T| > 4.303$ for an alpha equal to 0.05.

2.1.3. Paired T-Test: General Definition. A paired t-test is a statistical test that measures the population mean difference between two dependent populations [9]. In laymen's terminology, this means that population difference that is observed or analyzed should be taken from related or similar statistical units.

2.1.3.1 Assumptions. When performing a paired t-test, certain criteria should be met. First, one needs to assume that the paired differences follow a normal distribution. This can be tested by looking at a normal probability plot, and/or plotting a frequency histogram of the paired differences. Next, the matched differences must be obtained from the same statistical unit or a matched pair. This might mean looking at differences from the same person, plot of land, piece of machinery, etc. The key is that the two measurements should be related in some way or another. Finally, all of the paired differences should be independent of one another, meaning the difference of one statistical unit should not influence another.

2.1.3.2 Hypothesis tests. The hypotheses being tested in the paired t-test are given below.

Null Hypothesis: The population mean difference is zero or H$_o$: $\mu_d = 0$.

Alternative Hypothesis: The population mean difference is not zero or H$_a$: $\mu_d \neq 0$.

Here we define y$_{ij}$ to be the response value for the jth individual (j=1,…,n) under

treatment i (i=1,2). These can be used to calculate paired differences d$_j$ = y$_{1j}$ − y$_{2j}$ which

represent the difference in response of the jth individual between treatments 1 and 2. In

the hypothesis, μ_d represents the population mean of the differences. The paired *t*-test is

actually a one-sample *t*-test on the paired differences.

2.1.3.3 Test statistic. The t-statistic is computed by: $T = \dfrac{\bar{d}}{s_d / \sqrt{n}}$. Where $\bar{d}$ is the

sample mean of the paired differences, computed by $\bar{d} = \dfrac{\sum\limits_{j=1}^{n} d_j}{n}$, s2d is the sample

variance of the paired difference, sd the sample standard deviation of the paired

differences, calculated by $s_d = \sqrt{\dfrac{\sum\limits_{j=1}^{n}(d_j - \bar{d})^2}{n-1}}$ and n is the sample size.

2.1.4. Paired T-Test: Microarray Data. For this analysis, differences in gene or

exon expression were examined between three time intervals. Expression measures were

obtained from the beginning of the trial (baseline), after 24 hours of sleep deprivation,

and finally after each subject had gone 48 hours without sleep. A paired t-test is

performed for each gene/exon three times. The first test compared the mean expression of

the two subjects at baseline to that at 24 hours, the second test at baseline and 48 hours

and the last at 24 and 48 hours. Since this part of the analysis looks to examine any

differences between time points of the same subjects, a paired t-test was selected over an

independent, two-sample t-test. A degree of freedom is lost in performing this type of analysis, however if there is a positive correlation between the data collected on the same individual, there should also be less variance at the different time points. Because of the reduction in variance, the paired t-test can be more powerful than the independent t-test, especially in larger sample sizes when the correlation is moderately high. It is unclear from the data whether there will be such an improvement in variance reduction or not. The paired t-test will be compared to the independent t-test to examine this.

2.1.4.1 Assumptions. For the microarray data, it is assumed that the paired differences follow a normal distribution. Since the sample size is only two, a test for normality cannot be conducted. However previous studies with larger sample sizes report that the paired differences follow this normal distribution [10]. This study examines expression measures for two individuals over a 48 hour time period of sleep deprivation. The data was analyzed looking at differences from the same individual amongst the times measured. The time points baseline, 24 and 48 hours should have some correlation in expression measure for the same individual. Lastly, one of the individual's expression measures should not influence the other over the different times.

2.1.4.2 Hypothesis test. The specific hypotheses being tested for each gene or exon are given below. The expression measure is represented by y_{ij} for the jth subject (j=1, 2) at time i (i=1, 2, 3) for an individual gene or exon. The paired differences are calculated by $d_{ij} = y_{ij} - y_{i'j}$, which is the expression difference for the jth subject at time i and i' (where $i \neq i'$). In the hypotheses below, $\mu_{dii'}$ represents the population mean of the differences at times i and i'.

Null Hypothesis:

1.) There is no difference in mean expression measure between the baseline and 24 hours

of sleep deprivation: $\mu_{d12} = 0$.

2.) There is no difference in mean expression measure between the baseline and 48 hours

of sleep deprivation: $\mu_{d13} = 0$.

3.) There is no difference in mean expression measure between 24 and 48 hours of sleep

deprivation: $\mu_{d23} = 0$.

Alternative Hypothesis:

1.) There is a difference between the two means at baseline and 24 hours: $\mu_{d12} \neq 0$.

2.) There is a difference between the two means at baseline and 48 hours: $\mu_{d13} \neq 0$.

3.) There is a difference between the two means at 24 and 48 hours: $\mu_{d23} \neq 0$.

      2.1.4.3 Test statistic. A t-statistic is calculated multiple (17,874 times for the

gene level analysis and 284,805 times for the exon level analysis). The sample mean

difference is calculated by summing the differences across time points from the

individuals and dividing by the sample size. In our study the sample size is two, or $n = 2$.

This leaves us with only one degree of freedom. A t-distribution with only one degree of

freedom is also known as a Cauchy distribution. A comparison of the Cauchy distribution

to a normal distribution is shown in Figure 2.1. Notice how the bell curve captures more

observations around the mean while the Cauchy distribution has many observations

falling under the thick right and left tails. After the t-statistics are all computed, they are

compared to the t-critical value $T\sim t(\alpha/2,1)$, or $|t| > 12.706$ for an alpha equal to 0.05.
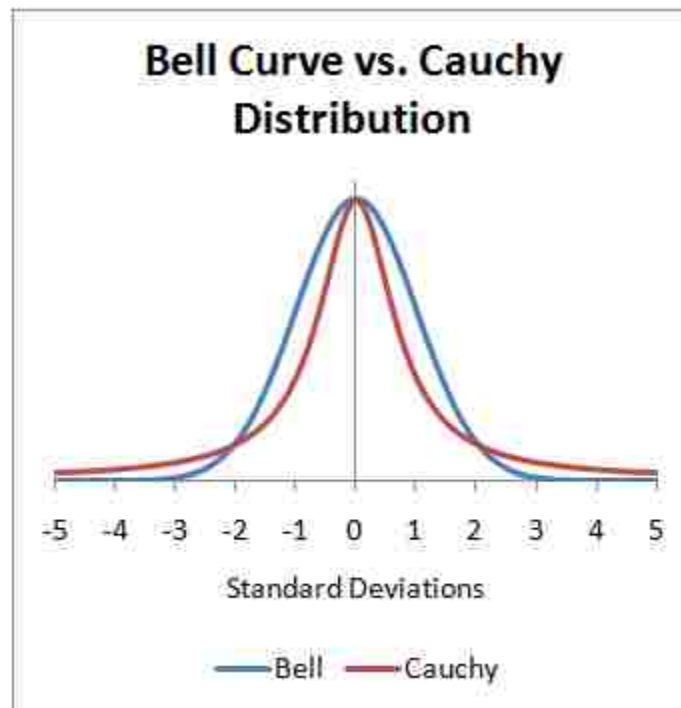
Figure 2.1 Normal vs. Cauchy Distribution

2.2 ANALYSIS OF VARIANCE

2.2.1. General Definition. A one-way analysis of variance (ANOVA) is similar to an independent two-sample t-test in that the test compares the population means for different treatment groups. A key difference between the two is that a one-way ANOVA allows for testing differences in population means when there are more than two treatment groups, and they are compared all at once. In general, ANOVA can test both main effects as well as interaction effects between the experimental variables [9]. The interaction effects looks at and tests a combination of variables, this is a multiplicative effect. In general, ANOVA allows for multiple main effects. These effects are additive. However for a one-way ANOVA, only one main effect is considered. A main effect measures the effect of one treatment variable against the response variable. It is different from the t-test because

here we can compare more than two means at one time. Also, the error degrees of
freedom will increase and allows a better estimate of the error variance. Two different
one-way ANOVA models can be considered:

Factor Effects Model:

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

Y_{ij} = Response for i^{th} treatment, j^{th} replicate i = 1, 2, …, a and j = 1, 2, …, n

μ = Overall mean

τ_i = Treatment effect for i = 1, 2, …, a

ε_{ij} = Error for i^{th} treatment, j^{th} replicate $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

and $\sum_{i=1}^{a} \tau_i = 0$ since the model is over parameterized.

Cell Means Model:

$Y_{ij} = \mu_i + \varepsilon_{ij}$

Y_{ij} = Response for i^{th} treatment, j^{th} replicate i = 1, 2, …, a and j = 1, 2, …, n

μ_i = Overall mean of treatment i, i = 1, 2, …, a

ε_{ij} = Error for i^{th} treatment, j^{th} replicate $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$

Table 2.1 ANOVA Table

Anova					
Source	Degrees of Freedom	Sums of Squares	Mean Squared	F Statistic	Critical F Value
Treatment	(a-1)	$SS_{Treatment} = \sum_{i=1}^{a} \frac{Y_{i\bullet}^2}{n} - \frac{Y_{\bullet\bullet}^2}{N}$	$MS_{Treatment} = SS_{Treatment}/(a-1)$	$F = MS_{Treatment}/MS_{Error}$	$F_{a-1,N-a}$
Error	(N-a)	$SS_{Error} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\bullet})^2$	$MS_{Error} = SS_{Error}/(N-a)$		
Total	(N-1)	$SS_{Total} = SS_{Treatment} + SS_{Error}$			

Table 2.1 is summarized below:

N is equal to $a \cdot n$ or the total number of observations. The total sum of squares can also

be computed by: $SS_{Total} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \overline{Y}_{\bullet\bullet})^2$ and can be thought of as the sums of squares

composed of both error and the treatment. It is used to measure the overall variability of

the data. Dividing the SS_{Total} by the total degrees of freedom, you obtain the sample

variance of the y_{ij}'s. The total sum of squares can be thought of as being partitioned

between a sum of squares differences between the treatment averages and grand mean

(Sum of Squares Treatment) plus the sum of squares of the differences of observations

within the treatments from the treatment average (Sum of Squares Error). In the SS_{Error}

term, if we divide $\sum_{i=1}^{n}(y_{ij} - \overline{y}_{i\bullet})^2$ by $(n - 1)$, this is the sample variance for the ith

treatment. If we look at a (treatment levels) sample variances and combine them all

together we obtain a single estimate of the population variance. This is equal to

$$\frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{Y}_{ij} - \overline{Y}_{i\bullet})^2}{N - a} \quad \text{or} \quad \frac{SS\ Errror}{N - a}, \text{ the Mean Squared Error, a pooled estimate of the variance}$$

across all of the treatments. If there aren't differences between the treatment means, use

the variation of the treatment averages subtracted from the grand mean to estimate the

variance $\dfrac{n\sum_{i=1}^{a}(\overline{Y}_{i\bullet} - \overline{Y}_{\bullet\bullet})^2}{a - 1} \quad \text{or} \quad \dfrac{SS\ Treatment}{a - 1}$, the Mean Squared Treatment. This is a valid

estimate for the variance when the treatment means are equal. There are now two

different ways to estimate the variance. One is based on variation within the treatments

(MSE), the other looks at variability between the treatments ($MS_{Treatments}$). When there is

no difference between the treatment means (the null hypothesis is true), the two mean squares should be very similar. If that is the case, the ratio between the two will elicit an F value near 1. Only if the F statistic is large enough relative to noise and signal in the data will we reject the null hypothesis in favor of the alternative.

   2.2.1.1 Assumptions. Some assumptions need to be met for the ANOVA model to be valid, mainly concerning the residuals of the error term, ε_{ij}. These residuals are assumed to be normally, randomly distributed with a mean of zero and a common variance σ^2. The residuals should be independent of one another, meaning the error from one observation should not influence another. Common types of residual plots are shown in Figure 2.2 The common variance assumption means that the residuals for the error term should be equal among all treatment groups, or the residuals are homogeneous.
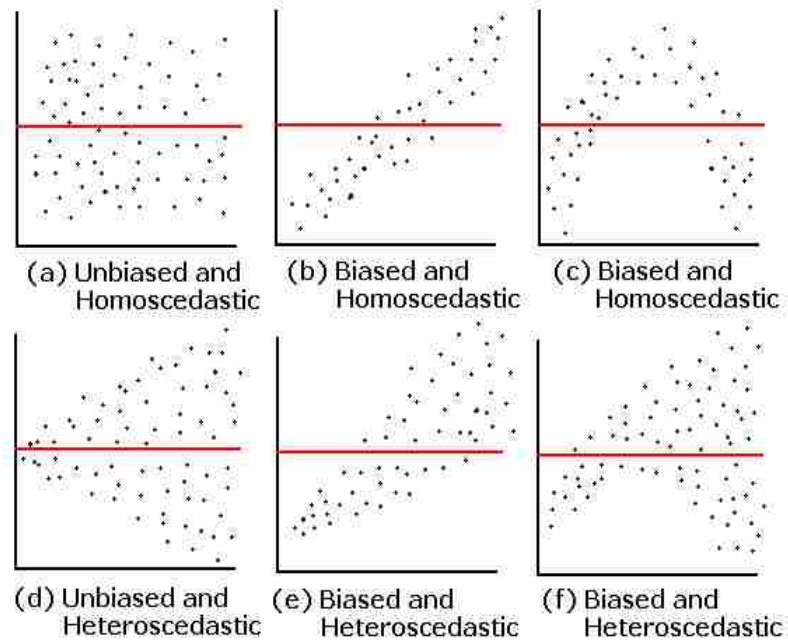
Figure 2.2 Diagnostic Plots for Residuals

Diagnostic tests are employed to evaluate these assumptions. One diagnostic test is a plot of the residuals versus the predicted values. Also the residuals can be plotted against the different treatment levels. The residual plots should look random and equally varied. If there is a clear trend in the plot of the residuals the model may need to be re-evaluated. Another item to observe is the normal probability plot. This will test to see if the data is normally distributed.

2.2.1.2 Overall f-test for equality of treatment means. The hypothesis being tested for the overall ANOVA is as follows (for the Factor Effects Model):

Null Hypothesis: There is no difference between any of the population means for the treatment effects, or $\tau_1 = \tau_2 = \ldots \tau_a = 0$ (where a = number of treatment effects).

Alternative Hypothesis: At least one of the population means for the treatment effects differs from one of the others, or $\tau_i \neq 0$ for at least one *i*.

2.2.1.3 F-statistic. The F-Statistic is computed by:

$$F_0 = \frac{MS_{Treatments}}{MS_{Error}} \text{ or } \frac{SS_{Treatment}/(a-1)}{SS_{Error}/(N-a)}$$ where a = the number of treatment groups and N =

the total number of observations, or y_{ij}. Commonly, studies chose alpha to be 0.05 or 0.1 and obtain an F critical value by $F_{\alpha, a-1, N-a}$ as shown in Figure 2.3. This F critical value is compared to F_0 and we reject the null hypothesis when $F_0 > F_{\alpha, a-1, N-a}$. After F_0 is compared to the F critical value, if one decides to reject the null hypothesis that one of the treatment means is indeed different than the others, there are a few other tests that can be conducted to make statistical inferences regarding the differences in treatment means.
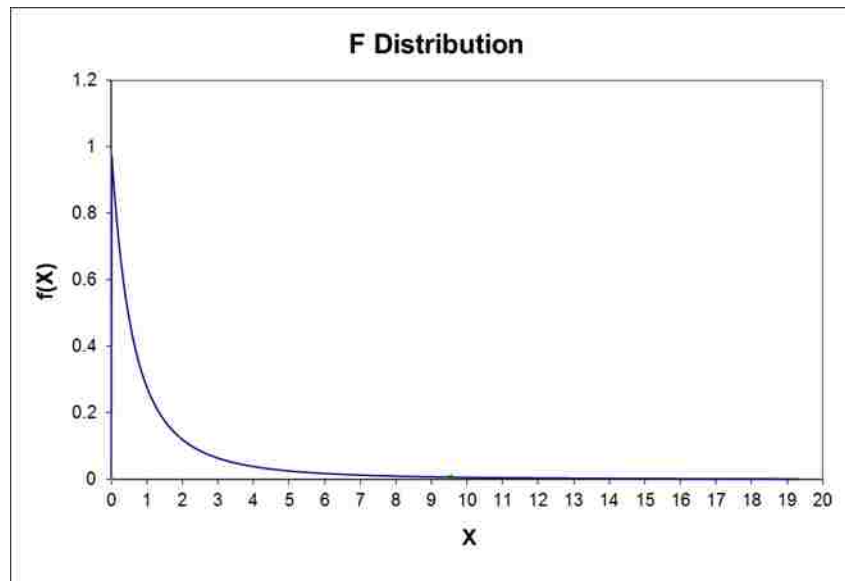
Figure 2.3 F Distribution with 2 Numerator and 3 Denominator Degrees of Freedom

2.2.1.4 Contrasts for the ANOVA model. A contrast is a linear combination of

the factor level population means defined as $L = c_1\mu_1 + ... + c_a\mu_a$ or $\sum_{i=1}^{a} c_i\mu_i$ where

$\sum_{i=1}^{a} c_i = 0$. An orthogonal contrast is a contrast whose sum of the coefficients sum to zero

[9]. Also the cross product of the coefficients should equal zero as well. The numbers of

different orthogonal contrasts that can be observed are equal to the degrees of freedom

for the treatment group, or the numerator degrees of freedom from the overall F test.

