# Abstract

**CHU, TZU-MING.** STATISTICAL NONPARAMETRIC AND LINEAR MIXED MODEL ANALYSES OF OLIGONUCLEOTIDE DNA CHIPS DATA. (Under the directions of Professors Bruce Weir and Russ Wolfinger)

Scientists investigate the dynamic relationships among genes and the associated phenotypes through gene expression array (microarray) studies. An essential step in the tasks is to identify the genes that actually interact with the phenotypic outcomes. This dissertation focuses on the selection of informative genes with statistical approaches.

In chapter one, a nonparametric approach that combines the Bootstrap resampling method and the Kruskal-Wallis test (the BKW test) for gene selection is discussed. Principal component and clustering analyses are performed for disease multi-type classification. In chapter two, steps are outlined and described for a statistically rigorous approach to analyzing probe-level GeneChip$^{TM}$ data. The approach employs classical linear mixed models and operates on a gene-by-gene basis. The method can accommodate complex experiments involving many kinds of treatments and can test for their effects at the probe level. Furthermore, mismatch probe data can be incorporated in

different ways or ignored altogether. In chapter three, an empirical comparison of the linear mixed model and the Li-Wong's multiplicative model is presented for a real data set, and it is found that the models perform quite similarly across most genes, but with some interesting and important distinctions. Results are also presented from a simulation study designed to assess inferential properties of the models, and a modified test statistic is presented for the Li-Wong model that provides an improvement in Type I error control.

The analysis approaches discussed here are applied to the data from oligonucleotide DNA chips. However, the concepts are also applicable to the data from cDNA microarrays.

# STATISTICAL NONPARAMETRIC AND LINEAR MIXED MODEL ANALYSES OF OLIGONUCLEOTIDE DNA CHIPS DATA

by

**TZU-MING CHU**

A dissertation submitted to the Graduate Faculty of
NORTH CAROLINA STATE UNIVERSITY
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

**STATISTICS**

RALEIGH
2002

**APPROVED BY:**

| | |
|---|---|
| Bruce S. Weir<br>Co-chair of Advisory Committee | Russell D. Wolfinger<br>Co-chair of Advisory Committee |
| Roger L. Berger | Jeffrey L. Thorne |

Anastasios Tsiatis

*To my parents and my wife*

# Biography

Tzu-Ming Chu was born in Taipei, Taiwan, to parents Peng-Cheng Chu and Su-Mei Tsao on May 31, 1968. He received a B.S. in Mathematics from National Tsing Hua University, Taiwan, in 1990 and a Master in Statistics from Michigan State University in 1996. Since August 1996, he has studied for the doctoral degree in the Department of Statistics with concentration in Bioinformatics at North Carolina State University, under the supervision of Drs. Bruce Weir and Russell Wolfinger.

# Acknowledgements

My gratitude goes to my co-advisors, Drs. Bruce Weir and Russell Wolfinger, for their guidance, support, and mentoring in the beginning of my research journey. Dr. Weir's enthusiasm in Bioinformatics inspired me to enter the field of genomic research. Dr. Wolfinger's thorough insights in our discussions not only provided direction for my dissertation work but also broadened my knowledge in both of the fields of Statistics and Bioinformatics. It is truly my honor and pleasure to work with them. I would also like to express my appreciation to my dissertation committee, Drs. Roger Berger, Jeffery Throne, and Anastasios Tsiatis, for their helpful suggestions.

There are many people who have helped me during my dissertation work. Dr. John Brocklebank, the former director of the Data Mining group at SAS Institute Inc., gave me an internship and encouraged me to work on microarray data analysis in 2000. Wendy Czika, my colleague in Genomics department in SAS Institute Inc., spent many hours helping me debug my C programs. Dr. Taiyeong Lee, my former colleague in the SAS Data Mining group, had many useful discussions with me. I had a very enjoyable time working with them. In

addition, the irradiation GeneChip$^{TM}$ data from Stanford University applied in this dissertation were provided by Virginia Tusher, Gilbert Chu, and Robert Tibshirani. I thank them for allowing me to use their data.

I am grateful to the former and current faculty, students and staff in the Statistics department and Bioinformatics Research Center for their valuable discussions and help. Special recognition goes to Drs. Ying-Hsuen Sun and Pei-Yun Chen, former students in the Forestry and Statistics departments, respectively. Dr. Sun is one of my good friends and was the first person who set up a microarray lab at NCSU. He provided me with my first experience with microarrays. Dr. Chen and I have known each other since we were research assistants in Academia Sinica in Taiwan in 1997. Dr. Chen is often the first audience for my thoughts and ideas and always listens with patience and interest.

Finally, I greatly appreciate my parents and my wife with all my heart. My parents, Peng-Cheng and Su-Mei, always support me unconditionally. My wife, Fang, always understands me, believes in me, and stands by me. Their inspiration and full support have enabled me come this far.

# Contents

# List of Tables

# List of Figures

# Chapter 1

# Nonparametric Filtering of Gene Expression Data for Multi-Type Disease Classification

## 1.1 Introduction

The technology of gene expression arrays, also known as microarrays, was a breakthrough developed at the end of last century that allows researchers to perform a single experiment on thousands of genes simultaneously (Schena et al., 1995; Lockhart et al., 1996). Major applications of expression array technology include monitoring transcriptional activity throughout the cell cycle and finding the signature genes characterizing different cell types. The results of studies using expression array technology enable us to explore the functionalities and interactions of genes on a genome scale. A tremendous amount of data, containing information from thousands of genes, is generated from the expression array studies. To convert the data into statistically significant evidence and then to provide biological meanings are the primary goals of gene expression studies.

Identifying signature genes plays an important role in disease type classification, which in turn is crucial for diagnosis and effective treatment for complex diseases. In general, the first problem encountered is how to extract sufficient information from all the genes in the experiments. In microarray studies, the proportion of genes relevant for the biological event of interest is often low. This leads to issues about setting a good filtering rule in order to select informative genes. Some methods (Golub et al., 1999; Dudoit et al., 2000a) have been discussed for the selection of informative genes. However, discussions about the number of informative genes to be selected have been somewhat arbitrary. Here, we propose a filtering method that combines bootstrap resampling and the Kruskal-Wallis test (the BKW test) to select informative genes without restricting the total number of genes to be selected. The next steps are to group informative genes into clusters, and to classify array samples into different disease types. Clustering analysis is one of the most general approaches for classification. We perform k-means clustering analyses on both the disease samples and the selected genes. Results of clustering analyses by using the informative genes selected from the BKW test as inputs show correct classification rates of at least 92% for the data sets used in this study.

Two data sets, which were collected using Affymetrix Hu6800 GeneChips (Lipshutz et al., 1999; Lockhart et al.,1996; McGall et al., 1996), from Golub et al. (1999) are studied here for classifying different subtypes of acute leukemia. Acute leukemia is usually a fatal disease in which white blood cells may

proliferate in an excessive amount and lose their normal functionality within a short period of time. Acute leukemia can generally be divided into two categories: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Moreover, ALL is divisible into two subtypes, B-lineage ALL and T-lineage ALL (McConkey, 1993). The initial steps for successful and effective treatment for acute leukemia are to differentiate ALL from AML and to recognize the relevant subtypes (Cripe, 1997). Patients with different subtypes of acute leukemia need different therapies (Bishop, 1997; Wrzesien-Kus, 1997). Therefore, accurate classification of leukemia subtypes, or other diseases, using expression array data would be a substantial achievement in biomedical studies.

In our study, the data set with 38 array samples is used as the training data set, and the other set with 34 array samples is used as the test data set. Each of the array samples corresponds to one leukemia patient. Table 1.1 displays the number of each subtype for both data sets. Simulations were conducted to compare Kruskal-Wallis test and its normal-theory competitor, the t-test, considering two subtypes in data. The Type I error and power of the two tests will be discussed. Data with or without outlier are considered in simulations.

## 1.2 Filtering Method

Obtaining a good filtering rule is the first and crucial step prior to further data analyses, such as clustering analysis, for disease type distinction and disease gene(s) investigation. A good filtering rule enables researchers to reduce the

number of genes to be investigated and to increase the accuracy of disease type distinction. For this task, researchers wish to select genes with expression profiles correlated to the disease types as highly as possible. However, they tend to increase the number of selected genes in order to reduce the risk of losing information from potential disease genes. Golub et al. (1999) suggested a t-test-like neighborhood analysis for two-type disease classification. In their study, 1,100 genes were highly correlated with ALL-AML class distinction, and the 50 most closely correlated genes were selected for further classification. Several questions arise regarding the number of genes used for classification analysis, the selection criterion of genes, and whether the selection method is applicable to multi-type disease classification. Will we lose some information by using the 50 most correlated genes for ALL-AML class distinction instead of using all 1,100 significant genes? How many genes should be included for succeeding classification? In addition, is the 50-gene selection rule applicable to other classification for other diseases? Furthermore, the micro-level sensitivity of the GeneChip platform often results in data containing outlying observations due to such factors as dust and tiny scratches. It is also not obvious that the data follow standard parametric assumptions used, for example, with a t-test.

To address these issues, we consider the well-known Kruskal-Wallis test statistic

$$KW_i = \frac{12}{N(N+1)} \sum_{l=1}^{L} n_l \left(\overline{R_{i(l)}} - \frac{N+1}{2}\right)^2$$

Here, $KW_i$ is the Kruskal-Wallis statistic for the $i^{th}$ gene, $L$ is the number of disease subtypes, $\overline{R_{i(l)}}$ is the average rank for the $i^{th}$ gene in the $l^{th}$ subtype, and $n_l$ is the sample size for subtype $l$. The standard, asymptotically valid, $p$-value for this test is computed as

$$p_i = \Pr(\chi^2_{L-1} > KW_i),$$

where $\chi^2_{L-1}$ is a chi-square random variable with $L$-1 degrees of freedom.

To take a more modern approach and adjust for potential small-sample biases, we compute $p$-values based on the bootstrap. Specifically, suppose that

$$\begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1N} \\ Y_{21} & Y_{22} & \dots & Y_{2N} \\ . & . & \dots & . \\ . & . & \dots & . \\ Y_{p1} & Y_{p2} & \dots & Y_{pN} \end{bmatrix}$$

is the original data set in which $Y_{ij}$ is the expression measurement of the $i^{th}$ gene in the $j^{th}$ array, $N$ is the number of arrays, and $p$ is the number of genes. From these data, we generate $K$ bootstrap resamples, and denote the $k^{th}$ resampled data set as

$$\begin{bmatrix} Y_{11}^{(k)} & Y_{12}^{(k)} & \dots & Y_{1N}^{(k)} \\ Y_{21}^{(k)} & Y_{22}^{(k)} & \dots & Y_{2N}^{(k)} \\ . & . & \dots & . \\ . & . & \dots & . \\ Y_{p1}^{(k)} & Y_{p2}^{(k)} & \dots & Y_{pN}^{(k)} \end{bmatrix},$$

where $Y_{ij}^{(k)}$ has been sampled with replacement from $\{Y_{i1}, Y_{i2}, \dots, Y_{iN}\}$. For each resampled set, we rank data across arrays for each gene to give the rank data set,

$$\begin{bmatrix} R_{11}^{(k)} & R_{12}^{(k)} & \dots & R_{1N}^{(k)} \\ R_{21}^{(k)} & R_{22}^{(k)} & \dots & R_{2N}^{(k)} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ R_{p1}^{(k)} & R_{p2}^{(k)} & \dots & R_{pN}^{(k)} \end{bmatrix},$$

where $\{R_{i1}^{(k)}, R_{i2}^{(k)}, ..., R_{iN}^{(k)}\}$ are the ranks of $\{Y_{i1}^{(k)}, Y_{i2}^{(k)}, ..., Y_{iN}^{(k)}\}$ for the $i^{th}$ gene. We then calculate the Kruskal-Wallis statistic for each gene in each resampled data set according to its associated rank data set:

$$KW_i^{(k)} = \frac{12}{N(N+1)} \sum_{l=1}^{L} n_l (\overline{R_{i(l)}^{(k)}} - \frac{N+1}{2})^2$$

Here $KW_i^{(k)}$ is the Kruskal-Wallis statistic for the $i^{th}$ gene in the $k^{th}$ resampled set, $L$ is the number of disease subtypes, $\overline{R_{i(l)}^{(k)}}$ is the average rank for the $i^{th}$ gene in the $l^{th}$ subtype in the $k^{th}$ resample, and $n_l$ is the sample size for subtype $l$.

We can thus simulate from the bootstrap sampling distribution of the Kruskal-Wallis statistic for each gene, and can compare $KW_i$ to this distribution to assess its significance. In particular, the fraction of $KW_i^{(k)}$ greater than $KW_i$ is an approximate $p$-value.

To select a filtered set of genes for clustering, one is faced with the classical multiple testing problem. Many methods for dealing with this are addressed in the literature, ranging from simple Bonferroni adjustments to complicated ones based on bootstrap or permutation resampling; Dudoit et al (2000b) provide a nice review. Westfall et al. (2001) provide some evidence that

the Bonferroni adjustment is sufficient, and so we use it here. Applied to the leukemia data with a familywise Type 1 error rate of 0.05, the bootstrap Kruskal-Wallis test with Bonferroni adjustment selects 291 out of 7129 genes.

Before using this set of genes for clustering analyses, we consider two additional issues. The first is the idea of using some characteristic of the bootstrap distribution itself to assess significance, rather than just using it as a reference distribution for the observed statistic in the original sample. Specifically, consider the collection of the standard Kruskal-Wallis *p*-values for the $k^{tk}$ resampled set, calculated as follows:

$$p_i^{(k)} = \Pr(\chi^2_{L-1} > KW_i^{(k)}),$$

where $\chi^2_{L-1}$ is a chi-square random variable with $L$-1 degrees of freedom. Under the null hypothesis of no differences among the subtypes, we would expect the $p_i^{(k)}$ to follow a uniform distribution. Thus, departures from a uniform could be construed as evidence against the null. For example, a test could be constructed by comparing the $95^{th}$ percentile of the $p_i^{(k)}$ to some cutoff much less than 0.95. We are currently investigating such a testing procedure and preliminary empirical results are very similar to those we present later from the standard method. We do not include any additional details here.

The second issue we address before considering clustering is how the nonparametric Kruskal-Wallis test compares with the parametric *t*-test in our context. This is discussed in the next section.

## 1.3 Simulation Study - Parametric vs. Nonparametric Tests

The intensity measurements of gene expression naturally have a right-skewed distribution of positive measurements with a small proportion in the right tail. Intuitively, the lognormal and gamma distributions are appropriate. In most microarray articles, the expression intensities are log-transformed (base 2) prior to the analysis, and the transformed values are assumed to be normally distributed. An early statistical analysis by Chen et al. (1997) for microarray data proposed that, despite the expression variability from gene to gene across microarrays, the expression measurements of genes have equal coefficients of variation. This motivated Newton et al. (2001) to consider gamma distributions that have a constant shape parameter to govern measurements from different groups of tissue samples. In this section, we select a candidate gene from the leukemia data and fit its expression data by normal, lognormal, and gamma distributions in order to obtain the necessary parameter estimates to conduct the simulations. The comparable parametric test to the Kruskal-Wallis test is the F-test in an ANOVA setting. For simplicity of comparison, the data are simulated under two groups, with or without mean differences, and the equivalent tests for comparison are the Wilcoxon rank sum test and the t-test. The Wilcoxon rank sum test is a version of Kruskal-Walllis test, so the latter term is retained. We are interested in comparing the Type I error and the power between the t-test and the Kruskal-Wallis test under different scenarios as will be explained in more detail in the following sections.

### 1.3.1 Candidate Genes Selection

To obtain better estimates, data from the 47 GeneChips from ALL patients in the training and test data sets are pooled together. The first two moments, sample means and variances, of all genes are calculated, and a gene with sample mean and variance not far from the median among all genes is selected. In order to fit lognormal and gamma distributions with positive support, those genes with any negative measurements will not be selected. A Gaba (Gamma-Aminobutyric Acid) gene, marked HG3255-HT3432 in Affy's Hu6800 GeneChip$^{TM}$, is chosen for simulation studies. The Gaba gene has a sample mean of 375.09 (the 70$^{th}$ percentile) and standard deviation of 149.73 (the 44$^{th}$ percentile). Figures 1.1 and 1.2 show the histograms of sample means and standard deviations. In these figures, there are many genes having negative means due to the algorithm Affymetrix used to measure gene expression. Therefore, it is not possible to directly transform data by taking the logarithm as the standard process for cDNA microarray. (The new Affymetrix algorithm, MAS5.0, avoids this issue of negative expression measurement. (www.affymetrix.com/products/)) We have further and detailed discussion about this problem in Chapter 2.

### 1.3.2 Simulation by the Normal, Lognormal, and Gamma Distributions without Outliers

We fit the 47 observations of the candidate gene to normal, lognormal, and gamma distributions with the SAS procedure CAPABILITY. Figure 1.3 shows

the histogram with three fitted distribution curves. The parameter estimates of the three distributions are listed in the figure legend and summarized in Table 1.2. "Theta" is the threshold parameter for both lognormal and gamma distributions. "Theta = 0" implies positive support for these two distributions. With the parameter estimates, the simulated data are created by the following formulas:

$$Y_n(i, j_i, t) = \hat{\mu} + \hat{\sigma} * Normal_t(0,1) + \delta * I_{\{i=2\}},$$
$$Y_l(i, j_i, t) = \exp\{\hat{\eta} + \hat{\gamma} * Normal_t(0,1)\} + \delta * I_{\{i=2\}},$$
$$Y_g(i, j_i, t) = \hat{\beta} * Gamma_t(\hat{\alpha}, 1) + \delta * I_{\{i=2\}},$$

where $i=1,2$; $j=1,2,\ldots J_i$; and $t=1,2,\ldots T$. $Y$ is the simulated expression intensity of the $j_i^{th}$ chip for the $i^{th}$ subtype disease in the $t^{th}$ time of simulation and the subscript of $Y$ indicates the distribution applied. $Normal_t(0,1)$ and $Gamma_t(\hat{\alpha}, 1)$ represent the random values drawn from normal and gamma distributions in the $t^{th}$ simulation, respectively. $\hat{\mu}$, $\hat{\sigma}$, $\hat{\eta}$, $\hat{\gamma}$, $\hat{\alpha}$, and $\hat{\beta}$ are parameter estimates listed in Table 1.3. $\delta$ represents the artificial mean expression difference between two subtype disease groups. $I_{\{i=2\}}$ is an indicator function. Two scenarios regarding the sample size are considered. We generate the data with the same sample size as in the original training data ($J_1$=27, $J_2$=11) and with a small size ($J_1$=5, $J_2$=5) in the first and the second scenarios, respectively. The mean differences between groups are obtained by setting $\delta$ to a series of values ranging from 0 to 2 standard deviations of original data ($0\hat{\sigma}, 1.1\hat{\sigma}, 1.2\hat{\sigma}, \ldots, 2\hat{\sigma}$). Also, the simulation size T is set to 10,000.

