A Comparative Study of Methods of Analyzing Gene Expression Data*

Joseph DeCampo, Kristin Harp, and Christopher H. Morrell†

Mathematical Sciences Department, Loyola College in Maryland, 4501 North Charles Street, Baltimore, MD, 21210-2699

Abstract

In analyzing microarray data it is often necessary to detect genes that are differentially expressed between two or more samples. This project aims to apply two methods to address statistical issues that arise when identifying differentially expressed genes. The first is a gene-by-gene analysis that attempts to overcome the small sample size issue that is often present in microarray data sets. By averaging the variances of genes with similar expression levels, we are able to stabilize the test statistics used in determining significant genes and obtain more powerful tests. When looking at thousands of tests, one for each gene, problems arise involving the type I error rates. The leads to multiple testing issues that must be addressed. We applied many methods of correcting or adjusting the p-values for multiple testing. Based on this study, the false discovery rate method appears to provide a reasonable balance between the type I error rate and allowing sufficient power to detect differential expression when present.

A second approach is to use an overall model for the entire data set. A mixed-effects model is fit to a subset of 100 genes. The estimates of random-effects may be examined to identify differentially expressed genes. An alternative overall model uses a hierarchical Bayesian model to analyze the entire data set using BUGS. This method was also applied to a subset of 100 genes.

1. Introduction

In recent years there has been rapid progress made in mapping the human genome and numerous other genetics-related projects. Many of these advances have been made possible by the use of microarrays. A microarray is a slide or membrane containing numerous probes that represent various genes of some biological specimen. Probes are either oligo-nucleotides that range in length from 25 to 60 bases, or cDNA clones with length from a hundred to several thousand bases. Microarrays are hybridized with labeled cDNA synthesized from a mRNA-sample of some tissue. The intensity of label (radioactive or fluorescent) of each spot on a microarray indicates the expression of each gene. One-dye arrays (usually with radioactive label) show the absolute expression level of each gene. Two-dye arrays (fluorescent label only) can indicate relative expression level of the same gene in two samples that are labeled with different colors and mixed before hybridization. One of these samples can be a universal reference which helps to compare samples that were hybridized on different arrays.

---

* Technical Report 2004-01, Mathematical Sciences Department, Loyola College in Maryland, 4501 North Charles Street, Baltimore, MD 21210-2699

† Corresponding author. e-mail: chm@loyola.edu

Microarray experiments try to simultaneously measure the expression levels of thousands of genes. In many situations, it is of interest to determine which genes are *differentially expressed* between two or more types of tissue or between subjects with or without some disease process.

There are many statistical issues that need to be overcome in identifying differentially expressed genes and many papers have been written in recent years proposing a variety of methods of analysis (a selection is provided in the reference section). These methods fall into two main classes: (i) methods that compare the groups gene-by-gene and make corrections to the p-values provided by each test; and (ii) methods that identify differentiably expressed genes by modeling the entire data set. The aims of this project are to compare a number of types of analyses from both classes.

In the first class of methods, test statistics are used to test the equality of expression levels across the groups being compared. These tests produce p-values that may be used to assess the statistical significance of the test. However, there are many p-values – one per gene. Since one p-value is calculated for every gene in the data set, multiple testing is an issue that needs to be addressed. If a number of statistical tests are applied (in the microarray case, thousands of tests), each with a specified probability of a type I error, then the overall probability of at least one type I error is much higher. A number of methods will be investigated to address this problem: a Bonferroni correction, a modified Bonferroni method, Sidak's method, Holm's method, and the False Discovery Rate approach.

Due to the large expense, frequently only a few microarray replicates are run. Consequently, microarray data sets typically have small sample sizes on many variables. In this case, the estimates of the variance that are used to compute the test statistics can be unstable. As a result, the tests will have few degrees of freedom for the error term and will have low power to detect real differences between samples. One way of overcoming this problem is to pool the variances over a number of genes (variance averaging) with a similar mean expression level to obtain a variance estimate with more degrees of freedom, thereby giving the tests more power to detect differences.

As mentioned above, a second class of methods seeks to provide a model for the entire data set. Wolfinger et al. (2001) developed a mixed-effects model that can be used to identify differentiably expressed genes while controlling the percent of false positives. This method improves on existing methods with respect to the number of false negatives. They use two interconnected mixed-effects models that account for the variability across and within genes. They provide example SAS code that can be modified to apply their methods to other data sets. Their method is appropriate for two-dye arrays. Since we have data available for single-dye arrays, we will fit a related mixed-effects model to our microarray data. Finally, a Bayesian analysis of the same model is fit using the software BUGS (Bayesian inference Using Gibbs Sampling).

Section 2 will discuss variance averaging to obtain improved variance estimates for use in the test statistics. Section 3 defines and investigates the use of a number of methods to control for multiple testing. Section 4 proposed a mixed-effects model for microarray data and fits the model to a reduced set of data. Section 5 fits a similar model using a Bayesian formulation. Finally, in Section 6 we present some conclusions of this study and areas for additional work.

2. Variance Averaging

When performing gene-by-gene analyses, the mean expression level (radioactive or fluorescent intensity) for each gene is calculated along with estimates of the variance. These statistics are then used to calculate the t statistic that is used to determine statistical significance. In this case, rejecting the null hypothesis means that that the gene is differentially expressed between the samples.