Contrasts are generally useful in analyzing pre-planned comparisons for the data, or

comparing treatment groups that the experimenters might already suspect to have

differences before analyzing data.

There are many different hypotheses to test when examining contrasts. Generally,

assume that the coefficients, c_i sum to zero, a general way to write the hypothesis tests

are as follows:

$$H_0: L = \sum_{i=1}^{a} c_i \mu_i = 0$$

$$H_a: \sum_{i=1}^{a} c_i \mu_i \neq 0$$

The test statistic for the different contrasts of a one-way ANOVA model each

follows a t-distribution. The contrast is written in terms of the sample treatment averages,

$\sum_{i=1}^{a} c_i \overline{y_{i\bullet}}$ and is compared to the mean square error. The test statistic is as follows:

$$t_0 = \frac{\sum_{i=1}^{a} c_i \overline{y_{i\bullet}}}{\sqrt{\frac{MS_E}{n} \sum_{i=1}^{a} c_i^2}}$$. This test uses the same degrees of freedom as the denominator

degrees of freedom (i.e. the error degrees of freedom) as the overall F test and the null

hypothesis is rejected if $|t_0| > t_{\alpha/2, N-a}$.

2.2.1.5 Pairwise comparisons for the ANOVA model. In some cases,

researchers might be interested in only comparing two treatments, or pairs of means to

one another. These are called pairwise comparisons and can be essentially thought of as a

special type of contrast where only two of the treatment means have coefficients that are

not equal to zero. One of the $c_{i's}$ of interest will be equal to one and the other negative

one. These pairwise comparisons are orthogonal because the coefficients sum to zero.

The pairwise comparisons are similar to an independent t-test in that we are comparing

two of the treatment means at one time, however the variance is computed differently and

degrees of freedom are gained. If the variance estimate is decreased then this test should

detect more differences than what would be found in a two-sample independent t-test.

The hypothesis test for the pairwise comparisons are similar again to the hypothesis for the contrasts only more specific because it only tests the difference of two population means. The null and alternative hypothesis tests are as follows:

H_o: μ_i - $\mu_{i'}$ = 0 or the two population treatment means are equal to one another.

H_a: μ_i - $\mu_{i'}$ $\neq$ 0 or the two population treatment means are not equal to one another.

Again, the test statistic for the pairwise comparisons is similar to that of the contrasts. It also follows a t-distribution and is written in terms of the treatment averages. However since it is only testing the difference between two means, the general form is $t_0 =$

$\dfrac{\overline{y}_{i\bullet} - \overline{y}_{i\bullet'}}{\sqrt{\frac{2}{n} MS_E}}$ Again this test statistic is compared to a critical t-value that is equal to $t_{\alpha/2, N\text{-}a}$

and we reject the null hypothesis when $|t_0| > t_{\alpha/2, N\text{-}a}$.

2.2.2. Microarray Data. The microarray data in this experiment was gathered at several different time points, so it makes sense to test differences between the means of the different times. Again, this is different from the independent t-test because multiple means at one time as opposed to two. Table 2.2 gives a general overview of the expression measures for each gene/exon.

Table 2.2 Observations and Means for Microarray Data

Microarray Data			
	Baseline	24 Hours	48 Hours	Subject Means
Subject 1	Y_{11}	Y_{21}	Y_{31}	$\overline{Y}_{\bullet 1}$
Subject 2	Y_{12}	Y_{22}	Y_{32}	$\overline{Y}_{\bullet 2}$
Treatment Means	$\overline{Y}_{1\bullet}$	$\overline{Y}_{2\bullet}$	$\overline{Y}_{3\bullet}$	$\overline{Y}_{\bullet\bullet}$

Each subject has expression measures collected over the different time points, Y_{ij},

$\overline{Y}_{i\bullet}$ is the treatment mean for the ith time point, $\overline{Y}_{\bullet j}$ is the mean for the jth subject and $\overline{Y}_{\bullet\bullet}$

is the overall mean across all subjects and time points. These treatment means are

compared to one another and tested to see if there are any statistically significant

differences at the population level. The two alternate formulations of the one-way

ANOVA here are:

Factor Effects Model:

$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

Y_{ij} = Expression for ith treatment, jth replicate i = 1, 2, 3; j = 1, 2

μ = Overall mean

τ_i = Treatment effect, i = 1, 2, 3

ε_{ij} = Error for ith treatment, jth replicate i = 1, 2, 3; j = 1, 2 $\varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$

and $\sum_{i=1}^{3} \tau_i = 0$ since the model is over parameterized.

Means Model:

$Y_{ij} = \mu_i + \varepsilon_{ij}$

Y_{ij} = Expression for ith treatment, jth replicate i = 1, 2, 3 j = 1, 2

μ_i = Overall mean of treatment i, i = 1, 2, 3

ε_{ij} = Error for ith treatment, jth replicate i = 1, 2, 3 j = 1, 2 $\varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$

Table 2.1 shows the general ANOVA table, but the degrees of freedom can be specifically updated given the number of different treatments and replicates in the study. The treatment degrees of freedom are generally computed as (a-1). For the microarray data this is equal to two. The error degrees of freedom are calculated as (N-a) where N is the total number of observed expressions. For this data set, N = 6 so the error degrees of freedom is 3. The total degrees of freedom can be figured by (N-1) or (a-1) + (N-a), which is equal to 5 total degrees of freedom.

    2.2.2.1 Assumptions. While certain assumptions should be met for the ANOVA model to be valid, just like in the independent *t*-test we assume that these assumptions hold true. It is difficult to test the residuals of the error term simply because we have a small sample size. We assume that the error terms are normally, randomly distributed around zero and that they are independent of one another. We will account for the potential dependence between observations on the same subject in a repeated measures model. The results of these two models will be compared to examine the potential correlation in the data.

    2.2.2.2 Overall f-test for equality of treatment means. The specific hypotheses

being tested for each gene or exon are given below. The treatment effect mean is represented by τ_i at time i (i=1, 2, 3) for an individual gene or exon when using the factor effects model. Alternatively, if the cell means model is employed, the treatment mean is represented by μ_i at time i (i=1, 2, 3). The null and alternative hypothesis are listed below:

Null Hypothesis: $\tau_1 = \tau_2 = \tau_3 = 0$ or all of the treatment effects are equal to zero or

$\mu_1 = \mu_2 = \mu_3$ the treatment means are all equal.

Alternative Hypothesis: At least one of the τ_i's is not equal to zero or not all μ_i's are equal.

2.2.2.3 F-statistic. The F-statistic is computed for each gene or exon after the ANOVA table from Table 2.1 is calculated for each individual biological unit. It is figured by $F_0 = \dfrac{SS_{Treatment}/2}{SS_{Error}/3}$ and is then compared to the critical f-value or $F_{0.05,\,2,\,3,}$ which is 9.55. When F_0 is greater than 9.55 the null hypothesis is rejected because there is sufficient evidence to suggest that the population treatment means/effects are not all equal to one another.

2.2.2.4 Contrasts and pairwise comparisons. For the microarray data all of the orthogonal contrasts were tested. Since there are three different treatment groups there are two degrees of freedom associated with the treatment effect. Therefore there are two different orthogonal contrasts to test. The first contrast compares the mean of one treatment group to that of the other two or $C_1 = \mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$ and this is done three times so that each mean is compared to the other two. The other orthogonal contrast compares only two of the means to each other. Again, this contrast can be thought of as a

pairwise comparison because it is comparing a set of paired data. $C_2 = \mu_1 - \mu_2$ is tested

so that the each treatment mean is compared to the other.

The hypothesis test for the first contrast, C_1 for the microarray data is as follows:

$$H_{o1}: 2\mu_1 - \mu_2 - \mu_3 = 0 \quad H_{a1}: 2\mu_1 - \mu_2 - \mu_3 \neq 0$$

$$H_{o2}: 2\mu_2 - \mu_1 - \mu_3 = 0 \quad H_{a2}: 2\mu_2 - \mu_1 - \mu_3 \neq 0$$

$$H_{o3}: 2\mu_3 - \mu_1 - \mu_2 = 0 \quad H_{a3}: 2\mu_3 - \mu_1 - \mu_2 \neq 0$$

The hypothesis that is being tested for C_1 is whether the treatment means of group i is

different than the average of the other two treatment means. Each null hypothesis posits

that there is no difference between the population treatment means while the alternative

states that there is difference. The hypothesis test for C_2 or the pairwise comparisons are

as follows:

$$H_{o1}: \mu_1 - \mu_2 = 0 \qquad\qquad H_{a1}: \mu_1 - \mu_2 \neq 0$$

$$H_{o2}: \mu_1 - \mu_3 = 0 \qquad\qquad H_{a2}: \mu_1 - \mu_3 \neq 0$$

$$H_{o3}: \mu_2 - \mu_3 = 0 \qquad\qquad H_{a3}: \mu_2 - \mu_3 \neq 0$$

These hypotheses are testing to see if there is a difference between two of the treatments.

Again, each null hypothesis states that there is no difference between the two population

treatment means, the alternative stating that there is a difference in the means. For the

contrast C_1 a t-statistic is calculated for each gene and exon. For the pairwise

comparisons, each gene and exon has a test statistic. After the t-statistics are all

computed, they are compared to the t-critical value T~$t_{(\alpha/2,3)}$, or $|T| > 3.187$.

2.3 MIXED MODEL

2.3.1. General Definition. A mixed model examines a relationship between a response variable along with predictor variables that are observed or measured during the experiment [9]. A key difference between a mixed model and a fixed effect model such as the one way ANOVA previously described is a mixed model will treat at least one of the variables as fixed while another is treated as random. The fixed effects have levels that are of primary interest and could possibly be used again if the experiment were repeated. Random effects have levels that are generally not of primary interest, but rather are thought of as a random selection from a population. In the case of a repeated measures experiment where the same subjects are measured across treatments or times, generally subject effects will be the random effects, while different treatment levels are the fixed effects. The idea behind the repeated measures mixed model is that the fixed effects should tell how the population means differ between the set of treatments, while the random effects represent the variability among measurements collected on the same subjects or units. Another key difference between the ANOVA model and the repeated measures mixed model is the way each computes the variance. The ANOVA model uses moment estimators of the data to calculate the variance. It cannot work with missing data, which can often be a problem when a study looks at data over time. This is an advantage that the mixed model has over a general, one-way ANOVA. By default, the mixed model uses a restricted maximum likelihood to calculate variance. This is helpful when there are noise parameters or missing data. However, like the ANOVA model, the mixed model allows testing multiple means at once. It is particularly useful when repeated measurements are made on the same statistical units, or where measurements are made on clusters of related statistical units since it incorporates any correlation between data

collected on the same subject into the variance calculation. Because of this, it draws similarities from both the paired t-test and the ANOVA model. The repeated measures mixed model can be defined as:

$Y_{ij} = \mu + \alpha_i + \rho_j + \varepsilon_{ij.}$ Where:

Y_{ij} is the observed response at fixed effect level i = 1, 2, …, a and random effect level j = 1, 2, …, n.

μ is the overall mean response.

α_i is the fixed main effect of level i = 1, 2, …, a and $\sum_{i=1}^{a} \alpha_i = 0$

ρ_j is the random effect of level j, $\rho_j \overset{iid}{\sim} N(0, \sigma_\rho^2)$

ε_{ij} is a random error term $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$

The variance of the Y_{ij} is the total variance from the model and can be thought of as a combination of the variance from both the error in the model summed with the variance from the random effect. The proportion of the total variation in the response due to within subject variation is the correlation coefficient and is defined as

$\omega = \dfrac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma_\varepsilon^2}$. The ANOVA table for a mixed model is similar to a one way ANOVA table, however rather than only having a treatment effect, both the random and fixed effect are included.

2.3.1.1 Assumptions. Since the repeated measures mixed model draws similarities from both the ANOVA model and the paired t-test, there is a combination of assumptions that should be met. First like the paired t-test, the data should be collected on

the same statistical unit, or one should expect some correlation among the data. The observations between the fixed effects should be independent, or one should not influence another. The random subject should be normally distributed with a mean centered at zero and with an unknown, similar variance. Also, the residuals should be normally distributed, equally varying, and independent among one another. The same diagnostic tests can be conducted as from the ANOVA model. One assumption that largely differs from the general linear model is the sphericity assumption [9]. When the mixed model is conducted and a repeated measures analysis is appropriate, the data does not need to be assumed to have similar correlations between the treatments. While compound symmetry (the sphericity assumption) is an option, there are also other options for choosing a covariance structure that can improve the data analysis.

2.3.1.2 Compound symmetry vs. autoregressive model. If a mixed model is an appropriate analysis for the data in question, it can be more powerful than a general linear model because it can better account for missing data and also we can vary the covariance matrix. Whenever a repeated measures mixed model is appropriate (that is, more than one observation has been collected from a subject) choosing a within subject covariance matrix is an important part of the analysis. The size of this matrix is determined by the number of observations collected from the data. For example, if one measurement is obtained at two different times, this will generate a 2x2 covariance matrix. If measurements from 4 different treatments are applied to each subject, this will elicit a 4x4 matrix. The structure of this matrix is also crucial. The variance of the difference in two within-subject means is calculated by:

$Var(\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}) = Var(\overline{Y}_{1\bullet}) + Var(\overline{Y}_{2\bullet}) - 2Cov(\overline{Y}_{1\bullet}, \overline{Y}_{2\bullet})$. The covariance term will make the

within subject comparisons more powerful than the between subject comparisons, since

this term is generally positive. The value of the covariance is determined by which

covariance matrix chosen. This is why it is very important to choose the correct

covariance structure for the analysis. The mixed model allows for many different types of

covariance matrices, but one common, simple type is called compound symmetry. This

covariance structure implies that each measured statistical unit is equally correlated with

one another and the total variation, $\sigma_y^2 = \sigma_\rho^2 + \sigma_\varepsilon^2$, can be partitioned into the shared,

within unit component, σ_ρ^2, and the unshared, or error component, σ_ε^2. The correlation

coefficient $\omega = \dfrac{\sigma_\rho^2}{\sigma_\varepsilon^2 + \sigma_\rho^2}$ is assumed to be equal across all fixed effects and they all share a

common error variance σ_ε^2. The compound symmetry within subject variance covariance

matrix is given below:

$$\sigma_\varepsilon^2 \begin{bmatrix} 1 & \omega & \omega & \omega \\ \omega & 1 & \omega & \omega \\ \omega & \omega & 1 & \omega \\ \omega & \omega & \omega & 1 \end{bmatrix}$$

One other type of covariance structure that can be used in a mixed model is called the

autoregressive model, or AR(1). Autoregressive is a term from times series analysis

which assumes observations are related to their own past values through the number of

lags between them [11]. An autoregressive structure implies that two observations closer

to each other over time or space should be more highly correlated than two observations

that are farther apart. The correlation coefficient between observations decreases (or

regresses) as difference between the observations become further apart. The AR(1) within

subject covariance matrix is given below:

$$\sigma_\varepsilon{}^2 \begin{bmatrix} 1 & \omega & \omega^2 & \omega^3 \\ \omega & 1 & \omega & \omega^2 \\ \omega^2 & \omega & 1 & \omega \\ \omega^3 & \omega^2 & \omega & 1 \end{bmatrix}$$

There are tests that can be conducted to help choose the best covariance structure for the data. A likelihood ratio test can be conducted as well as observing two different information criteria tests. The statistic that is observed for the likelihood ratio test is called a -2log Likelihood statistic. This ratio test is used to calculate two information criteria tests called the AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) statistics. Smaller AIC and BIC scores are ideal, so choosing the structure that lowers these is a good idea.

2.3.1.3 Hypothesis tests for overall effects. The hypothesis test for the mixed model is similar of that for the ANOVA model. The hypothesis being tested for the fixed effects is:

Null Hypothesis: There is no difference between any of the population means for the
fixed effect levels, or $\alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$ (where a = number of levels of the
fixed effect)

Alternative Hypothesis: At least one of the population means for the fixed effect levels
differs from the others, or $\alpha_i \neq 0$ for at least one *i*.

The mixed model can also test whether there is significant within subject variation. The tests are as follows:

Null Hypothesis: $\sigma_\rho^2 = 0$

Alternative Hypothesis: $\sigma_\rho^2 > 0$

2.3.1.4 Test statistic. The test for the fixed effect is typically the primary test of

interest and the test statistic for this test is: $F_0 = \dfrac{MS_{Fixed}}{MS_{Error}}$ or $\dfrac{SS_{Fixed}/(a-1)}{SS_{Error}/(a-1)(n-1)}$ where a

is the number of levels of the fixed effects and n is the total number of levels of the

random effect. Commonly, studies chose alpha to be 0.05 and obtain an F critical value

by $F_{\alpha,\,a-1,(a-1)(n-1)}$. This F critical value is compared to F_0 and we reject the null hypothesis

when $F_0 > F_{\alpha,\,a-1,(a-1)(n-1)}$. After F_0 is compared to the F critical value, if one decides to

reject the hull hypothesis that one of the treatment means is indeed different than the

others, there are a few other tests that can be conducted to make statistical inferences

regarding the differences in treatment means.