#### 1.3.2.1 Scenario 1 - Same Sample Size as Training Data Set

We consider $J_1 = 27$ and $J_2 = 11$ in the first scenario. Table 1.3 lists the rejection rates (power) of the t- and the Kruskal-Wallis test with nominal 0.05 significance level. The McNemar test is used to assess the testing agreement of the t- and the Kruskal-Wallis tests in the simulation. When the true mean difference is zero, the rejection rates indicate the Type I error. In this scenario, the McNemar's *p*-values indicate the Type I errors are similar for the two tests. This implies the powers of both tests are comparable. As might be expected, the Kruskal-Wallis test has lower power under the normal distribution but higher power under the lognormal and gamma distributions. Figure 1.4 shows power curves of all six test-distribution combinations.

#### 1.3.2.2 Scenario 2 - Small Sample Size

Microarray studies are often based on a small number (less than 10) of arrays because of cost concerns. In the second scenario, we repeat the first scenario except for having only five simulated samples for the two disease subtypes. Table 1.4 lists the rejection rates of the t- and Kruskal-Wallis tests. Examining the rejection rates when no mean difference exists, the t-test controls the nominal .05 Type I error under all three distributions, and the Kruskal-Wallis test has slightly liberal error rate for the three distributions. The powers in this scenario are not exactly comparable since the tests are not based on the same significance level. However, the t-test has better rejection rates when the mean difference is large

($1.9\hat{\sigma}$ and $2\hat{\sigma}$). This implies that, compared to the Kruskal-Wallis test with a small sample size, the t-test is more powerful for more significant mean differences.

### 1.3.3 Simulations with an Outlier Involved

The presence of non-biologically-relevant outliers in raw data is a common issue in microarray experiments (Schadt et al., 2000 and Li and Wong, 2001a). Figure 2.1 in Chapter 2 also shows an example. Here, we arbitrarily make one observation an outlier in the second group with the same sample sizes ($J_1$=27 and $J_2$=11) as the first scenario in the previous simulation to investigate the control of Type I error. The outlier is fixed at $\hat{\mu} + k\hat{\sigma}$ and the parameter k is varied from 3 to 10. Table 1.5 and Figure 1.5 displays the simulation results and the simulated Type I error curves, respectively. The simulated Type I errors for the Kruskal-Wallis test are slightly larger than 0.05, the nominal significance. These results are expected since the group sums of rankings are affected very little by the magnitude of the outlier. However, the simulated Type I errors for the t-test are all significantly higher than 0.05. Therefore, the Kruskal-Wallis test is much more robust than the t-test in the presence of outliers.

## 1.4 Clustering Analyses

Once the significant genes are identified, more questions will arise. Is there any similarity among those significant genes? Can different disease types be

differentiated based on the expression level of the significant genes? Typically, clustering analysis plays a key role in answering these questions. For the data set containing information about significant genes, we can either consider genes as observations and samples (arrays) as variables or conversely. In the first case, we perform a clustering analysis to classify the significant genes into groups according to their expression profiles across the samples. In the second case, we perform a clustering analysis to group the array samples based on the expression profiles across the significant genes.

### 1.4.1 Clustering Analysis on Genes

First, we standardize the measurements of expression level for each significant gene across all arrays by subtracting the average from the measurements and then dividing the difference by the standard deviation. A k-means clustering analysis (Johnson and Wichern, 1992) is implemented with an automatic search for the best number of clusters determined by the cubic clustering criterion (CCC) (Sarle, 1983). For two-type disease classification, the 291 significant genes selected by the BKW test are automatically divided into two clusters, of 163 and 128 genes. For three-type disease classification, the 402 significant genes are automatically grouped into three clusters, of 171, 93, and 138 genes. Figures 1.6 and 1.7 show the average standardized expression profiles for each cluster.

For two-type disease classification, the genes classified into Cluster1 show consistently lower average levels for the AML disease group. As compared to the

AML group, the average levels of Cluster1 are higher in the ALL disease group (except the ALL-B12), but fluctuate over a wider range. The genes in Cluster2 behave in the opposite manner.

For three-type disease classification, the higher expression levels of genes in Cluster2 and Cluster3 are observed in the T-lineage and AML samples, respectively. For B-lineage samples, apart from ALL-B12 and ALL-B17, the expression levels of genes in Cluster1 are comparatively higher than those in other clusters.

Figures 1.8 and 1.9 display grid plots of the expression levels for all significant genes. The order of array samples in Figures 1.8 and 1.9 are the same as those in Figures 1.6 and 1.7.

### 1.4.2 Clustering Analysis on Array Samples

We also classify the array samples into different groups based on the similarity among the expression levels of those significant genes. Array samples grouped in the same cluster tend to have the same disease type. In general, all the selected informative genes can be used as inputs directly for clustering analysis. However, when the number of the genes that are actually dominant in disease type classification is relatively small in the data, the contribution of these genes to the computed distances (or similarities) between clusters might be small. Therefore, disease types are not easily differentiated. This situation usually occurs when the numbers of informative genes among gene-clusters are highly unbalanced. An

example is given in section 1.4.2.1. To improve the clustering results in the situation above, we suggest a principal component analysis on the informative genes and performing clustering analysis on array samples based on the major principal components. In our study, we choose the smallest number of principal components (17 for two-type disease classification and 20 for three-type disease classification) with the cumulative proportion of variation at least 0.95. A succeeding clustering analysis is applied and the generated classification rule is applied to the test data set. In two-type disease classification, the correct classification rates are 95% (with samples ALL-B12 and ALL-B25 missed) for training data and 94% (with sample numbers 66, 67 missed) for test data. In three-type disease classification, they are 92% (with samples ALL-B12, ALL-B17 and ALL-B25 missed) for the training data and 94% (with sample numbers 66, 67 missed) for the test data. The classification results, along with the results of Golub et al. (1999), are summarized in Table 1.6.

#### 1.4.2.1 Example of Highly Unbalanced Numbers of Genes among Gene Clusters

The situation of unbalanced numbers of genes among gene clusters may occur when stimulating or starving experiments are applied. Large numbers of genes can be induced or depressed in such experiments. Although the unbalanced situation does not appear in leukemia data used here, for demonstration purposes we modify the list of informative genes by removing the 128 least significant

genes from the third gene cluster (Cluster3). Therefore, the gene clusters have 152, 104, and 10 genes, respectively. Based on the new set of informative genes, the k-means clustering analysis without a principal component analysis shows that 42.1% (8/19) of the ALL-B samples are grouped together with the AML samples. When a principal component analysis is applied, only 10.5% (2/19) of the ALL-B samples are in the same cluster as the AML samples.

## 1.5 Discussion

We have applied Bootstrap Kruskal-Wallis filtering as a preliminary selection for all the genes that show statistical significance. The Kruskal-Wallis test in the BKW is interchangeable with alternative parametric approaches such as the ANOVA F-test or the t-test. We have conducted simulations to compare the t- and the Kruskal-Wallis tests under various scenarios. Under the most widely used distributions, lognormal and gamma, for microarray raw data, the Kruskal-Wallis test is slightly more powerful when the sample size is not small. Also, the Kruskal-Wallis test is much more robust than the t-test in the presence of outliers. However, the t-test works better for a small sample size without outliers. Outlier detection is desired for more comprehensive studies in microarray analysis.

The informative genes selected can further be classified into groups by using a k-means clustering analysis that automatically selects the best number of clusters. The genes within each cluster show coherent expression activities and significantly different behavior between disease subtypes. Also, we combined

these significant genes by their expression level principal components and classified the samples into different disease subtypes. We applied these methods to a training data set and tested them on an independent test data set. This approach can be applied to classify multi-type disease. For the leukemia data sets, the classification rates are 95% (36/38) and 92% (35/38) for the training data set; and 94% (32/34); and 94% (32/34) for the test data set, when performing two-type and three-type disease classification, respectively.

Although the BKW method enables us to select statistically significant genes, those selected genes may not be biologically related to the disease subtypes. Genes within a cluster show statistical similarity in transcription level, but this does not address whether they have biological interactions. Studies to search for common motifs in the promoter regions of the clustered genes, and to mine databases for the structures and functions of proteins provide important information to further investigate the mechanism of gene interaction. However, a rigorous and thorough analysis in the first step to select significant genes ensures that later investigations more efficient and accurate.

**Table 1.1:** Numbers of array samples for various disease types

| | B-lineage ALL | T-lineage ALL | ALL | AML |
|---|---|---|---|---|
| Training Data | 19 | 8 | 27 | 11 |
| Test Data | 19 | 1 | 20 | 14 |

**Table 1.2:** Parameter estimates for three distributions

| Normal | | Lognormal | | Gamma | |
|---|---|---|---|---|---|
| Mean( $\hat{\mu}$ ) | Std Dev( $\hat{\sigma}$ ) | Scale( $\hat{\eta}$ ) | Shape( $\hat{\gamma}$ ) | Scale( $\hat{\alpha}$ ) | Shape( $\hat{\beta}$ ) |
| 375.085 | 149.734 | 5.856 | 0.375 | 7.222 | 51.940 |

**Table 1.3:** Simulation results for scenario 1

| Mean Diff. | Normal | | | Lognormal | | | Gamma | | |
|---|---|---|---|---|---|---|---|---|---|
| $(\delta/\hat{\sigma})$ | t | KW | p | t | KW | p | t | KW | p |
| 0 | .0442 | .0455 | .36 | .0477 | .0451 | .13 | .0529 | .0513 | .31 |
| 0.1 | .0604 | .0581 | .14 | .0621 | .0561 | 0 | .0635 | .0582 | 0 |
| 0.2 | .0846 | .0808 | .02 | .0929 | .0943 | .53 | .0889 | .0856 | .09 |
| 0.3 | .1257 | .1176 | 0 | .1392 | .1438 | .07 | .1429 | .1386 | .07 |
| 0.4 | .1923 | .1833 | 0 | .2008 | .2173 | 0 | .2177 | .2101 | .01 |
| 0.5 | .2767 | .2585 | 0 | .2908 | .3206 | 0 | .3071 | .3046 | .41 |
| 0.6 | .3789 | .3569 | 0 | .3991 | .4469 | 0 | .4273 | .4330 | .07 |
| 0.7 | .4813 | .4509 | 0 | .5115 | .5776 | 0 | .5328 | .5463 | 0 |
| 0.8 | .5904 | .5626 | 0 | .6217 | .7004 | 0 | .6417 | .6602 | 0 |
| 0.9 | .6869 | .6593 | 0 | .7145 | .7947 | 0 | .7445 | .7668 | 0 |
| 1 | .7763 | .7481 | 0 | .8002 | .8688 | 0 | .8351 | .8572 | 0 |
| 1.1 | .8450 | .8227 | 0 | .8630 | .9246 | 0 | .8964 | .9123 | 0 |
| 1.2 | .9064 | .8861 | 0 | .9146 | .9604 | 0 | .9377 | .9500 | 0 |
| 1.3 | .9442 | .9277 | 0 | .9499 | .9815 | 0 | .9670 | .9767 | 0 |
| 1.4 | .9668 | .9532 | 0 | .9671 | .9875 | 0 | .9844 | .9895 | 0 |
| 1.5 | .9814 | .9778 | 0 | .9840 | .9960 | 0 | .9918 | .9947 | 0 |
| 1.6 | .9920 | .9883 | 0 | .9902 | .9983 | 0 | .9972 | .9987 | 0 |
| 1.7 | .9960 | .9935 | 0 | .9958 | .9995 | 0 | .9992 | .9997 | .10 |
| 1.8 | .9986 | .9970 | 0 | .9985 | .9998 | 0 | .9996 | .9999 | .08 |
| 1.9 | .9995 | .9990 | .03 | .9991 | 1 | . | .9998 | .9999 | .32 |
| 2 | .9999 | .9997 | .16 | .9995 | 1 | . | 1 | 1 | . |

"t" and "KW" columns indicate the rejection rate of 10,000 simulations from the t- and the Kruskal-Wallis test under the three distributions listed above, respectively. "p" Columns indicate the *p*-values of the McNemar test for the agreement of the t- and Kruskal-Wallis tests. The red and blue colors of these *p*-values indicate the t- and KW tests have larger rejection rate, respectively.

**Table 1.4:** Simulation results for scenario 2

| Mean Diff. | Normal | | | Lognormal | | | Gamma | | |
|---|---|---|---|---|---|---|---|---|---|
| $(\delta/\hat{\sigma})$ | t | KW | Sign | t | KW | Sign | t | KW | Sign |
| 0 | .0494 | .0539 | 0 | .0483 | .0574 | 0 | .0514 | .0596 | 0 |
| 0.1 | .0469 | .0539 | 0 | .0531 | .0645 | 0 | .0532 | .0596 | 0 |
| 0.2 | .0616 | .0677 | 0 | .0586 | .0701 | 0 | .0560 | .0632 | 0 |
| 0.3 | .0684 | .0732 | 0 | .0744 | .0853 | 0 | .0738 | .0825 | 0 |
| 0.4 | .0883 | .0960 | 0 | .0943 | .1100 | 0 | .0943 | .0996 | 0 |
| 0.5 | .1069 | .1153 | 0 | .1176 | .1352 | 0 | .1251 | .1395 | 0 |
| 0.6 | .1350 | .1426 | 0 | .1523 | .1741 | 0 | .1540 | .1659 | 0 |
| 0.7 | .1601 | .1675 | 0 | .1887 | .2114 | 0 | .1975 | .2117 | 0 |
| 0.8 | .1947 | .2001 | .02 | .2359 | .2601 | 0 | .2382 | .2484 | 0 |
| 0.9 | .2363 | .2505 | 0 | .2818 | .3029 | 0 | .2769 | .2902 | 0 |
| 1 | .2886 | .2950 | .01 | .3339 | .3588 | 0 | .3397 | .3513 | 0 |
| 1.1 | .3328 | .3378 | .07 | .3988 | .4129 | 0 | .3925 | .4010 | 0 |
| 1.2 | .3808 | .3855 | .09 | .4507 | .4677 | 0 | .4539 | .4547 | .77 |
| 1.3 | .4380 | .4372 | .77 | .5166 | .5286 | 0 | .5086 | .5183 | 0 |
| 1.4 | .4901 | .4926 | .38 | .5608 | .5719 | 0 | .5674 | .5708 | .23 |
| 1.5 | .5552 | .5519 | .24 | .6155 | .6163 | .78 | .6218 | .6237 | .49 |
| 1.6 | .6040 | .6010 | .29 | .6609 | .6582 | .33 | .6810 | .6786 | .39 |
| 1.7 | .6614 | .6566 | .09 | .7119 | .7090 | .27 | .7313 | .7193 | 0 |
| 1.8 | .6978 | .6945 | .23 | .7566 | .7450 | 0 | .7681 | .7562 | 0 |
| 1.9 | .7442 | .7373 | .01 | .7885 | .7722 | 0 | .8110 | .7947 | 0 |
| 2 | .7873 | .7767 | 0 | .8184 | .7978 | 0 | .8412 | .8217 | 0 |

**Table 1.5:** Type I error with one outlier

| | Normal | | Lognormal | | Gamma | |
|---|---|---|---|---|---|---|
| Outlier | t | KW | t | KW | t | KW |
| None | .044 | .046 | .048 | .045 | .053 | .051 |
| $\hat{\mu}+3\hat{\sigma}$ | .079 | .060 | .081 | .054 | .085 | .055 |
| $\hat{\mu}+4\hat{\sigma}$ | .091 | .057 | .097 | .055 | .105 | .057 |
| $\hat{\mu}+5\hat{\sigma}$ | .101 | .054 | .110 | .057 | .111 | .057 |
| $\hat{\mu}+6\hat{\sigma}$ | .110 | .057 | .122 | .057 | .117 | .054 |
| $\hat{\mu}+7\hat{\sigma}$ | .122 | .059 | .125 | .059 | .128 | .060 |
| $\hat{\mu}+8\hat{\sigma}$ | .115 | .057 | .126 | .061 | .127 | .064 |
| $\hat{\mu}+9\hat{\sigma}$ | .118 | .058 | .121 | .053 | .118 | .058 |
| $\hat{\mu}+10\hat{\sigma}$ | .118 | .056 | .115 | .055 | .114 | .059 |

**Table 1.6:** Classification results of Golub et al. (1999) and the BKW method

| Disease Classification | Golub et al. (1999)* | BKW | |
|---|---|---|---|
| | Two-type | Two-type | Three-type |
| No. of Informative Genes | 50 | 290 | 402 |
| No. of Principal Components | N/A | 17 | 20 |
| Correct Classification Rate | | | |
| --- training data | 37/38 | 36/38 | 35/38 |
| --- test data | 32/34 | 32/34 | 32/34 |