One limitation that occurs in microarray analyses is that multiple replications can be expensive and thus it is typical for a microarray sample to consist of only 3 or 4 slides. This small sample size is problematic because the variance estimates used to calculate test statistics have few degrees of freedom for the error term. As a result, the power to detect differentially expressed genes between two types of tissues decreases. Variance averaging or variance pooling attempts to address this problem. The method collects the variances for genes with a similar expression level (measured by the mean intensity) and computes an average variance in groups of 100 to 500 genes. This average variance replaces the actual variance used in computing the test statistic. By averaging the variances, we increase the degrees of freedom in our hypothesis testing, which lends more power to our results. This method of variance averaging can be applied because, in general, a relationship exists between the mean expression level of a gene and its variance estimate (See Figure 1). Having more degrees of freedom also allows us to obtain p-values using the standard normal distribution rather then the t-distribution.

To implement the variance averaging method, we used SAS software for all computations and Excel to produce the graphical results. First, the genes are sorted by mean expression level and the average variance is estimated using a sliding window of 100 or 500 genes. We were able to create this sliding window using the following SAS procedure:

```
proc expand data=finalsd out=cumvar method=none;
convert var = avgvar / transform = (movave 500);
run;
```

The average variance calculated for each group of genes is assigned to the gene in the center of the window, i.e. position 50 or 250. We were able to assign the average variances to this center position using the **lag** statement in SAS. For the first 50 (or 250) and last 50 (or 250) genes in the set, we use the first (last) average variance computed.

The next step in the process is to compare p-values calculated for each gene using the pooled and unpooled variances. For each set, genes with p-values less than or equal to 0.05 and 0.10 are examined to determine the effects of the variance averaging method on the results. We are particularly interested in genes that "switch" between the two methods. These are genes that were significant under the usual test statistic and became insignificant when using the average variance method, or conversely. These genes will enable us to visualize the impact the average variance method has on determining differentially expressed genes.

We worked with two data sets for the first phase of our analysis. The first data set contained data from 16 microarrays - 8 from patients who died with heart disease and 8 from patients who died without heart disease. This set contained approximately 9,000 genes to be analyzed (Boheler et al. (2003)). For this data set the expression levels have been normalized to account for background and eliminate any within-array problems. The resulting data are z-scores representing the expression levels. Our second data set had a far smaller sample size; it contained only 6 microarrays - 3 from mouse placentas and 3 from mouse embryos. This set however, contained approximately 12,600 genes (Tanaka et al. (2000)). In this case we analyzed the log-transformed expression levels.

First, to visualize the relationship between the variance and mean expression level, we graphed the variance versus the mean expression level for each gene (see Figure 1).

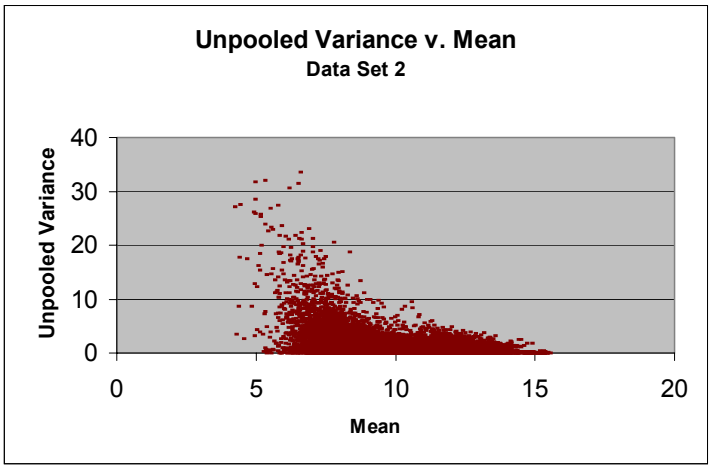

Figure 1. Variance vs. mean for each gene in the second data set.

Figure 1 shows that genes with higher expression levels tend to have a lower variance than genes with low expression levels. This supports our assumption that a relationship exists between expression level and variance and enabled us to proceed with the variance averaging method.

We used both a 100 and a 500-sliding window for both data sets. In each case, the 500 windows significantly reduced the spread in the variances. We observed interesting relationships between the averaged variances and the mean expression levels in both data sets. Figures 2 and 3 show the relationship between the average variance and mean expression level for both the 100 and 500 window for the first data set. Figures 4 and 5 show the same for the second data set.

For the first data set, we compared the number of significant genes before and after using the average variance method. We looked at genes that were significant for p=0.05 and p=0.10. Table 1 summarizes our findings. The first data set contained 9,182 genes, of which 22 tested significant under both methods. Thirty-five genes tested significant using the average variance method, which is considerably less than the 102 the tested significant using the actual variances.

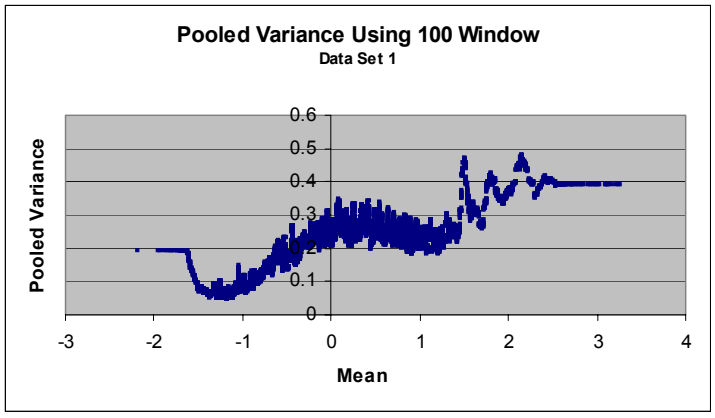

Figure 2. Pooled variance vs. mean for each gene in the first data set. Window = 100 genes.