2.3.1.5 Contrasts and pairwise comparisons. A contrast is a preplanned

comparison of the treatment means for some data set where $L = c_1\mu_1 + ... + c_a\mu_a$ or $\sum_{i=1}^{a} c_i\mu_i$

and L is a linear combination of the treatment means, μ_i. An orthogonal contrast is a

contrast whose coefficients sum to zero. The sum of these cross products should also

equal zero. The numbers of different orthogonal contrasts that can be observed are equal

to the degrees of freedom for the treatment group, or the numerator degrees of freedom

from the overall F test for fixed effects. Contrasts are generally useful in analyzing pre-

planned comparisons for the data, or comparing treatment groups that the experimenters

might already suspect to have differences before analyzing data. The contrasts for the

mixed model are very similar to the contrasts for the ANOVA model. The difference

between the two lies in the degrees of freedom. The degrees of freedom for the two tests are different which results in the two tests having a different mean squared error as well.

There are many different hypotheses to test when examining contrasts. Assume that the coefficients, c_i sum to zero, a general way to write the hypothesis tests are as follows:

$$H_0: \sum_{i=1}^{a} c_i \mu_i = 0$$

$$H_a: \sum_{i=1}^{a} c_i \mu_i \neq 0$$

The test statistic for the different contrasts of a one-way ANOVA model each follows a t-distribution. The estimate for the contrast is written in terms of the sample treatment averages, $\sum_{i=1}^{a} c_i \overline{y_{i\bullet}}$ and is compared to the mean square error. The test statistic is

as follows: $t_0 = \dfrac{\sum_{i=1}^{a} c_i \overline{y_{i\bullet}}}{\sqrt{\dfrac{MS_E}{n} \sum_{i=1}^{a} c_i^2}}$. This test uses the same degrees of freedom as the

denominator degrees of freedom from the overall F test for fixed effects and the null hypothesis is rejected if $|t_0| > t_{\alpha/2,(a-1)(n-1)}$. The main difference in the contrasts and pairwise comparisons from the ANOVA model and the mixed model is in the number of degrees of freedom that are available. This changes the MSE as well as the entire ANOVA table. The test statistic for the pairwise comparisons is similar to that of the contrasts. It also follows a t-distribution and is written in terms of the treatment averages. However since it

is only testing the difference between two means, the general form is $t_0 = \dfrac{\overline{y_{i\bullet}} - \overline{y_{i\bullet'}}}{\sqrt{\frac{2}{n} MS_E}}$. Again

this test statistic is compared to a critical *t*-value that is equal to $t_{\alpha/2,(a-1)(n-1)}$ and we reject

the null hypothesis when $|t_0| > t_{\alpha/2,(a-1)(n-1)}$.

    2.3.2. Microarray Data. The microarray data involves both a random (subject)

and fixed (time) effect, so a mixed model will hopefully elicit the best results due to the

correlation between the subjects over time. Because a new parameter is introduced into

the model (the random subject effect), the mixed model will have different degrees of

freedom than the one-way fixed effect ANOVA, particularly in the error term. One

degree of freedom is lost to the subject effect. The subject effect will have (n-1), or 1

degree of freedom. The fixed time effect still has 2 degrees of freedom, or (a-1). This

only leaves 2 degrees of freedom left for the error or (a-1)(n-1). This leaves less degrees

of freedom from the one-way ANOVA model, but since the data is measured over time,

the assumption that the data has some correlation over time will hopefully overcome that.

The repeated measures mixed model for the microarray data can be defined as:

    $Y_{ij} = \mu + \alpha_i + \rho_j + \varepsilon_{ij.}$ Where:

        Y_{ij} is the expression level at time point i = 1, 2, 3 and subject j = 1, 2.

        μ is the overall mean expression.

        α_i is the fixed effect, time at level i = 1, 2, 3.

        ρ_j is the random subject effect of level j = 1, 2 $\rho_j \overset{iid}{\sim} N(0,\sigma_\rho^{\ 2})$

        ε_{ij} is a random error term $\varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma_\varepsilon^{\ 2})$

Normally there is an interaction term between the fixed and random variable that could

be considered in the model. However since our sample is very small, we could not afford

the degrees of freedom to include this in our analysis. Therefore we only examine main

effects of subject and time.

     2.3.2.1 Assumptions. For the mixed model to be a valid option for the microarray data, several assumptions must be met. The data is collected on the same two individuals over time since we are expecting some correlation or dependencies over time. However the two individuals should not bear any influence on each other's biological expression, the different subjects should be independent of one another. The measured expressions and the error are assumed to be normally distributed with a mean of zero and an unknown variance. This assumption should also hold for the subject effect as well.

     2.3.2.2 Hypothesis tests for overall effects. The hypotheses being tested for the microarray data for the repeated measures mixed model for the fixed effects is as follows:

Null Hypothesis: There is no difference between any of the population means for the

     fixed effect time, or $\alpha_1 = \alpha_2 = \alpha_3 = 0$

Alternative Hypothesis: At least one of the population means for the fixed effect time,

     differs from the others, or $\alpha_i \neq 0$ for at least one i where i = 1, 2, 3.

     2.3.2.3 Test statistic. $F_0 = \dfrac{MS_{Fixed}}{MS_{Error}}$ or $\dfrac{SS_{Fixed}/2}{SS_{Error}/2}$. The F critical value is

represented by $F_{\alpha, 2, 2}$ and is equal to 19.0. This F critical value is compared to F_0 and we reject the null hypothesis when $F_0 > 19$. After F_0 is compared to the F critical value and a rejection decision has been made the contrasts and pairwise comparisons are conducted.

     2.3.2.4 Contrasts and pairwise comparisons. For the microarray data all possible orthogonal contrasts were tested. Since there are three different treatment groups being considered, this leaves two degrees of freedom associated with the treatment effect.

Therefore there are two different orthogonal contrasts to test. The first contrast for the

microarray data compares the mean of one treatment group to that of the other two or

$C_1 = \mu_1 - \frac{1}{2}(\mu_2 + \mu_3)$. This is done three times so that each mean is compared to the other

two. The other orthogonal contrast compares only two of the means to each other. Again,

this contrast can be thought of as a pairwise comparison because it is comparing a set of

paired data. $C_2 = \mu_1 - \mu_2$ is an example of one of the three paired comparisons for the

microarray data. A general form for the second orthogonal contrast is $C_2 = \mu_i - \mu_{i'}$.

The hypothesis test for the first contrast, C_1 for the microarray data is as follows:

$$H_{o1}: 2\mu_1 - \mu_2 - \mu_3 = 0 \quad H_{a1}: 2\mu_1 - \mu_2 - \mu_3 \neq 0$$

$$H_{o2}: 2\mu_2 - \mu_1 - \mu_3 = 0 \quad H_{a2}: 2\mu_2 - \mu_1 - \mu_3 \neq 0$$

$$H_{o3}: 2\mu_3 - \mu_1 - \mu_2 = 0 \quad H_{a3}: 2\mu_3 - \mu_1 - \mu_2 \neq 0$$

The hypothesis that is being tested for C_1 is, whether the treatment means of group i

differs from the average of the other two treatment means. Each null hypothesis posits

that there is no difference between the population treatment means while the alternative

states that there is difference. The hypothesis test for C_2 or the pairwise comparisons are

as follows:

$$H_{o1}: \mu_1 - \mu_2 = 0 \qquad\qquad H_{a1}: \mu_1 - \mu_2 \neq 0$$

$$H_{o2}: \mu_1 - \mu_3 = 0 \qquad\qquad H_{a2}: \mu_1 - \mu_3 \neq 0$$

$$H_{o3}: \mu_2 - \mu_3 = 0 \qquad\qquad H_{a3}: \mu_2 - \mu_3 \neq 0$$

These hypotheses are testing to see if there is a difference between two of the treatments.

Again, each null hypothesis states that there is no difference between the two population

treatment means, the alternative stating otherwise.

For the contrast C_1 a t-statistic is calculated for each gene and exon. For the pairwise comparisons, each gene and exon has a test statistic value. After the t-statistics are all computed, they are compared to the t-critical value T~$t_{(\alpha/2, 2)}$, or $|t| > 4.303$. The null hypothesis is rejected when the t-statistic is greater than the t-critical value.

2.4 EMPIRICAL BAYES

2.4.1. General Definition. The empirical Bayes or moderated t-test has similarities to the independent t-test. Both of the tests analyze response measures from two different populations. Just like comparing the paired and independent t-tests, inherent differences exist for the empirical Bayes in how the degrees of freedom are calculated as well as the variance for the data. The moderated t-test draws information from all of the genes or exons contained in the analysis to improve the power of the t-test [12]. Rather than estimating the variance for each gene or exon, the moderated t uses all genes/exons to estimate the variance. The power of the test will increase because often times we are able to gain a substantial amount of degrees of freedom as compared to both the independent and paired t-test.

2.4.1.1 Assumptions. The assumptions for the moderated t-test are similar to that of the independent t-test. The data collected should be from two different populations should follow a normal distribution with equal variances. The populations should be

independent of one another and the observations within the populations should also be independent.

2.4.1.2 Hypothesis test. The moderated *t*-test has similar hypothesis tests to the

independent *t*-test as well. The test that is being conducted is still testing to see if there is

a difference between the two different population means.

H_o: There is no difference between the two population means or $\mu_i - \mu_{i'} = 0$

H_a: There is a difference between the two population means or $\mu_i - \mu_{i'} \neq 0$

2.4.1.3 Test statistic. The moderated *t*-test is available in the Limma package

from Bioconductor in R. The Limma package is used for analyzing expression measures

for microarray data, or specifically experiments involving comparisons of many RNA

samples at the same time. Limma uses linear models to analyze the microarray

experiments.

This approach requires two matrices to be specified; a design matrix and a

contrast matrix. The design matrix provides a representation of the different RNA targets

hybridized to the arrays. The contrast matrix allows the coefficients defined by the design

matrix to be combined into contrasts of interest. Each contrast corresponds to a paired

comparison of interest between the RNA targets [12]. The expected value or mean as

well as the variance for each gene are computed via these matrices. A weight of the

matrix is used in conjunction with the variance of the biological expression. From these

two measurements the test statistic is calculated by $t_{gj} = \dfrac{\beta_{gj}}{s_g \sqrt{\upsilon_{gj}}}$ where g = the gene or

exon index and j = the specific contrast of interest. This test statistic is compared to the t

critical value and we reject the null hypothesis if $|T| > t_{(\alpha/2, dg + do)}$. The β_{gj} is equivalent to

$\bar{y}_i - \bar{y}_{i'}$ from the regular, independent *t*-test and the υ_{gj} is $\frac{2}{n}$ when the sample sizes are

equal. If the gene or exon is not differentially expressed, the difference between the two

population means will be zero. The s_g is estimated via empirical Bayes method. From

Bayes' theorem we assume a prior probability, $p(\theta|\alpha)$ for σ_g^2 that is based off of some

prior knowledge before any data has been collected. The parameters of the prior

probability are called hyperparameters, for σ_g^2 these are d_o and s_o^2 (degrees of freedom

and variance respectively). Some information is obtained strictly from the data as well.

This information is based off the likelihood, $p(x|\theta)$ and marginal likelihood, $p(x|\alpha)$ from

the data. The posterior probability $p(\theta|x,\alpha)$ is a conditional probability of the parameters

given the prior knowledge and updated knowledge given from the data being analyzed.

The posterior distribution is calculated by $p(\theta \,|\, x, \alpha) = \dfrac{p(x\,|\,\theta)\,p(\theta\,|\,\alpha)}{p(x\,|\,\alpha)}$. Assuming the

posterior distribution of σ_g^2 follows some distribution, a mean and variance can be

estimated. The common variance estimated from the mean of the posterior σ_g^2 is

calculated by $s_g^{-2} = \dfrac{d_o s_o^2 + d_g s_g^2}{d_o + d_g}$ where d_o, s_o^2 are the hyperparameters and s_g^2 and d_g are

the sample variance and degrees of freedom equivalent to the one-way ANOVA. The

square root of this is used as the standard deviation in the moderated *t*-test. Truly

Bayesian statistics would allow the user to define the values of the hyperparameters d_o

and s_o^2. However empirical Bayes methods estimate the hyperparameters from the data

and ultimately draw information from all of the genes or exons in doing so.

   2.4.2. Microarray Data. The data being analyzed from this experiment does

come from a microarray experiment. RNA was collected from saliva samples resulting in

numerous genes and exon expression measurements. This should be beneficial to the analysis because the variance will be computed differently based on all of the genes and exons as opposed to treating each one as its own entity. Also recalling the independent and paired t-test, simply adding in one degree of freedom changed the critical t-value immensely, allowing for more power in the test. We can expect to see an increase in the degrees of freedom by using information from all of the genes or exons which should allow for more of the tests to become significant when there are true differences in expression.

  2.4.1.1 Assumptions. As mentioned previously, the data does come from a microarray experiment. The data comes from two normal populations, and the observations are assumed to be independent from one another.

  2.4.1.2 Hypothesis test. Again, the hypothesis test is similar to the independent *t*-test.

Null Hypothesis:

1.) There is no difference in expression measure between the baseline

   and 24 hours of sleep deprivation: $\mu_1 - \mu_2 = 0$

2.) There is no difference in expression measure between the baseline

   and 48 hours of sleep deprivation: $\mu_1 - \mu_3 = 0$

3.) There is no difference in expression measure between 24 and 48 hours of

   sleep deprivation: $\mu_2 - \mu_3 = 0$

Alternative Hypothesis

1.) There is a difference between the means at baseline and 24 hours: $\mu_1 - \mu_2 \neq 0$

2.) There is a difference between the means at baseline and 48 hours: $\mu_1 - \mu_3 \neq 0$

3.) There is a difference between the means at 24 and 48 hours: $\mu_2 - \mu_3 \neq 0$

      2.4.1.3 Test statistic. The test statistic from Formula (?) is calculated for each gene and exon. The d_g and s_g are actually calculated the same way as we calculated the degrees of freedom and MSE from the paired comparison in the ANOVA model. This gives us three degrees of freedom for d_g and a way to calculate s_g. The d_o and s_o^2 are a bit more challenging to calculate however. The $z_j = \log(s_j^2)$ follows a Fishers Z Distribution, with mean and variance functions $E(z_j) = \log s_o^2 + \psi(d_g/2) - \psi(d_o/2) + \log(d_o/d_g)$ and $Var(z_j) = \psi'(d_g/2) + \psi'(d_o/2)$ where ψ and ψ' are the first and second derivatives of the log of the gamma function. The final estimates for d_o and s_o are:

$$d_o = 2\psi'^{-1}\left(\frac{1}{J-1}\sum_{j=1}^{J}(z_j - \bar{z})^2 - \psi'(d_g/2) \right) \text{ and}$$

$$\hat{s}_o^2 = \exp\left\{ \bar{z} - \psi(d_g/2) + \psi(d_o/2) - \log(d_o/d_g) \right\}. \text{ For the microarray data gene level}$$

analysis, $d_o \approx 2.15$ and in the exon level analysis $d_o \approx 4.88$. For the microarray data in question, the total degrees of freedom for the gene level analysis is equal to 5.15. For the exon level data the total degrees of freedom is equal to 7.88. The *t*-statistic is compared to the critical t value and the null hypothesis is rejected when $| t_{gi} | > T \sim t_{\alpha/2, do+dg}$ or 2.571 for the gene level analysis and 2.365 for the exon level analysis.

3. REPEATED MEASURES VS INDEPENDENT TESTS

When conducting the above tests, there are several costs and benefits to choosing to use an independent test versus a dependent test. In the independent tests, one assumes that the data are completely independent within the subjects. The dependent tests assume that there is some correlation in the data within the subjects. One thing that changes when we make these different assumptions is the degrees of freedom. When the assumption is that there is some correlation over time for the subjects, we lose a degree of freedom to estimating this. However, the thought is that this degree of freedom can be made up for by hopefully observing a smaller variance in the data when we assume this correlation. Table 3.1 summarizes the analysis considered for these data.

Table 3.1 Dependent vs. Independent Measures

Independent Tests			Dependent Tests		
Test	DF_E	Critical Values	Test	DF_E	Critical Values
Independent T-Test	2	4.303	Dependent T-Test	1	12.706
One-Way ANOVA	3	3.182	Repeated Measures Mixed	2	4.303
Moderated T-Test	5.15 (gene) 7.87 (exon)	2.571 2.365			

In observing the degrees of freedom from each test, it is very apparent that adding degrees of freedom immensely improves the power analysis. However, what cannot be seen in the tests above is what the variance estimates look like. For the *t*-tests the variance is calculated in pairs for each comparison, or only two of the times are used to compute the variance. Both the one-way ANOVA and repeated measures mixed model draw information from all three time points to calculate the variance for each individual gene or exon. The moderated *t*-test uses all of the time effects as well as all of the genes or exons to calculate the variance. Tables 3.2 and 3.3 detail the variance parameter estimates as well as the tradeoffs between the models.

