MIXED ANOVA MODEL ANALYSIS OF MICROARRAY EXPERIMENTS WITH LOCALLY POOLED ERROR

Yuan Liu

A Thesis Submitted to the
University of North Carolina at Wilmington in Partial Fulfillment
Of the Requirements for the Degree of
Master of Arts

Department of Mathematics and Statistics

University of North Carolina at Wilmington

2004

Approved by

Advisory Committee

_____ _____

_____

Chair

Accepted by

_____

Dean, Graduate School

TABLE OF CONTENTS

ABSTRACT

The determination of a list of differentially expressed genes is a basic objective in many cDNA microarray experiments. Combining information across genes in the statistical analysis of microarray data is desirable because of relatively small number of data points obtained for each individual gene. Our LPE approach finds a middle ground between global F test and gene-specific F test by pooling the information across a group of genes that have similar variance estimates and shrinks the within-gene variance estimate towards an estimate including more genes. This method provides a powerful and robust approach to test differential expression of genes but does not suffer from biases of the global F test and low power of gene-specific F test. In our approach the two-stage Mixed ANOVA model provides a conceptually and computationally efficient means to analyze the microarray data.

DEDICATION

This thesis is dedicated to my parents, my husband and all statistics faculty at department of Mathematics and Statistics of UNCW. Thanks for their encouragement, love and support.

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Background

The advent of the genome project has vastly increased our knowledge of the genomic sequences of humans and other organisms. Various techniques such as cDNA microarrays and high-density oligonucleotide arrays have been developed to exploit this growing body of science and promise a wealth of data that can be used to develop a more complete understanding of gene function, regulation and interaction [1].

In this paper, our discussion is mainly relevant to cDNA microarrays. Spotted cDNA microarrays are a tool for high-throughput analysis of gene expression which provides rapid, parallel surveys of gene expression patterns for hundreds or thousands of genes in a single array. In the first step of the technique, DNA is "spotted" and immobilized on glass slides or other substrate, the microarrays. Each spot on an array contains a particular sequence, although a sequence may be spotted multiple times per array. Next, mRNA from cell population under study is reverse-transcribed into cDNA and one of two fluorescent dye labels, Cy3 (green) and Cy5 (red), is incorporated. Two pools of differently-labeled cDNA are mixed and washed over an array. Dye-labeled cDNA can hybridize with complementary sequences on the array, and any unhybridized cDNA is washed off. The array is then scanned for Cy3 and Cy5 fluorescent intensities. Although there are many unknown quantities in a microarray hybridization, such as the sizes and densities of the probe spots, and the hybridization and labeling efficiencies of different sequences, the basic principle is the following: for a given sequence spotted on the array, if one sample contains more of the corresponding transcript, the signal intensity for the dye used to label that sample should be higher than the other dye. Aside from the enormous scientific potential of microarrays to help in understanding gene regulation and interactions,

they have very important applications in pharmaceutical and clinical research. By comparing gene expression in normal and disease cells, microarrays may be used to identify disease genes and target for therapeutic drugs [2].

Any microarray experiment involves a number of distinct phases. Table 1 gives a schematic view of these phases of microarray experimentation that involve data-analytic steps [3]. In this paper, we focus on the identification of differentially expressed genes across experimental conditions in Data Analysis step by exploring a statistical approach to improving estimates of variability of differential expression.

Microarray experiments generate large and complex multivariate data sets. On a single glass slide, 10,000 to 20,000 cDNA probes can be spotted [4]. The current bottleneck in the processing of microarray data occurs after the data are generated. The difficulties stem primarily from myriad potential sources of random and systematic measurement error in the microarray process and from the small number of replications (both biological and technical replications) relative to the large number of variables (probes)[5]. Statistical methods have been used as a way to systematically extract biological information and to assess the associated uncertainty.

The simplest statistical method for detecting differential expression is the t test, which can be used to compare two conditions when there is replication of samples, based on the the fold change or the base 2 log of the expression ratio. Since gene-specific t test (use an estimate of error variance from one gene at a time) and global t test (assume the homogeneous variance between different genes and use an estimate of pooled error variance across all genes) are subject to low power and bias, respectively; while modified versions of t test find a middle ground. One version is regularized t test proposed by Baldi and Long [6] replaces the denominator for

Table 1: Phases of Microarray Experimentation

Experimental Design
Choice of sample size
Assignment of experimental conditions to arrays
Signal Extraction
Image analysis
Gene filtering
Probe level analysis of oligonucleotide arrays
Normalization and removal of artifacts for comparisons across arrays
Data Analysis
Selection of genes that are differentially expressed across experimental conditions
Clustering and classification of biological samples
Clustering and classification of genes
Validation and Interpretation
Comparisons across platform
Use of multiple independent datasets

a gene-specific t test with a Bayesian estimator based on a hierarchical prior distribution. In SAM (significance analysis of microarrays) version of t test, a small positive constant is added to the denominator of the gene-specific t-test to stabilize the small variances. When it comes to more than two conditions or more complex (multi-factor) experimental design, it is not enough simply to compute ratios. The analysis of variance (ANOVA) model can be applied to cDNA microarray data from any experimental design; however, microarray ANOVA models are not based on ratios but are applied directly to relative expression values. [7]

1.2 Mixed ANOVA Model

Every measurement in a microarray experiment is associated with a particular combination of an array in the experiment, a dye (red or green), a variety (can be treatment or experimental conditions), and a gene. Let y_{ijkgr} be the log fluorescent intensity from the r^{th} spot for gene g on array i for dye j and variety k [8]. A typical ANOVA model for a micorarray experiment can be the form of

$$y_{ijkgr} = \mu + A_i + D_j + (AD)_{ij} + G_g + (AG)_{igr} + (DG)_{jg} + (VG)_{kg} + \varepsilon_{ijkgr} \quad (1)$$

Here, μ is the overall mean expression level; the array effects A_i account for differences between arrays averaged over all genes, dyes, and varieties; the dye effects D_j account for differences between the average signal form each dye; $(AD)_{ij}$ is the term accounting for effects of the interaction between the array and the dye. These three kinds of effects account for overall variation in array and dyes and also are considered as "global" effects. They are not of interest, but accounting for them amounts to data normalization. In addition to these "global" normalization terms, there are source of variation to consider at the level of individual genes. They are the

terms of G_g,$(AG)_{igr}$,$(DG)_{jg}$ and $(VG)_{kg}$ in ANOVA model (1) and are considered as "gene-specific" effects. The gene effects G_g account for the expression level of genes averaged over the other factors. The term $(AG)_{igr}$ account for the average effect of the spot on array i for gene g . The $(DG)_{jg}$ terms account for the effect of dye j on gene g. The variety-by-gene terms $(VG)_{kg}$ represent levels of signal intensity for genes that can specifically be attributed to the RNA varieties under study which is the term that is of primary interest in our analysis.

Similarly model (1) can be specified in two stages [9]: normalization model and gene-specific model:

$$\text{Normalization model: } y_{ijkgr} = \mu + A_i + D_j + (AD)_{ij} + \gamma_{ijkgr} \tag{2}$$

$$\text{Gene-specific model: } \gamma_{ijkgr} = G_g + (AG)_{igr} + (DG)_{jg} + (VG)_{kg} + \varepsilon_{ijkgr} \tag{3}$$

Normalization model (2) mainly contains the "global" terms in model (1). The residuals from the normalization model are the input data for the gene-specific model which is applied one gene at a time.

Normalization plays an important role in the first stage of microarray data analysis which removes experiment-wide systematic effects that could bias inferences made on the data from the individual genes. Most common is red-green bias due to differences between the labeling efficiencies and scanning properties of the two fluoresces. Global normalization and locally weighted scatterplot smoothing (LOWESS) are the two widely used normalization methods. The normalization methods of using statistical models (model (1) and (2)) assume various effects are additive, and hence they are similar to the global normalization which subtracts a global constant to get normalized values [10].

A practical advantage of two-stage model is computational feasibility. The method of two-stage modeling can be viewed as a computationally tractable reformulation of model (1). General statistical software cannot handle model(1) because of the large number of parameters [11]. Especially when considering some effects as random and apply a mixed model, more complex computational methods are involved making it very difficult to fit a full mixed model in a single attempt.