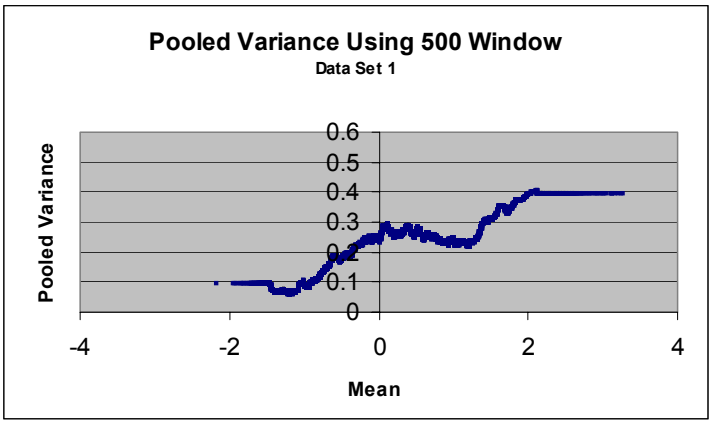

Figure 3. Pooled variance vs. mean for each gene in the first data set. Window = 500 genes.

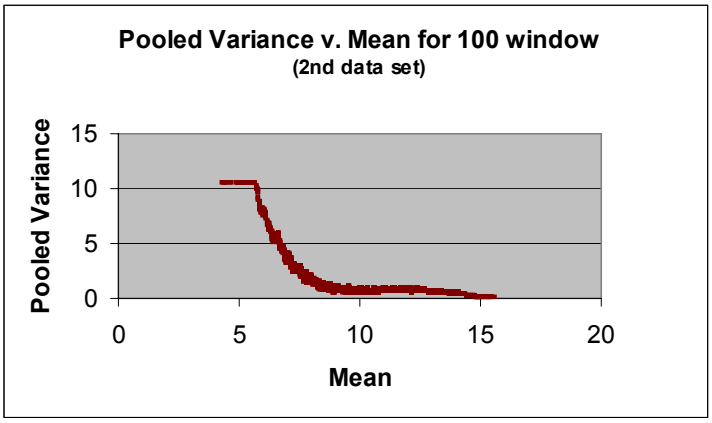

Figure 4. Pooled variance vs. mean for each gene in the second data set. Window = 100 genes.

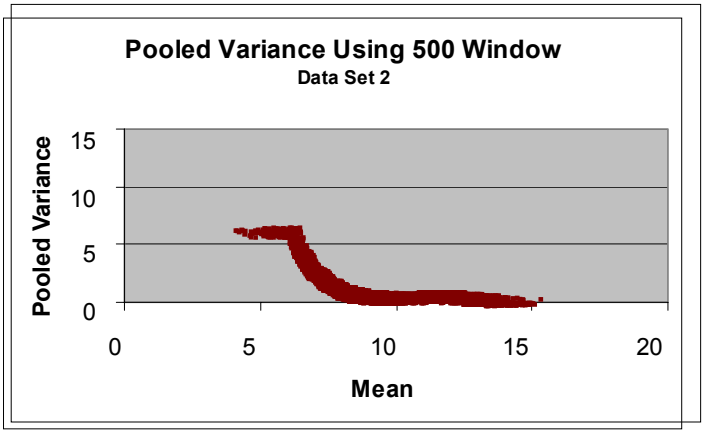

Figure 5. Pooled variance vs. mean for each gene in the second data set. Window = 500 genes.

Table of pt05 by pz05			
pt05	pz05		
Frequency Percent	0	1	Total
0	9067 98.75%	13 0.14%	9080 98.89%
1	80 0.87%	22 0.24%	102 1.11%
Total	9147 99.62%	35 0.38%	9182 100%

Table 1. Statistical significance for data set 1. Rows represent the t-test (no variance averaging) while columns are for the pooled variance z-test.
0 = Not statistically significant. 1 = Statistically significant.

Table of pt05 by pz05			
pt05	pz05		
Frequency Percent	0	1	Total
0	11526 90.82%	405 3.19%	11931 94.01%
1	506 3.99%	254 2.00%	760 5.99%
Total	12032 94.81%	659 5.19%	12691 100%

Table 2. Statistical significance for data set 2. Rows represent the t-test (no variance averaging) while columns are for the pooled variance z-test.
0 = Not statistically significant. 1 = Statistically significant.

Table 2 presents the same results for the second data set. Our second data set contained 12,691 genes, of which 254 tested significant under both methods. 659 genes were significant using the average variance method versus 760 that were significant using the actual variances.

3. Multiple Testing

A type I error is the probability of rejecting a single null hypothesis that is, in fact, true. The probability of committing a type I error is denoted by α. It is desirable to keep α small. In general, a null hypothesis is rejected if the calculated p-value is less than some predetermined acceptable α-level, often 0.05.