* Not available for three-type classification

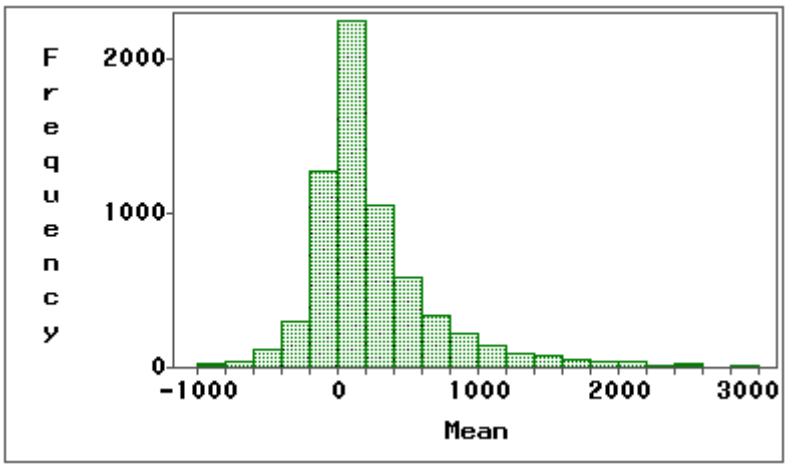



**Figure 1.1:** Histogram of sample means of all genes in ALL groups.

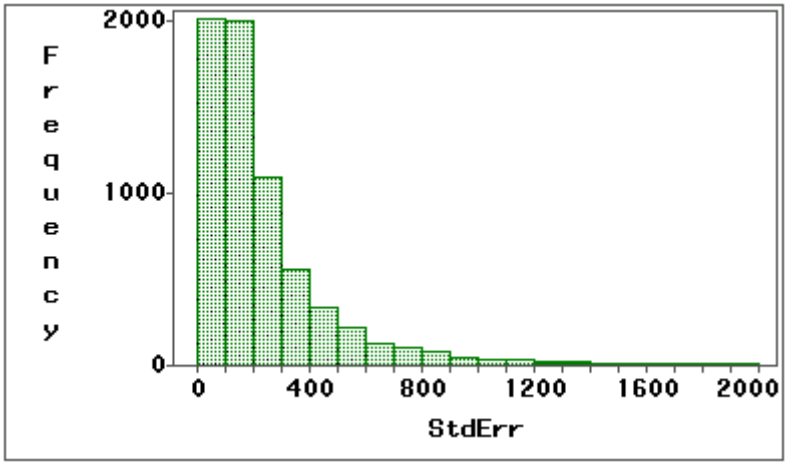



**Figure 1.2:** Histogram of sample standard deviations of all genes in ALL groups.

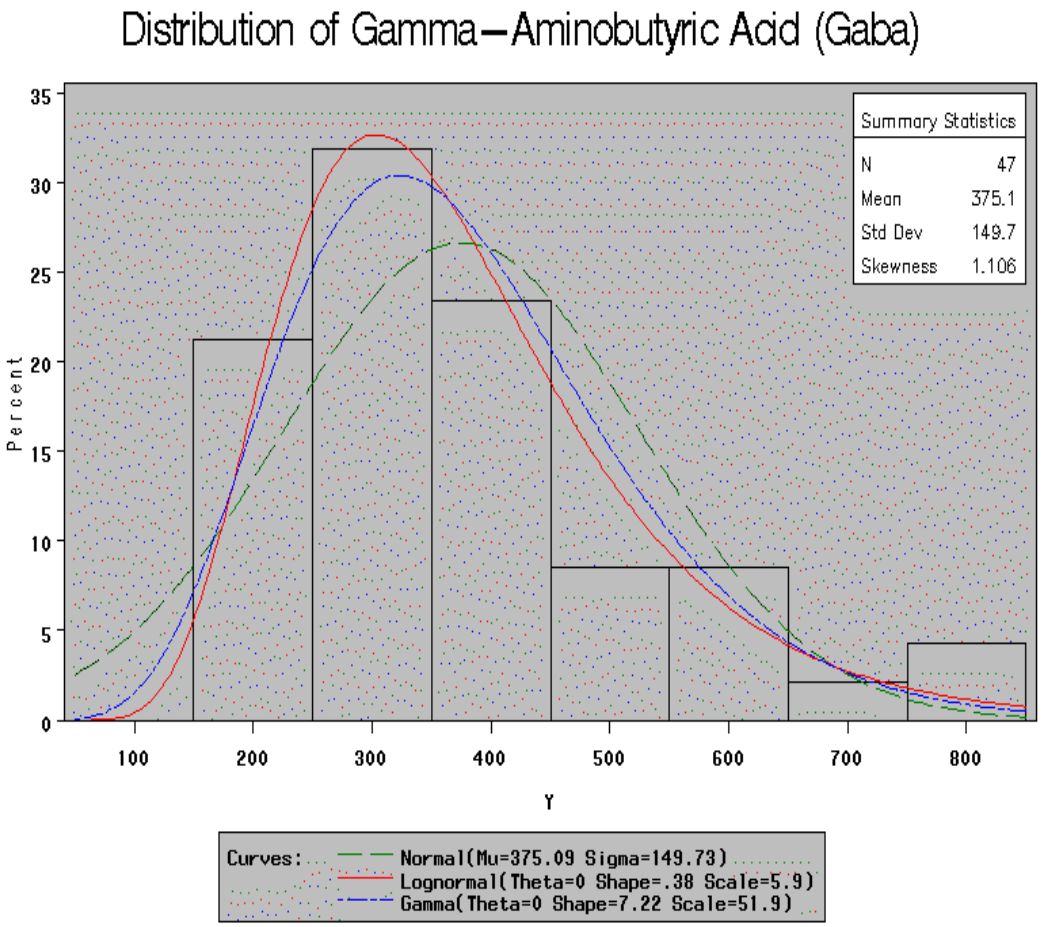



**Figure 1.3:** Histogram of candidate gene with fitted Normal, Lognormal, and Gamma distribution curves.

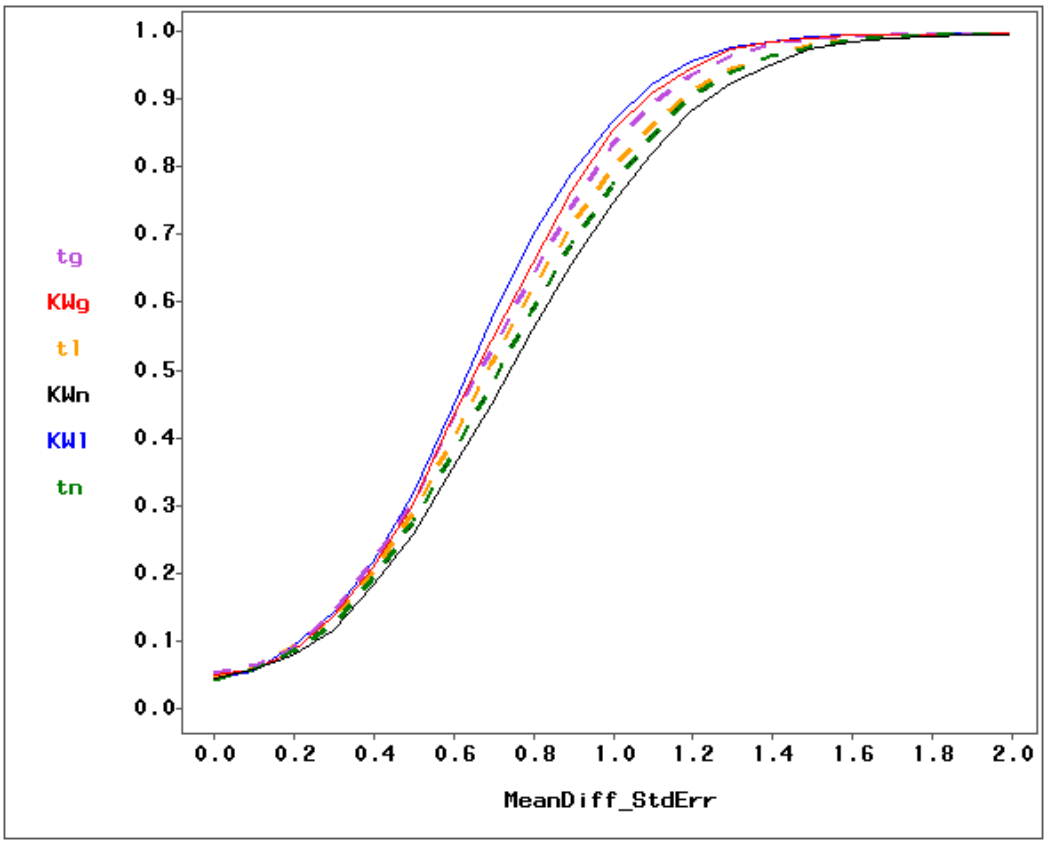



**Figure 1.4:** Power curves of t- (t, dashed lines) and Kruskal-Wallis (KW, solid lines) tests under Normal (n), Lognormal (l), and Gamma (g) distributions. The testing powers from best to worst are KWl, KWg, tg, tl, tn, and KWn.

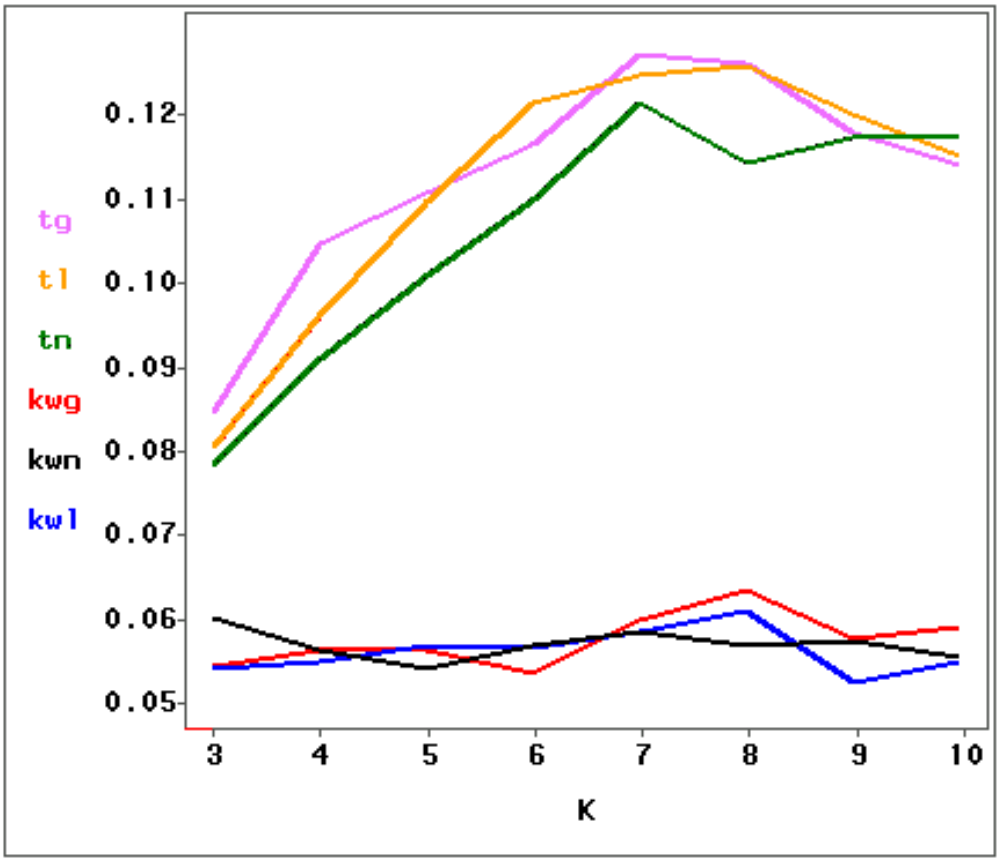



**Figure 1.5:** Simulated Type I error curves with one outlier in data.

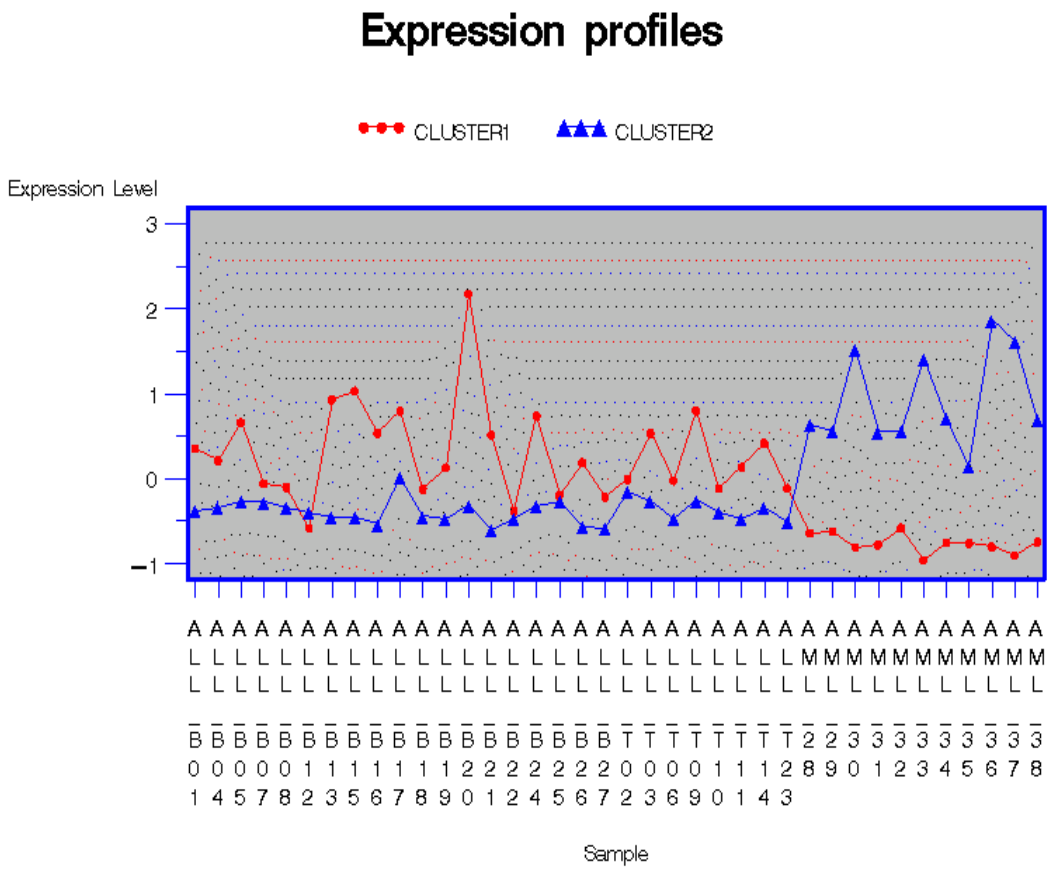



**Figure 1.6:** Line plot of average standardized expression levels for clusters of genes in two-type disease classification.

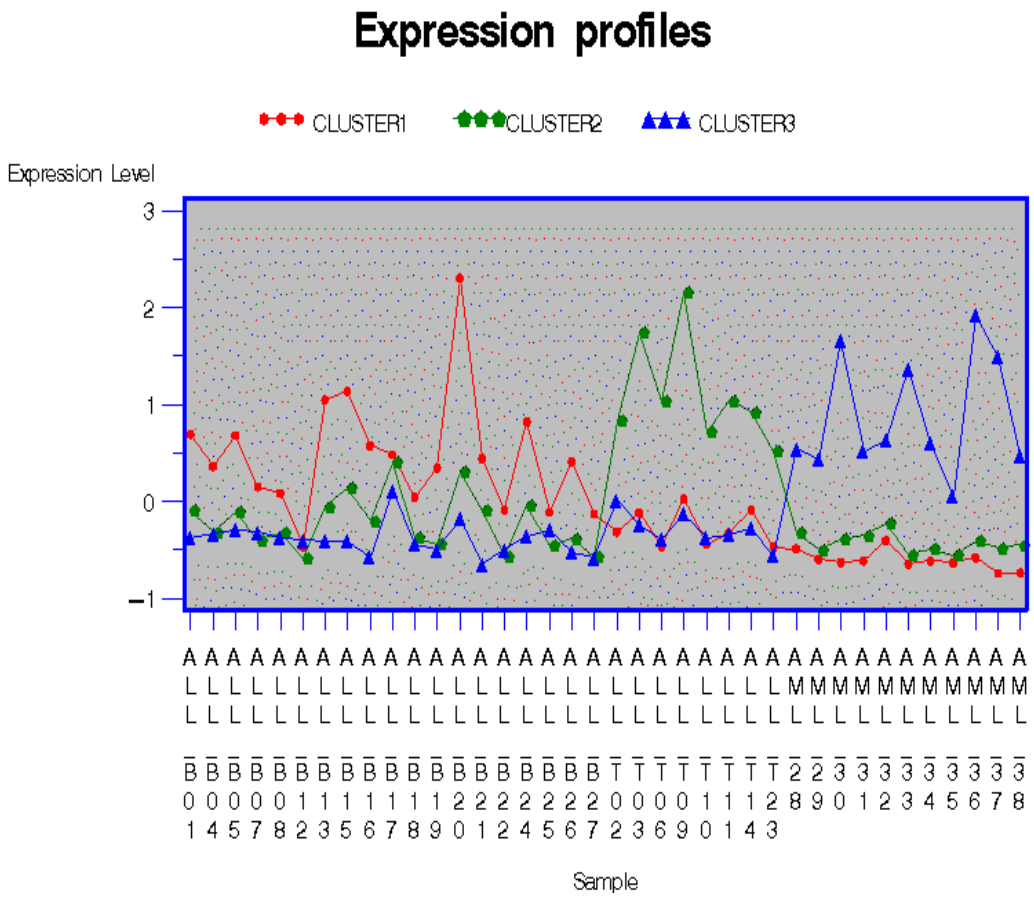



**Figure 1.7:** Line plot of average standardized expression levels for clusters of genes in three-type disease classification.