Models (1),(2) and (3) may be taken as fixed-effects ANOVA model which assumes independence among all observations and only one source of random variation ε_{ijkgr}. Although it is applicable to many microarray experiments, the fixed-effects model does not allow for multiple sources of variation, nor does it account for correlation among the observations that arise as a consequence of different layers of variation. For example, measurements obtained on the same spot (one green and one red) will be correlated because they share common variation in the spot size. Failure to account for these correlations can result in underestimation of technical variance and inflated assessments of statistical significance [7]. Mixed ANOVA models [9] are appropriate to deal with these kinds of problems. The mixed ANOVA model has the same structure as the fixed-effects model; the difference is in the interpretation of terms that are treated as random effects. Typically, the array effects A_j and those terms involve array effects such as $(AD)_{ij}$ and $(AG)_{igr}$ in the fixed-effects model will be modeled as random and are assumed to have normal distribution with mean of zero and variance components $\sigma^2_A, \sigma^2_{AD}, \sigma^2_{AG}$ respectively. The mixed ANOVA model provides a general and powerful approach to allow full utilization of the information available in microarray experiments with multiple factors and a hierarchy of sources of variation.

Note that the properties of ANOVA estimates are tied to the experimental design and the model discussed here can be modified based on the specific experimental design.

1.3 Local Pooling of Errors (LPE)

The concept of local pooling of errors (LPE) was first proposed by Jain et al[12]. The approach is one to improve estimates of variability and statistical tests of differential expression. This method is based on the assumption that the variance of individual gene expression measurements is a (non-linear) function of gene intensity or is intensity-dependent in some way. By pooling the errors across the genes that have similar expression intensity values, it shrinks the within-gene variance estimate towards an estimate including more genes, and leads to more powerful signal-to-noise tests using this shrunken variance. The process for the oligonucleotide data is as follows: First, each baseline-error distribution (MA plot) is derived from all the replicated arrays under each condition. Let x_{ijk} be the observed expression intensity at gene j for array k and condition i. for duplicate arrays, $k = 1, 2$, MA plot is M ($= log_2(x_{ij1}/x_{ij2})$ and $log_2(x_{ij2}/x_{ij1})$) versus A ($= log_2\sqrt{x_{ij1}x_{ij2}}$). (1) estimation of error of M within quantiles of A (containing equal numbers of genes) and (2) non-parametric fit to the quantile error estimates. This two-stage error estimation approach is adopted because direct non-parametric estimation often leads to extreme estimates of error when only a small number of observations are available at a fixed-width intensity range. The LPE statistic for the median (log-intensity) difference for each gene under the two compared the conditions is then calculated as:

$$Z = \frac{Med_1 - Med_2}{\sigma_{pooled}}$$

Where $Med_i, i = 1, 2$ is the median log-intensity of the i^{th} condition;

$$\sigma^2_{pooled} = \frac{\pi}{2}[\sigma^2_1(Med_1)/n_1 + \sigma^2_2(Med_2)/n_2]$$

Where n_1 and n_2 are number of replicates in the two array samples being compared; $\sigma^2_i(Med_i), i = 1, 2$ is the estimate of variance of X (or Y) from the i^{th} LPE baseline-error distribution at each median log-intensity Med_i.

The similar idea can be seen in the work of Kerr et al [13]. Rather than making the extreme assumption that each gene has its own error distribution, they assume that the magnitude of the error is intensity-dependent. By pooling the information about genes with similar average level of expression, they get around the problem of few observations per gene. The method involves several steps. First, plot the standard deviation of the residuals per gene against the estimated gene effect $\widehat{G}_g$, and then fit a LOESS (locally weighted scatterplot smoothing) curve through the plot, so that for each gene, the standard deviation of its residual distribution is estimated by pooling the information across the genes locally within a fixed-width intensity range which is determined by loess span parameter. Next, re-scale the studentized residuals by dividing each residual by the estimated standard deviation of the associated gene (denoted $\widehat{SD}_g$). Let $\widehat{\epsilon}$ and $\widehat{e}$ denote the studentized and the rescaled residuals respectively, so that $\widehat{e}_{ijkg} = \widehat{\varepsilon_{ijkg}}/\widehat{SD}_g$. Third, create B bootstrap datasets

$$y^*_{ijkg} = \widehat{\mu} + \widehat{A}_i + \widehat{D}_j + (AD)_{ij} + \widehat{G}_g + (AG)_{ig} + (DG)_{jg} + (VG)_{kg} + \widehat{SD}_g * e^*_{ijkg}$$

Where e^*_{ijkg} is drawn with replacement from the $\widehat{e}_{ijkg}$. And finally get the bootstrap confidence intervals for relative expression of each gene.

The idea of LPE we discussed above can be summarized like this: Based on the assumption of intensity-dependent variance, locally combine the information across genes at similar intensity level for estimating error variances and making inferences. This method produces more robust inference than modeling error separately for every gene. However, the assumption of intensity-dependent variance need to be evaluated before using this method.

2 METHOD

The F test based on the ANOVA model has the null hypothesis of no differential expression and an alternative hypothesis with differential expression among the conditions. Cui and Churchill reviewed three flavors of F test for testing differential expression of gene [7]. One is called gene-specific F test which assumes heterogeneity in variance across genes and the F test of each gene is based on its own variance parameter. Practically in microarray experiments, each gene has only a small number of observations and this can cause the test for differential expression to have low power. In contrast the second one assumes homogeneity in variance across all genes and arrives the global variance F test, which uses only one estimate of error variance by pooling across all genes. Although it is the most powerful among the three F tests, it may suffer from bias if the error variance is not truly constant for all genes. Other than making those two extreme assumptions, a middle ground is achieved by the third F test, analogous to the regularized t test; it uses a weighted combination of global and gene-specific variance estimates in the denominator. The third F test improves estimates of variance and has power comparable to the second one but has a lower FDR (false discovery rate) than the global F test. Just like the third F test, a number of approaches to improving estimates of variability and statistical tests of

differential expression have been proposed and our approach is one of them. Inspired by LPE method of Jain [12], and our approach also finds a middle ground between global F test and gene-specific F test by pooling the information across a group of genes who have relatively close variance estimates, thus shrinking the within-gene variance estimate towards an estimate including more genes. Our approach is not based on any specific intensity-dependent variance assumption, but directly groups genes locally by their variance estimates so that a relative homogeneity of each group comes as a direct consequence.

The two-stage Mixed ANOVA model provides the computational feasibility and efficiency to apply this idea in practice.

$$\text{Normalization model: } y_{ijkgr} = \mu + A_i + D_j + (AD)_{ij} + \gamma_{ijkgr} \tag{4}$$

$$\text{Gene-specific model: } \gamma_{ijkgr} = G_g + (AG)_{igr} + (DG)_{jg} + (VG)_{kg} + \varepsilon_{ijkgr} \tag{5}$$

First, after a base-2 logarithmic transform, the data from the microarray experiment are fitted to normalization model (4) of the two-stage mixed ANOVA model, then the gene-specific model (5) is applied to γ_{ijkgr} one gene a time to get estimated error $\widehat{\sigma}_g$ for each gene, which is estimated by standard deviation of ε_{ijkgr} for each gene. Next, sort the genes in ascending order of their variance estimates $\widehat{\sigma}_g$, and slice the entire set of genes into groups by the quantiles of their $\widehat{\sigma}_g$ (containing equal numbers of genes for each group). Third, apply the residual from the normalization model (4) γ_{ijkgr} to the gene-specific model (5) again, one group of genes a time. Within each group, based on a common pooled-error term, for each specific gene an F-test or t-test is conducted to test the effect of variety-by-gene $(VG)_{kg}$, which is same as testing simple main effect of variety for each specific gene in the group. By this step a list of genes which are differentially expressed across experimental condi-

tions can be generated. To deal with the multiple testing problems, the Bonferroni rule is used to get the adjusted p-value for each gene (see Appendix A to refer the SAS code for this procedure).