Table 3.2 Variance Estimators

Test	Variance Parameter Estimate	Brief Explanation
Independent ***T*-Test**	$$\frac{(n-1)s_1^2 + (n-1)s_2^2}{n+n-2}$$	Variance is likely larger since we are not accounting for repeated measures. Variance estimates come from only the two treatments being tested.
Paired ***T*-Test**	$$\frac{\sum_{j=1}^{n}(d_j - \bar{d})^2}{n-1}$$	Variance is likely smaller since data should have within subject correlation. Variance estimates come from only the two treatments being tested.
One-Way **ANOVA**	$$\frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{Y_{ij}} - \bar{Y}_{i\bullet})^2}{N-a}$$	Variance estimate should be larger since this test does not account for within subject correlation. Variance estimates come from all treatments as opposed to two.
Repeated Measures Mixed Model	$$\frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(\overline{Y_{ij}} - \bar{Y}_{i\bullet})^2}{(a-1)(n-1)}(1-\rho^{|i-j|})$$	Variance estimate should be smaller because this test accounts for within subject correlation. Also the variance estimate uses information from all treatments as opposed to two.
Moderated **T-Test**	$$\exp\left\{\bar{z} - \psi(d_g/2) + \psi(d_o/2) - \log(d_o/d_g)\right\}$$	Variance estimate does not account for within subject correlation. However the estimate uses all of the data to generate a variance estimate. This estimate could be larger or smaller than the estimate calculated in the independent two sample *t*-test because the variance is normalized.

Table 3.3 Model Tradeoffs

Test	Variance Parameter Estimate	Variance Cost/Benefit Ratio	True Variance
Independent T-Test	$\dfrac{(n-1)s_1^2 + (n-1)s_2^2}{n+n-2}$	Variance is likely larger since we are not accounting for repeated measures. Variance estimates come from only the two treatments being tested.	$2\sigma^2 \big/ n$
Paired T-Test	$\dfrac{\sum\limits_{j=1}^{n}(d_j - \bar{d})^2}{n-1}$	Variance is smaller when data has positive within subject correlation. Variance estimates come from only the two treatments being tested.	$\dfrac{2\sigma^2}{n}(1-\rho)$
One-Way ANOVA	$\dfrac{\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{n}(\overline{Y_{ij}} - \overline{Y}_{i\bullet})^2}{N-a}$	Variance estimate should be larger since this test does not account for within subject correlation. Variance estimates come from all treatments as opposed to two.	$2\sigma^2 \big/ n$
Repeated Measures Mixed Model	$\dfrac{\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{n}(\overline{Y_{ij}} - \overline{Y}_{i\bullet})^2}{(a-1)(n-1)}(1-\rho^{|i-j|})$	Variance is smaller when data has positive within subject correlation. Also the variance estimate uses information from all treatments as opposed to two.	$\dfrac{2\sigma^2}{n}(1-\rho^{|i-j|})$

Table 3.3 details the variance parameter estimates and true variance as well. It can be seen that when there is indeed a positive correlation in the data, the repeated measures models are beneficial. However when there is a negative correlation in the data, a repeated measures analysis would not benefit the model. The variance for independent two-sample *t*-test and paired *t*-test are plotted against each other below. The x-axis is the pooled standard deviation (independent) and the y-axis is the paired standard deviation (dependent).

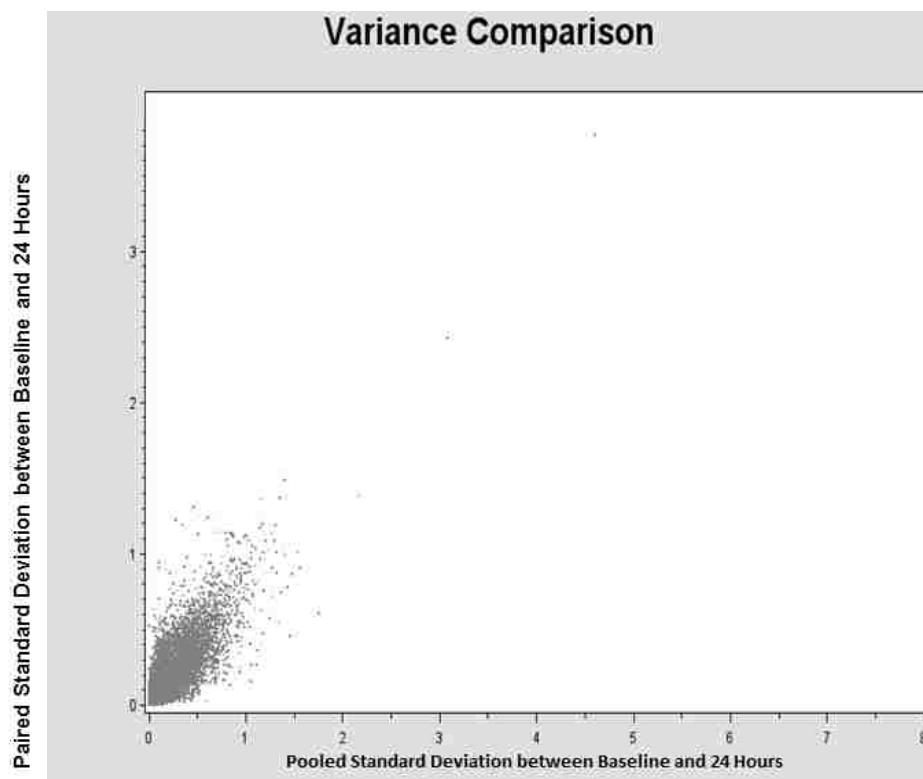Figure 3.1 Variance Comparison from Baseline to 24 Hours

Figure 3.1 compares the variance between baseline and 24 hours for the dependent and independent measures *t*-tests. The variance is smaller in 62.7% of the data for the dependent measures tests.
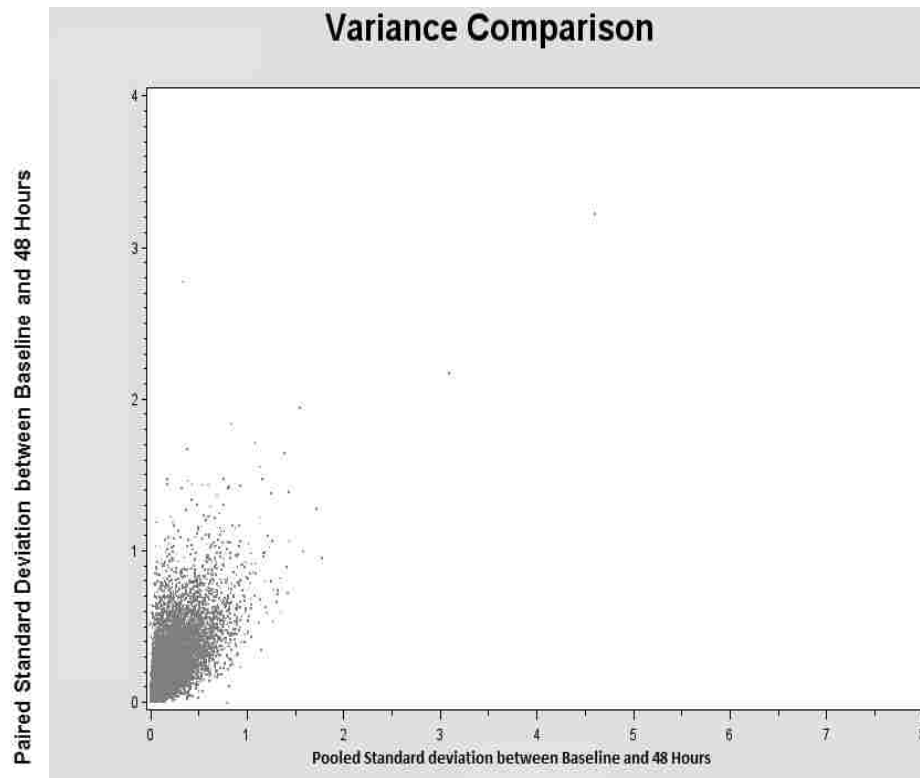
Figure 3.2 Variance Comparison from Baseline to 48 Hours

Figure 3.2 compares the variance between baseline and 48 hours for the dependent and independent measures *t*-tests. The variance is actually smaller for the independent measures tests in 51.8% of the data. For the test comparing these time points, the model is not improved at all using a dependent measures test.
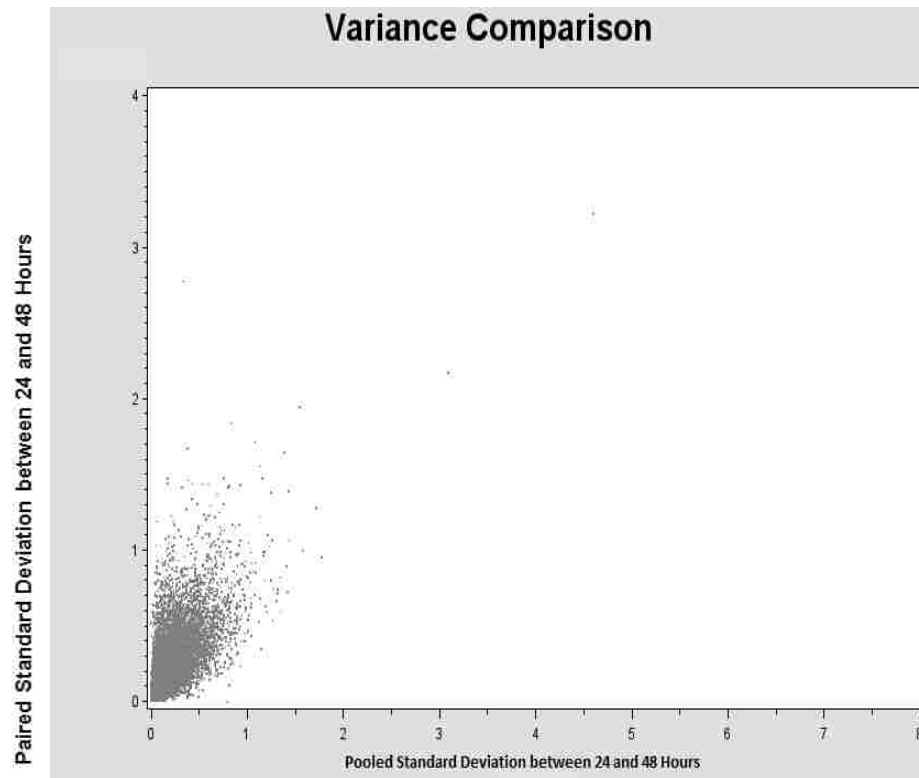
Figure 3.3 Variance Comparison from 24 to 48 Hours

Figure 3.3 compares the variance between 24 and 48 hours for the dependent and independent measures *t*-tests. The variance is smaller for the dependent measures test for 61.6% of the data.

These graphs shown above in Figures 3.1-3.3 are shown simply to illustrate how the variance changes when independent versus repeated measures are considered. In tests that are closer in time the data has slightly less variation when a repeated measure is used. For the test that compares times that are farther apart there is more variation when a repeated measure is used.

4. MULTIPLE TESTING CORRECTIONS

4.1 GENERAL DEFINITION

In statistics there are two types of errors that can be committed, type I and type II error, both detailed in Figure 4.1.

Decision

		Accept	Reject
Null Hypothesis	True	Correct Decision Probability is 1-α, called confidence level	TYPE I ERROR probability of making error is α (always known) minimize by decreasing α (the significance level)
	False	TYPE II ERROR probability of making error is β (rarely known) minimize β by increasing difference of alternative hypothesis, increasing sample size, increasing α, or by choosing a different test	Correct Decision probability is 1-β, called power of test power is determined by signficance level, alternative hypothesis, sample size, and nature of test

Figure 4.1 Types of Statistical Errors

Type I error can be defined as rejecting the null hypothesis when it is indeed true. In statistics, we assume that there will be some error in our analysis, but we can control this error by setting a cutoff for the amount of error we are comfortable allowing. We call this an alpha or significance level and it is thought of as the probability of making a type I error. Type II error is the error that is committed when we fail to reject a false null

hypothesis. This can be better controlled for by decreasing the noise or variability in the data and increasing the sample size. The alpha level (from the type I error) can be kept constant when conducting only one test. A problem occurs when more than one test is being conducted on the data and the probability of making at least one type I error (across all tests) is considered. This is called a multiple testing problem and it needs to be controlled. The multiple corrections procedure looks at the issue of multiple testing by adjusting the p-value from a family of hypothesis tests. An adjusted p-value is obtained and is defined as the smallest significance level where the null hypothesis would be rejected when the entire set of tests is considered [13]. After controlling for a family wise error rate an alpha value is obtained and will be rejected if the adjusted p-value is less than that alpha. The family wise error rate can be obtained by $\bar{\alpha} = 1 - \left(1 - \alpha_{\{per\ comparison\}}\right)^n$.

A popular test that uses FWER (family wise error rate) is a Bonferroni Correction [13]. This examines the probability of committing a type I error for any one of the m tests being conducted on the data. This ends up resulting in the probability of committing a type I error being less than or equal to the number of tests being conducted, m times the alpha chosen (generally 0.05). We adjust the individual alpha, let this be α'. α' is equal to the original, raw alpha over m or α'/m. This makes the resulting FWER extremely conservative. Another popular type of corrections test that can be used is called False Discovery Rate. The False Discovery Rate controls the expected proportion of false positive among the significant tests [13]. For this type of correction one would compile all of our paired hypothesis tests (H_1, H_2,…, H_m) with their corresponding raw p-values (P_1, P_2,…, P_m) into one file, and arrange them in order from smallest to largest ($P_{(1)} \leq P_{(2)}$

$\le \ldots \le P_{(m)})$s. We let *k* equal the largest *i* where $P_{(i)} < \dfrac{i}{m}\alpha$, and reject all $H_{(i)}$ i = 1, 2, …, k.

The FDR test assumes that all of the tests being conducted are independent of one another.

4.2 MICROARRAY DATA

For the microarray data, thousands of genes and exons are being tested. For all of the analysis above (t-tests, ANOVA, mixed model, and the moderated t-test) only raw p-values were computed. Since so many analysis are being conducted, the type I error cannot be kept at a constant level of alpha equal to 0.05. If we consider the gene level analysis where over 17,000 genes made it into the final analysis, this makes the probability of committing at least one type I error approximately equal to (1 − (1 − 0.05)17,000) ≈1. This makes the probability of rejecting the null hypothesis when it is actually true extremely likely for at least one of our tests. If we consider the Bonferroni Correction, the p-value will need to be less than 0.000003. The raw p-value for the exon level analysis would be even smaller. For most microarray data there are generally thousands of tests being conducted and this experiment is not an exception. Therefore using the Bonferroni Correction can be considered a little bit too conservative. We opt for using the False Discovery Rate option for the multiple testing corrections problem. All of the hypothesis tests are aligned with their corresponding raw p-values. The p-values are sorted from smallest to largest. The smallest raw p-value will be the first p-value given and the largest raw p-value is the mth p-value (m tests are conducted). Somewhere in

these thousands of tests, there is an i such that $P_{(i)} < \frac{i}{m}\alpha$. When this i is found, all of the

p-values that are smaller than this value are declared significant. For the microarray data,

we chose the corrected p-value to reject the null hypothesis to be 0.1 since this is

common in microarray studies.

The false discovery rate multiple corrections test assumes that all of the tests for

the genes or exons are independent of one another. This can be difficult because some of

the genes/exons can be dependent while others are independent, but is commonly

assumed anyway.