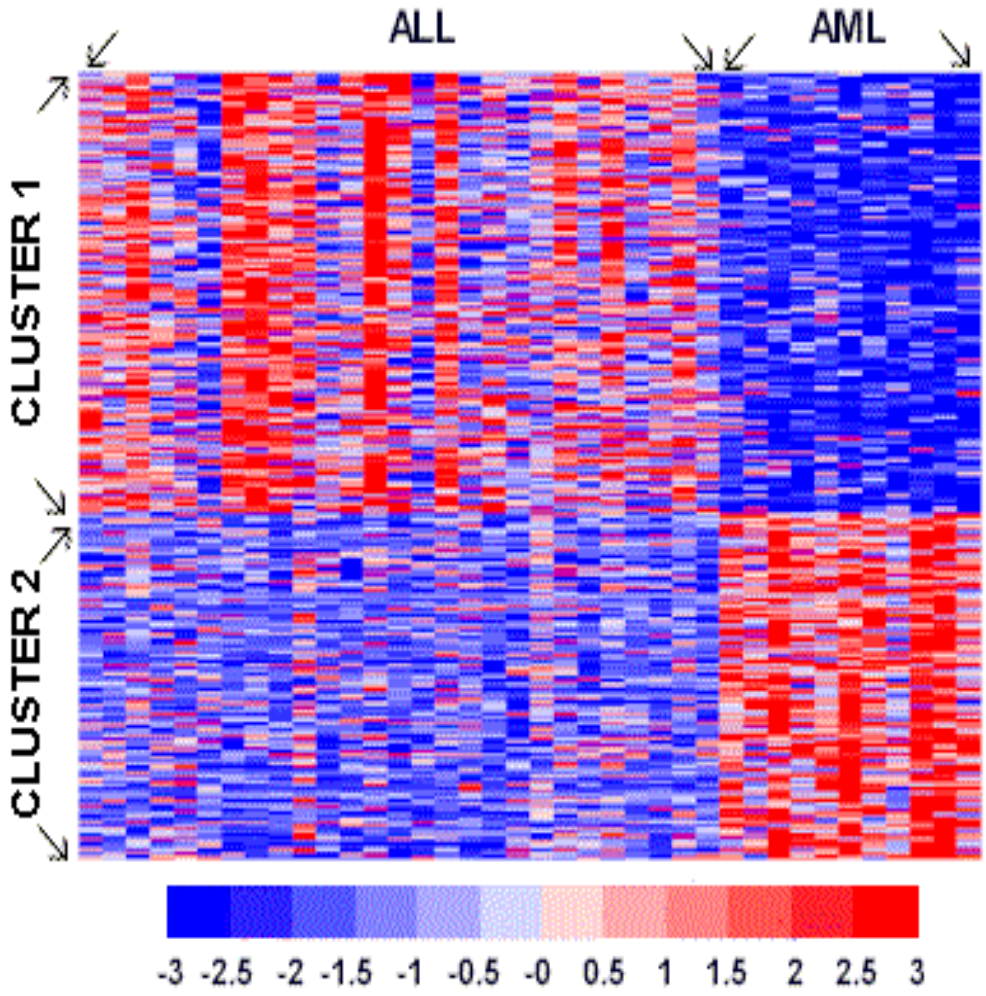



**Figure 1.8:** Grid plot of expression levels for significant genes in two-type disease classification

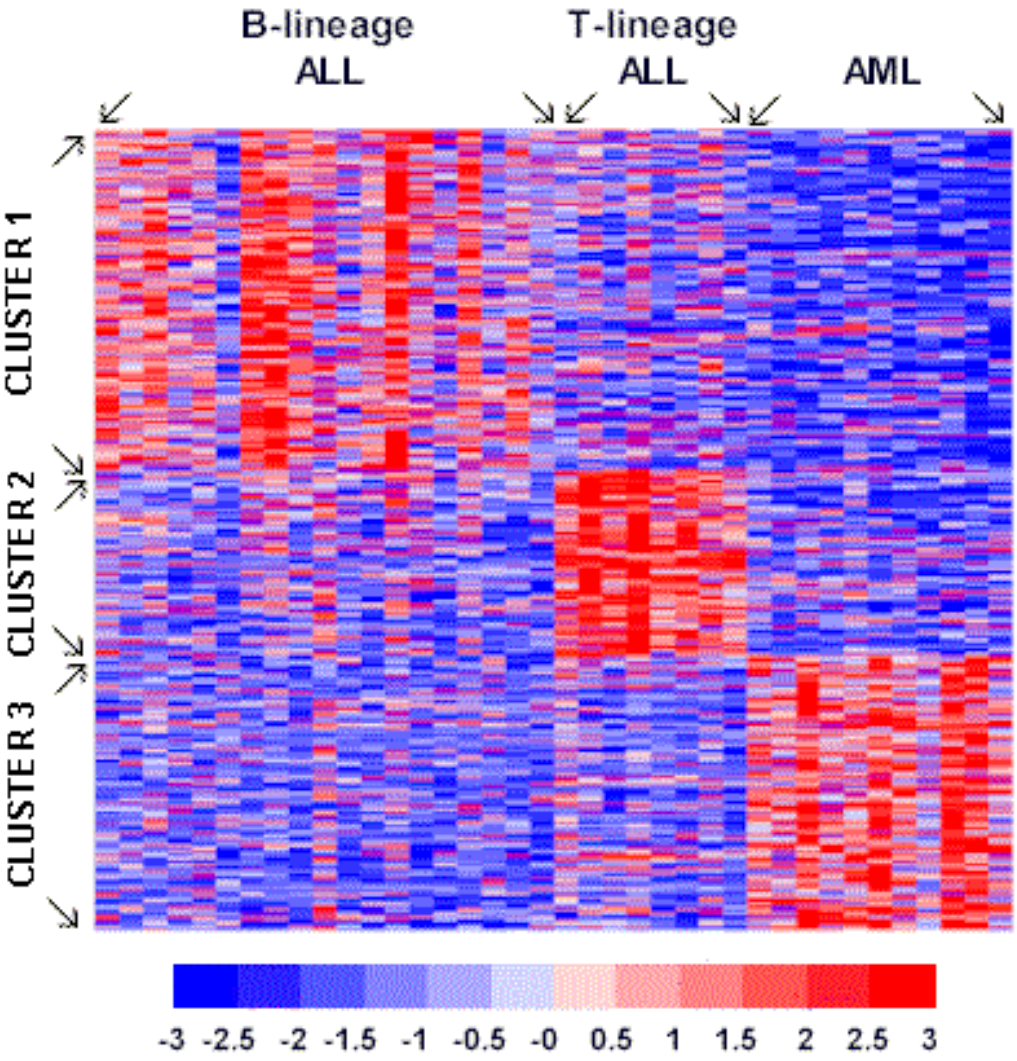



**Figure 1.9:** Grid plot of expression levels for significant genes in three-type disease classification

# Chapter 2

# A Systematic Statistical Linear Modeling Approach to Oligonucleotide Array Experiments

## 2.1 Introduction

The GeneChip$^{TM}$ from Affymetrix is currently a popular commercial oligonucleotide technology for studying gene expression. Although scientific progress with this remarkably miniaturized platform has been considerable, many subtle issues remain regarding the proper analysis and interpretation of the data it produces. Many investigators struggle with such issues as the reliability of a "present" or "absent" call, proper incorporation of mismatched probe information, and appropriate comparisons across many chips. In this paper, we outline a systematic approach for handling data generated from GeneChip$^{TM}$ experiments, with a particular emphasis on statistical model selection and gene significance testing.

The GeneChip contains a probe set representing unique genes, and each probe set consists of 20 probe pairs. Each probe pair consists of a perfect match (PM) oligonucleotide probe, which is designed exactly complementary to a preselected 25-mer of the target gene, and a mismatch probe (MM), which is identical to PM except for one single nucleotide difference at position 13. According to Lockhart et al. (1996), the purpose of the mismatch probe is to serve as an internal control of hybridization specificity. Affymetrix provides basic software to summarize the expression information of the probe set measurements by averaging the difference or log ratio of PM and MM after deleting those extreme measurements which exceed three standard deviations from the mean. Schadt et al. (2000) address many of the important issues and provide useful extensions to the Affymetrix methods.

While summary methods for one or two chips are certainly useful, a statistically optimal approach for experimental data involving many chips requires that we consider all PM and MM data simultaneously. This provides 40 times the data compared to traditional summary methods and gives more power for statistical inference. However, several questions arise regarding the statistical relationships of PM and MM within a probe pair, between probe pairs within a probe set, and between probe sets across arrays. Do they have a linear relationship? Are the amounts of cross-hybridization similar for PM and MM probes? To what extent does MM serve as an internal quality control? Li and Wong (2001a) investigate these and other questions in the context of a

multiplicative model for the measurements, whereas Efron et al. (2000) consider scaled logarithms. Lazaridis et al. (2001) suggest using PM information only. In this chapter, we draw on the rich tradition of statistical linear models and propose methods for potentially complex experiments involving many chips. Our methods are also related to Kerr et al. (2000) and Wolfinger et al. (2001) that propose analysis of variance (ANVOA) applying on cDNA spotted microarrays. The former uses only fixed effects and has all genes being incorporated in one large model. The latter suggest a two-step mixed models to normalize data in array level and, then, to analyze the residuals from first model by forming "single-gene" mixed model for finding significant differential genes and extracting those interesting effects. Our methods are closer to Wolfinger et al. (2001). We apply regression to normalized data and propose a template mixed model for each single gene for further analysis.

In the next section, we outline and discuss a step-by-step approach to handling oligonucleotide array data. We then use the ionizing radiation response data from Tusher et al. (2001) as an example to illustrate these steps and provide a detailed comparison to results of their “significance analysis of microarrays” (SAM) method.

## 2.2 Analysis Steps

### 2.2.1 Identify the Experimental Design

Prior to any formal data analysis, an early and detailed understanding of the statistical experimental design is crucial for maximizing information gain from the data. This entails identification of all real and potential effects impacting both the location and dispersion of the data, how these effects interrelate, and how they affect the experimental units. The effects most commonly of interest are those changed experimentally, and can involve treatment and genotype (cell line) effects. In addition to the experimental factors, the designs we consider also include broad effects on entire arrays and probe-specific effects for each gene. Experimental design has a long and successful history in the statistics literature, and we employ traditional designs such as the split plot used in agricultural field trials (Steel et al., 1997).

### 2.2.2 Extract Numerical Data from the Image

This is obviously the first and one of the most critical steps to properly investigating array data. For sake of brevity and emphasis in this chapter, we assume that this step has been completed in a satisfactory fashion and those reliable numerical intensities corresponding to each PM and MM probes are available.

### 2.2.3 Formulate and Fit a Statistical Model

This is the key step in our approach and requires careful consideration. The goal is to derive a statistical model that adequately accounts adequately for all aspects

of the experimental design yet is simple enough to be interpretable. Doing this allows the researcher to make rigorous quantitative assessments about effects influencing the data that properly separate true signals from experimental and biological noise. As a reasonable starting point, we recommend the classical mixed linear model as a suitably flexible framework; refer to Littell et al. (1996), Verbeke and Molenberghs (2000), and McCulloch and Searle (2001) for theoretical background, examples, and references.

An important initial decision in formulating a linear model involves determining precisely what data values will be modeled for each gene. This usually involves a transformation to make the statistical modeling assumptions reasonable and an adjustment for gross chip-wide effects. As a default method, we recommend a log base 2 ($\log_2$) transformation for individual PM and MM measurements. If PM-MM differences are desired, accommodations must be made for negative values before applying the log transform. Since our proposed statistical model will be additive, using a log transformation on the response can be interpreted as fitting a multiplicative model on the original scale (Li and Wong, 2001a), and resulting statistical estimates are interpretable as fold changes. To adjust for gross array-level effects, we also recommend centering the logged values so that they have mean 0. We find that this can be the simplest and reasonable way for normalization (Figure 2.1). However, this adjustment involves an assumption that the within-chip averaged logged expression levels are the same among chips. More complicated smoothing spline normalizations are

also possible (Dudoit et al., 2000b; Li and Wong, 2001b; Schadt et al., 2000; Yang et al., 2001), although we are not sure they add much value if all of the experimental data for a gene are considered together.

Once a response value is chosen, additive analysis-of-variance effects are specified to partition its variability. In the mixed model setting, one must decide whether these effects are "fixed" or "random". Fixed effects are those effects with a well-defined, finite number of levels and only these finite levels are of interest in the experiment. For fixed effects, we estimate each level and do testing among all levels or comparisons between levels to see if they are significantly different. Random effects are those effects considered to be drawn from an infinite population having some probability distribution, usually normal. For random effects, we estimate the parameters of this probability distribution (variance components in the normal case) and possibly also individual effect estimates properly shrunken towards zero. Inclusion of random effects also allows inferences about the fixed effects to be made to broader populations.

For GeneChip experiments, we typically consider cell line, treatment, and probe effects to be fixed, and because of potentially complex experimental sources of variation such as cross-hybridization, it is typically sensible to include two-way interactions of these effects as well. Effects impacting arrays can be considered random, reasoning that they are the accumulation of small experimental sources of noise. Putting these all together, the following linear mixed model serves as an initial template for the data from a single gene:

$$Y_{ijkl} = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{l(ij)} + \varepsilon_{ijkl} \qquad (2.1)$$

Here, $Y_{ijkl}$ is the transformed and centered expression measurement of the $i^{th}$ cell line applying the $j^{th}$ treatment at the $k^{th}$ probe in the $l^{th}$ replicate. $Y$ can be the centered $\log_2$(PM) measurements if we do not wish to incorporate any MM information, or $Y$ can be the centered $\log_2$ differences of PM-MM pairs (suitably adjusted for negative values) if we believe that MM serves directly as an additive internal control on the original scale. A somewhat intermediate position explored by Efron et al. (2000) is to let $Y$ take the form log(PM) – 0.5 log(MM), and this can be used directly or generalized by including log(MM) as a covariate in the right-hand side of the model.

The symbols *L, T, LT, P, LP, TP* and *A* in (2.1) represent cell line, treatment, cell line-treatment interaction, probe, cell line-probe interaction, treatment-probe interaction, and array effects, respectively. The $A_{l(ij)}$'s are assumed to be independent and identically distributed normal random variables with mean 0 and variance $\sigma_a^2$. The $\varepsilon_{ijkl}$'s are assumed to be independent identically distributed normal random variables with mean 0 and variance $\sigma^2$, and are independent of the $A_{l(ij)}$'s. We will elaborate on these effects and on variations of the model in the context of our example in the next section. For fitting the model, standard maximum likelihood methods are usually best and can be accessed through software like Proc Mixed (SAS Institute Inc., 1999b).

### 2.2.4 Check Assumptions, Remove Outliers, Reformulate and Refit the Model if Necessary

Because we make probabilistic assumptions in the preceding model, it is wise to perform some diagnostic checking on results of the model to verify that it adequately represents the data. In (2.1), the randomness in each observation $Y_{ijkl}$ is represented by two terms, $A_{l(ij)}$ and $\varepsilon_{ijkl}$. According to our normality assumptions in (2.1),

$$A_{l(ij)} + \varepsilon_{ijkl} \sim N(0,\ \sigma_a^2 + \sigma^2).$$

$$Cov(A_{l(ij)} + \varepsilon_{ijkl},\ A_{l'(i'j')} + \varepsilon_{i'j'k'l'}) = \begin{cases} \sigma_a^2 + \sigma^2 & \text{if } (i,j,k,l) = (i',j',k',l') \\ \sigma_a^2 & \text{if } (i,j,l) = (i',j',l') \quad \text{but } k \neq k'. \\ 0 & \text{otherwise} \end{cases}$$

A standard way of checking these assumptions is to examine the residuals from the fitted model. The nature and definition of residuals are more complicated in mixed models than in standard linear models because there are multiple sources of random error. For simplicity, and as a first step, we recommend inspecting residuals formed by subtracting the fitted fixed effects from the observed data and then standardizing these values by an estimate of their variance, $\sigma_a^2 + \sigma^2$. This allows the residuals from all genes to be plotted as groups or together; see Figure 2.2 in the next section. Model departures are often apparent in such plots by a nonrandom scatter around the zero horizontal line. Standardized residuals with magnitudes larger than 3 are potential outliers and can be eliminated from the

analysis. Li and Wong (2001a) discuss systematic ways to eliminate outliers using a multiplicative model.

### 2.2.5 Perform Basic Statistical Inference and Filter Out Insignificant Genes

After suitable model determination and validation, an investigator can study the results of the fitted model for each gene. General statistical tests for the individual fixed effects in the model are available, as are custom estimates of the fold change and significance for specific treatment comparisons. A useful graphical display is the "volcano plot", which plots negative log (base 10) p-values on the y-axis versus estimated $\log_2$ fold change on the x-axis; see Figure 2.4 in the next section and Wolfinger et al. (2001) and Jin et al. (2001) for examples. The plot takes the shape of a "V" because larger fold changes tend to be more significant, although usually the most significant genes do not exhibit the greatest fold change.

A simple procedure for statistically filtering genes is to draw a horizontal cutoff line on a volcano plot to represent a desired false positive rate for the test under consideration. Genes corresponding to points below this line are not considered in subsequent analyses. This method can produce dramatically different results from filtering genes on the basis of fold change alone.

### 2.2.6 Perform Additional Analyses of Statistically Filtered Data

Again, for sake of brevity, we do not elaborate or discuss this step, which usually involves analyses such as clustering and principal components, except to say that appropriate statistical filtering and inference must be done prior to this step to help ensure its validity.

## 2.3 Ionizing Radiation Example

We now illustrate the preceding steps using the ionizing radiation data from Tusher et al. (2001). The data arise from eight Affymetrix GeneChips with 7,129 probe sets (genes) in each and were designed to study transcriptional responses of human cells to ionizing radiation.

### 2.3.1 Identify the Experimental Design

There are two experimental effects, treatment and cell line, with two levels each, and two replicate arrays for each effect combination. At the array level, this is a $2 \times 2$ experiment with 2 replicates. Combined with 20 probes for a probe set, there are 4 factors as stated below.

1. Two levels of radiation treatment (irradiated, unradiated).
2. Two levels of cell line (line I and line II).
3. Twenty PM-MM probe pairs in a probe set (P1 to P20).
4. Two replicate arrays (array I and array II).

This results in a total of 160 observations for each gene, and Table 2.1 shows an example design layout. This is a split plot design (refer to Littell et al., 1996 and

Steel et al., 1997), and the whole plot units are the arrays. Radiation treatment and cell line are the whole-plot effects, and probe is the sub-plot effect.

### 2.3.2 Extract Numerical Data from the Image

As indicated in the previous section, we are skipping this step and assuming that reliable numerical data are available.