Some alternatives to the grouping method listed previously may be considered as well. Referring to the previous method as the residual variance method, then what follows will be defined as the expression variance method. First, fit the data to the normalization model (4) and get the residuals γ_{ijkgr} which can be considered as normalized intensity observations. Then compute the standard deviation of γ_{ijkgr} for each gene, say $\widehat{SD}_g$. Next, just as the second step of the residual variance method, sort and slice the entire set of genes by quantiles of their $\widehat{SD}_g$. The third step is the same as the third step of the residual variance method, fit the model by gene groups. Since the expression variance method is less computationally involved than the residual variance method, we may prefer it provided the results of the two method are similar. These two approaches will be compared in the next section.

We will also compare our approaches to the gene-specific method and the expression mean method based on the mixed ANOVA model. The expression mean method is similar to Jain's LPE method. For the mixed model, the expression mean method is very similar to the expression variance method except that we compute the mean of γ_{ijkgr} (from model 4) for each gene, which is also the estimated gene effect $\widehat{G}_g$, and then sort and slice genes by the quantiles of $\widehat{G}_g$.

In our approaches, we slice and group the genes by quantiles of their standard deviations or intensity mean. The way of slicing and grouping is arbitrary, but the purpose of doing so is try to pool the genes who share the similar variance estimates. In the discussion section other methods will be proposed for slicing and grouping

genes.

3 DATA ANALYSIS

3.1 Yeast Data Background

The data is from the study of the Saccharomyces cerevisiae swi/snf mutation of Sudarsanam et al. (2000) [14] which investigates mutants deleted for a gene encoding on conserved (Snf2) or on unconserved (Swil) component, each in either rich or minimal media. The data are available at *http://genome-www.stanford.edu/swisnf* as ScanAlyze files [15].

The experimental design is "Reference" design [16](see Figure 1). The same wild-type strain is used as reference sample in all twelve arrays and is labeled with Cy5 (red) in channel 2, while the experimental strains (snf2-rich, snf2-mini, swi1-rich, and swi1-mini) are labeled with Cy3 (green) in channel 1. Each array was spotted with same set of 6,917 genes and no replication of spots within an array. The replication was achieved by using three arrays to study the two samples.

3.2 Mixed ANOVA Model

$$\text{Normalization model: } y_{gij} = \mu + T_i + A_j + (TA)_{ij} + \gamma_{gij} \tag{6}$$

$$\text{Gene-specific model: } \gamma_{gij} = G_g + (GT)_{gi} + (GA)_{gj} + \varepsilon_{gij} \tag{7}$$

Let y_{gij} be the base-2 logarithm of the background-corrected measurement from gene g $(g = 1, ..., 6917)$, treatment $i(i = 1, ..., 5)$, and array $j(j = 1, ..., 12)$. "Treat-

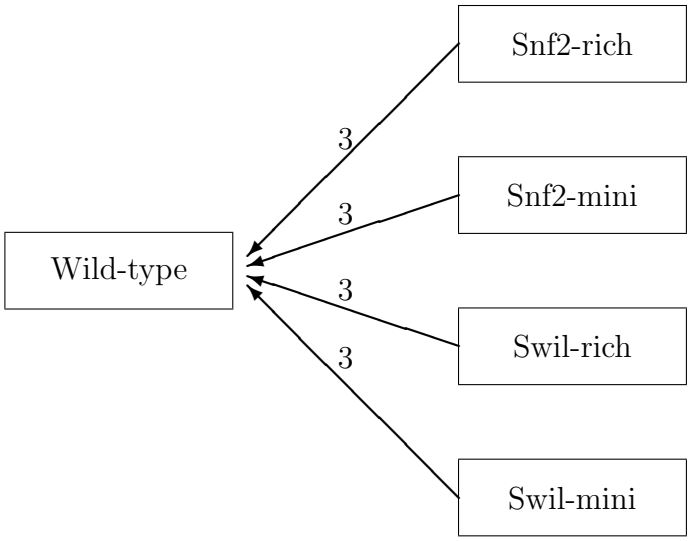

Figure 1: Experimental Design

ment" here signifies the type of cDNA samples (snf2-rich, snf2-mini, swi1-rich, swi1-mini, and wild-type). No dye effect was included in these models because in this experiment wild-type was always labeled with Cy5 and therefore the treatment effect T was already accounting for differences between dyes. In particular, the effects A_j, $(TA)_{ij}$, γ_{gij}, $(GA)_{gi}$ and ε_{gij} are all assumed to be normally distributed random variables with zero means and variance components σ^2_A,σ^2_{TA},σ^2_{GA},σ^2_γ,σ^2_ε, respectively. These random effects are assumed to be independent both across their indices and with each other. The remaining terms in the models are assumed to be fixed effects. Restricted maximum likelihood (REML) is used to estimate the variance components as well as produces estimates of all effects in the model along with appropriate standard errors.

3.3 Results

We applied this data set to our approaches as well as the gene-specific method and the expression mean method and compared these 4 methods.

Table 2 compares the numbers of significant genes, according to a 0.05 significance level, found by the four methods. First, the number of significant genes claimed by the gene-specific method is much less than the other methods. This is consistent with our suspicion that by pooling the information about genes leads to potentially more powerful F tests. And among the three pooling-gene methods, the residual variance method claims much more genes as significant than the other two methods even though they each use the same number of groups. the residual variance method groups genes by their estimated errors directly and leads to homogeneous variances inside each group while for the expression mean method and the expression variance method the homogeneity inside the group is not guaranteed. Next, comparing

the gene-specific method and the residual variance method, the genes claimed as significant by the gene-specific method are all identified by the residual variance method, but there are 1181 more genes claimed by the residual variance method. The two methods agree on 79.7%. Third, The results from the expression mean method and the residual variance method only agree on 82.4%. 718 more genes are claimed as significant by the residual variance method than the expression mean method. Fourth, the expression variance method detects 650 less genes than the residual variance method and the two methods agree on 83.8%. Not as we expected before, the expression mean method may not be a proper alternative to the residual variance method. Finally, the expression mean method and the expression variance method agree on 91.1%.

In this experiment, the gene-specific method failed to evaluate 474 genes due to the lack of replications within some individual genes. In table 2 they are taken as nonsignificant, but in table 3 those 474 genes are excluded and only those genes that have been evaluated by all four methods have been tabulated. Comparing the two tables we find that 46 genes among those 474 genes are detected as significance by the expression mean method and 42 genes by the expression variance method and 50 genes by the residual variance method. Thus, another advantage of pooling the information about genes is the ability to get around the problem of insufficient observations per gene.

Figure 3 shows a plot between the negative base-10 log of the adjusted p-value computed by the residual variance method and maximum base-2 log fold change among 5 treatments which is $Max\ \{|\gamma_{gi_1.} - \gamma_{gi_2.}|\}$ ($i_1 \neq i_2$, i=1,...5). The horizontal line represents the 0.05 significance level and vertical line the 2 fold-change. In order to make the figure readable, we treat those p-value less than 1×10^{-10} as 1×10^{-10}

(condensing some spots at the top of the graph). Notice that at the lower right section of the graph, many genes with high fold-change fail to be identified as significant, while at the upper left section many of low fold-change genes can be detected. This situation shows differential-expression discovery based on fold-change alone is misleading. Looking at Figure 4, we can see more genes rise above the horizontal line in the residual variance method, showing the residual variance method is more sensitive in detecting differential expression. The same result is noted in the comparisons for the gene-specific method vs the expression mean method and the gene-specific method vs the expression variance method (see Figure 5 and Figure 6). Also, the similar result is shown in Figure 7 and Figure 8 when comparing the expression mean method to the residual variance method and the expression variance method to the residual variance method respectively.

Genes with high fold-change often attract the greatest interest from biologists. At the right side of two fold-change vertical line in Figure 4, 696 genes are detected by the residual variance method, 558 of which are not declared significant by the gene-specific method. Of the 138 genes identified by the gene-specific method, all of them are also identified by the residual variance method (also see Table 4). In this experiment, for those genes whose maximum fold-change are greater than 2, the residual variance method identifies 27.5% more genes than the gene-specific method, 7.3% more than the expression mean method and 10.0% more than the expression variance method. Genes with lower fold-change may often be less interesting to the biologists. From Figure 4 , at low fold-change (less than 0.5), the residual variance method and the gene-specific method are fairly consistent agreeing on 90.9% of the genes. Five of 1181 genes at this low fold change area are claimed as significant by the gene-specific method, while 106 of them are claimed by the residual variance method(see table 5). At this low fold change area, the residual variance method

claims only 9.1% more genes than the gene-specific method. It seems the residual variance method is more sensitive for detecting high fold change genes than low fold change genes. For low fold change genes, the residual variance method identifies 8.8% more genes than the expression mean method and 7.9% more genes than the expression variance method.