5. GENE LEVEL ANALYSIS RESULTS

Figures 5.1-5.3 show volcano plots that are a common graphical summary for microarray data. Data between the different time points are plotted using results from the paired *t*-test without multiple testing corrections. The x-axis is the mean difference of the time points while the y-axis is the standard deviation between the two times. The red markers indicate that the gene was significantly differentially expressed from the two time points while the grey shows non-significance. These graphs represent a way to visualize how the size of the mean difference and the variation in the data affect significance. Similar plots could be created for all of the analyses but these are shown for demonstrative purposes.
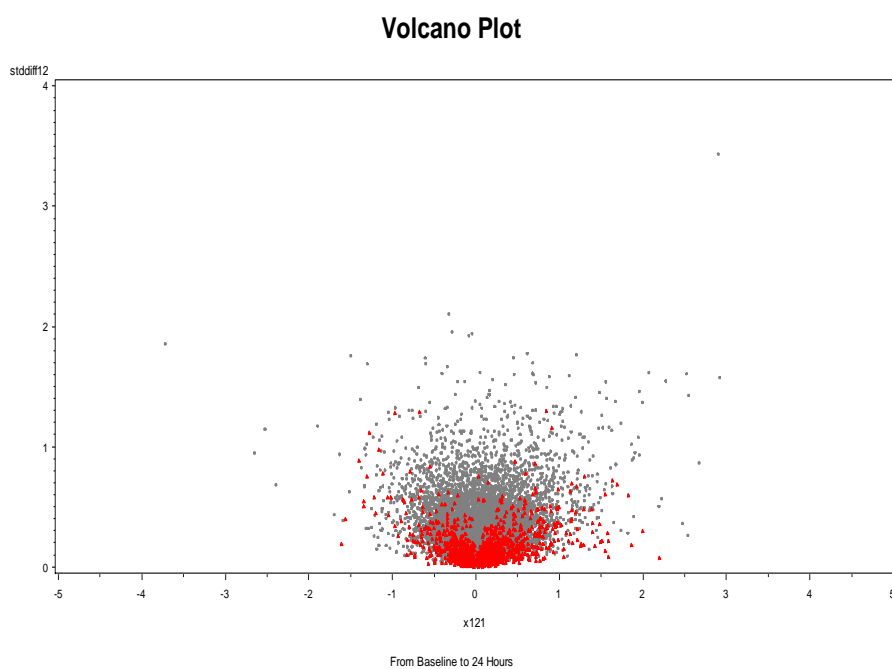

Figure 5.1 Volcano Plot from Baseline to 24 Hours
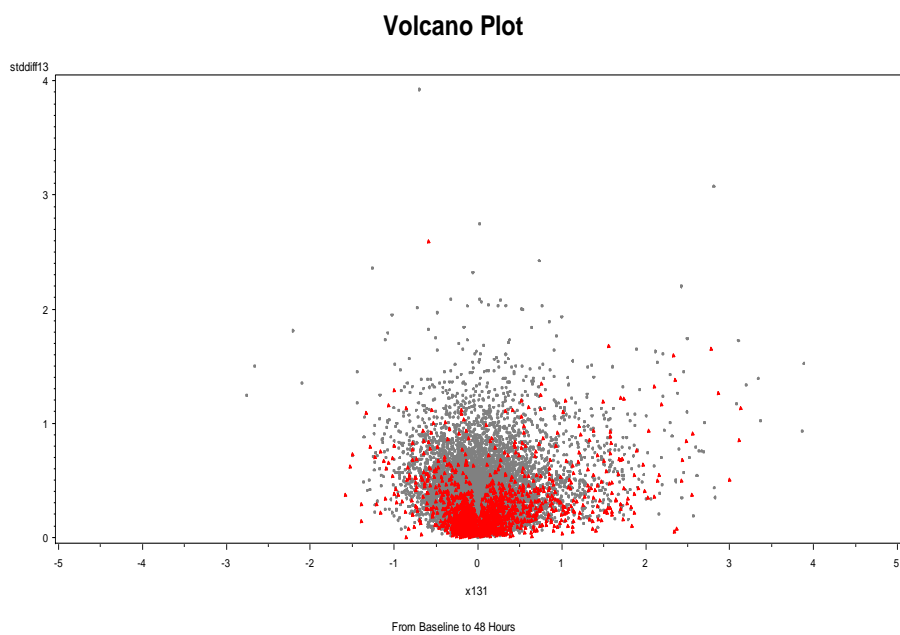
Volcano Plot

Figure 5.2 Volcano Plot from Baseline to 48 Hours
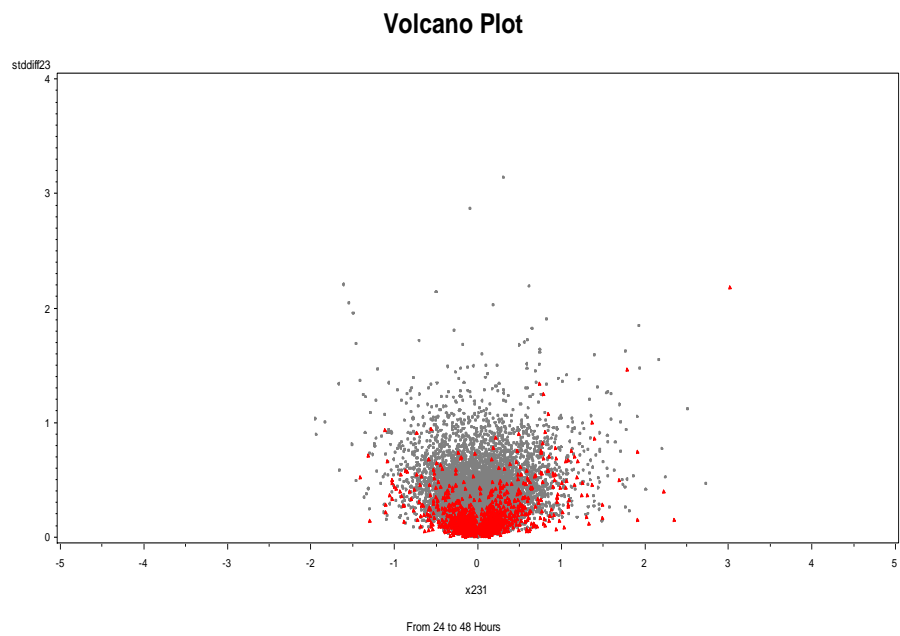
Volcano Plot

Figure 5.3 Volcano Plot from 24 to 48 Hours

Figures 5.4-5.6 are another way to visualize the results by showing the log fold change for the genes compared at the two time points using the paired t-test. The x-axis is an ID variable that lists all of the genes. The y axis is the logfold change between the two time points given. Again, the red points indicate that the gene was significantly differentially expressed (for the raw p-value). We could have constructed these graphs for other analyses too. These give an idea about logfold change and variance related to significance.

Figure 5.4 Log Fold Change from Baseline to 24 Hours
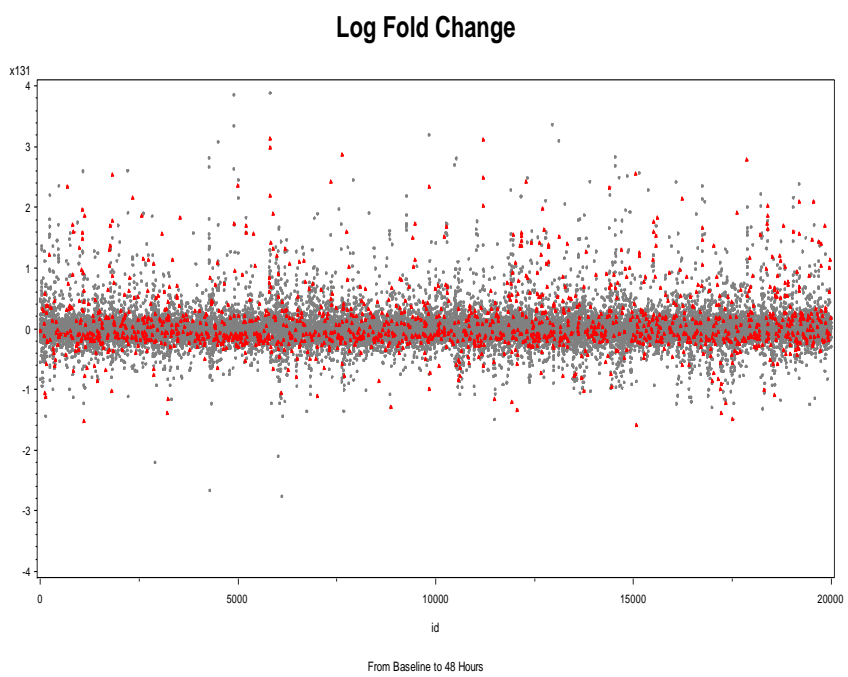
Figure 5.5 Log Fold Change from Baseline to 48 Hours
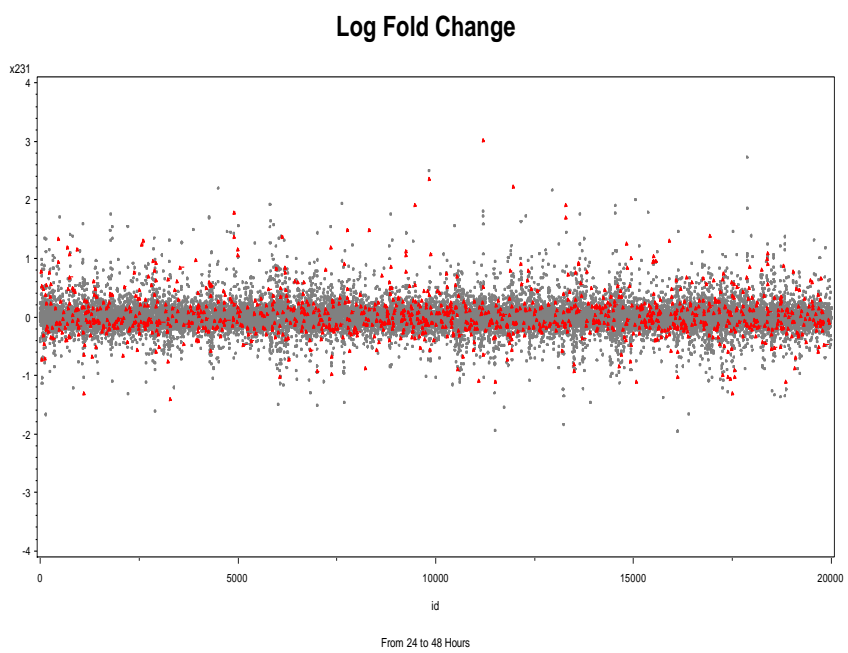
Figure 5.6 Log Fold Change from 24 to 48 Hours

5.1 INDEPENDENT T-TEST

Table 5.1 gives the number of genes significantly, differentially expressed along with the whether or not the gene was up or down regulated for the 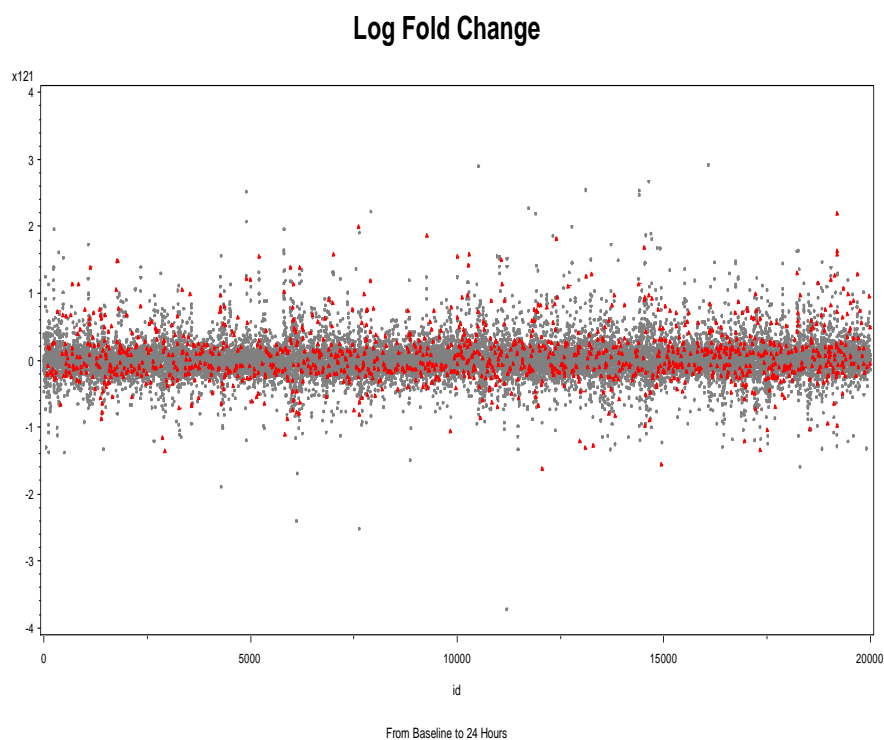paired *t*-test without multiple testing corrections. If a gene is up regulated it means that the expression measure started lower and was measured significantly higher at the later time point.

Table 5.1 Significant Genes from the Independent T-Test

Independent T-Test Gene Level Data without Multiple Testing Corrections			
Test	Number Significant	Number Down Regulated	Number Up Regulated
Baseline to 24 Hours	1127	597	530
Baseline to 48 Hours	1475	720	755
24 to 48 Hours	912	505	407

Figure 5.7 is a venn diagram showing the number of significant genes at the different testing times for the independent *t*-test. Here all genes are considered from the analysis.

Baseline to 24 Hours

735

14,962

230   100

62

1035   602

148

Baseline to 48 Hours   24 to 48 Hours
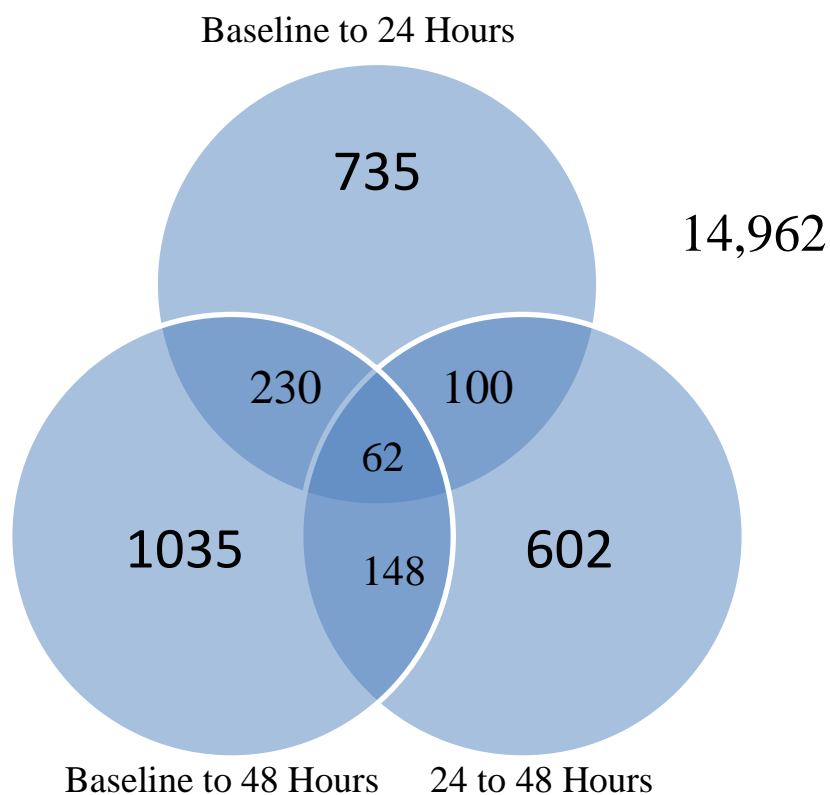
Figure 5.7 Venn Diagram for Independent T-Test

After running the independent two sample t-tests, 3,514 significant differences were

obtained among the three different comparisons. Of these 1,822 were down regulated and 1,692

were up regulated. A total of 62 genes were significant in all three comparisons and 478 of them

were significant in two. However since so many (17,874) comparisons are performed on the same

data set, there is a greater chance of finding false significant differences. To address this problem,

the raw p values were entered into SAS and a multiple corrections procedure was performed.

Two of the results were significant after controlling for the multiple tests that were run on

the data at the alpha = 0.05 level. The genes that were significant were gene ID 3,130,823 and

2,470,165 and they were both significantly down regulated from 24 to 48 hours of sleep

deprivation. There was also one additional test that was significant at the alpha =0.1 level. Gene

ID 2,551,924 was up regulated from the test measuring from baseline to 48 hours.

5.2 PAIRED T-TEST

Table 5.2 gives the number of genes significantly, differentially expressed along with the whether or not the gene was up or down regulated for the paired *t*-test without multiple testing corrections. If a gene is up regulated it means that the expression measure started lower and was measured significantly higher at the later time point.

Table 5.2 Significant Genes from the Paired T-Test

Paired T-Test Gene Level Data without Multiple Testing Corrections			
Test	Number Significant	Number Down Regulated	Number Up Regulated
Baseline to 24 Hours	1003	474	529
Baseline to 48 Hours	1244	662	582
24 to 48 Hours	987	489	498

Figure 5.8 shows the number of significant genes significant at the different testing times for the paired *t*-test. Here all genes are considered from the analysis.

Baseline to 24 Hours
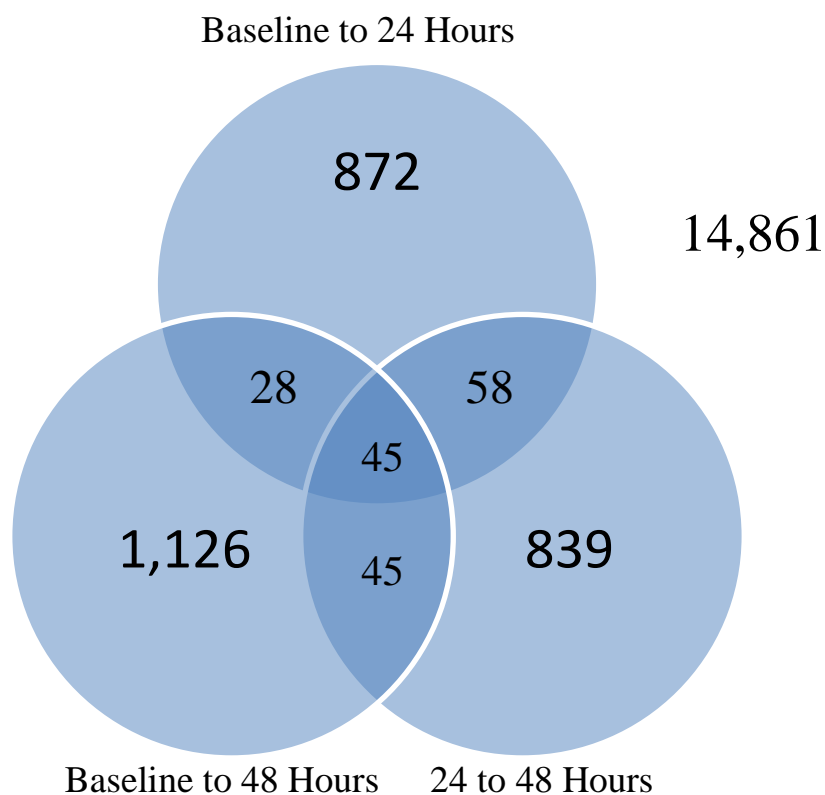
Figure 5.8 Venn Diagram for Paired T-Test

After running the three paired t-tests, 3,234 significant differences among the three different tests were found. Of these 1,625 were down regulated and 1,609 were up regulated. A total of 45 genes were significant in all three comparisons and 131 of them were significant in two. However since so many (17,874) tests were run on the same data set, there is a greater chance of finding false significant differences. To address this problem, the raw p values were entered into SAS and a multiple corrections test was conducted.