In microarray experiments we simultaneously test null-hypotheses for all genes. The problem that arises from trying to test thousands of hypotheses is that if the significance level is controlled for each individual test, then the overall probability of rejecting at least one true null

hypothesis increases with the number of tests performed. For example, if a single microarray contains 10,000 genes and α is set to the 0.05 level, then we expect the experiment to result in 500 genes testing significant, even if all null hypotheses are in fact true. This example shows that the p-values must be used with care for multiple hypotheses testing with microarrays. In general, if we have N tests, the probability of rejecting at least one true null hypothesis is given by $1-(1-\alpha)^N$. Table 3 shows the relationship between the number of simultaneous hypotheses tested and α .

Table 3: Type I Error Rates for Various Numbers of True Null Hypotheses (N)

N	$\alpha = 0.1$	.05	.01
1	.1	.05	.01
5	.41	.23	.049
10	.65	.40	.096
50	.995	.923	.39
100	.99997	.994	.634
1000	1	1	.99996

As the overall probability of a type I error becomes unacceptably high, it is expected that many false positives will result. Therefore, it is necessary to control the type I error rate. There are two types of rates that may be controlled in some way. The Comparison-Wise Error Rate is the probability of a type I error for each hypotheses tested. The Family-Wise Error Rate (FWER) is the overall probability of at least one type I error among all hypotheses tested.

In 1995, Benjamini and Hochberg introduced a new multiple hypothesis testing error measure with a different goal in mind – to control the proportion of type I errors among all rejected null hypotheses. The FDR (false discovery rate) is the proportion of false positives among all genes that we consider significant. If a FDR of 5% is used, it is expected that 5% of all rejected hypotheses are, in fact, falsely identified as such.

There are many methods of adjusting p-values in order to control either the FWER or the FDR. The first and simplest method of controlling the FWER is the Bonferroni method; the adjusted p-value = $\min(N \times \text{p-value}, 1)$. This method ensures an overall type I error rate of at most 5% (if a 5% cutoff is used to determine significance). However, this method is very conservative, especially when the number of tests is large as it is with microarray experiments.

The Sidak method also controls the FWER. The adjusted p-value = $1 - (1 - \text{p-value})^N$. This method results in slightly smaller adjusted p-values than the Bonferroni method, but very similar ones, especially in cases with very large N. A third, related, method is the Holm Step-down method. Here the adjusted p-value = $\text{Min}((\text{N-rank}+1) \times \text{p-value}, 1)$. This method takes into account the rank of each p-value (when the p-values are sorted in order from smallest (or most significant) to largest (closest to one)). The purpose for including the rank of each gene (in attempting to control the FWER) is that if the smallest p-value is rejected, then the number of genes being considered is one less (for the second gene, say). Step-down methods modify their adjustments accordingly. Holm-adjusted p-values are generally less conservative than the Bonferroni adjusted p-values. However, the improvement is only slight when N is very large.

The FDR is an intermediate method between unadjusted p-values and the Bonferroni correction method. The FDR is the proportion of false positives among all genes that we consider significant. Benjamini and Hochberg (1995) define FDR = min(p-value$\times$N/rank, 1)). Recently, Storey (2002) and Storey and Tibshirani (2003) consider an alternative definition of the FDR. They provide code written for the statistical software package R, that calculates this alternative version of the false discovery rate, which they call "q-values."

In 1993, Westfall and Young developed a permutation method for adjusting p-values (nonparametric re-sampling). First, test statistics are calculated to obtain unadjusted p-values using usual techniques. Then the data is pooled and re-sampled as new test statistics are computed for each and every possible one of these new permuted samples. The p-value for the original test statistic is then computed as the tail area from the distribution of these new re-sampled test statistics. The major benefit of this method is that it does not assume independence between the tests, however it requires relatively large samples to be effective. Microarray studies frequently have small samples (3-8) on many genes.

Results for Adjusted p-values

Standard analysis of variance (ANOVA) techniques were performed on the first data set, which contained 9182 genes with 16 replications (8 diseased and 8 non-diseased tissue samples) using the program SAS. This is equivalent to performing a two-sample t-test for each gene. Before adjustment, 222 genes (of the original 9182) appeared to be significant at the 10% level ($\alpha = 0.1$). After the methods described in section 3were implemented, only one gene proved to be significant. Table 4 shows the results of the various adjustment methods in comparison with the unadjusted p-values for alpha levels 0.01, 0.05, and 0.1.

The second data set contained 15,123 genes and 6 replications (3 mouse placenta and 3 mouse embryo tissue samples). Before adjustment, over 1000 genes appeared to be significant at the 10% level ($\alpha = 0.1$). After these methods were implemented, it became apparent that the different adjustment methods yielded different numbers of statistically significant genes. Table 5 illustrates this for alpha levels 0.01, 0.05, and 0.1.

Table 4: Number of Significant Genes at Various Significance Levels for Data Set 1

Cutoff	Raw p-vals	Bonferroni	Sidak	Holm	Permute	FDR	Q-vals
0.01	25	0	0	0	0	0	0
0.05	102	0	0	0	0	0	0
0.1	222	1	1	1	1	1	1

Table 5: Number of Significant Genes at Various Significance Levels for Data Set 2

Cutoff	Raw p-vals	Bonferroni	Sidak	Holm	FDR	Q-vals
0.01	60	1	1	1	1	1
0.05	299	1	1	1	10	15
0.1	1059	3	3	3	17	22

It should be noted that the permutation method was not run for this data set. Due to the low number of replications, there were not enough possible permutations of the data to apply this method.