4 DISCUSSION

Variance components in microarray experiments display varying degrees of heterogeneity, across experiments, across variance components and across gens within a variance component [17]. Assumptions of variance heterogeneity lead to the use of individual gene specific tests, but these tests often suffer from low power due to small degrees of freedom. On the other hand, the assumption of common variance leads to powerful tests, but at risk of bias in the event that the common variance assumption is not true. The residual variance method finds a middle ground between the gene specific method and the common variance method. This approach locally combines information across a group of genes who have similar gene specific variance estimates, which gains power by utilizing more information for testing but avoids bias by reaching a relatively homogenous variance within each group.

While the method proposed to slice and group genes is based on quantiles creating groups with equal numbers of genes, it is not the only method for grouping. Another approach is equal-space slice method, which slices the genes into equal intervals of the estimated values. Figure 9 shows the distribution of gene-specific estimated error in this microarray experiment. Here, equally spaced vertical slices correspond to the quantile slice method. Equally spaced horizontal slices correspond to the equal-space slice method. In this experiment, the variances of genes in groups at the low quantiles seems homogenous by the quantile slice method, but at high

quantiles it does not appear to be true. The equal-space slice method appears to do a better job of avoiding the problem of non-homogenous variance at high quantiles. However, a large amount of genes are condensed in the same group at low quantiles, with only a few genes per group at larger quantiles. When the number of groups is 200, the equal-space slice method identifies 45 more genes than the quantile slice method. It takes more than 6 hours to compute by the equal-space slice method while only 9 minutes by the quantile slice method, and it is because of the large amount of observations at lower quantiles groups by equal-space slice method.

When the number of groups, N, increase and each group contains less genes, our approach may suffers less from the bias of non-homogeneity inside each group (checking homogeneity for each group needs to be developed in the future) but will lose power and finally becomes the gene-specific method (N= number of genes in the experiment). On the other hand, if the N is small, then our approach will gain power but at risk of bias when the variance in some groups are not homogenous. It would seem there is an optional balance to be achieved, which is the goal of future investigation. To achieve this, simulated data will be generated and analyzed to evaluate the method's ability to create homogenous groups and compare its power and false discovery rates to other methods.

Table 2: Pairwise comparisons of four methods. 'yes' means the estimate of differential expression was statistically significant across five treatments; 'no' means it was not.

		gene-specific method		expression mean method		expression variance method		residual variance method	
		no	yes	no	yes	no	yes	no	yes
gene-specific method	no	6578		5887	691	5826	752	5201	1377
	yes		196	32	164	25	171	0	196
expression mean method	no			5919		5584	335	4964	955
	yes				855	267	588	237	618
expression variance method	no					5851		4977	874
	yes						923	224	699
residual variance method	no							5201	
	yes								1573

Table 3: Pairwise comparisons of four methods without the 474 genes not evaluated by gene-specific method. 'yes' means the estimate of differential expression was statistically significant across five treatments; 'no' means it was not.

		gene-specific method		expression mean method		expression variance method		residual variance method	
		no	yes	no	yes	no	yes	no	yes
gene-specific method	no	6104		5459	645	5394	710	4777	1327
	yes		196	32	164	25	171	0	196
expression mean method	no			5491		5175	316	4561	930
	yes				809	244	565	216	593
expression variance method	no					5419		4571	848
	yes						881	206	675
residual variance method	no							4777	
	yes								1423

Table 4: Pairwise comparisons of four methods for those genes whose fold-change ≥ 2. 'yes' means the estimate of differential expression was statistically significant across five treatments; 'no' means it was not.

		gene-specific method		expression mean method		expression variance method		residual variance method	
		no	yes	no	yes	no	yes	no	yes
gene-specific method	no	1892		1476	416	1525	367	1334	558
	yes		138	6	132	11	127	0	138
expression mean method	no			1482		1389	93	1197	285
	yes				548	147	401	137	411
expression variance method	no					1536		1231	305
	yes						494	103	391
residual variance method	no							1334	
	yes								696

Table 5: Pairwise comparisons of four methods for those genes whose log2(fold-change) ≤ 0.5. 'yes' means the estimate of differential expression was statistically significant across five treatments; 'no' means it was not.

		gene-specific method		expression mean method		expression variance method		residual variance method	
		no	yes	no	yes	no	yes	no	yes
gene-specific method	no	1160		1154	6	1142	18	1054	106
	yes		5	2	3	4	1	0	5
expression mean method	no			1156		1141	15	1051	105
	yes				9	5	4	3	6
expression variance method	no					1146		1044	102
	yes						19	10	9
residual variance method	no							1054	
	yes								111

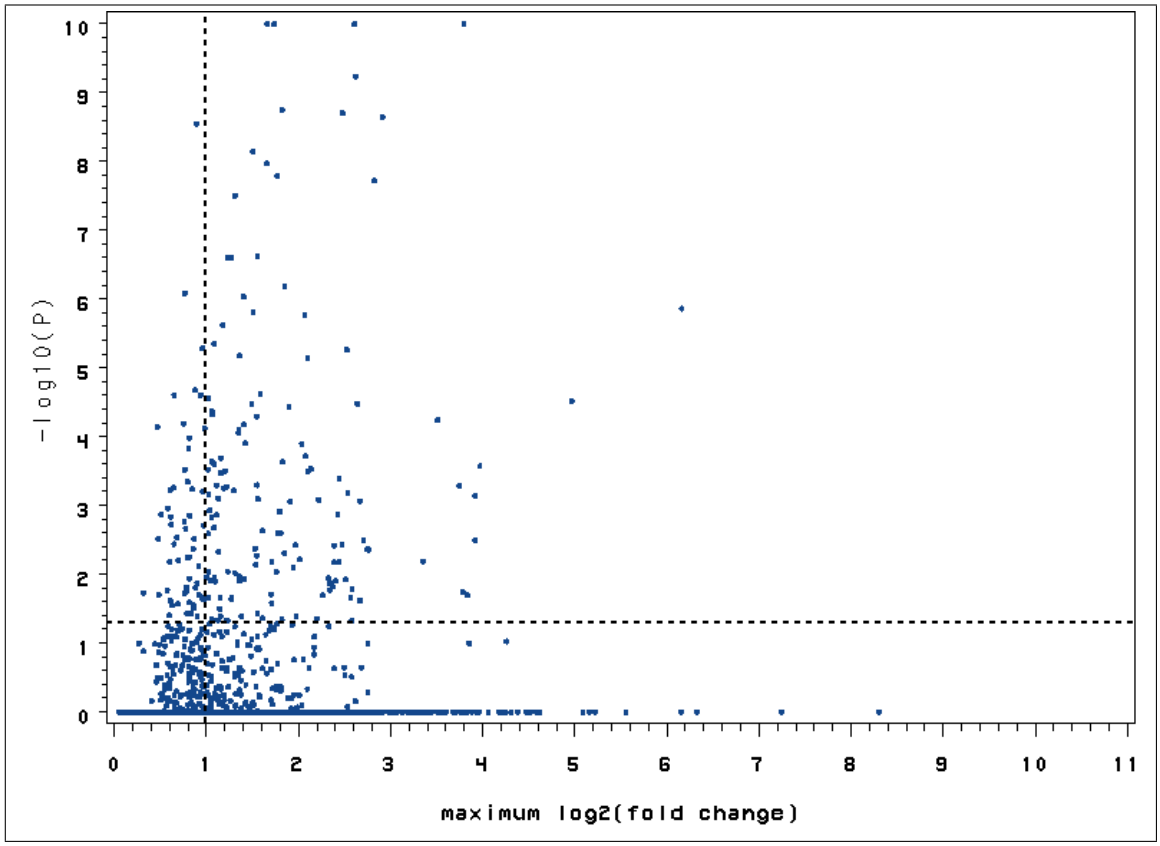

Figure 2: Gene significance results by Gene specific method.