None of the results were significant after controlling for the multiple tests that were run on the data at the alpha = 0.05 level or the alpha = 0.1 level.

5.3 ANALYSIS OF VARIANCE

Table 5.3 gives the number of genes significant for the one-way ANOVA

analysis. It is important to note that the tests for the pairwise comparisons only include

the data from the tests that were significant in the overall f test.

Table 5.3 Significant Genes from ANOVA Model

ANOVA Gene Level Data without Multiple Testing Corrections			
	Number Significant		
Overall F	1,257		
Contrast Baseline (different from 24 or 48)	930		
Contrast 24 Hours (different from 24 or 48)	638		
		Up Regulated	**Down Regulated**
Contrast 48 Hours (different from 24 or 48)	875		
BS is Different than 24	46	20	26
BS is Different than 48	66	35	31
24 is Different than 48	25	14	11

Figure 5.9 shows a Venn Diagram comparing the tests for the pairwise

comparisons for the one-way ANOVA model. Only the genes that were significant for

the overall f test were considered for the comparisons.

Baseline to 24 Hours
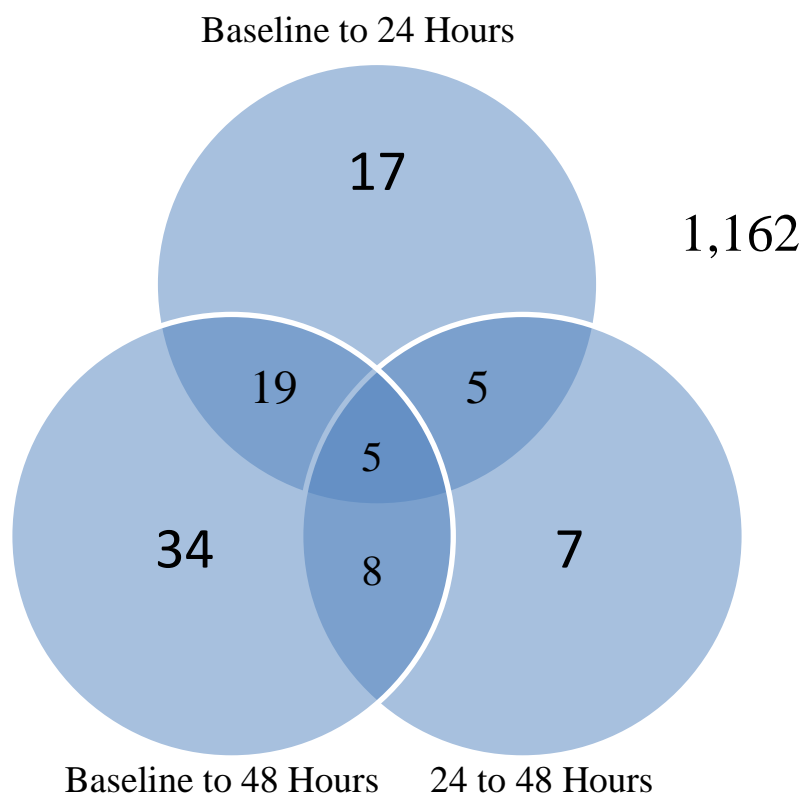
Figure 5.9 Venn Diagram for ANOVA Model

From the Table 5.3 it is seen that there were 1,257 genes were significant for the overall f test. Therefore, only the results for those 1,257 genes are listed for the pairwise comparisons above in Table 5.3. The pairwise mean comparisons for the ANOVA model show that of the 1,257 genes from the overall f test, 17 genes were differentially expressed from baseline to 24 hours, 34 from baseline to 48 and 7 genes are significantly different from 24 to 48 hours. Figure 5.9 shows the break down of how these genes were differentially expressed. Five of the genes have significant paired comparisons for all three of the comparisons while 1,162 were not significant for any of the paired comparisons.

After the multiple testing corrections, there weren't any siginificant genes at the

alpha equals 0.05 level for the pairwise comparisons. Only one gene was significant at

the alpha = 0.1 level. Gene ID 2,844,453 was significantly down regulated from baseline

to 48 hours.

5.4 REPEATED MEASURES MIXED MODEL

Table 5.4 gives the number of genes significant for the repeated measures

analysis. It is important to note that the tests for the pairwise comparisons only include

the data from the tests that were significant in the overall f test.

Table 5.4 Significant Genes from Mixed Model

Mixed Model Gene Level Data without Testing Corrections		Up Regulated	Down Regulated
	Number Significant		
Compound Symmetry	1,688		
Autoregressive Model	2,092		
Contrast Baseline (AR1)	1,366		
Contrast 24 Hours (AR1)	1,150		
Contrast 48 Hours (AR1)	1,225	Up Regulated	Down Regulated
BS is Different than 24 (AR1)	892	470	422
BS is Different than 48 (AR1)	1,021	593	428
24 is Different than 48 (AR1)	728	362	366

Table 5.4. gives some information about the significant genes from the mixed

model analysis. Overall, when the compound symmetry covariance matrix was

implemented in the analysis, 1,688 genes had significant findings. The autoregressive

model found more significant genes, however 1,464 were significant for both analyses as shown in Figure 5.10.

224   1464   628

Compound Symmetry   Autoregressive
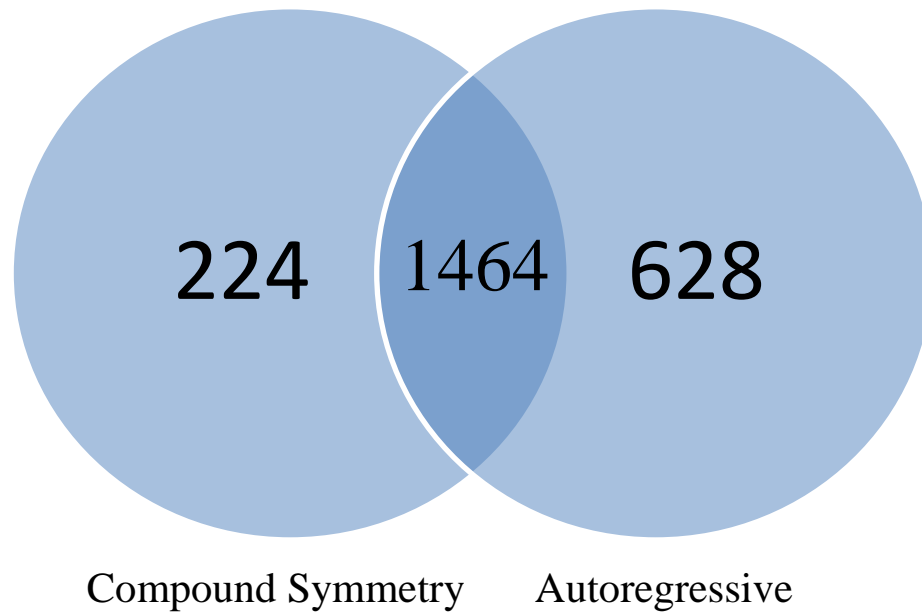
Figure 5.10 Covariance Matrix Comparison

Baseline to 24 Hours

160

625

375   159

198

204   127

244

Baseline to 48 Hours   24 to 48 Hours
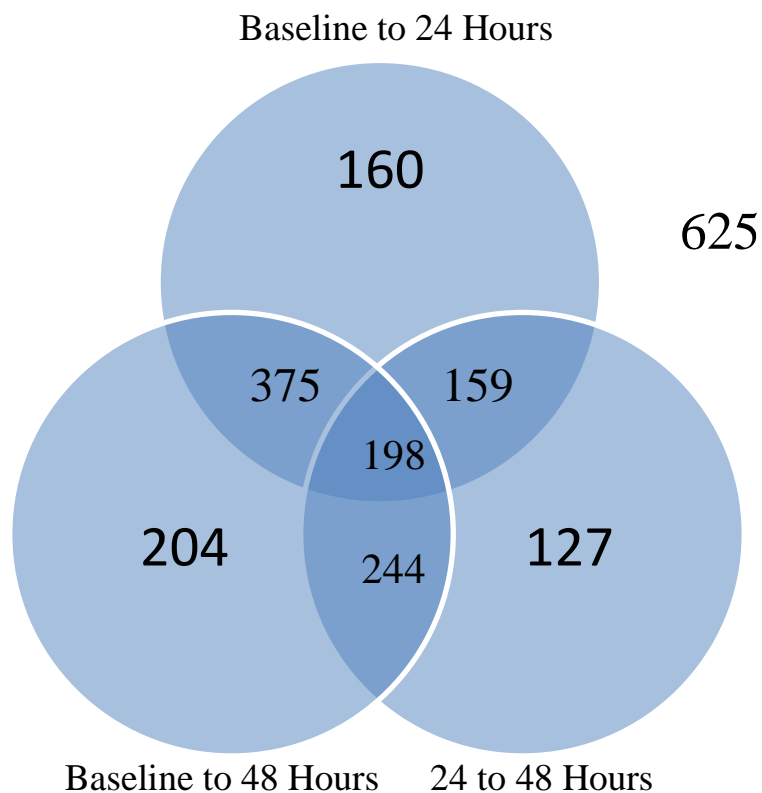
Figure 5.11 Venn Diagram for Repeated Measures Mixed Model

Since the autoregressive model found 2,092 significant genes for the overall f test and the AIC found it to be a better fit for a larger percentage of the genes, only these genes were examined for the pairwise comparisons. Of those 2,092 genes, 198 were significant for all comparisons and 625 were not significant for any of the tests. After the multiple testing corrections, none of genes were significant at the alpha equals 0.05 level for the pairwise comparisons. When alpha was raised to 0.1, the test comparing baseline to 24 hours found two genes significantly differentially expressed. The test comparing baseline to 48 hours found eight significant genes. The test comparing 24 to 48 hours found 3 significant genes.

5.5 EMPIRICAL BAYES

Table 5.5 shows the number of genes that were significantly, differentially expressed at the different testing times along with whether the gene was up or down regulated.

Table 5.5 Significant Genes for Moderated T-Test

Moderated T-Statistic Gene Level Data without Multiple Testing Corrections			
	Number Significant	Number Down Regulated	Number Up Regulated
Baseline to 24 Hours	1,579	940	639
Baseline to 48 Hours	1,986	1,202	784
24 to 48 Hours	1,205	553	652

Figure 5.12 is a Venn diagram that shows the number of genes significantly differentially expressed for the moderated *t*-test.
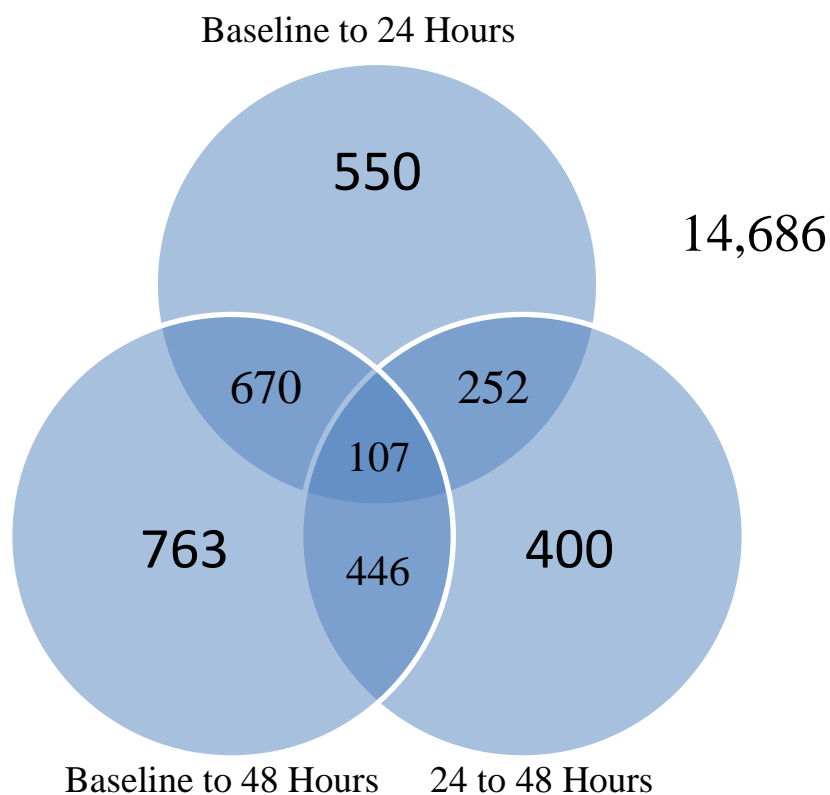
Figure 5.12 Venn Diagram for Moderated T-Test

As seen in Figure 5.12, 107 of the genes were significant for all three of the moderated *t*-tests. 14,686 of the genes were not significant for any of the tests. 1,475 of the tests were significant for at least two of the comparisons. To address the issue of a multiple testing problem, a multiple corrections procedure is performed on the raw p-values.

After running a multiple testing correction on the raw p-values for the moderated *t*-statistic, there were 131 genes significant after allowing the alpha level to reach 0.1 from baseline to 48 hours. Of these 113 were down regulated and 18 were up regulated. None of the other comparisons showed significant results for the multiple testing corrections.

6. EXON LEVEL ANALYSIS RESULTS

6.1 INDEPENDENT T-TEST

Table 6.1 gives the number of genes significantly, differentially expressed along with the whether or not the gene was up or down regulated for the paired t-test without multiple testing corrections. If a gene is up regulated it means that the expression measure started lower and was measured significantly higher at the later time point.

Table 6.1 Significant Exons for Independent T-Test

Independent T-Test Exon Level Data without Multiple Testing Corrections			
Test	Number Significant	Number Down Regulated	Number Up Regulated
Baseline to 24 Hours	19,460	8,975	10,485
Baseline to 48 Hours	22,129	8,937	13,192
24 to 48 Hours	15,594	7,171	8,423

Figure 6.1 shows the number of significant genes significant at the different testing times for the independent t-test.

Baseline to 24 Hours
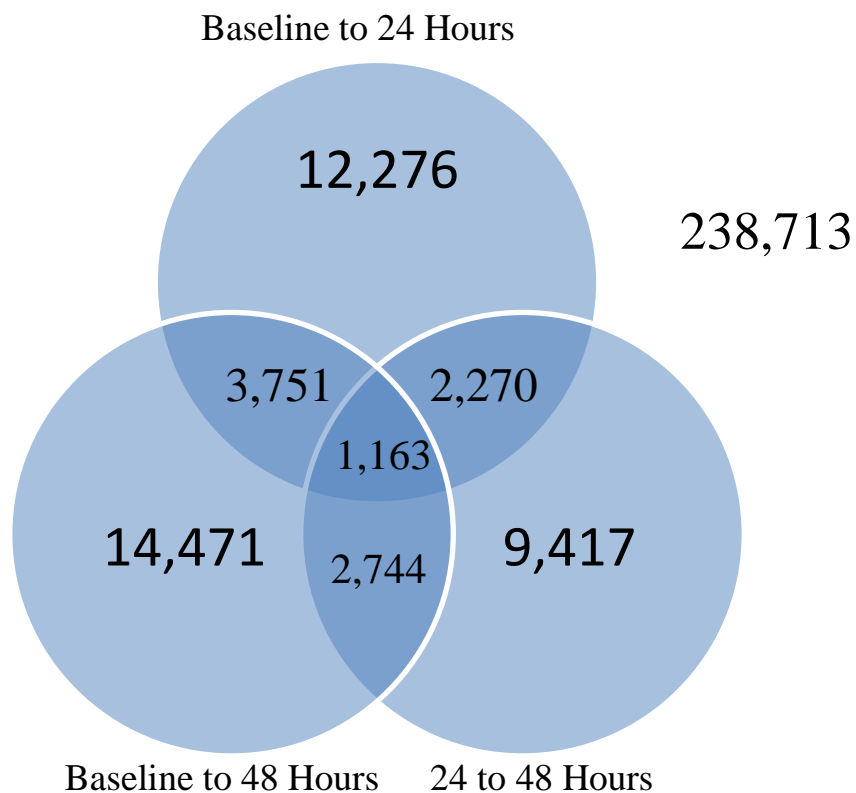
Figure 6.1 Venn Diagram for Independent T-Test

After running the three independent t-tests, 57,183 significant differences were found among the three different tests. Of these 25,083 were down regulated and 32,100 were up regulated. A total of 1,163 exons were significant in all three comparisons and 8,765 of them were significant in two. However since so many tests are performed on the same data set, there is a greater chance of finding false significant differences. To address this problem, the raw p values were entered into SAS and a multiple corrections test was conducted.

None of the results from the multiple testing corrections were significant at either the alpha equals 0.05 or alpha equals 0.1 significance levels.

6.2 PAIRED T-TEST

Table 6.2 shows the number of exons both up and down regulated for the paired *t-*

test.

Table 6.2 Significant Exons for Paired T-Test

Paired T-Test Exon Level Data without Multiple Testing Corrections			
Test	Number Significant	Number Down Regulated	Number Up Regulated
Baseline to 24 Hours	16,963	7,851	9,112
Baseline to 48 Hours	18,329	8,383	9,946
24 to 48 Hours	15,960	7,433	8,527

Figure 6.2 is a venn diagram that shows the number of significant genes

significant at the different testing times for the paired *t*-test.