### 2.3.3 Formulate and Fit a Statistical Model

The linear mixed model is a "perfect match" for data arising from a split-plot design. As is common practice, we consider both the whole- and sub-plot effects as fixed and the whole-plot experimental units (arrays) as random. Working from the basic model template described previously in (2.1), we consider the following three models:

Model I :

$$\log_2(PM_{ijkl}) = L_I + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + \beta \log_2(MM_{ijkl}) + A_{l(ij)} + \varepsilon_{ijkl.}$$

Model II :

$$\log_2(PM_{ijkl}) = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{l(ij)} + \varepsilon_{ijkl.}$$

Model III :

$$\log_2(D_{ijkl}) = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{l(ij)} + \varepsilon_{ijkl.}$$

In model I, we use array-centered $\log_2$(PM) as the response variable and array-centered $\log_2$ (MM) as a covariate, with $\beta$ as its coefficient. In model II, we only use the perfect match probe data and no mismatch probe data. In model III, we

use the centered $\log_2$ of the difference (D) of the PM-MM pair as the response variable, and before doing the log transformation on D, we truncate those measurements less than 10 to 10 to avoid the negative value problem. Over one third (38.5%) of the data are truncated in this fashion.

### 2.3.4 Check Assumptions, Remove Outliers, Reformulate and Refit the Model if Necessary

As a quick check on data standardization requirements, Figure 2.1 shows plots of pairs of replicated $\log_2$ intensities for the four combinations of cell line and irradiation treatment. The fitting regression equations are shown above each plot. The $R^2$ values from left to right are 0.9794, 0.9801, 0.9879, 0.9819 and are not changed by standardization. The linearity of the plots indicates that a simple mean standardization is suitable. Those points far from diagonal line indicate that there may be undesirable outliers in data, but we retain them for subsequent comparison purposes.

Using Model I as an illustrative example, Figure 2.2A displays standardized residual plots of genes 1000, 2000, 3000, 4000, 5000, and 6000. These plots exhibit random scatter around the zero horizontal line and indicate no significant departures from model assumptions. Figure 2.2B, 2.2C, and 2.2D plot all of the standardized residuals together for the three models in what we have nicknamed a “submarine plot”. These "submarine plots" are not regular residual plots, as the residuals are pooled from different models. However, they can be

useful for observing genome-wide features of gene variability. The “bubbles” in these plots represent potential outliers with large positive or negative standardized residuals. A rough rule of thumb is to eliminate observations with standardized residuals having magnitude larger than three, but doing this results in little to no changes in the most highly significant genes considered later. For subsequent comparison purposes, therefore, we filter no outliers. The absence of points in the lower left portion of Figure 2.2D is an artifact of the truncation at 10 rule we used to analyze the data in Model III.

Also, it is interesting that while residual plots of single gene show no evidence of pattern, the "submarine plot" suggests heteroscedacity. We try to test for normal distribution by Kolmogorov-Smirnov statistic with Bonferroni's correction on standardized residuals for each gene excluding those residuals larger than 3 or less than -3. Almost all genes (98.51 %) pass the normal distribution test in Model I. Background correction may be a reason for observing a "submarine" that has a bigger head (larger dispersion for small measurement) and a small tail (smaller dispersion for large measurement). According to Affymetrix's algorithm, the probe intensities are calculated by averaging out the pixel intensities and subtracting the background correction term, which averages out the lowest 2% pixel intensities. Those small measurements are expected to be more sensitive to background corrections than are large measurements.

For a quick assessment of goodness-of-fit, we calculate $R^2$ values of each gene and draw a histogram for Model I in Figure 2.3. The definition of $R^2$ in the

mixed model is somewhat ambiguous, so here we apply an ordinary model concept of $R^2$, and define it as

$$R^2 = 1 - \frac{\sum R_{ijkl}^2}{\sum (Y_{ijkl} - \overline{Y})^2},$$

where $\overline{Y}$ is the average of all $Y_{ijkl's.}$ and $R_{ijkl}$ is the residual term of $Y_{ijkl}$ from the model. This histogram excludes those 160 genes having 1/5 of probe data missing. The first percentile is 0.86, indicating that Model I fits the data from almost all genes very well.

### 2.3.5 Perform Basic Statistical Inference and Filter Out Nonsignificant Genes

We first consider significance tests for all of the fixed effects in Figure 2.4. The "volcano" plots of Figure 2.4A illustrate the relationship of the estimates of three whole-plot effects (cell line, treatment, cell line-treatment interaction) and their significance compared among all three models for all genes. The horizontal lines indicate p-values equal to 0.001 and the vertical lines indicate 2-fold changes of effect estimates. In Figure 2.4A, we can see that there are only a few points outside the vertical lines and many points above the horizontal lines, especially, in the cell line and treatment volcano plots. The small magnitudes of the fold change estimates are remarkable and appear to represent excellent sensitivity in this experiment; see also the fold-change estimates in Tables 2.2 and 2.3. In

addition, estimates from Model III are much more variable, indicating that direct subtraction of MM adds a lot of noise to the PM data.

The three plots in Figure 2.4B show the significance of probe, cell line-probe interaction, and treatment-probe interaction effects for Model I, each of which results from F-tests with (19 ,94) degrees of freedom. Results from Models II and III are similar. In Figure 2.4B, the probe main effects are highly significant, and even after a Bonferroni adjustment, 7088 (99.42%) are significant at the 5% level. This indicates huge variability in probe effectiveness. There is also evidence of some large interactions with the probe effect and a few examples are depicted in Figure 2.5. In Figure 2.5A, PM curves (red) show significant different in probes 9 and 14 comparing top (cell line I) and bottom (cell line II) rows; whereas, MM curves (green) show significant different in probe 9. In Figure 2.5B, PM curves (red) show significant different in probes 11,12,13, and 14 comparing top (irradiation treatment) and bottom (no radiation treatment) rows; whereas, no significant different in MM curves (green). After a Bonferroni adjustment, 666 (9.34%) and 42 (0.59%) of the genes have significant (cell line)-probe and treatment-probe interactions, respectively. Causes for these interactions such as cross-hybridization may warrant further investigation.

Figure 2.4C shows the significance of using $\log_2$ (MM) as a covariate in Model I. The estimates of $\beta$ range from -1 to 2, and can be viewed as gene-specific generalizations of the constant 0.5 value considered by Efron et al.

(2000). However, only 422 (5.92%) of these coefficients are significantly different from zero at the 5% level with a Bonferroni correction.

Tables 2.2 and 2.3 list the top induced and repressed genes according to our three models along with a comparison to the top genes from the SAM method of Tusher et al. (2001). The methods generally agree, although there are a few key differences that we discuss below. Tusher et al. (2001) highlight twelve genes that were previously reported in the literature to respond transcriptionally to ionizing radiation. Nine of these are included in the top eighteen of Model I's induced or repressed genes. The other three (X62048, S78187, X63717) are also marginally significant with negative log p-values larger than 2 but not as significant as the result of Tusher's SAM.

One likely reason for these discrepancies is that the SAM results make use of the Affymetrix summary measures across the probe level data whereas the mixed model results here delete no outliers. Figure 2.6A uses gene X62048 as an example. Affymetrix applies a "three standard deviation rule" for outlier deletion; that is, if the difference of probe pair exceeds three standard deviations within a probe set, it will be excluded for averaging. In Figure 2.6A, probe pair 10 in array "2ir1" is an outlier according to the "three standard deviation rule". Although it is questionable to say this probe pair is an outlier, excluding that point when calculating the average difference will lower the single expression measurement of gene X62048 in array "2ir1", and this may increase the significance of repression when comparing the irradiated group to the untreated group. We also

inspected several genes which were highly significant according to SAM but not significant for Models I, II, and III. Outliers appear to be at the root of these differences as well, and another example is in Figure 2.6B. The difference of probe pair 14 in arrays "1ir2", "1un1", and "1un2" are outliers identified by "three standard deviation rule" in this case. Resolutions of issues like these almost surely need to be done at the probe level, and indeed, independent analyses applied at the probe level produce almost identical results to ours [V. Tusher, personal communication].

Comparing results from Models I, II and III, we see those from Models I and II are similar whereas some of the results from Model III are very different. The overall behavior of the treatment effects from these three models is displayed in Figure 2.7A. The correlation coefficients of the negative log p-values of the model pairs (I, II), (I, III), and (II, III) are 0.95, 0.53, and 0.45 respectively. Here many of the discrepancies appear to be related to the truncation at 10 used in Model III to avoid negative difference values when taking logarithms, and Figure 2.7B gives an example. Results like these shed doubt on the practical usefulness of trimming the data in this fashion. Also, including MM as a covariate does not change the results very much for this example, suggesting a lack of need for MM.

Fold-change estimates of the Model I treatment effects for genes are included in the next-to-last columns of Tables 2.2 and 2.3. Most are surprisingly small (between 1.1 and 1.3) and illustrate the potential of statistical methods to detect subtle changes.

### 2.3.6 Perform Additional Analyses on Statistically Filtered Data

This step is omitted, except to say that clustering could now be performed on the treatment estimates from the previous step. Estimated treatment means appropriately adjusted for other effects in the model are very useful quantities for subsequent analyses like clustering and principal components, and can provide more accurate results than the raw data themselves.

## 2.4 Discussion

The differences noted above between SAM and our linear models are typically resolvable and of an order of magnitude smaller than differences between these methods and ones that make decisions based purely on fold change. It is absolutely critical to accommodate important sources of experimental variability when assessing GeneChip$^{TM}$ data, and failure to do so will surely result in higher rates of both false positives and false negatives. The systematic approach detailed here provides a good outline on how to analyze these data in a statistically sensible fashion, and the linear mixed modeling framework can flexibly accommodate most any kind of experimental design.

**Table 2.1:** Design layout (within a gene): There are 4 experimental effects (2 lines and 2 treatments) applied on 8 whole plot units as boxes (probe sets) below. Each whole plot unit has 20 sub-plot units (probes).

Replicate A I

| | Treatment I (irradiated) | | | | | Treatment II (not irradiated) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Line I | P1 | P2 | … | P20 | Line I | P1 | P2 | … | P20 |
| Line II | P1 | P2 | … | P20 | Line II | P1 | P2 | … | P20 |

Replicate A II

| | Treatment I (irradiated) | | | | | Treatment II (not irradiated) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Line I | P1 | P2 | … | P20 | Line I | P1 | P2 | … | P20 |
| Line II | P1 | P2 | … | P20 | Line II | P1 | P2 | … | P20 |

**Table 2.2:** Most significantly induced genes

| Gene No | Accession No. | Rank | | | | nlog | | | Fold | Gene Description |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Tusher | I[1] | II[2] | III[3] | I[1] | II[2] | III[3] | I[1] | |
| 2863 | U18300† | 4 | 1 | 2 | 6 | 4.26 | 4.43 | 3.16 | 1.31 | p48, xeroderma pigmentosum group E gene |
| 6684 | X83490† | 2 | 2 | 3 | 4 | 4.19 | 4.33 | 3.34 | 1.26 | Fas (alternate splice deleting exins 3 & 4) |
| 2357 | M92424† | 18 | 3 | 5 | 9 | 3.78 | 3.76 | 2.71 | 1.14 | mdm2 |
| 2715 | U09579† | 1 | 4 | 1 | 3 | 3.61 | 4.48 | 3.60 | 2.06 | p21 |
| 5800 | M58509 | NA | 5 | 15 | 14 | 3.51 | 3.18 | 2.42 | 1.15 | adrenodoxin reductase gene |
| 3850 | U82987 | 17 | 6 | 17 | 1 | 3.36 | 3.10 | 3.71 | 1.35 | bcl-2 binding component 3 (bbc3) |
| 1610 | L42176 | 26 | 7 | 8 | 24 | 3.32 | 3.37 | 2.04 | 1.11 | DRAL mRNA |
| 1453 | L29008 | NA | 8 | 6 | 100 | 3.31 | 3.51 | 1.30 | 1.10 | L-iditol-2 dehydrogenase |
| 170 | D00762 | NA | 9 | 7 | 88 | 3.30 | 3.44 | 1.32 | 1.09 | proteasome subunit HC8 |
| 5915 | L08096 | 29 | 10 | 12 | 2 | 3.29 | 3.25 | 3.69 | 1.32 | CD27 ligand mRNA |
| 1154 | J05614† | 13 | 11 | 4 | 5 | 3.26 | 3.83 | 3.19 | 1.56 | PCNA, proliferating cell nuclear antigen |
| 4598 | X77794† | 9 | 12 | 10 | 18 | 3.13 | 3.31 | 2.19 | 1.28 | cyclin G1 |
| 6683 | X83492† | 14 | 13 | 14 | 37 | 3.11 | 3.22 | 1.64 | 1.15 | Fas (alternate splice deleting exins 4 & 7) |
| 1395 | L20971 | 123 | 14 | 18 | 39 | 3.05 | 3.06 | 1.64 | 1.12 | phosphodiesterase mRNA |
| 3148 | U39400 | 8 | 15 | 20 | 12 | 3.02 | 2.92 | 2.46 | 1.15 | NOF1 |
| 5431 | U72649 | NA | 16 | 11 | 15 | 2.99 | 3.29 | 2.27 | 1.17 | BTG2 |
| 1883 | M28209 | NA | 17 | 23 | 28 | 2.96 | 2.72 | 1.92 | 1.09 | GTP-binding protein (RAB1) |
| 6089 | M60974† | 10 | 18 | 26 | 8 | 2.95 | 2.66 | 2.77 | 1.29 | gadd45 |
| 5469 | X63717† | 6 | 27 | 32 | 13 | 2.60 | 2.61 | 2.43 | 1.22 | APO-1 cell surface antigen |
| 276 | D21089 | 7 | 44 | 73 | 19 | 2.33 | 2.04 | 2.16 | 1.36 | XPC, xeroderma pigmentosum group C gene |
| 782 | D90224 | 11 | 80 | 114 | 58 | 1.94 | 1.78 | 1.45 | 1.14 | OX40 ligand, TNF ligand superfamily |
| 6539 | X85116 | 15 | 89 | 117 | 20 | 1.84 | 1.77 | 2.16 | 1.11 | EPB72, integral membrance protein |
| 3283 | U48296 | 5 | 99 | 128 | 337 | 1.80 | 1.71 | 0.93 | 1.14 | protein tyrosine phosphatase PTP(CAAX1) |
| 2946 | U25138 | 12 | 2286 | 3136 | 2636 | 0.22 | 0.13 | 0.04 | 1.00 | maxi K potassium channel beta subunit |
| 3320 | U50136 | 16 | 2725 | 2548 | 2622 | 0.12 | 0.17 | 0.16 | 0.99 | leukotriene C4 synthase (LTC4S) |
| 3270 | U47621 | 3 | 3054 | 3225 | 3859 | 0.05 | 0.01 | 0.03 | 1.00 | No 55 nucleolar autoantigen |

† Genes previously reported to respond transcriptionally to ionizing radiation. Tusher rank values from Tusher et al. (2001)
[1]Results from Model I. [2]Results from Model II. [3]Results from Model III. Fold_I indicates the treatment fold change according to Model I. NA indicates the gene has less than 1.5 fold change and is filtered out prior to the SAM analysis.

**Table 2.3:** Most significantly repressed genes

| Gene No. | Accession No. | Rank | | | | Nlog | | | Fold | Gene Description |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Tusher | I[1] | II[2] | III[3] | I[1] | II[2] | III[3] | I[1] | |
| 1860 | M25753[†] | 5 | 1 | 2 | 3 | 5.14 | 5.19 | 4.12 | 1.39 | cyclin B |
| 2645 | U05340 | 4 | 2 | 3 | 2 | 4.84 | 4.72 | 4.27 | 1.27 | p55cdc; present in dividing cells |
| 2576 | U01038 | 1 | 3 | 1 | 1 | 4.80 | 5.44 | 4.43 | 1.37 | PLK, polokinase homolog |
| 2789 | U14518 | 19 | 4 | 4 | 7 | 4.28 | 4.11 | 3.06 | 1.22 | centromere protein-A (CENP-A) |
| 5136 | Z36714 | 29 | 5 | 16 | 114 | 4.22 | 2.99 | 1.43 | 1.15 | cyclin F |
| 4153 | X14850 | 58 | 6 | 6 | 17 | 4.15 | 4.06 | 2.33 | 1.19 | histone H2A.X |
| 3682 | U73379 | 20 | 7 | 5 | 5 | 3.98 | 4.09 | 3.57 | 1.35 | cyclin-selective ubiquitin carrier protein |
| 6702 | X97267 | 6 | 8 | 7 | 16 | 3.90 | 3.54 | 2.36 | 1.19 | lymphosphatase assoc phosphoprotein |
| 6815 | HG1980 | NA | 9 | 8 | 37 | 3.49 | 3.53 | 1.97 | 1.16 | "Tubulin" Beta 2 |
| 4214 | X51688 | 62 | 10 | 10 | 39 | 3.26 | 3.30 | 1.96 | 1.15 | cyclin A |
| 3535 | U63743 | 9 | 11 | 14 | 42 | 3.15 | 3.04 | 1.93 | 1.10 | MCAK, mitotic centromere-associated kinesin |
| 2353 | M91670 | 2 | 12 | 9 | 12 | 3.14 | 3.30 | 2.46 | 1.20 | ubiquitin carrier protein (E2-EPF) |
| 4273 | X54942 | 8 | 13 | 11 | 4 | 3.14 | 3.14 | 3.88 | 1.41 | ckshs2, cks1 protein homolog |
| 1612 | L42324 | 67 | 14 | 13 | 28 | 3.13 | 3.12 | 2.08 | 1.09 | G protein-linked receptor gene (GPCR) gene |
| 203 | D13633 | 38 | 15 | 15 | 35 | 3.07 | 3.04 | 1.99 | 1.11 | KIAA0008 gene |
| 5409 | U37426 | 116 | 16 | 19 | 10 | 3.05 | 2.95 | 2.48 | 1.10 | kinesin-like spindle protein HKSP (HKSP) |
| 6377 | X62534 | NA | 17 | 17 | 24 | 3.03 | 2.96 | 2.19 | 1.16 | HMG-2 |
| 2306 | M86699 | 33 | 18 | 18 | 20 | 2.93 | 2.95 | 2.27 | 1.12 | kinase (TTK) mRNA |
| 4453 | X67155 | 16 | 19 | 23 | 8 | 2.92 | 2.91 | 2.91 | 1.14 | MKLP-1, mitotic kinesin-like protein-1 |
| 2988 | U28386 | 13 | 34 | 33 | 50 | 2.51 | 2.50 | 1.88 | 1.25 | hSRP1alpha, NLS receptor |
| 5063 | Z15005 | 18 | 36 | 40 | 27 | 2.43 | 2.30 | 2.14 | 1.10 | CENP-E putative kinetochore motor |
| 4370 | X62048[†] | 11 | 46 | 48 | 43 | 2.22 | 2.19 | 1.93 | 1.04 | wee1 kinase |
| 4039 | X02910 | 14 | 49 | 60 | 23 | 2.19 | 2.03 | 2.22 | 1.09 | tumor necrosis factor (TNF-alpha) |
| 2511 | S78187[†] | 7 | 56 | 74 | 52 | 2.09 | 1.93 | 1.85 | 1.11 | cdc25 phosphatase |
| 5847 | HG3523 | 17 | 71 | 139 | 67 | 1.96 | 1.54 | 1.69 | 1.08 | c-Myc, alternate splice form 3 |
| 2245 | M80359 | 12 | 106 | 330 | 79 | 1.70 | 1.16 | 1.61 | 1.05 | C-TAK1, cdc25c associated protein kinase |
| 361 | D31764 | 15 | 136 | 297 | 119 | 1.59 | 1.19 | 1.39 | 1.05 | hEphB1b, Eph-like receptor tyrosine kinase |
| 674 | D86973 | 10 | 2581 | 3082 | 935 | 0.16 | 0.05 | 0.57 | 1.00 | GCN1, translational regulator of GCN4 |
| 3615 | U68233 | 3 | 805 | 1285 | 1208 | 0.86 | 0.69 | 0.47 | .98 | HRR-1 farnesol receptor |

[†] Genes previously reported to respond transcriptionally to ionizing radiation. [1] Results from Model I. [2] Results from Model II. [3] Results from Model III. Fold I indicates the reciprocal of fold change for comparison purpose.