In studies such as these, some type I errors cannot be avoided, but it is still desirable to keep false positives to a minimum. Based on our results, we recommend that the FDR should be controlled before the FWER. Based on this and on Tables 4 and 5, the two methods used for controlling the False Discovery Rate (FDR and Q-value) yielded the most appropriate results. Figure 6 shows the relationship between the FWER-controlling Bonferroni correction and the FDR method for the second data set. The horizontal line drawn in the scatter plot corresponds to the 0.05 cutoff value for the FDR. It can be seen that even as Bonferroni adjusted p-values approach 1.0, FDR adjusted values remain close to 0.05. FDR corrected p-values increase at a slower rate, thereby yielding more significant results.

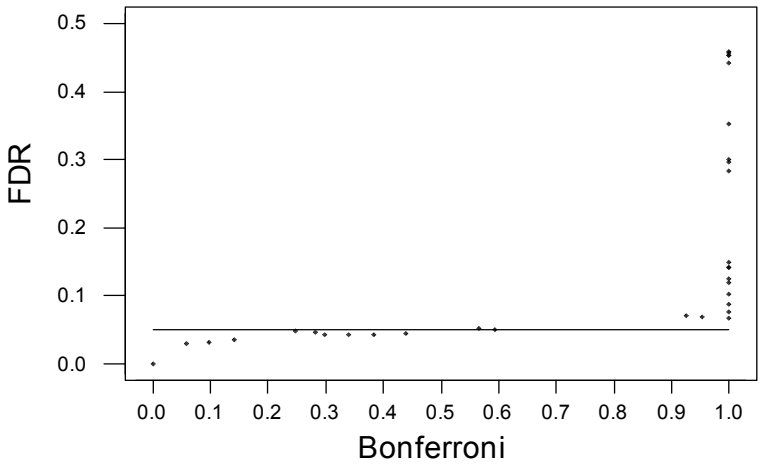

Figure 6: FDR vs. Bonferroni adjusted p-values

4. A Mixed-Effects Model for Detecting Differential Expression

Wolfinger et al (2001) proposed a mixed model for analyzing all the data from the microarray. We had hoped to apply this methodology to our data. However, on closer inspection, this model assumed a different experimental design than the one we used in our microarrays. In particular, their model was for two-dye microarrays. The data we had available was from single-dye microarrays. Consequently, we could not apply this method to our data.

As an alternative approach, we decided to use a mixed model approach to try to determine differential expression. We fit the model:

$$Y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})X_{ij} + \varepsilon_i, \quad i = 1,\ldots, N, \; j = 1, .., n_i. \qquad (1)$$

$$\text{where } X_{ij} = \begin{cases} 0 & \text{if gene i is in condition } 0 \\ 1 & \text{if gene i is in condition } 1 \end{cases}$$

where β_0 and β_1 are fixed-effects and represent population parameters within the model. β_0 is the mean expression for condition 0 and β_1 is the difference in mean expression level between the two conditions. b_{i0} and b_{i1} are random effects. b_{i0} is the difference in expression for gene *i* from the mean of *all genes* for condition 0, and b_{i1} is the *differential expression* for gene *i* between condition 1 and condition 0. Consequently, by examining estimates of the random effects, b_{i1}, one can identify genes that are differentially expressed.

We attempted to fit this model, but due to numerical problems with SAS we were not able to get sensible results for the entire data set, The variance components estimates all converged to 0 – an unreasonable result given the observed variation between and within genes.. When we restricted attention to 100 genes the method appeared to work quite well. For future work, one could derive the formulas for this special case of the mixed model and compute the estimates using another software program or by writing a program in another programming language.

5. Estimation of the Mixed-Effects Model via Bayesian Methods

As a second approach to modeling the entire data set, we developed a hierarchical Bayesian model similar to the mixed-effects model described in section 4. We used a software package called "BUGS" (Bayesian inference Using Gibbs Sampling) to fit this mixed effects model. The Bayesian method models the data using a probability model. The parameters of this model are described by a set of prior distributions. These prior distributions have parameters and these prior parameters may also have (hyper) prior distributions themselves. We used Model (1) above in this context but estimate the parameters of the model using Gibbs sampling. As mentioned above, the random effects are the primary interest as they will determine which genes are differentially expressed between the samples. To set up the model for use in the BUGS software, we rewrite y as:

$$y_{ij} = \mu_{ij} + \varepsilon_{ij} \qquad (2)$$

where k = 0 or 1 for the two samples and

$$\mu_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})\, X_{ij} \qquad (3)$$

The model is then translated into the following BUGS code:

```
model
{
for (i in 1:N) {                                  // N total genes
   for (k in 1:(2*n)) {                           // n replications per gene
   muy[i,k]<- beta0 + b1[i] + (beta1 + b2[i])*X[i,k]
   Y[i,k] ~ dnorm(muy[i,k],pe)
}
b1[i] ~ dnorm(0, d11)                             //prior for b_i 0
mub2[i] <- rho*d22/d11*b1[i]
b2[i] ~ dnorm(mub2[i], varb2)                     //prior for b_i 1
}
varb2 <- d22*(1-rho*rho)
d11 ~ dgamma(.001, .001)
d22 ~ dgamma(.001, .001)
rho ~ dunif(-1,1)
pe ~ dgamma(.001, .001)
beta0~ dnorm( 93984.88, 0.000001)                 //prior for β0
beta1~dnorm( -32223.6, 0.000001)                  //prior for β1
}
```

The priors for β_0 and β_1 are normal distributions where the prior means are obtained from the sample mean calculated from the data set. BUGS parameterizes the normal distribution using the mean and precision (1/variance). The small prior precision leads to a relatively flat and non-informative prior distribution.