Baseline to 24 Hours

9,987

244,078

3,167   2,781

1,028

11,613   2,521   9,630

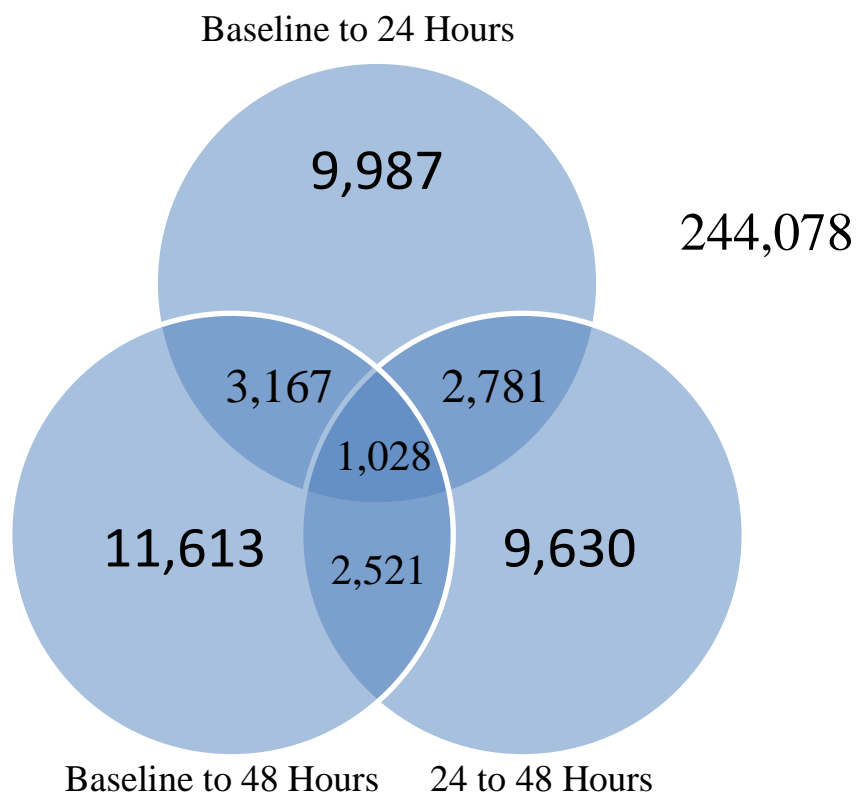Baseline to 48 Hours   24 to 48 Hours

Figure 6.2 Venn Diagram for Paired T-Test

After running the three paired t-tests, 51,252 significant differences were found among the three different tests. Of these 23,667 were down regulated and 27,585 were up regulated. A total of 1,028 exons were significant in all three comparisons and 8,469 of them were significant in two. However since so many tests are being ran on the same data set, there is a greater chance of finding false significant differences. To address this problem, the raw p values were entered into SAS and a multiple corrections test was conducted.

Several of the results were significant after controlling for the multiple tests that were run on the data at the alpha = 0.05 level. The exons that were significant and up regulated for the comparison from baseline to 24 hours were probe set ID 3,031,827 and 2,640,916. None of the exons were significantly differentially expressed from baseline to 48 hours. Five exons were significantly down regulated from 24 to 48 hours, they are probe set ID's 3,525,655 2,665,526 3,718,502 2,443,537 and 3,130,823. Probe set ID 3,642,707 was significantly up regulated for this comparison as well. When alpha is allowed to equal 0.1, two more exons are differentially expressed for that comparison as well. They are probe set ID's 3,922,793 and 2,995,476. No other tests were significant at this level.

6.3 ANALYSIS OF VARIANCE

Table 6.3 gives the number of exons significantly expressed for each test

Baseline to 24 Hours

Figure 6.3 Venn Diagram for ANOVA Model

As seen in Figure 6.3, 83 of the exons were significant for all of the pairwise comparisons. 722 of the comparisons were significant in two of the tests. 20,403 were not significant in any of the comparisons. It is likely that these had other significant contrasts. Because the pairwise comparisons were tested so many (284,805) times, a multiple corrections test was conducted.

None of the pairwise comparisons were significant when for the multiple testing corrections when alpha was equal to 0.05 or 0.1.

6.4 MIXED MODEL

Table 6.4 gives the number of exons significantly expressed for each test conducted for the repeated measures mixed model. The only exons whose pairwise

comparisons were conducted were the 25,739 significant exons from the overall f test.

Table 6.4 Significant Exons for Mixed Model

Mixed Model Exon Level Data without Multiple Testing Corrections			
	Number Significant		
Autoregressive Model	25,739		
Contrast Baseline (AR1)	18,920		
Contrast 24 Hours (AR1)	19,583		
Contrast 48 Hours (AR1)	17,061	**Up Regulated**	**Down Regulated**
BS is Different than 24 (AR1)	1,332	679	653
BS is Different than 48 (AR1)	1,340	701	639
24 is Different than 48 (AR1)	1,007	518	489

Figure 6.4 shows a Venn diagram of the pairwise comparisons for the mixed

model analysis.

Baseline to 24 Hours
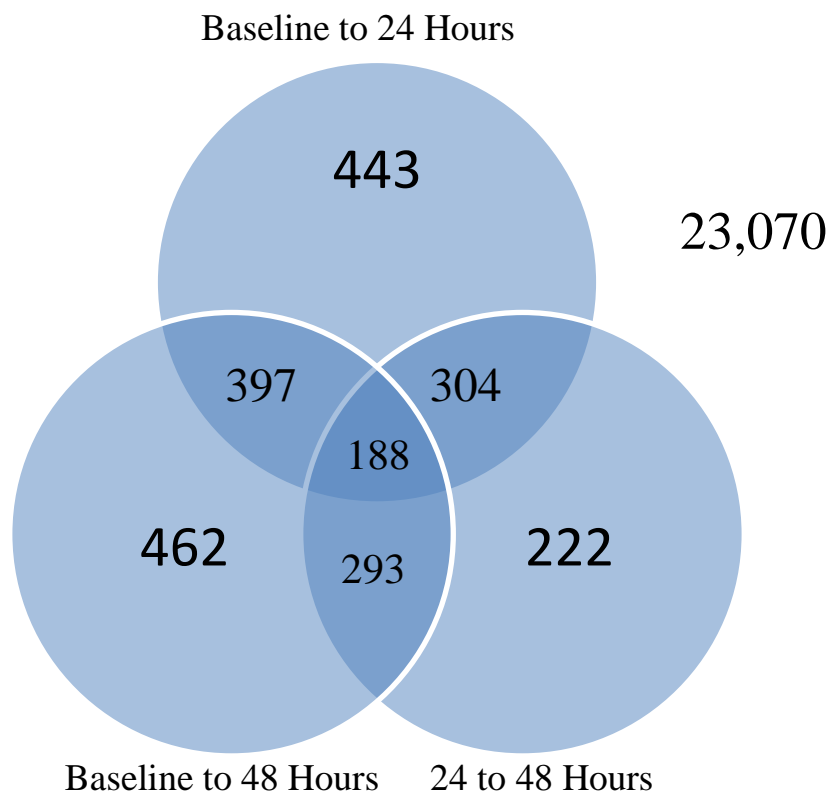
Figure 6.4 Venn Diagram for Repeated Measures Mixed Model

Figure 6.4 shows that 188 of the exons were significant for all three of the pairwise comparisons. 23,070 were not significant for any of the pairwise comparisons. While there were 2,309 significant pairwise comparisons, a multiple corrections test should be conducted since there are 284,805 tests being conducted on the data.

After running the multiple testing corrections on the data from the repeated measures mixed model analysis, none of the exons were significant at either the 0.05 or 0.1 alpha levels.

6.5 EMPIRICAL BAYES

Table 6.5 gives the number of exons significant for the moderated *t*-test for each

comparison. Also in the table is whether or not the exon was down or up regulated.

Table 6.5 Significant Exons for Moderated T-Test

Moderated T-Statistic Exon Level Data without Multiple Testing Corrections			
	Number Significant	**Number Down Regulated**	**Number Up Regulated**
Baseline to 24 Hours	24,385	14,257	10,128
Baseline to 48 Hours	27,591	17,793	9,798
24 to 48 Hours	20,210	8,563	11,647

Figure 6.5 shows a Venn diagram comparing the moderated *t*-tests different

comparisons.

Baseline to 24 Hours
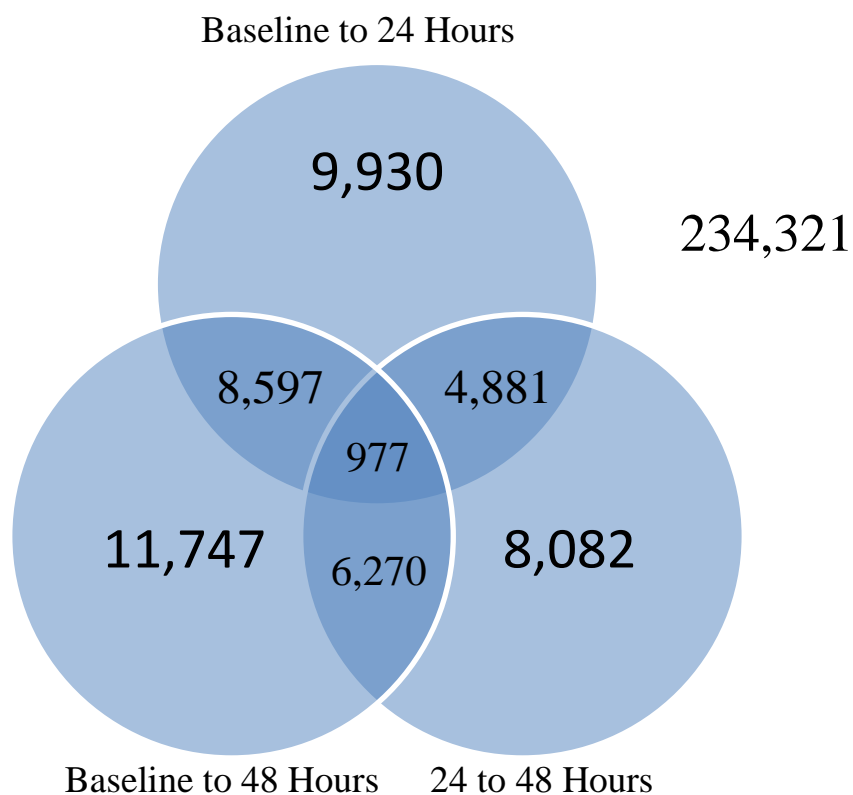
Figure 6.5 Venn Diagram for Moderated T-Test

As seen in Figure 6.5, 977 of the exons were differentially expressed in all of the tests. Tens of thousands of the tests were significant for more than one of the comparisons. Because so many tests are being conducted on the same data set, a multiple testing corrections test is ran to control for type I error.

When the alpha level is set at 0.05 for the multiple testing corrections one exon was significant. Probe set ID 2,614,423 was significantly up regulated from baseline to 24 hours. When the alpha level increases to 0.1, hundreds of exons are significant. 739 are significant when comparing baseline to 48 hours. 608 of them were down regulated and 131 were up regulated. Three additional exons were up regulated from baseline to 24 hours when alpha equals 0.1. Figure 6.6 compares the exons that are significant after multiple testing corrections.
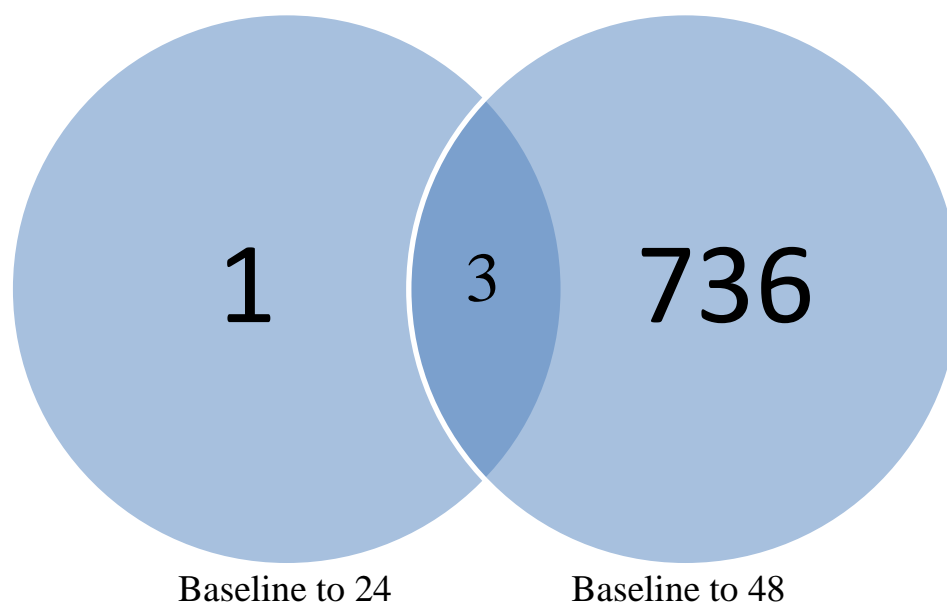
Figure 6.6 FDR Significant Exons for Moderated T-Test

7. CLUSTERING RESULTS

After running the different analyses, it was of interest to observe the relationship pattern of the significantly expressed genes. Table 7.1 lists the genes that were significantly expressed according to whether or not they were up or down regulated from the paired *t*-test analysis. For example, gene ID 2,374,956 was significantly down regulated from baseline to 24 hours and then again down regulated from 24 to 48 hours. 13 genes followed the down, down regulated pattern, 28 followed the up, up pattern, 25 were down regulated then up regulated and 31 went up and then down.

Table 7.1 Significant Clustering Patterns for Genes

Down, Down	Up, Up	Down, Up	Up, Down
2374956	2356115	2497018	2344393
2403099	2362394	2536476	2438125
2404377	2363689	2643217	2499158
2527196	2391840	2646327	2565119
2844453	2434575	2674335	2840768
3065546	2526980	2841472	2949043
3135156	2527606	2844335	3159330
3212350	2599303	2876046	3165957
3284882	2738928	2945518	3272106
3403539	2982319	3091848	3301609
3737677	3114649	3161167	3316987
3753220	3167220	3186207	3361031
3980964	3372174	3256279	3364747
	3377044	3364739	3413456
	3377892	3455426	3453384
	3392924	3474940	3464417
	3393670	3477967	3525655
	3569754	3620276	3556214
	3635125	3757970	3592109
	3676674	3764103	3638607
	3728889	3841777	3672059
	3739859	3861243	3683050
	3819745	3958329	3696666
	3837895	3959654	3699133
	3844978	4016572	3727499
	3854066		3749546
	3860137		3800834
	3869237		3822049
			3918574
			3959744
			4052881

Table 7.2 shows a truncated table of the clustering patterns for the significant exons from the paired *t*-test. 112 followed a down, down pattern, 204 followed the up, up pattern, 551 followed a down up pattern and 494 went up and then down.

Table 7.2 Truncated Significant Clustering Patterns for Exons

Down Down	Up Up	Down Up	Up Down
2331704	2317447	2316462	2319149
2348680	2331556	2319855	2319451
2348905	2334705	2320641	2321257
2348965	2355671	2321044	2323094
2377042	2356117	2322293	2326361
2384716	2356127	2322883	2327552
2385261	2356136	2324400	2330120
2389037	2360040	2326860	2331474
2406985	2361423	2327020	2331683
2408012	2371023	2327348	2333266
2408546	2372879	2327632	2335173
2409978	2375675	2327831	2347177
2434978	2377271	2327836	2348096
2436304	2404161	2331426	2350505
2454210	2408269	2333654	2352313
2469571	2435866	2338529	2359038
2476245	2440478	2343100	2360114
2494777	2447461	2344435	2360713
2527977	2464529	2345007	2361425
2581019	2486986	2345102	2361847
2585364	2492026	2347098	2363542
2585370	2501229	2353507	2365876
2598264	2522646	2358972	2379904
2618530	2524019	2361845	2381044

8. CONCLUSIONS

Since several genes and exons were significant after multiple testing corrections, an obvious question is, whether any of the genes and exons are significant in more than one test. A few of the genes were significant for two tests. Table 8.1 shows genes and exons that are significant after multiple testing corrections, the comparison that it was significant for, which test found it significant and whether it was up or down regulated.

Table 8.1 FDR Significant Genes and Exons

Exons								
BS - 24	**Test**	**Regulation**	**BS - 48**	**Test**	**Regulation**	**24 - 48**	**Test**	**Regulation**
3,031,827	PAIRED	UP				3,525,655	PAIRED	DOWN
2,640,916	PAIRED	UP				2,665,526	PAIRED	DOWN
						3,718,502	PAIRED	DOWN
						2,443,537	PAIRED	DOWN
						3,130,823	PAIRED	DOWN
						3,642,707	PAIRED	UP

Genes								
BS - 24	**Test**	**Regulation**	**BS - 48**	**Test**	**Regulation**	**24 - 48**	**Test**	**Regulation**
3,373,420	RMMM	UP	2,551,924	INDEPENDENT T	UP	3,130,823	INDEPENDENT T	DOWN
4,047,607	RMMM	UP	2,844,453	ANOVA	DOWN	2,470,165	INDEPENDENT T	DOWN
			3,167,220	RMMM	DOWN	3,166,718	RMMM	UP
			3,166,718	RMMM	DOWN	3,942,007	RMMM	UP
			3,556,556	RMMM	DOWN	3,840,194	RMMM	DOWN
			3,564,790	RMMM	DOWN			
			2,847,229	RMMM	DOWN			
			3,592,023	RMMM	DOWN			
			3,854,066	RMMM	DOWN			

*Denotes Significant in Moderated t-test

The 739 exons that were significant from the moderated *t*-test belonged to 568 different genes. Of those 568 genes, 61 of those were significant from the gene level moderated *t*-test analysis. All of them were down regulated from baseline to 48 hours and each gene is listed below in Table 8.2.