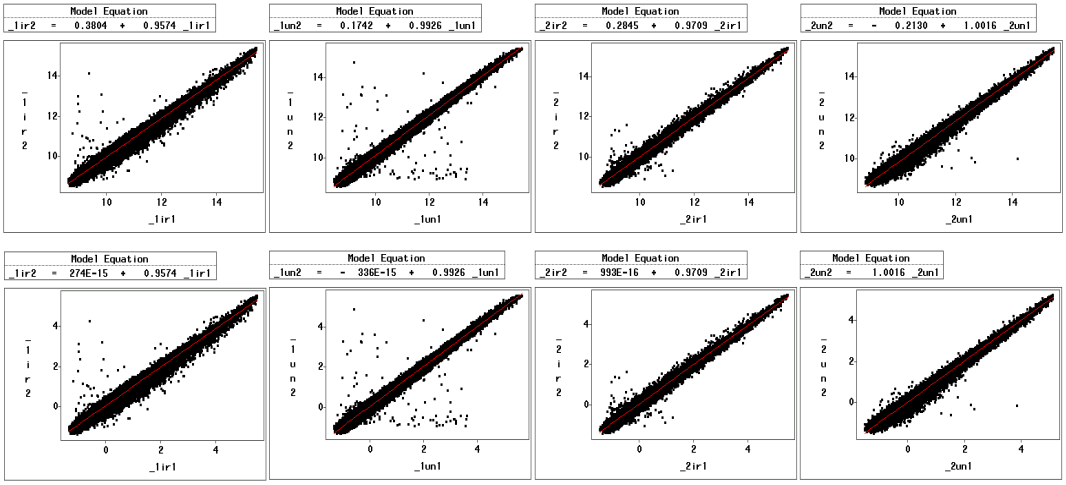



**Figure 2.1:** Scatter plots of four experimental effects between two replicate arrays before (top four) and after (bottom four) standardization. Here, each X or Y variable identifies those eight arrays. "_1ir2" indicates the array was applied by cell line I with irradiation treatment in replicate array II.

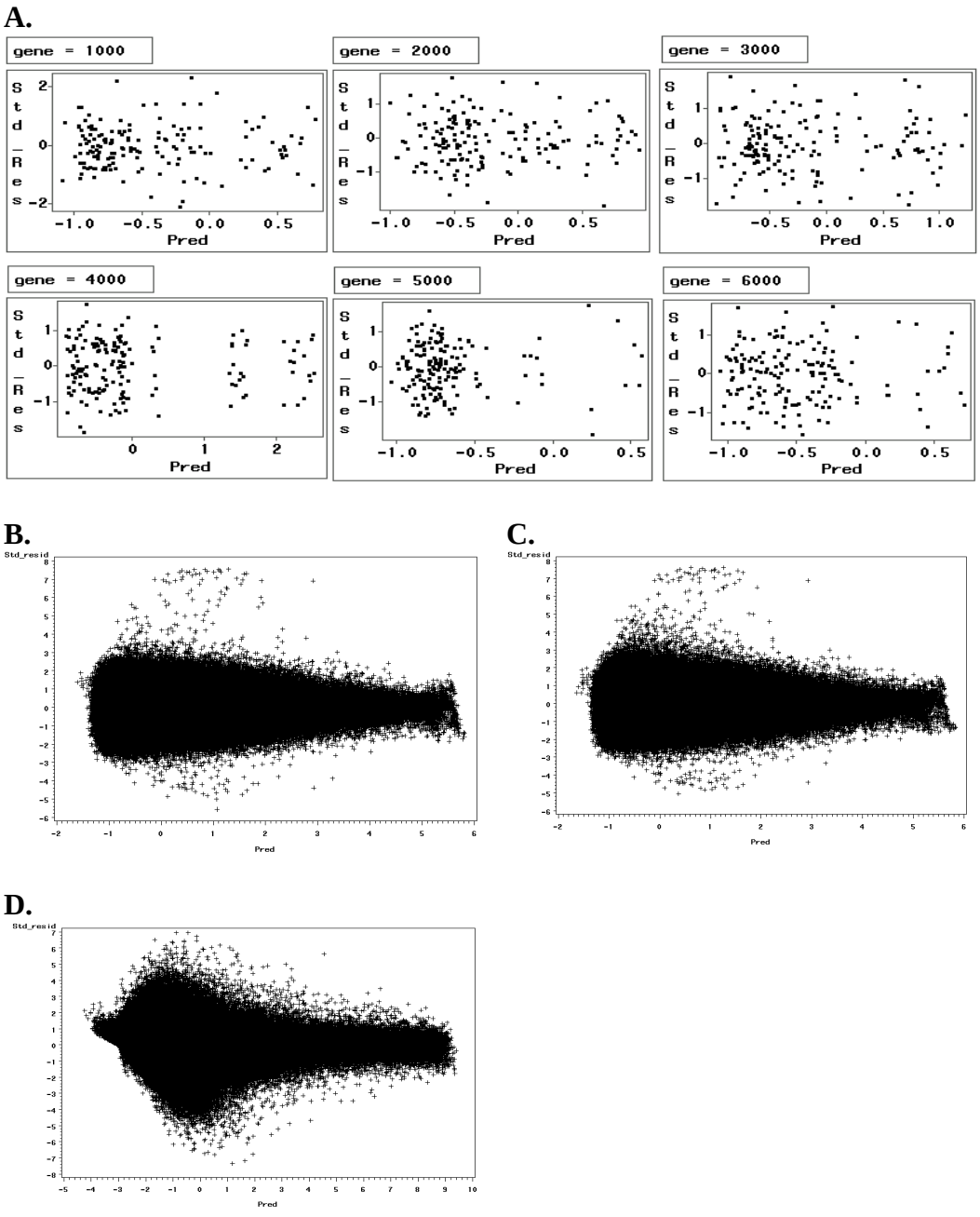



**Figure 2.2:** **A.** Standardized residual plots for gene 1000, 2000, 3000, 4000, 5000 and 6000 from Model I. **B.C.D.** "Submarine plots" of standardized residuals of all genes from Model I, II, and III.

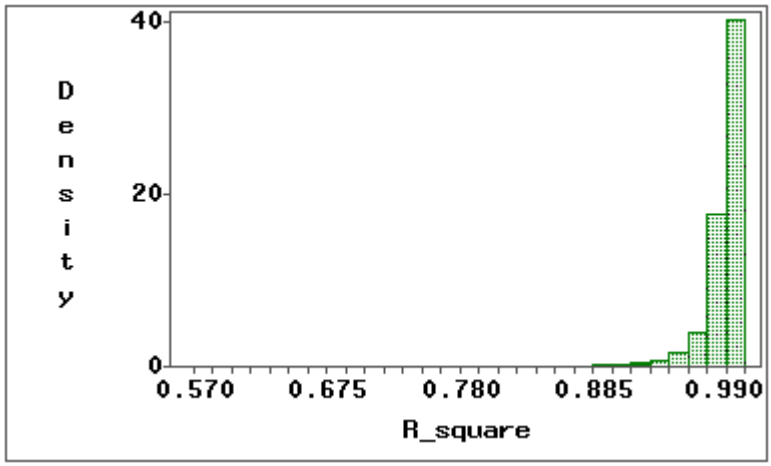



**Figure 2.3:** Histogram of $R^2$ values from Model I for all genes excluding 160 genes that have 1/5 data missing.

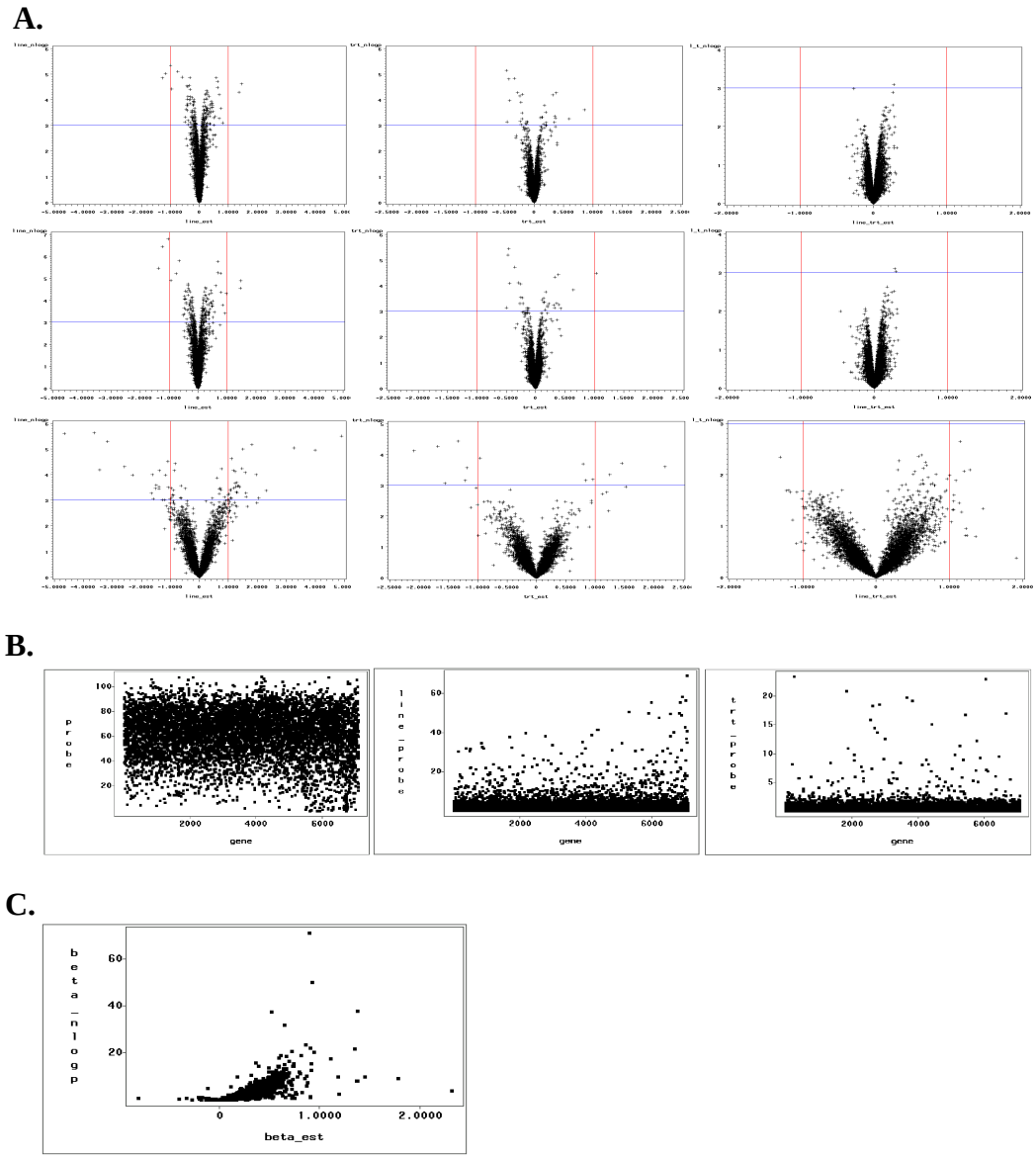
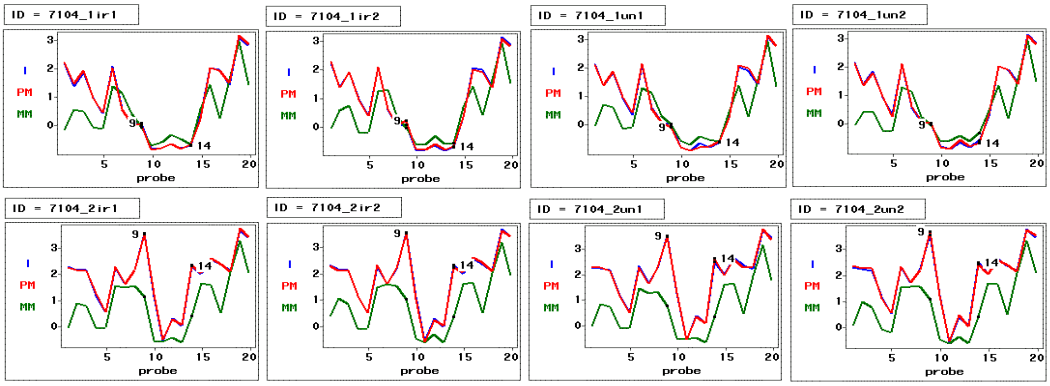



**Figure 2.4:** **A.** "Volcano" plots of cell line, treatment, and cell line-treatment interaction effects among Models I (top row), II (middle row), and III (bottom row) for all genes. The X-axis is the $\log_2$ estimate of the difference of levels in the effect. The Y-axis is the negative log p-value. The horizontal blue lines indicate the testing p-value equal to 0.001 and the vertical red lines indicate 2-fold change of effect estimates. **B.** Significance plots of the probe and its interaction effects applying Model I for all genes. The same plots applying Model II and III are similar to these three plots. The X-axis is the gene number. The Y-axis is the negative log p-value. **C.** Significance plot of the covariate effect applying Model I. The X-axis is the estimate of the parameter, $\beta$, of $\log_2$(MM) covariate.

**A.**




**B.**

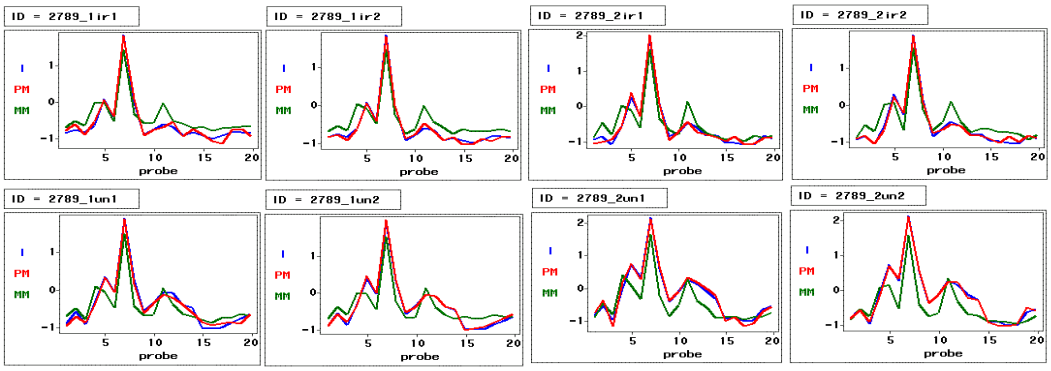
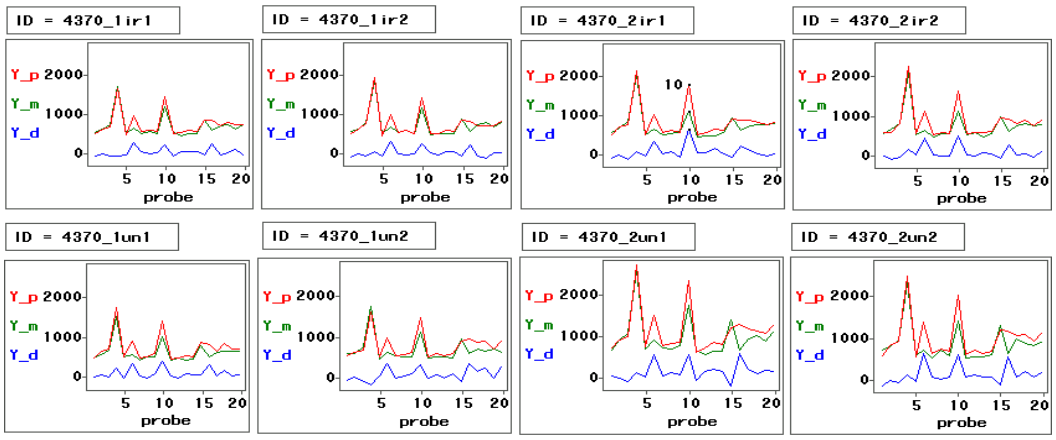



**Figure 2.5: A.** Significant cell line-probe interaction in gene 7104 (X3068). The blue, red, and green lines indicate the prediction of model I, standardized PM, and standardized MM respectively. **B.** Significant treatment-probe interaction in gene 2789 (U14518).