The following sections of code were used to input the data and initialize the variables used in the analysis. In our work with the BUGS model, we were only able to use a small subset

of genes in the analysis due to limitations in our ability to load the entire data set. The first one hundred genes were selected and analyzed.

```
data:
list(Y=structure(.Data = c(20667.24556,403.04722,2486.24222,2293.441111,4050.49,1,
10345.36556,1535.86722,189.66222,1679.241111,2333.21,1,
23604.53556,4236.29722,5316.44222,2822.561111,4155.22,4542.45389,
…                                                    // Expression levels
68367.27556,32782.00722,32700.31222,7903.151111,67915.97,15540.69389,
29362.08556,36110.55722,41499.54222,2158.031111,25105.25,8234.52389
), .Dim = c(100,6)),
X=structure(.Data = c(
0,0,0,1,1,1,
0,0,0,1,1,1,
…                                    // 0 and 1 differentiate between the samples
0,0,0,1,1,1,
0,0,0,1,1,1), .Dim = c(100,6)), N = 100, n = 3)

inits:
list(d11 = 0.000001, d22 = 0.000001, rho =0, pe = 0.000001,
beta0 = 93984.88, beta1=-32223.6)
```

The model is fit using BUGS. We tracked four parameters in the model: β_0 - the mean of condition 0, β_1 - the difference in means between the two conditions, b_{i0} - the difference in expression for gene *i* from the mean of *all genes* for condition 0, and b_{i1} - the *differential expression* for gene *i* between condition 1 and condition 0. Given the starting values used for the Gibbs samples, the estimates of both β_0 and β_1 are expected to remain relatively constant over the iterations by the BUGS software. The following BUGS output supports this expectation.

Estimate of β_0:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	94030.0	993.6	37.24	91960.0	93980.0	96020.0	501	500

Here we see that the average of the estimates of β_0 is 93,980. This is consistent with the sample mean of 93,984.88, which was the starting estimate. We also graphed the sampling distribution and iteration history for β_0 to illustrate its constancy. Figure 7 displays the sampling distribution of β_0 and Figure 8 shows the value of β_0 for each iteration.

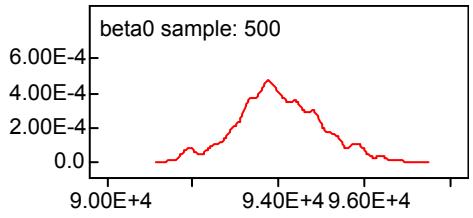

Figure 7: Sampling distribution from the Gibbs sampler for β_0

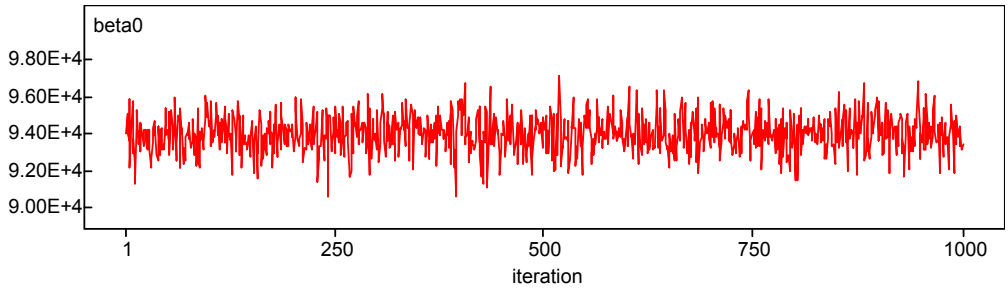

Figure 8: Estimates of β_0 for each iteration from the Gibbs sampler

The following output also shows that the β_1 estimate from the model did not differ significantly from the sample mean of -32,223.6. The parameter estimates exhibit a stationary pattern indicating that the Gibbs sampler has reached a steady state. Consequently, the results from the iterations can be used to construct the sampling distributions which allow us to perform inferences on the various parameters.

Estimate of β_1:

Node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta1	-32310.0	1019.0	58.84	-34330.0	-32350.0	-30310.0	501	500