Table 8.2 Mapping Significant Exons to Significant Genes

2356115	3445786	3126087	3680953
2394680	3453837	3146433	3684486
2512930	3475794	3153328	3688311
2599303	3488985	3166718	3734379
2682271	3557947	3167220	3752258
2705706	3563317	3175119	3757078
2730173	3564790	3212294	3766893
2730194	3580179	3238761	3815243
2903401	3592023	3293762	3816380
2948887	3592511	3315675	3846831
2948926	3603408	3316208	3847005
2974635	3608113	3385175	3850261
3057955	3638411	3400625	3861948
3098977	3678462	3410384	3862018
3443891	3907190	3417457	3883971
			4035833

When only examining raw p-values, thousands of genes and exons are significantly differentially expressed. However because thousands of tests are conducted on the same data set, the type I error rate must be controlled. Several of these were significant after a multiple testing procedure as well. These genes and exons should be examined closely to see if they play any role in sleep patterns or cycles. In order to claim there is a causal relationship between these genes and exons and sleep, more microarray

experiments should be conducted. Ideally the data would be collected on more participants, possibly at smaller time intervals.

Finally for future analysis on this microarray data, perhaps it is not advantageous to assume that the data is dependent. The dependence assumption forfeits a degree of freedom and doesn't show a large reduction in variance. It is important to note that this is not the case for all microarray data. With a larger sample size, it would be possible to investigate the correlation between expression measurements on the same individual more in depth. However, for the data available in this study, it might be of interest to conduct an analysis that treats some of the data as dependent and others as independent. Perhaps conducting an analysis like this might result in more significant findings after multiple testing corrections.

APPENDIX

COMMENTED SAS CODE

```
/* importing gene level csv file into SAS obtained from RMA method */

PROC IMPORT OUT= SASUSER.GENESHORTDATA
            DATAFILE= "home\sb6xb\geneleveldata_short.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

/* transforming data into a different format */

data repeated;
set SASUSER.GENESHORTDATA;
exprs = exprsbsa; time = 1; sub = 1; output;
exprs = exprsbsb; time = 1; sub = 2; output;
exprs = exprs24a; time = 2; sub = 1; output;
exprs = exprs24b; time = 2; sub = 2; output;
exprs = exprs48a; time = 3; sub = 1; output;
exprs = exprs48b; time = 3; sub = 2; output;
drop exprs24a exprs24b exprs48a exprs48b exprsbsa exprsbsb;
proc sort data = repeated;
by gene;
run;

/* creating data sets by time points; baseline, 24 hours, and 48 hours
*/

data time1;
/* baseline */
set repeated;
if time = 1;
drop obs;
rename time=time1;
rename exprs=exprs1;
run;

data time2;
/* 24 hours */
set repeated;
if time = 2;
drop obs;
rename time=time2;
rename exprs=exprs2;
run;

data time3;
/* 48 hours */
set repeated;
if time = 3;
drop obs;
```

```
rename time=time3;
rename exprs=exprs3;
run;

/* merging time point data together, necessary to perform t tests */

data ttestdata;
merge time1 time2 time3;
by gene;
run;

/* turning off the output window in SAS */

ODS LISTING CLOSE;

/* creating a csv file path for t test comparing baseline to 24 hours
*/

ODS HTML FILE="E:\Thesis\exprs1exprs2.csv";

/* running t test comparing baseline to 24 hours */

proc ttest data=ttestdata;
paired exprs1*exprs2;
by gene;

/* creating ODS table name for t test data /*

ods output Ttests=ttest_output;
run;

/* printing the t test output to the csv file */

proc print data=ttest_output;
run;

/* creating a csv file path for t test comparing baseline to 48 hours
*/

ODS CSV FILE="E:\Thesis\exprs1exprs3.csv";

/* running t test comparing baseline to 48 hours */

proc ttest data=ttestdata;
paired exprs1*exprs3;
by gene;

/* creating ODS table name for t test data /*

ods output Ttests=ttest_output;
run;

/* printing the t test output to the csv file */

proc print data=ttest_output;
```

```
run;

/* creating a csv file path for t test comparing 24 to 48 hours */

ODS CSV FILE="E:\Thesis\exprs2exprs3.csv";

/* running t test comparing 24 to 48 hours */

proc ttest data=ttestdata;
paired exprs2*exprs3;
by gene;

/* creating ODS table name for t test data /*

ods output Ttests=ttest_output;
run;

/* printing the t test output to the csv file */

proc print data=ttest_output;
run;

/* creating ODS table names for GLM procedures */

ods output overallANOVA=anova Contrasts=contr lsmeans=means diff=diff;

/* running proc glm in SAS */

proc glm data = repeated;
class sub time;
model exprs = time sub;
by gene;

/* creating contrasts comparing the time points */

contrast 'Baseline VS 24 and 48' time 2 -1 -1 / est;
contrast '24 VS Baseline and 48' time -1 2 -1 / est;
contrast '48 VS Baseline and 24' time -1 -1 2 / est;

/* creating comparison matrices for the time points */

lsmeans exprs / pdiff;
means exprs / tukey lsd;
run;

/* creating file path for contrast data */

ods csv file='E:\Thesis\contrasts.csv';

/* printing contrast table to csv file */

proc print data=contr;
run;

/* creating file path for p values and anova table data */
```

```
ods csv file='E:\Thesis\anovafvalues.csv';

/* printing p values and anova table data to csv file*/

proc print data=anova;
run

/* creating file path for difference matrix data */

ods csv file='E:\Thesis\pdiff.csv';

/* printing difference matrices to csv file */

proc print data=diff;
run

/* creating ODS table names for mixed procedures */

ods output covparms=cp asycov=asy infocrit=aic tests1=tests1
contrasts=contr diff=diff;

/* running the mixed procedure in SAS */

proc mixed data=repeated covtest asycov ic;
class sub time;
model exprs = time / s htype=1;
by gene;

/* creating contrasts comparing the time points */

contrast 'BS VS 24 and 48' time 2 -1 -1 / est;
contrast '24 VS BS and 48' time -1 2 -1 / est;
contrast '48 VS BS and 24' time -1 -1 2 / est;

/* creating comparison matrices for the time points */

lsmeans time / pdiff=all;
repeated time / TYPE = AR(1) subject=sub r;
run;

/* specifying file path for ODS table data, difference matrices */

ods csv file="E:\Thesis\pdiffmixed.csv";

/* printing csv file, difference matrices */

proc print data=diff;
run;

/* specifying file path for ODS table data, AIC scores */

ods csv file="E:\Thesis\geneaicscorear1.csv";

/* printing csv file, AIC score */
```

```
proc print data=aic;
run;

/* specifying file path for ODS table data, p values and anova tables
*/

ods csv file="E:\Thesis\genemixedpvaluesar1.csv";

/* printing csv file, pvalues and anova tables */

proc print data=tests1;
run;

/* specifying file path for ODS table data, residual variances */

ods csv file="E:\Thesis\mixedwresidualar1.csv";

data res;
set cp;
if subject = 'sub' then
delete;
run;

/* printing csv file, residual variances */

proc print data=res;
run;

/* specifying file path for ODS table data, subject variances */

ods csv file="E:\Thesis\mixedwsubjectar1.csv";

data sub;
set cp;
if subject = ' ' then
delete;
run;

/* printing csv file, subject variances */

proc print data=sub;
run;

/* importing in raw p value data from t tests */

PROC IMPORT OUT= SASUSER.GENEMULTTEST
            DATAFILE= "E:\Thesis\genemulttest.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

/* cutting data to only include necessary p values from first t test */
```

```
data multtest1;
set sasuser.genemulttest;
rename probt12=Raw_P;
drop prob13;
drop prob23;
run;

/* specifying file path for output table to print to */

ods csv file="E:/Thesis/ttestgenemulttest12.csv";

/* running the multiple testing corrections adjustment in SAS using fdr
*/

proc multtest inpvalues=multtest1 fdr;
run;

/* cutting data to only include necessary p values from second t test
*/

data multtest2;
set sasuser.genemulttest;
rename prob13=Raw_P;
drop probt12;
drop prob23;
run;

/* specifying file path for output table to print to */

ods csv file="E:/Thesis/ttestgenemulttest13.csv";

/* running the multiple testing corrections adjustment in SAS using fdr
*/

proc multtest inpvalues=multtest2 fdr;
run;

/* cutting data to only include necessary p values from third t test */

data multtest3;
set sasuser.genemulttest;
rename prob23=Raw_P;
drop prob13;
drop probt12;
run;

/* specifying file path for output table to print to */

ods csv file="E:/Thesis/ttestgenemulttest23.csv";

/* running the multiple testing corrections adjustment in SAS using fdr
*/

proc multtest inpvalues=multtest3 fdr;
run;
```

```
/*Multiple testing corrections file for ANOVA, MIXED, and MOD T*/

/* importing data into SAS on the NIC Cluster */

PROC IMPORT OUT= SASUSER.EXONSHORTdata
            DATAFILE= "home\sb6xb\exonleveldata_short.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;

/* transforming data into a different format */

data ttest;
set sasuser.exonshortdata;
exprs = bsa_cel; time = 1; sub = 1; output;
exprs = bsb_cel; time = 1; sub = 2; output;
exprs = x24a_cel; time = 2; sub = 1; output;
exprs = x24b_cel; time = 2; sub = 2; output;
exprs = x48a_cel; time = 3; sub = 1; output;
exprs = x48b_cel; time = 3; sub = 2; output;
drop bsa_cel bsb_cel x24a_cel x24b_cel x48a_cel x48b_cel;
proc sort data = ttest;
by probesetid;
run;

/* creating data sets by time points; baseline, 24 hours, and 48 hours
*/

data time1;
/* baseline */
set repeated;
if time = 1;
drop obs;
rename time=time1;
rename exprs=exprs1;
run;

data time2;
/* 24 hours */
set repeated;
if time = 2;
drop obs;
rename time=time2;
rename exprs=exprs2;
run;

data time3;
/* 48 hours */
set repeated;
if time = 3;
drop obs;
rename time=time3;
rename exprs=exprs3;
```

```
run;

/* merging time point data together, necessary to perform t tests */

/* creating a csv file path for t test comparing baseline to 24 hours
*/

ODS HTML FILE="E:\Thesis\exprs1exprs2.csv";

/* running t test comparing baseline to 24 hours */

proc ttest data=ttestdata;
paired exprs1*exprs2;
by probesetid;

/* creating ODS table name for t test data /*

ods output Ttests=ttest_output;
run;

/* printing the t test output to the csv file */

proc print data=ttest_output;
run;

/* creating a csv file path for t test comparing baseline to 48 hours
*/

ODS CSV FILE="E:\Thesis\exprs1exprs3.csv";

/* running t test comparing baseline to 48 hours */

proc ttest data=ttestdata;
paired exprs1*exprs3;
by probesetid;

/* creating ODS table name for t test data /*

ods output Ttests=ttest_output;
run;

/* printing the t test output to the csv file */

proc print data=ttest_output;
run;

/* creating a csv file path for t test comparing 24 to 48 hours */

ODS CSV FILE="E:\Thesis\exprs2exprs3.csv";

/* running t test comparing 24 to 48 hours */

proc ttest data=ttestdata;
paired exprs2*exprs3;
by probesetid;
```

```
/* creating ODS table name for t test data /*

ods output Ttests=ttest_output;
run;

/* printing the t test output to the csv file */

proc print data=ttest_output;
run;

/* creating ODS table names for GLM procedures */

ods output overallANOVA=anova Contrasts=contr lsmeans=means diff=diff;

/* running proc glm in SAS */

proc glm data = repeated;
class sub time;
model exprs = time sub;
by probesetid;

/* creating contrasts comparing the time points */

contrast 'Baseline VS 24 and 48' time 2 -1 -1 / est;
contrast '24 VS Baseline and 48' time -1 2 -1 / est;
contrast '48 VS Baseline and 24' time -1 -1 2 / est;

/* creating comparison matrices for the time points */

lsmeans exprs / pdiff;
means exprs / tukey lsd;
run;

/* creating file path for contrast data */

ods csv file='E:\Thesis\exoncontrasts.csv';

/* printing contrast table to csv file */

proc print data=contr;
run;

/* creating file path for p values and anova table data */

ods csv file='E:\Thesis\exonanovafvalues.csv';

/* printing p values and anova table data to csv file*/

proc print data=anova;
run

/* creating file path for difference matrix data */

ods csv file='E:\Thesis\exonpdiff.csv';
```

```
/* printing difference matrices to csv file */

proc print data=diff;
run;

/* creating ODS table names for mixed procedures */

ods output covparms=cp asycov=asy infocrit=aic tests1=tests1
contrasts=contr diff=diff;

/* running the mixed procedure in SAS */

proc mixed data=repeated covtest asycov ic;
class sub time;
model exprs = time / s htype=1;
by probesetid;

/* creating contrasts comparing the time points */

contrast 'BS VS 24 and 48' time 2 -1 -1 / est;
contrast '24 VS BS and 48' time -1 2 -1 / est;
contrast '48 VS BS and 24' time -1 -1 2 / est;

/* creating comparison matrices for the time points */

lsmeans time / pdiff=all;
repeated time / TYPE = AR(1) subject=sub r;
run;

/* specifying file path for ODS table data, difference matrices */

ods csv file="E:\Thesis\exonpdiffmixed.csv";

/* printing csv file, difference matrices */

proc print data=diff;
run;

/* specifying file path for ODS table data, AIC scores */

ods csv file="E:\Thesis\exongeneaicscorear1.csv";

/* printing csv file, AIC score */

proc print data=aic;
run;

/* specifying file path for ODS table data, p values and anova tables
*/

ods csv file="E:\Thesis\exonmixedpvaluesar1.csv";

/* printing csv file, pvalues and anova tables */
```

```
proc print data=tests1;
run;

/* specifying file path for ODS table data, residual variances */

ods csv file="E:\Thesis\exonmixedwresidualar1.csv";

data res;
set cp;
if subject = 'sub' then
delete;
run;

/* printing csv file, residual variances */

proc print data=res;
run;

/* specifying file path for ODS table data, subject variances */

ods csv file="E:\Thesis\exonmixedwsubjectar1.csv";

data sub;
set cp;
if subject = ' ' then
delete;
run;

/* printing csv file, subject variances */

proc print data=sub;
run;
```

BIBLIOGRAPHY

[1]   Alberts B., Johnson A., Lewis J., et al. *Molecular Biology of the Cell*, 4th ed., New York, Garland Science, 2002. "Studying Gene Expression and Function," NCBI website http://www.ncbi.nlm.nih.gov/books/NBK26818/ (June 3, 2013).

[2]   Griffiths A., Wessler S., Lewontin R., Gelbert W., et al. *Introduction to Genetic Analysis,* 8th ed., W.H. Freeman and Company, New York, (2005).

[3]   National Human Genome Research Institute, 2011 "DNA Microarray Technology," National Institute of Health website http://www.genome.gov/10000533 (June 6, 2013).

[4]   R Core Team, 2013. "R: A Language and Environment for Statistical Computing," *R Foundation for Statistical Computing*, Vienna, Austria. R Project website http://www.R-project.org/ (May 13, 2012).

[5]   Irizarry R. A., Bolstad B. M., Collin F., Cope L.M., Hobbs B. Speed T. P, "Summaries of Affymetrix GeneChip Probe Level Data," *Nucleic Acids Research* Vol. 31 No.4 p. 15 (2003).

[6]   Human Exon 1.0 ST Array - Support Materials, 2013. "Support by Product for GeneChip® Human Exon 1.0 ST Array," Affymetrix website http://www.affymetrix.com/support/technical/byproduct.affx?product=huexon-st (July 19, 2012).

[7]   The NCBI Handbook, 2002. "National Center for Biotechnology Information," NCBI website http://www.ncbi.nlm.nih.gov/books/NBK21101 (July 6, 2013).

[8]   GenBank Overview, 2013. "What is GenBank," NCBI website http://www.ncbi.nlm.nih.gov/genbank/ (July 6, 2013)

[9]   Montgomery D. C., *Design and Analysis of Experiments*, 7th ed., John Wiley & Sons, (2009).

[10]   Wang J, Jia M, Zhu L, Yuan Z, Li P, et al., "Systematical Detection of Significant Genes in Microarray Data by Incorporating Gene Interaction Relationship in Biological Systems," *PLoS ONE,* Vol. 5 No. 10 pp 1-13 (2010)

[11]   Brockwell P. J., Davis R. A., *Introduction to Time Series and Forecasting,* 2nd ed., Springer, (2002).

[12]   Smyth, G. K. (2005). "Limma: Linear Models for Microarray Data," *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, pp 397-420.

[13]   Benjamini Y., Hochberg Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological),* Vol. 57 No. 1 pp 289-300.

VITA

Stephanie Marie Berhorst was born on April 21, 1987 in Jefferson City, Missouri. She holds a B.Sc. Psychology degree from Lincoln University in Jefferson City, Missouri 2010. She will complete a M.Sc. in Applied Mathematics in December 2013 from Missouri University of Science and Technology.