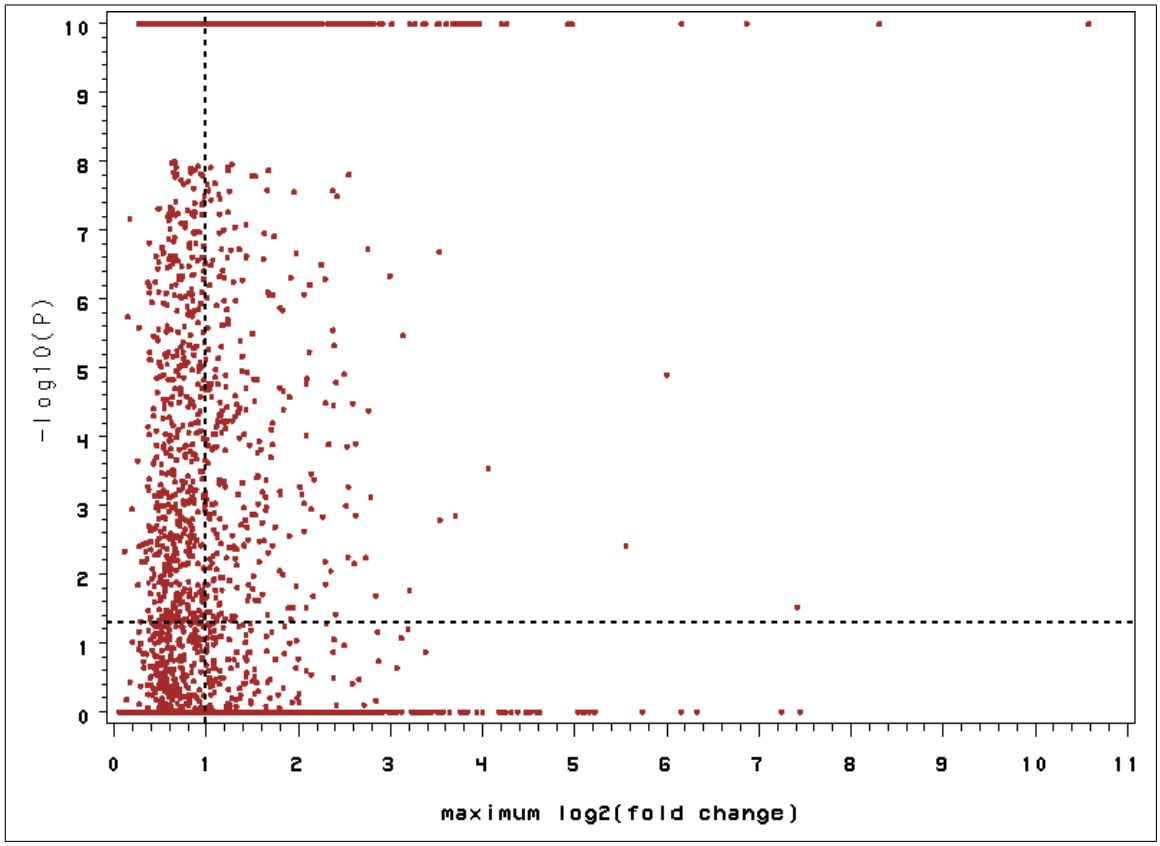

Figure 3: Gene significance results by the residual variance method.

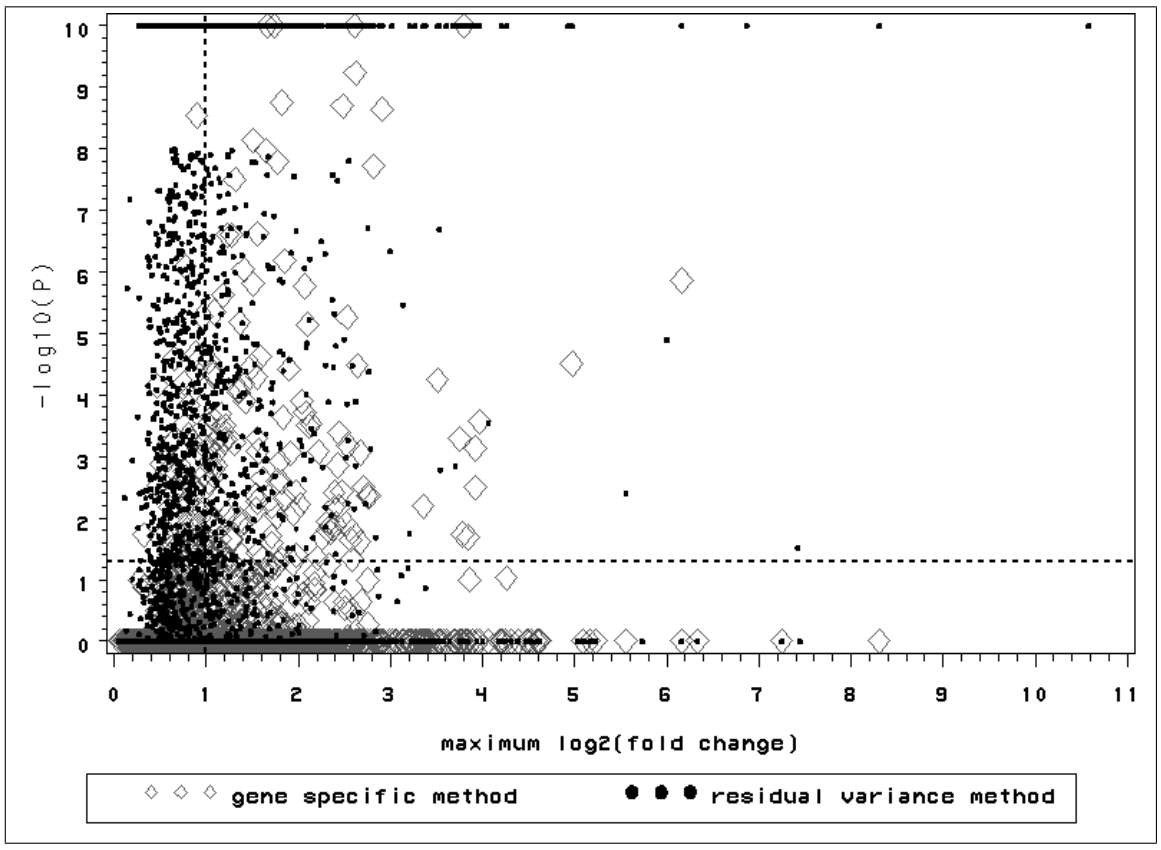

Figure 4: Gene significance results comparison for the gene-specific method and the residual variance method.