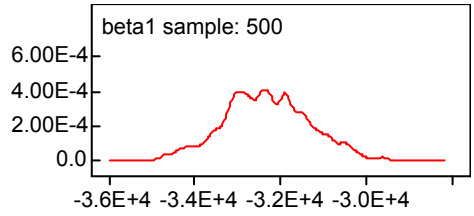

Figure 9. Sampling distribution from the Gibbs sampler for β_1

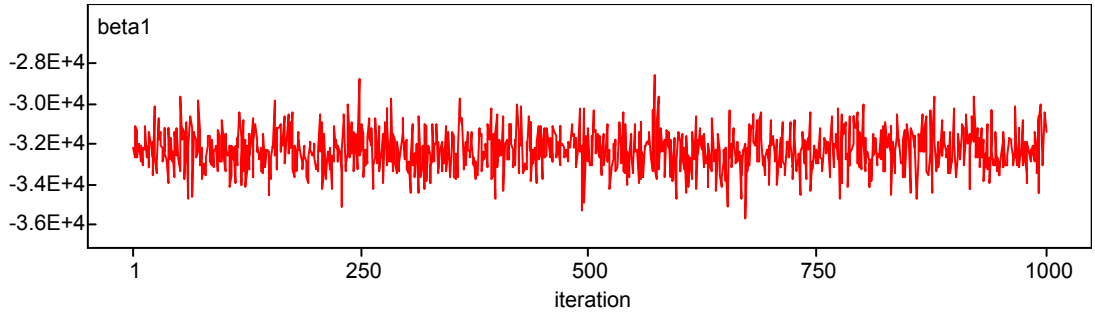

Figure 10. Estimates of β_1 for each iteration from the Gibbs sampler

Figure 9 shows the sampling distribution for β_1 and Figure 10 shows the value of β_1 for each iteration. For the random effects b_{i0} and b_{i1}, statistics were tracked for each individual gene. Criteria are determined to identify genes that are differentially expressed based on the results of the fitted model. In the code used to write the model, b_{i1} is called b2[i]. The graphical results of a selection of estimates per gene are presented below (Figure 11). The figures on the left are a

history of the value of b2[i] at each iteration. The figures on the right are the sampling distribution for the same selected genes.

In each of these genes, we see that b2 is centered on zero. The means of the b2[i] are our estimate of differential expression for gene i. Following the graphs, are the descriptive statistics for the means of the 100 b2[i]. A histogram of these 100 means shows that there are some outliers which may qualify as differentially expressed genes (Figure 11). Figure 12 shows the distribution of b2[i] without these "outliers."

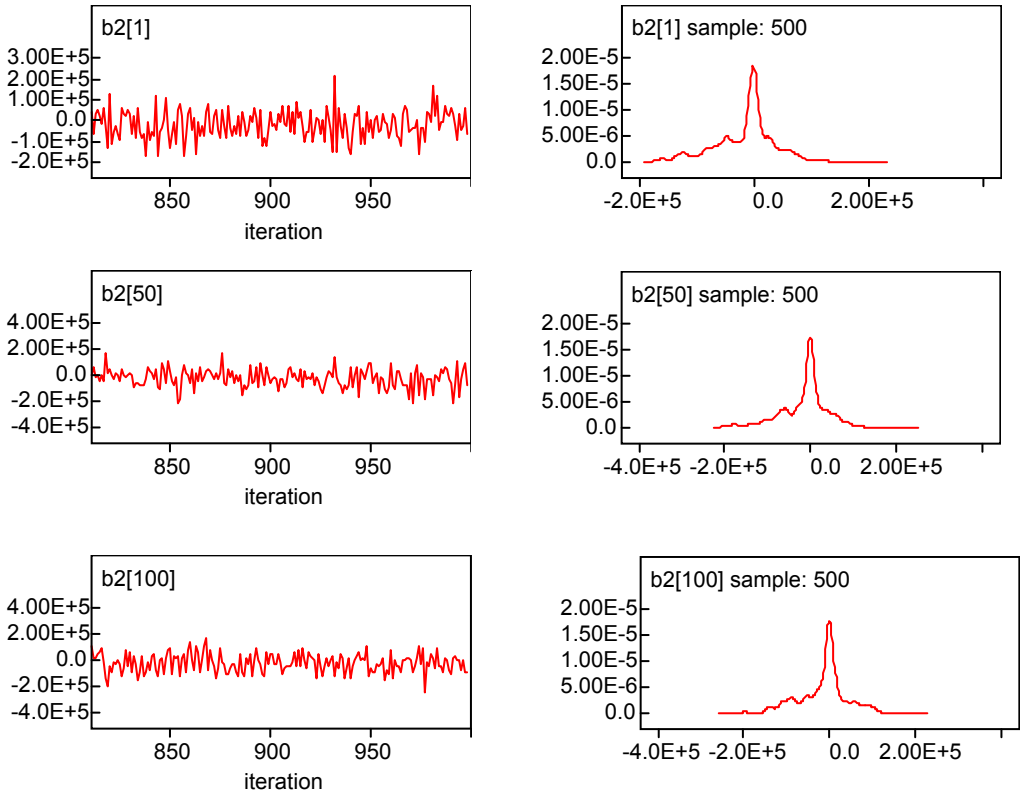

Figure 11. Gibbs sampler estimates and sampling distributions for selected random effects.

Descriptive Statistics: b2

Variable	N	Mean	Median	TrMean	StDev	SE Mean
b2	100	121	-11255	-5629	35392	3539

Variable	Minimum	Maximum	Q1	Q3
b2	-18110	265300	-13695	-2138

We identified thirteen genes with extremely high values for b2. These genes are candidates to be considered differentially expressed based on the model. Table 6 lists the mean expression level estimated by BUGS and the gene number for each of the thirteen genes.