**A.**




**B.**

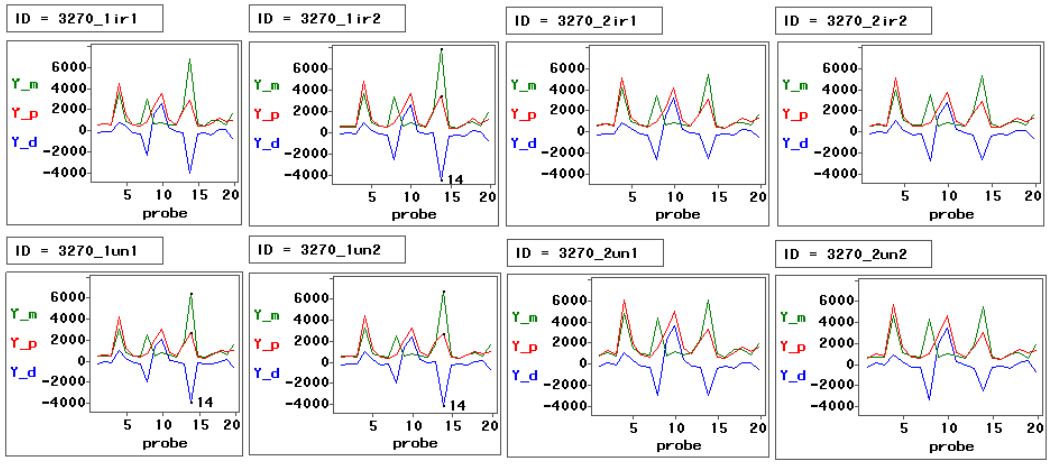
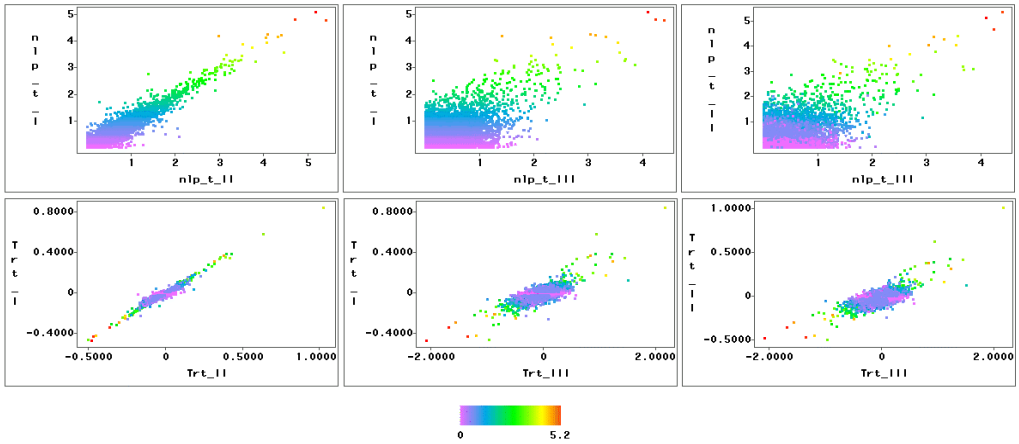



**Figure 2.6:** **A.** Probe profiles of gene 4370 (X62048). Red, green, blue lines indicate the original expression measurements of PM, MM, and difference of PM and MM pair. **B.** Probe profiles of gene 3270 (U47621).

**A.**




**B.**

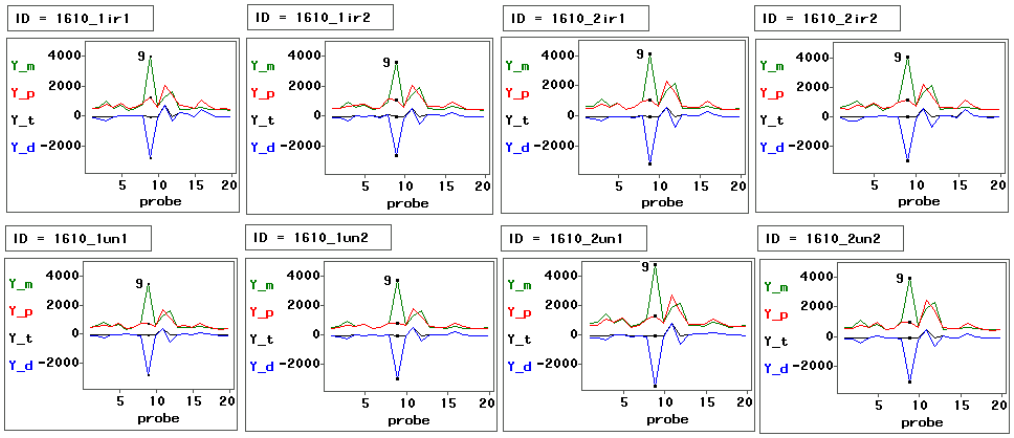



**Figure 2.7:** **A.** Scatter plots for comparing the negative log p-values (top row) and estimates (bottom row) of treatment effect among Model I, II, and III. Legend is scaled according to negative log p-value in Model I. Model I and II have very similar behavior whereas Model III has different behavior. **B.** Probe profile of gene 1610 (L42176). The green, red, black, and blue lines represent the expression profiles of MM, PM, truncated difference, and difference. The difference of probe pairs 2, 9, and 12 are converted to 10.

# Chapter 3

# Comparison of Li-Wong and Loglinear Mixed Models for the Statistical Analysis of Oligonucleotide Arrays

## 3.1 Introduction

The Affymetrix GeneChip$^{TM}$ is currently the most widely-used commercial expression array technology. A typical chip contains 16-20 oligonucleotide 25-mer probes for each of thousands of genes (Lockhart et al. 1996). Although each oligonucleotide within a probe set is designed to interrogate the same gene, there are well-known and very strong differences between the performance of individual probes. Statistical analysis and summarization of the probe-level data is therefore a critical challenge that must be addressed in order to effectively assess results from the chips.

Affymetrix itself has been responsive by providing a new summary measure based on Tukey's biweight function in the software accompanying the chips (www.affymetrix.com/product/). Although this certainly represents an

improvement over their earlier methods, by its nature the Affymetrix summary prevents analysts from making their own adjustments for individual probe effects. Research addressing this concern has arisen from numerous sources, including Teng et al. (1999), Schadt et al. (2000), Efron et al. (2001), Irizarry et al. (2001), Lemon et al. (2001), etc., and we forgo an attempt to provide a comprehensive review here.

What we do attempt to accomplish is to provide an empirical comparison between one of the first and most compelling approaches, that of Li and Wong (2001a,b), and the more recent recommendations of Chu et al. (2002). Li and Wong's models for the probe-level measurements are a combination of multiplicative and additive terms, and have been applied successfully in a number of contexts (references on http://www.biostat.harvard.edu/complab/dchip/references.htm). A key advantage of such an approach is a direct, parametric specification of important sources of variability in the chip data. This enables rigorous statistical quality control by automatically flagging observations that significantly deviate from the model, as well as quantified statistical inference about experimental effects. Chu et al. (2002) describe a systematic statistical linear modeling approach based on the well-known mixed linear model. These models have many similarities to those of Li and Wong (2001 a,b) and enjoy the same advantages, although they operate on the logarithms of the probe-level data.

So how do the Li-Wong and loglinear mixed models compare in practice? To investigate this question, we first consider a real data set from the literature

(Tusher et al., 2001) and select two specific forms of the models applicable for these data (Section 3.2). We then fit these two models to the data and compare them in terms of goodness-of-fit, normality diagnostics, model flexibility, and model robustness (Sections 3.3). We then conduct a simulation study to investigate statistical inference properties of the models, and propose a new test statistic for the Li-Wong model that provides better control of Type I error (Section 3.4). We conclude with some summarizing remarks (Section 3.5).

## 3.2 The Ionizing Radiation Data and Associated Li-Wong and Mixed Models

We consider the models in terms of the ionizing radiation data of Tusher et al. (2001). These data are from eight HUGeneFL GeneChips$^{TM}$, arising from two replicates of a 2 x 2 factorial design (two cell lines and two radiation treatments, see Table 3.1). Measurements for 6810 genes are available from each chip, and for each gene, 40 values corresponding to 20 perfect match and mismatch pairs arise from quantitation of a flourescently-derived image. The perfect match and mismatch values are typically subtracted to form a paired difference, although there is considerable debate about how the mismatch data should be incorporated. For simplicity of exposition and comparison, we consider only the 160 perfect-match data points for a single arbitrary gene, and do not perform any normalization on the data prior to analysis.

The Li-Wong model we consider is a combination of multiplicative and additive terms.

Model [LW] :

$$PM_{ijkl} = \nu_k + \theta_{ijl}\phi_k + \gamma_{ijkl}\ ,\ \Sigma\phi_k^2 = K.$$

Here, $PM_{ijkl}$ denotes the perfect match expression measurement of the $i^{th}$ cell line receiving the $j^{th}$ treatment at the $k^{th}$ probe in the $l^{th}$ replicate**.** The parameter $\nu_k$ is the baseline response of the $k^{th}$ probe, $\theta_{ijl}$ is an expression index for the different samples, $\phi_k$ is the multiplicative effect of the $k^{th}$ probe, and $\gamma_{ijkl}$ is the stochastic error term. The $\gamma_{ijkl}$'s are assumed to be independent and identically distributed normal random variables with mean 0 and variance $\omega^2$. A summation constraint is imposed for identifiability, with $K$ equals to the total number of probes within probe set. Nonlinear least square methods to fit [LW] are implemented in the dChip software (DNA-Chip Analyzer), although we use Proc NLP in SAS/OR (SAS Institute Inc. 1999a) for results in this report. Various hypothesis tests and confidence intervals can be constructed using the parameter estimates and their estimated standard errors (Li and Wong 2001b).

The mixed analysis-of-variance model we consider here, based on Chu et al. (2002), is as follows.

Model [MM]:

$$\log_2(PM_{ijkl}) = L_i + T_j + LT_{ij} + P_k + LP_{ik} + TP_{jk} + A_{l(ij)} + \varepsilon_{ijkl.}$$

The indices are the same as before, and the symbols *L, T, LT, P, LP, TP,* and *A* represent cell line, treatment, cell line-treatment interaction, probe, cell line-probe interaction, treatment-probe interaction, and chip effects, respectively. They are the analogs of the $\nu$, $\theta$, $\phi$ parameters in [LW]. The $A_{l(ij)}$'s are assumed to be independent and identically distributed normal random effects with mean *0* and variance $\sigma_a^2$, and induce a common correlation across all observations on the same chip. The $\varepsilon_{ijkl}$'s are assumed to be independent and identically distributed normal random variables with mean *0* and variance $\sigma^2$, and are independent of the $A_{l(ij)}$'s. We fit [MM] using the method of restricted maximum likelihood (REML) with Proc Mixed in SAS/STAT (SAS Institute Inc. 1999b).

Note that [MM] has 65 degrees of freedom in its mean model, whereas [LW] has 47. We could have easily adjusted either model to make these numbers closer to each other, but we reckoned the models as specified represent what analysts might use as typical starting points in practice.

## 3.3 Results from Ionizing Radiation Data

### 3.3.1 Goodness-of-fit

Figure 3.1A shows the regression of the predictions from [LW] on those from [MM] from the fits of both models to the data from each of the 6810 genes. Here, the predictions from [MM] are exponentiated by 2 in order to have the same scale

with the predictions from [LW]. Predictions from both models are consistently close except for a few points in the lower off-diagonal region in Figure 3.1A.

Figure 3.1B is a scatter plot comparison for the $R^2$ values from the model fits. The $R^2$ is defined as

$$R^2 = 1 - \frac{\sum R_{ijkl}^2}{\sum (PM_{ijkl} - \overline{PM})^2} \ ,$$

where $\overline{PM}$ is the average of all $PM_{ijkl}$'s and $R_{ijkl}$ is the residual term of $PM_{ijkl}$ from the model. The lower $5^{th}$ percentile of $R^2$ equals 0.96 for both models, indicating very good fit in most of cases. For a few genes the $R^2$ values of Model [MM] are much less than the $R^2$ values of [LW] for some genes. We inspected the expression profiles of those 66 genes that have $R^2$ values for [MM] less than 0.95 and at least 0.1 less than the respective $R^2$ values for [LW]. All of the 66 cases have outlying observations.

Figure 3.2A is an example for gene number 3096 (U35451), which has [LW] $R^2$=0.989 and [MM] $R^2$=0.171. The red, blue, and black-dotted curves represent the predictions from [LW], the exponentiated predictions from [MM], and the raw PM expression intensities, respectively. An apparent PM outlier is for probe 18 in array "112" (note the change in scale in the y-axis for this array's plot). [LW] fits this outlier very well whereas [MM] does not, leading to the discrepancies in $R^2$. This is a direct result of the model specifications, since [LW] has distinct parameters for individual replicates whereas [MM] averages across replicates. We consider that averaging is more appropriate in this case, as it

allows one to automatically detect this outlier and remove it from the analysis. Figure 3.2B for gene number 1860 (M25753) represents a converse situation, for which [LW] $R^2$=0.987 and [MM] $R^2$=0.991. Note the PM profiles of the probes are highly consistent throughout all eight arrays except for probe 1, which is relatively low in untreated arrays but relatively high in treated arrays. This is a treatment-by-probe interaction, and is captured by [MM] but not by [LW]. Using a conservative Bonferroni correction for multiple testing, 9.34% and 0.59% of the probe sets have 0.05-level significant cell-line-by-probe and treatment-by-probe interactions, respectively, according to [MM].

### 3.3.2 Normality Diagnostics and Outlier Detection

Figures 3.3A and 3.3B show standardized residual plots for all of the model fits of [LW] and [MM]**,** respectively. The standardized residuals are the residuals divided by the square root of estimated variances. For [LW], the adjustment for degrees of freedom, the number of parameters to be estimated, we apply here is larger than it is from conditional approach. This causes the unconditional estimate of variance to be larger. In Figure 3.3A, the plot looks like a flask along the X-axis. In Figure 3.3B, we see a "submarine" in the plot and a few bubbles on the top or bottom. Note we have left each plot in its original scale, because automatic outlier detection is typically done by selecting standardized residuals having a magnitude greater than 3. With this outlier criterion, the proportions of outliers for all genes are 0.567% and 0.077% for [LW] and [MM], respectively.

Figure 3.3C shows the single-gene standardized residual plots of eight selected genes from the 12 genes reported to respond transcriptionally to ionizing radiation (Tusher et al., 2001). These eight genes are numbers 1154 (J05614), 1860 (M25753), 2511 (S78187), and 2715 (U09579) from left to right on the top and numbers 2863 (U18300), 4370 (X62048), 4598 (X77794), and 5469 (X63717) from left to right on the bottom. The red and blue points represent the standardized residuals from [LW] and [MM] respectively. Note [MM] generally has smaller standardized residuals, an observation undoubtedly due in part to the fact that they are derived on the log scale.

We also applied the Komogorov-Smirnov test with Bonferroni adjustment to formally test the standardized residuals of each gene for normality. Almost all genes (97.47%) pass the test in [MM], but only three fourths of genes (74.24%) pass the test in [LW].

## 3.4 A Simulation Study

To investigate the inferential operating characteristics of [LW] and [MM], we conducted a simulation study using parameter values estimated from data for gene number 2863 (U18300) from the ionizing radiation data. This gene has $R^2$ values of 0.949 and 0.986 from [LW] and [MM], respectively. The corresponding log-transformed estimates of the random components are -1.12, -1.10, and 5.17 for $\hat{\sigma}^2, \hat{\sigma}_a^2,$ and $\hat{\omega}^2$, respectively. Figure 3.4A shows the histograms and Figure 3.4B

shows the expression profiles. We simulated data from both models and compared the results.

To conduct statistical analyses according [LW], we use the confidence interval approach as described in Li and Wong (2001b). As noted in Li and Wong (2001a), this conditional method should be valid for a large number of arrays since the probe-specific parameters, $\nu_j$'s and $\phi_j$'s, can be estimated accurately. However, for small experiments like the ionization radiation data, an unconditional approach may be more appropriate. To this end, we also investigate a Wald Z statistic (Wald 1943) for [LW]**,** as derived in the Appendix. To perform statistical analyses according to [MM], it is convenient to use the t-tests from the SAS Mixed procedure (Searle et al. 1993; Littell et al. 1996; SAS Institute Inc. 1999b).

### 3.4.1 Scenario 1 – [LW] is the True Model

The 10 simulation cases were created as follows:

1. Obtain the estimates of $\nu_k$'s, $\phi_k$'s, $\theta_i$'s, and $\omega^2$ by applying the NLP procedure to real data, and set $\theta_\mu$ equal to be the average of all the $\theta_i$ estimates.
2. Set the scale parameter $\delta$, which represents the fold change, to 10 different values from 1 to 2.
3. Set parameters $\theta_1$, $\theta_2$, $\theta_5$, and $\theta_6$ equal to $\theta_\mu$ and set parameters $\theta_3$, $\theta_4$, $\theta_7$, and $\theta_8$ equal to $\delta \times \theta_\mu$.

4. Use all estimates of $\nu_k$'s, $\phi_k$'s, $\omega^2$ and the new parameters $\theta_i$'s in step 3 as the true parameters for [LW].
5. Change the scale parameter $\delta$ and generate 2000 simulated datasets in each case.

These settings create 10 simulation cases of different fold changes, ranging from one to two, between treatment groups and no changes between cell line groups. Therefore, testing the cell line effects allows us examine how well the tests control Type 1 error (false positive rate). Testing the treatment effects allows us to simultaneously examine power. We set the nominal significance level for the tests to 0.05.