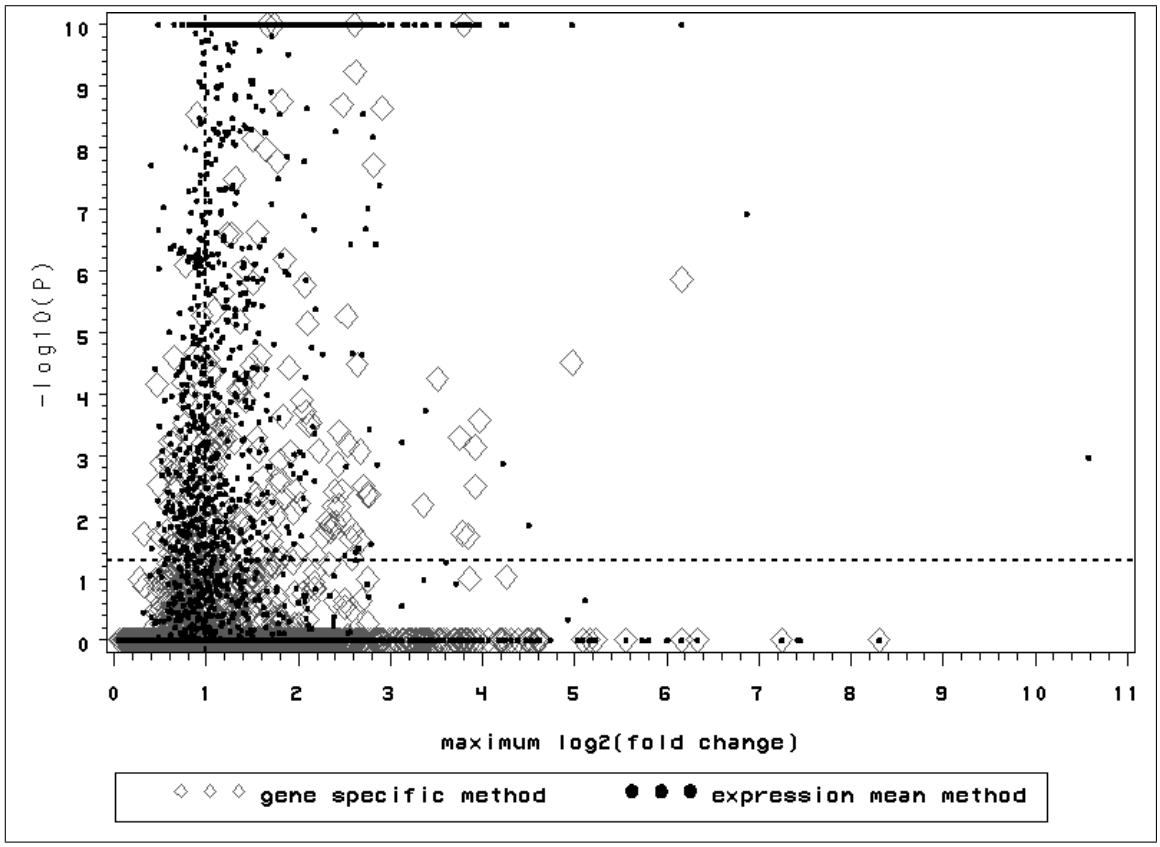

Figure 5: Gene significance results comparison for the gene-specific method and the expression mean method .

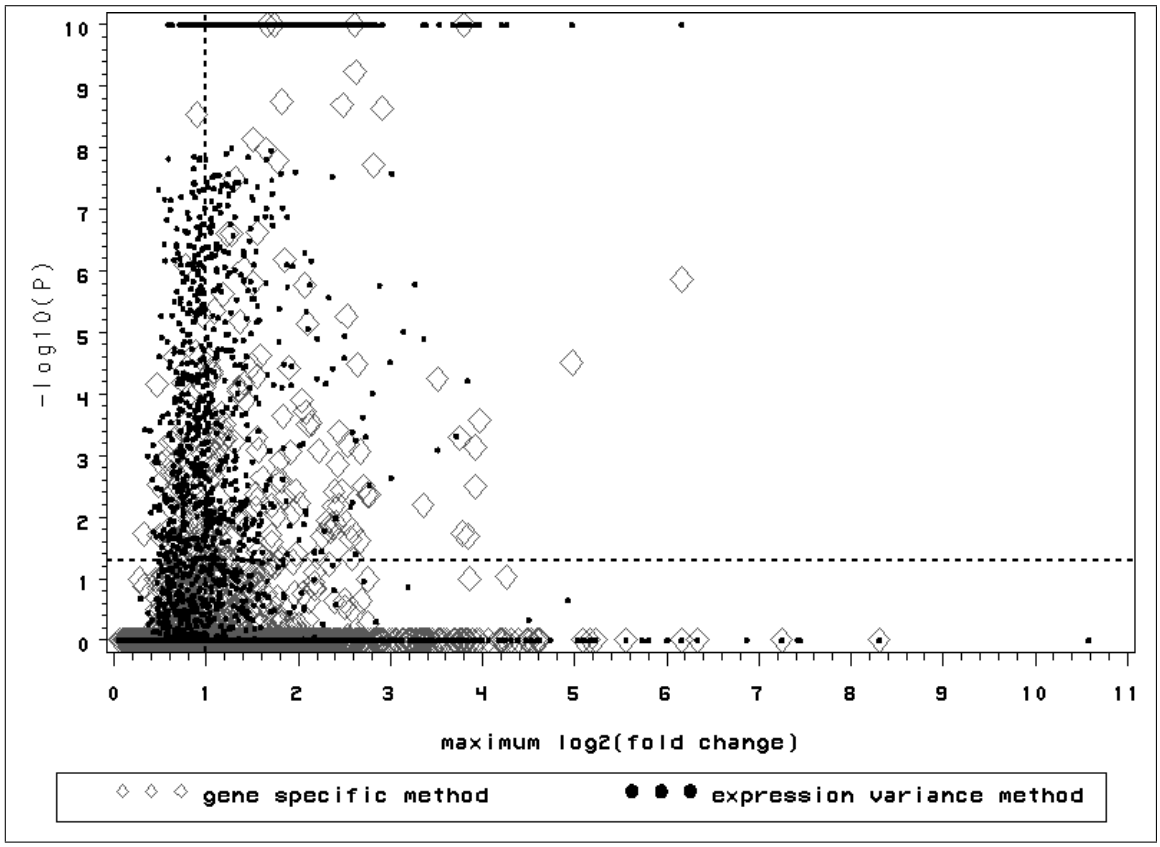

Figure 6: Gene significance results comparison for the gene-specific method and the expression variance method .

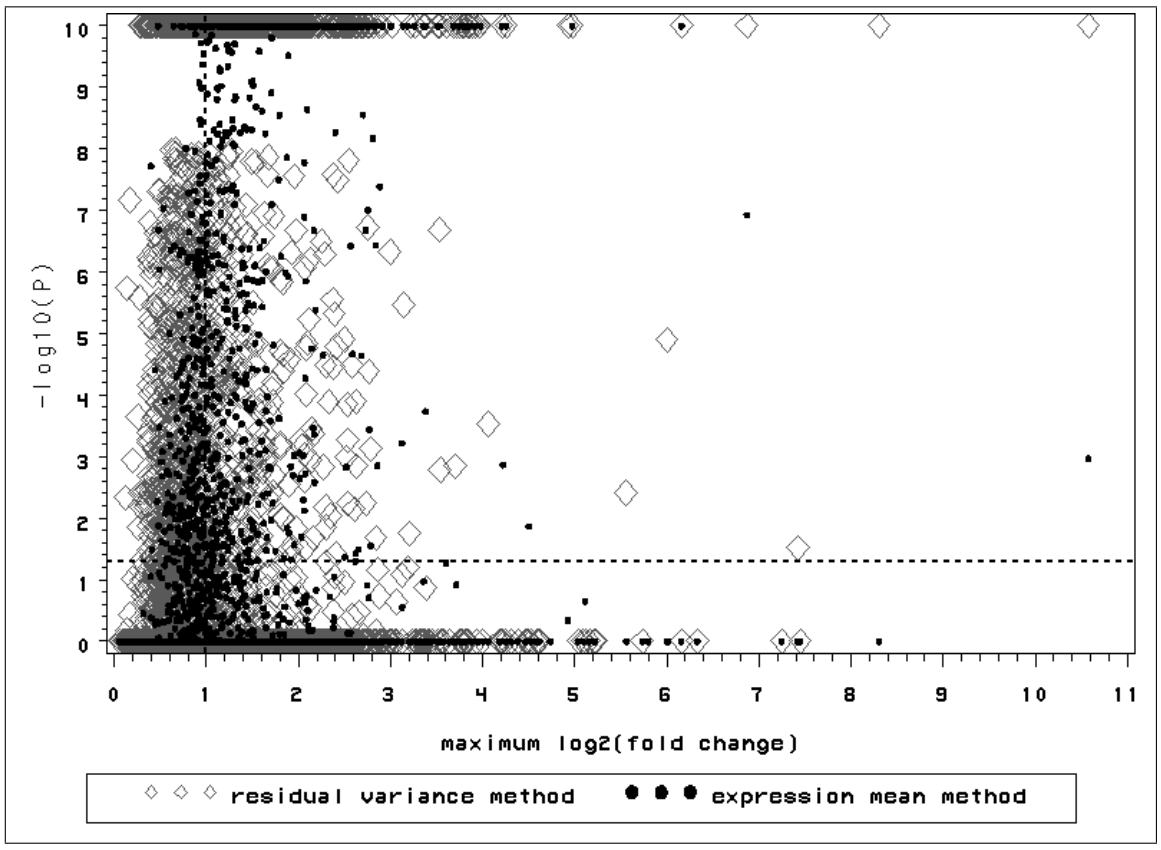

Figure 7: Gene significance results comparison for the expression mean method and the residual variance method .