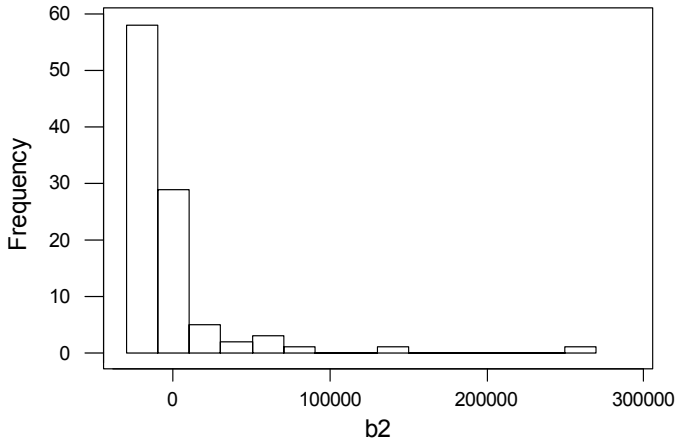

Figure 11. Distribution of the Bayesian estimates of differential expression for the 100 selected genes.

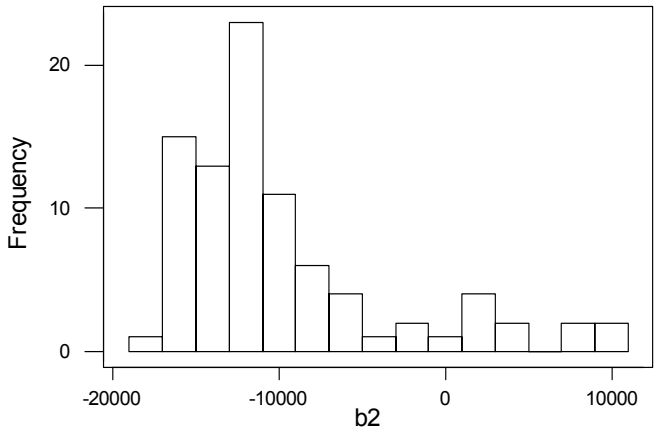

Figure 12. Distribution of the Bayesian estimates of differential expression after omitting the 13 large values that may indicate differential expression

Mean	Position
22400	b2[88]
24310	b2[34]
24730	b2[59]
24750	b2[44]
27620	b2[76]
30560	b2[30]
48640	b2[85]
59060	b2[89]
62500	b2[12]
63860	b2[53]
80940	b2[26]
130900	b2[39]
265300	b2[56]

Table 6. The 13 largest estimates of differential expression.

6. Conclusions

After exploring these methods, we recommend the use of test statistics for each gene that compare the expression levels among the varieties. These test statistics must be computed using some kind of variance averaging procedure that will provide better variance estimators for each test and consequently a higher power for each test of obtaining a small p-value. Some adjustment must be made to account for the multiple testing issue. We recommend the use of the false discovery rate (FDR) method. More work may be done to investigate the models of Kerr et al. (2000, 2000, 2001) and Wolfinger et al. (2001). Also the unified approach may be explored further to attempt to fit these models to larger sets of genes.

Acknowledgement: The work of Joseph DeCampo and Kristin Harp was supported by the Summer Hauber Research Program at Loyola College during the summer of 2003.

REFERENCES

K.A. Baggerly, K.R. Coombes, K.R. Hess, D.N. Stivers, L.V. Abruzzo, and W. Zhang (2001). 'Identifying Differentially Expressed Genes in cDNA Microarray Experiments,' *Journal of Computational Biology*, **8**, 639-.

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

Boheler, K.R., Volkova, M., Morrell, C.H., Garg, R., Zhu, Y., Margulies, K., Seymour, A., Lakatta, E.G. (2003) 'Sex and Age-dependent human trancriptome variability: Implications for chronic heart failure,' *Proceedings of the National Academy of Sciences*, 100, 2754-2759.

The BUGS Project (Bayesian inference Using Gibbs Sampling), http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

S. Dudoit, Y. H. Yang, T. P. Speed, and M. J. Callow (2002). 'Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.' *Statistica Sinica*, Vol. 12, No. 1, p. 111-139.

Kerr, Martin and Churchill(2000), 'Analysis of variance for gene expression microarray data,' *Journal of Computational Biology*, **7**:819-837.

Kerr and Churchill(2001), 'Statistical design and the analysis of gene expression microarrays,' *Genetical Research*, **77**:123-128.

Kerr and Churchill(2000), 'Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments,' *Proceedings of the National Academy of Sciences*, **98**:8961-8965.

Lee MLT, Lu W, Whitmore GA, Beier D. (2002) 'Models for microarray gene expression data,' *Journal of Biopharmaceutical Statistics*, 12: 1-19.

Mei-Ling Ting Lee, G. A. Whitmore, and Rus Y. Yukhananov, (2003) 'Analysis of Unbalanced Microarray Data,' *Journal of Data Science*, **1**:2, 103-121.

M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, and K.W. Tsui (2001). 'On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data,' *Journal of Computational Biology*, **8**, 37-52.

Storey JD. (2002) 'A direct approach to false discovery rates' *Journal of the Royal Statistical Society, Series B*, **64**: 479-498.

Storey JD and Tibshirani R. (2003) 'Statistical significance for genome-wide studies,' *Proceedings of the National Academy of Sciences*, in press.

Tanaka TS, Jaradat SA, Lim MK, Kargul GJ, Wang X, Grahovac MJ, Pantano S, Sano Y, Piao Y, Nagaraja R, Doi H, Wood WH 3rd, Becker KG, and Ko MSH. (2000). 'Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray,' *Proceedings of the National Academy of Sciences*. USA 97: 9127-9132.

R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules (2001) 'Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models,' *Journal of Computational Biology*, **8**, 625-637.