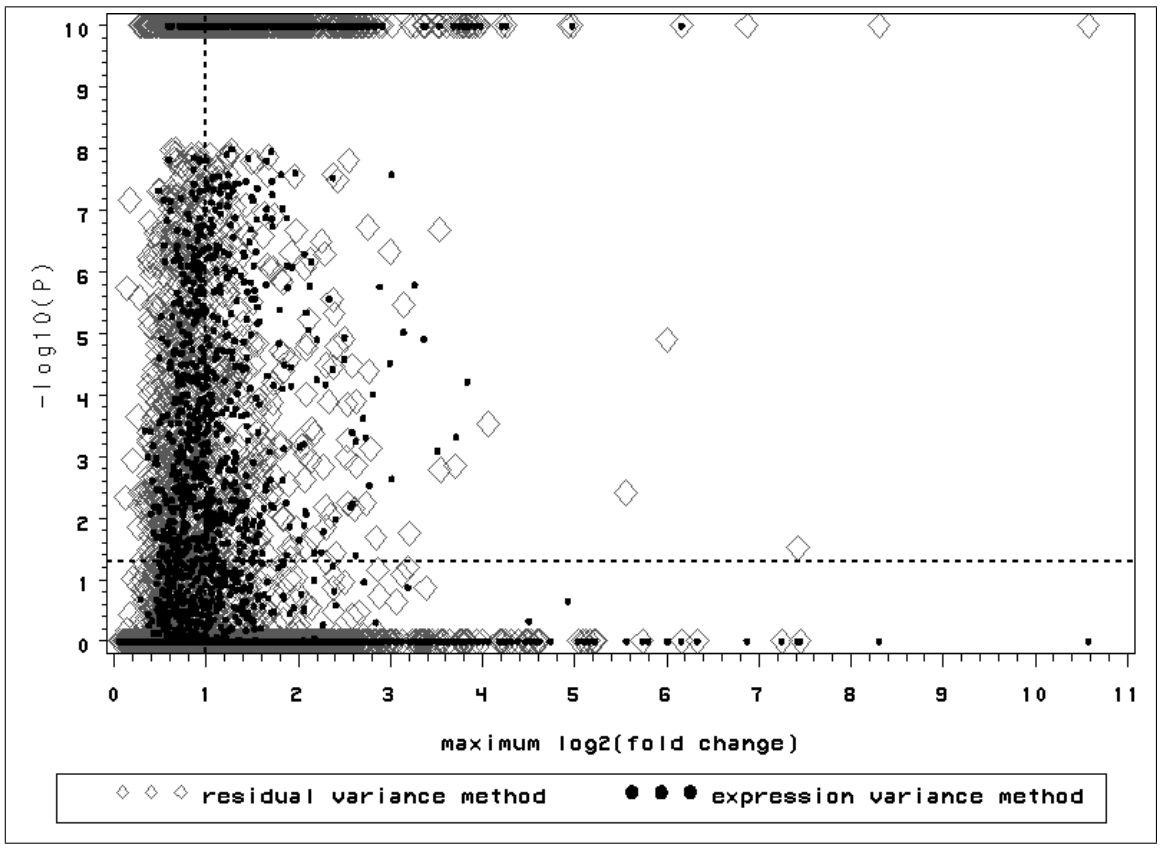

Figure 8: Gene significance results comparison for the expression variance method and the residual variance method .

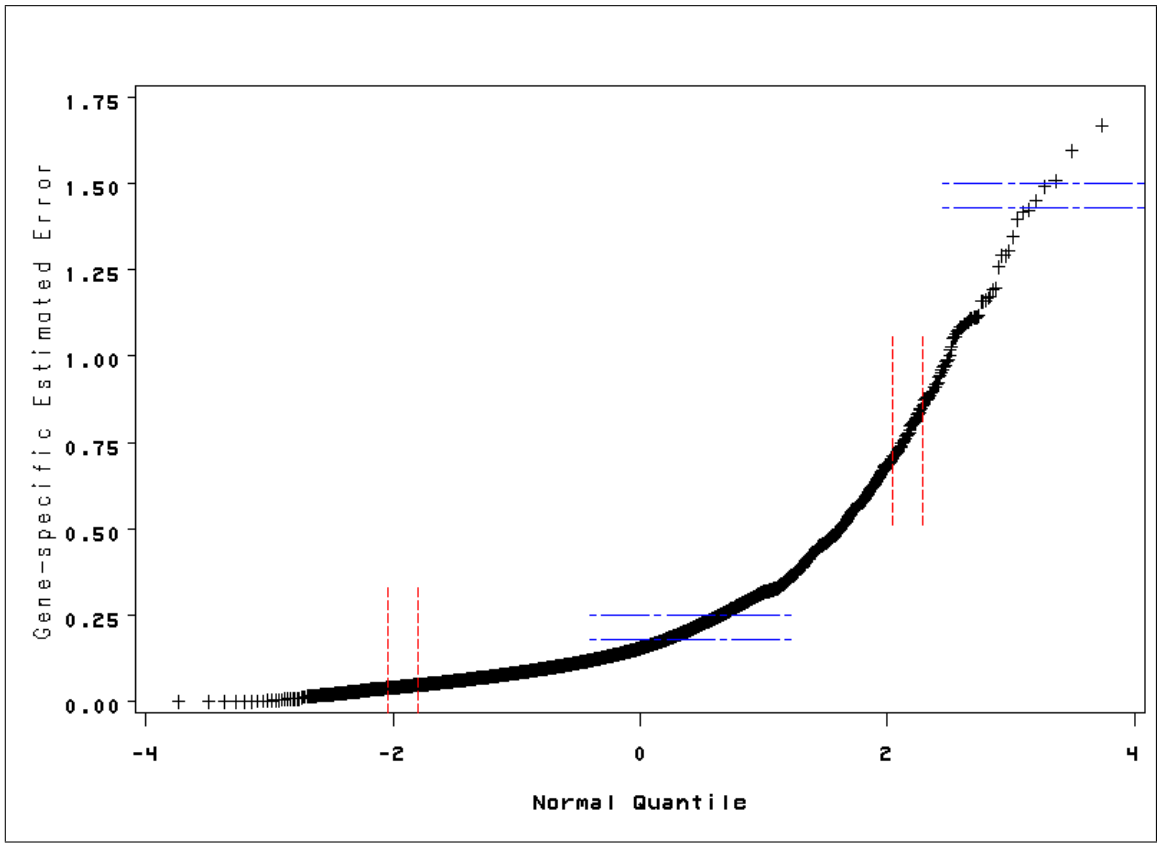

Figure 9: The distribution of $\widehat{\sigma}_g$.

REFERENCES

[1] John Quackenbush (2001). Computational Analysis of Microarray. *Nature reviews/genetics* 2:418-427.

[2] Sandrine Dudoit, Yee Hwa Yang, Matthew J. Callow, and Terence P.Speed (2002). Statistical methods for identifying differentially expressed genes in rplicated cDNA microarray experiments. *Statistic Sinica* 12:111-139.

[3] Giovanni Parmigiani, Elizabeth S.Garrett, Rafael A.Irizarry, Scott L.zeger (2003) "The Analysis of Gene Expression Data: Methods and Software".

[4] Yee Hwa Yang and Terry Speed (2002). Design issues for cDNA microarray experiments. *Nature reviews/genetics* 3:579-588.

[5] Robert Nadon and Jennifer Shoemaker (2002). Statistical issues with microarrays: processing and analysis. *TRENDS in Genetics* 18(5):265-71.

[6] Baldi, P & Long, A.D. (2001).A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17:509-19.

[7] Xiangqin Cui and Gary A.Churchill (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* 4:210.

[8] Kerr M.K., Martin M., and Churchill G.A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 7:819-37.

[9] Wolfinger R.D., Gisbon G., Wolfinger E.D., Bennett L., Hamadeh H,Bushel P,Afshari C, Paules RS (2001). Assessing gene significance form cDNA microarray expression data via mixed models. *J Comput Biol* 8:625-637.

[10] Gordon K.Smyth,Yee Hwa Yang and Terry Speed (2002). Statistical issues in cDNA microarray data analysis. *Methods Mol Biol* 224:111-36.

[11] Kerr M.K. (2003).Linear Models for Microarray Data Analysis: Hidden Similarities and Differences. *UW Biostatistics Working Paper Series Year 2003 Paper 190 http://www.bepress.com/uwbiostat/paper190.*

[12] Nitin Jain, Jayant Thatte, Thomas Braciale, Klaus Ley, michael O'Connel and Jae K.Lee (2003). Local-pooled-error test for identifying differentially expressed genes with a small number fo replicated microarrays. *Bioinformatics* 19:1945-1951.

[13] M.Kathleen Kerr, Cynthia A.Afshari, Lee Bennett, Pierre Bushel,Jeanelle Martinez, Nigel J.Walker and Gary A.Churchill (2002). Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinia* 12:203-217.

[14] Sudarsanam P.,Vishwanath R.I.,Brown P.O., and Winston F.,(2000). Whole-genome expression analysis of snf/swi mutants in Saccaromyces cerevisiae. *Proc. Natl. Acad. Sci.* 97:3364-9.

[15] Eisen M., Spellman P.T.,Brown P.O.,and Botstein D.(1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* 95:14863-8.

[16] M.Kathleen Kerr,Gary A.Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics*. To appear in *www.jax.org/research/churchill/pub/index.html*.

[17] Cui,X, Kerr, M.K, & Churchill, G.A. (2003).Improved Statistical Tests for Differential Gene Expression by Shrinking Variance Components. To appear in *http://www.jax.org/staff/churchill/labsite/*.

APPENDIX

Appendix A. Example SAS code for the residual variance method

```
/*---read in the data; assumes files are stored as text files
     with names sudarsanam1.txt - sudarsanam12.txt; change the
     datafile= pathname below to match your directory---*/

libname d '';
%macro readdata;

proc datasets library=d;
   delete s2;
run; quit;

%do a = 1 %to 12;

    data d.S ;
     %let _EFIERR_ = 0;
     /* set the ERROR detection macro variable */
     infile "sudarsanam&a..txt" delimiter='09'x
             MISSOVER DSD lrecl=32767 firstobs=2 ;
     format NAME $13. ;           format TYPE $7. ;
     format GENE $13. ;           format CH1I best12. ;
     format CH1B best12. ;        format CH1D best12. ;
     format CH2I best12. ;        format CH2B best12. ;
     format CH2D best12. ;        format CH2IN best12. ;
     format CH2BN best12. ;       format CH2DN best12. ;
     format RAT2 best12. ;        format RAT1N best12. ;
     format RAT2N best12. ;       format MRAT best12. ;
     format REGR best12. ;        format CORR best12. ;
     format FLAG best12. ;        format CAT $1. ;
     format DESC $1. ;            format CH1AB best12. ;
     format CH2AB best12. ;       format SPIX best12. ;
     format BGPIX best12. ;       format LFRAT best12. ;
     format CH1GTB1 best12. ;     format CH2GTB1 best12. ;
     format CH1GTB2 best12. ;     format CH2GTB2 best12. ;
     format BLAST $1. ;           informat NAME $13. ;
     informat TYPE $7. ;          informat GENE $13. ;
     informat CH1I best32. ;      informat CH1B best32. ;
     informat CH1D best32. ;      informat CH2I best32. ;
     informat CH2B best32. ;      informat CH2D best32. ;
```

```
      informat CH2IN best32. ;     informat CH2BN best32. ;
      informat CH2DN best32. ;     informat RAT2 best32. ;
      informat RAT1N best32. ;     informat RAT2N best32. ;
      informat MRAT best32. ;      informat REGR best32. ;
      informat CORR best32. ;      informat FLAG best32. ;
      informat CAT $1. ;           informat DESC $1. ;
      informat CH1AB best32. ;     informat CH2AB best32. ;
      informat SPIX best32. ;      informat BGPIX best32. ;
      informat LFRAT best32. ;     informat CH1GTB1 best32. ;
      informat CH2GTB1 best32. ;   informat CH1GTB2 best32. ;
      informat CH2GTB2 best32. ;   informat BLAST $1. ;
      input  NAME $ TYPE $ GENE $ CH1I CH1B CH1D CH2I CH2B
             CH2D CH2IN CH2BN CH2DN RAT2 RAT1N RAT2N MRAT
             REGR CORR FLAG CAT $ ESC $ CH1AB CH2AB SPIX
             BGPIX  LFRAT CH1GTB1 CH2GTB1 CH1GTB2 CH2GTB2
             BLAST $;
      if _ERROR_ then call symput('_EFIERR_',1);
       /* set ERROR detection macro variable */
run;

   data d.s;
      format gene $13. name $13.;
      set d.s;
      name = upcase(name);
      type = upcase(type);
      gene = upcase(gene);
      array = &a;
      if (name=" ") then name = type;
      if (gene=" ") then gene = name;
      spot = _n_;
      if (flag=0) then do;
         if (array <= 3) then strain = "snf2rich";
         else if (array <= 6) then strain = "snf2mini";
         else if (array <= 9) then strain = "swi1rich";
         else if (array <= 12) then strain = "swi1mini";
         diff = ch1i-ch1b;
         if (diff > 0) then logi = log2(diff);
         else logi = .;
         output;
         strain = "wildtype";
         diff = ch2i-ch2b;
         if (diff > 0) then logi = log2(diff);
         else logi = .;
         output;
```

```
      end;
      keep array gene name spot strain logi;
   run;

   proc append base=d.s2 data=d.s;
   run;

%end;
%mend;

%readdata
run;

/*---Normalization Model---*/
proc mixed data=d.s2 covtest cl;
   class array strain;
   model logi = strain /
         outp=d.sudarp(keep=array gene name spot strain resid);
   random array array*strain;
   lsmeans strain / diff cl;
run;

/*---removes some genes that slow the analysis---*/

data d.sudarp;
   set d.sudarp;
   where gene not in ('EMPTY', 'NORF');
run;

proc sort data=d.sudarp;
   by gene array spot;
run;

ods listing close; run;

data d.sudarp;
   set d.sudarp;
   rename resid=residual;
run;

/*---Gene Specific Models for each gene using residuals---*/

 proc mixed data=d.sudarp;
   by gene;
```

```
   class array spot strain;
   model resid = strain / outp=d.sudarr;
   random spot(array);
   lsmeans strain / diff;
   ods output covparms=d.sudarc tests3=d.sudart
      diffs=d.sudard(keep=gene effect strain df probt);
run;

/*----Group gene by their residual variance----*/

%macro groupdata(dataset=, option=,var=resid_mean,name=);

ods listing close;
proc means data=d.&dataset &option;
class gene;
var resid;
ods output summary=one;
run;

proc univariate data=one; var &var;
output out=group
       pctlpre=pctl_ pctlpts=1 to 100 by 1;
run;

proc transpose data=group out=group_transposed;
run;

data _null_;
set group_transposed;
call symput(_name_,col1);
run;

data d.&name;
  set one;
   if &var <= &pctl_1 then &name=1;
   if &var > &pctl_99 then &name=100;
   else %do a=1 %to 98;
          %let i=%eval(&a+1);
           if &&pctl_&a < &var <= &&pctl_&i then &name=&i;
        %end;
   keep gene &name;

 proc sort; by gene; run;
```

```
ods listing;

%mend groupdata;

%groupdata(dataset=sudarr, option=std,var=resid_stddev,name=groupbystd);

/*Merge the group information to Sudarp*/

data d.group;
    merge d.groupbystd
          d.sudarp;
    by gene;
run;

/*Apply Gene-specific model to a group of gene a time*/

%macro secondmodel(method);

proc sort data=d.group; by &method; run;

ods listing close;

proc mixed data=d.group;
   where &method ne .;
   by &method;
   class gene array spot strain;
   model resid = gene|strain;
   random spot(array);
   lsmeans strain*gene /slice=gene;
   ods output slices=one;
run;

ods listing; run;

/*Bonferroni procedure for adjusted P-value of each gene */

 data one;
 set one;
 rename probf=raw_p;
 run;

 proc multtest pdata=one bon out=outp;run;

 data outp; set outp; keep effect gene &method raw_p bon_p; run;
```

```
 proc sort data=outp out=d.multest_&method; by bon_p; run;

%mend;

%secondmodel(groupbystd);
```