### 3.4.2 Scenario 2 – [MM] is the True Model

In order to have 10 cases comparable with Scenario 1, we set the [MM] parameters as follows:

1. Obtain the estimates for $P_k$'s, $\sigma^2$, and $\sigma_a^2$ by applying the Mixed procedure to the logarithms (base 2) of the real data.
2. Set the scale parameter $\delta$ to 10 different values from 1 to 2.
3. Set the parameters $L_i$'s, $LT_{ij}$'s, $LP_{ik}$'s, and $TP_{jk}$'s to 0;
4. Set parameter $T_1$ to 0 and parameter $T_2$ to $\log_2(\delta)$;
5. Put all estimates and parameters in steps 1, 3, and 4 into [MM];
6. Change the scale parameter and generate 2,000 simulated data sets in each case.

Again, the line effects are tested to examine the false positive rate, and treatment effects are tested to examine power.

### 3.4.3 Simulation Results

Table 3.2 and the left three plots in Figure 3.5 show the results from Scenario 1. Table 3.3 and the right three plots in Figure 3.5 show the results from Scenario 2. The red, green, and blue curves in Figure 3.5 indicate the results from the confidence interval tests in [LW], the Wald Z tests in [LW], and tests in [MM], respectively. From the top two plots in Figure 3.5, the tests performed by [MM] almost exactly control the significance level (0.05) no matter which model is true. The [LW] confidence interval approach controls significance level poorly; its 95% confidence band does not include 0.05 in all 10 cases where [LW] is true. When [MM] is true, the simulated false positive rate is even worse and exceeds 0.2 for larger fold changes. The Wald $Z$ in [LW] appropriately controls the significance level under [LW], but creeps up above 0.15 when [MM] is true.

The middle two plots of Figure 3.5 show power curves for testing treatment fold change. These plots of the [LW] confidence interval tests is not comparable here because these tests are too liberal in their Type 1 error rates. The [MM] and [LW] Wald Z have similar power under the Scenario 1, with power near 100% for a small 1.3 fold change. This implies that [MM] can perform well as [LW] even when [LW] is the true model. In the middle right plot of Figure 3.5, the [MM] tests have the best power, as expected since [MM] is the true model,

but true even though the [LW] Wald $Z$ is liberal in this scenario. The bottom two plots in Figure 3.5 show the average $R^2$ values for the two models in the 10 simulation cases. Under the first scenario, the average $R^2$ value for [MM] is about 0.014 lower than the average $R^2$ value for [LW]. Under the second scenario, the average $R^2$ value for [MM] is about 0.005 higher than the average $R^2$ value for [LW]. This result emphasizes again that both models have very similar goodness-of-fit.

## 3.5 Discussion

We have empirically compared instances of two statistical modeling approaches, Li-Wong and log-linear mixed models, for oligonucleotide expression array data at the probe level. Both models fit our example data very well, and both appear to have great ability to capture the key measurable sources of variability of oligonucleotide arrays. In our real-data example, we did find genes that differentiated the two methods, and were able to provide reasonable explanations for the differences based on the particular models we selected. Of course both models can be adjusted to accommodate different kinds of data patterns, but the mixed model is operationally more convenient because of its linearity and rich tradition in handling sources of variation expressed as effects and their interactions. The mixed model is also easily applicable to experimental designs more complex than simple factorials, whereas the basic Li-Wong framework is nonlinear and works only with a one-way analysis-of-variance treatment structure.

The two models also naturally differ in their outlier selection criteria because they are on different scales. According to our simulations, the conditional Li-Wong confidence interval method has excessive Type 1 error rates, but these can be corrected by using larger, unconditional standard errors. The mixed model performed well in all cases, even when the simulated data arose from a Li-Wong model.

**Table 3.1:** Experimental design for the ionizing radiation data.

| Chip | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| Cell Line | I | I | I | I | II | II | II | II |
| Treatment | ir | ir | un | un | ir | ir | un | un |
| Replicate | 1 | 2 | 1 | 2 | 1 | 2 | 1 | 2 |

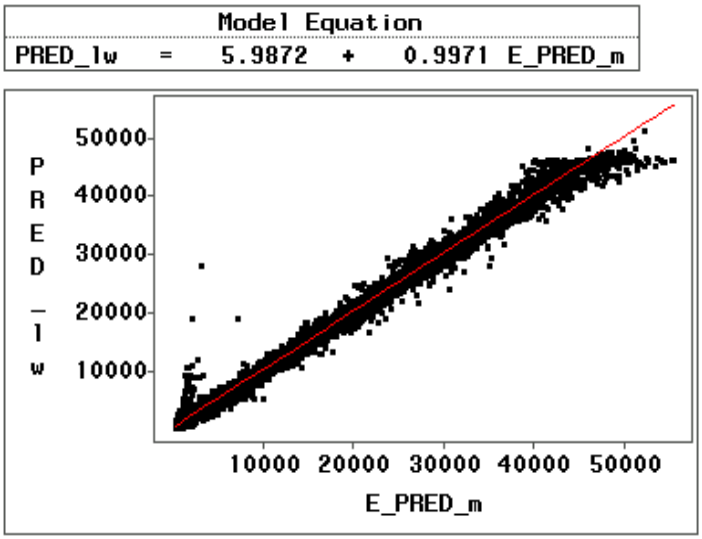
**Table 3.2:** Testing results assuming [LW] as true model - the rejection rates of 2,000 simulations for each case. The capital characters "M", "CI", and "W" indicate the test is perform by [MM], [LW] confidence interval, and [LW] Wald Z approaches, respectively. The characters "l", "t", "lt", "p", "lp", and "tp" indicate the main or interaction effects of line, treatment, line-treatment, probe, line-probe, and treatment-probe respectively. $R^2$_MM and $R^2$_LW indicate the average $R^2$ values for models [MM] and [LW] respectively.

| | [MM] | | | | | | [LW] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fold** | **M_l** | **M_t** | **M_lt** | **M_p** | **M_lp** | **M_tp** | **CI_l** | **CI_t** | **W_l** | **W_t** | **$R^2$_MM** | **$R^2$_LW** |
| 1 | .049 | .041 | .035 | 1 | .089 | .110 | .071 | .074 | .044 | .040 | .921 | .932 |
| 1.05 | .049 | .156 | .035 | 1 | .090 | .112 | .072 | .514 | .040 | .199 | .924 | .934 |
| 1.1 | .050 | .468 | .035 | 1 | .092 | .120 | .075 | .976 | .036 | .440 | .926 | .936 |
| 1.15 | .048 | .779 | .036 | 1 | .092 | .135 | .084 | 1 | .030 | .581 | .928 | .939 |
| 1.2 | .050 | .930 | .036 | 1 | .092 | .161 | .098 | 1 | .031 | .743 | .930 | .942 |
| 1.3 | .051 | .997 | .037 | 1 | .095 | .236 | .121 | 1 | .047 | .994 | .935 | .950 |
| 1.4 | .051 | 1 | .038 | 1 | .094 | .334 | .127 | 1 | .054 | 1 | .939 | .956 |
| 1.5 | .051 | 1 | .038 | 1 | .096 | .473 | .118 | 1 | .055 | 1 | .943 | .960 |
| 1.75 | .053 | 1 | .041 | 1 | .098 | .786 | .110 | 1 | .052 | 1 | .951 | .969 |
| 2 | .053 | 1 | .043 | 1 | .097 | .935 | .106 | 1 | .053 | 1 | .958 | .975 |

**Table 3.3:** Testing results assuming [MM] as true model - the rejection rates of 2,000 simulations for each case. Notation is the same as in Table 3.2.

| | [MM] | | | | | | [LW] | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Fold** | **M_l** | **M_t** | **M_lt** | **M_p** | **M_lp** | **M_tp** | **CI_l** | **CI_t** | **W_l** | **W_t** | **$R^2$_MM** | **$R^2$_LW** |
| 1 | .055 | .056 | .058 | 1 | .049 | .051 | .189 | .208 | .093 | .108 | .977 | .971 |
| 1.05 | .049 | .855 | .055 | 1 | .049 | .055 | .203 | .954 | .085 | .598 | .977 | .971 |
| 1.1 | .056 | 1 | .051 | 1 | .050 | .046 | .207 | 1 | .092 | .775 | .978 | .971 |
| 1.15 | .065 | 1 | .048 | 1 | .048 | .042 | .199 | 1 | .103 | .916 | .978 | .972 |
| 1.2 | .043 | 1 | .048 | 1 | .054 | .052 | .205 | 1 | .12 | .965 | .979 | .972 |
| 1.3 | .063 | 1 | .046 | 1 | .053 | .058 | .207 | 1 | .14 | .993 | .98 | .974 |
| 1.4 | .058 | 1 | .051 | 1 | .048 | .044 | .211 | 1 | .144 | 1 | .981 | .976 |
| 1.5 | .046 | 1 | .057 | 1 | .051 | .045 | .247 | 1 | .173 | 1 | .982 | .978 |
| 1.75 | .043 | 1 | .043 | 1 | .052 | .05 | .224 | 1 | .154 | 1 | .985 | .981 |
| 2 | .055 | 1 | .06 | 1 | .054 | .043 | .241 | 1 | .166 | 1 | .987 | .984 |

**A.**




**B.**

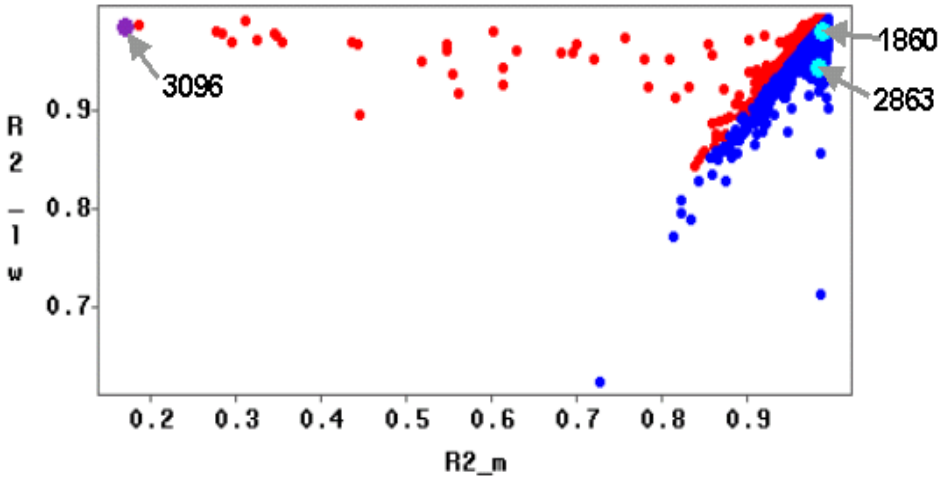



**Figure 3.1: A.** Scatter and regression plot of predictions from [LW] and [MM]. "E_PRED_m" represents exponentiated prediction ($2^{PRED\_m}$) from [MM]. "PRED_lw" represents prediction from [LW]. **B.** Scatter plot of $R^2$ comparison. Blue indicates [MM] has better $R^2$ (49.9%). Red indicates [LW] has better $R^2$ (50.1%).

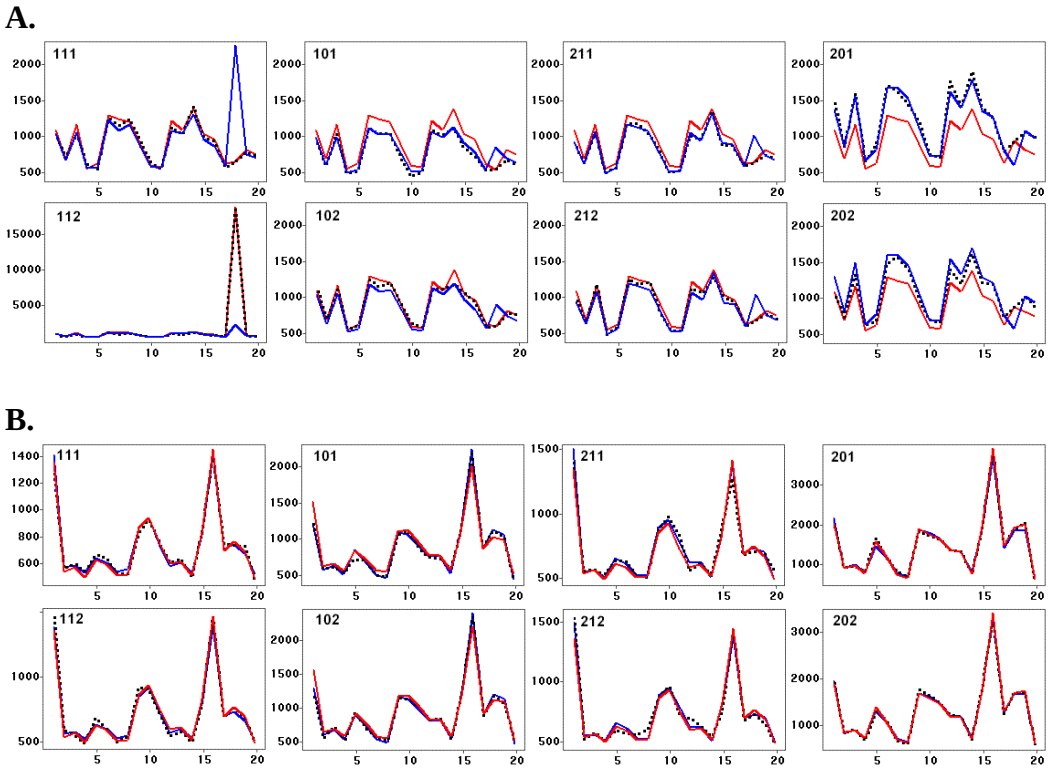
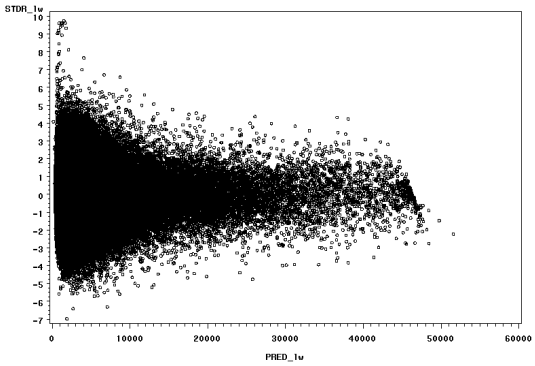
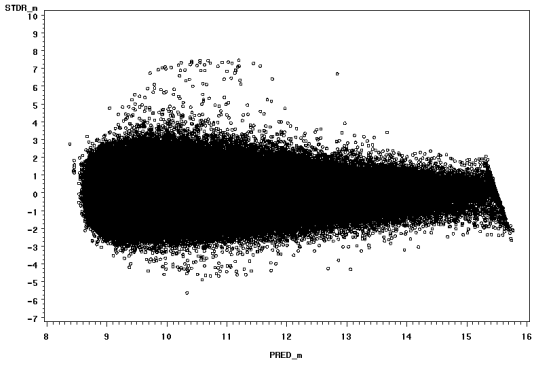



**Figure 3.2:** **A.** Expression profiles from gene 3096 (U35451). Red, blue, and black-dotted curves indicate the predictions from [LW], exponentiated predictions from [MM], and PM, respectively. X-axis represents for probe no. The 3 digits on each plot indicate the experimental condition applied to GeneChip. The first digit indicates the 2 different cell lines (1 or 2). The second digit indicates the treatment condition (1: treated, 0: untreated). The third digit indicates the replicate number (1 or 2). Note the change in scale in the y-axis for array 112; others are consistent. **B.** Profiles from gene 1860 (M25753).

**A.**




**B.**




**C.**

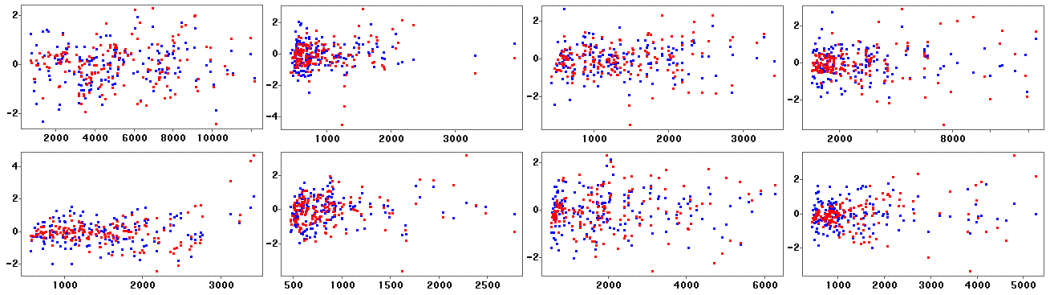
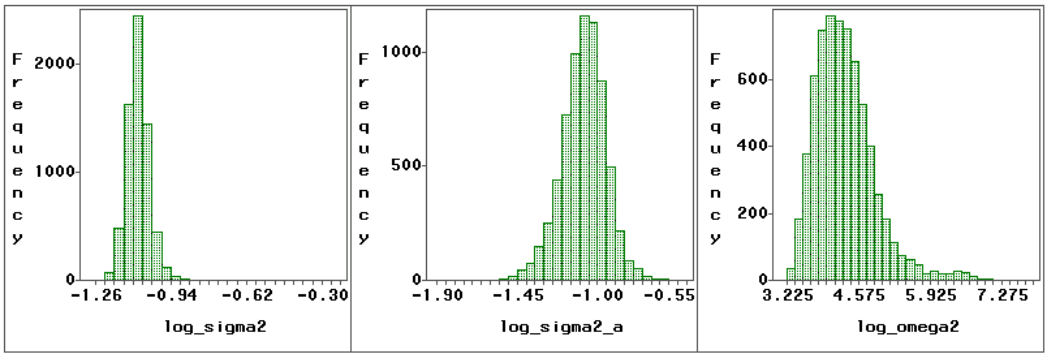


**Figure 3.3: A.** Pooled standardized residual plots of [LW]. X-axis represents for predictions. **B.** Pooled standardized residual plots of [MM]. **C.** Standardized residual plots of 8 genes. Genes (Accession Numbers) on the top row are 1154 (J05614), 1860 (M25753), 2511 (S78187), and 2715 (U09579), and those on the bottom row are 2863 (U18300), 4370 (X62048), 4598 (X77794), and 5469 (X63717). Red and blue dots indicate residuals from [LW] and [MM] respectively.

**A.**




**B.**

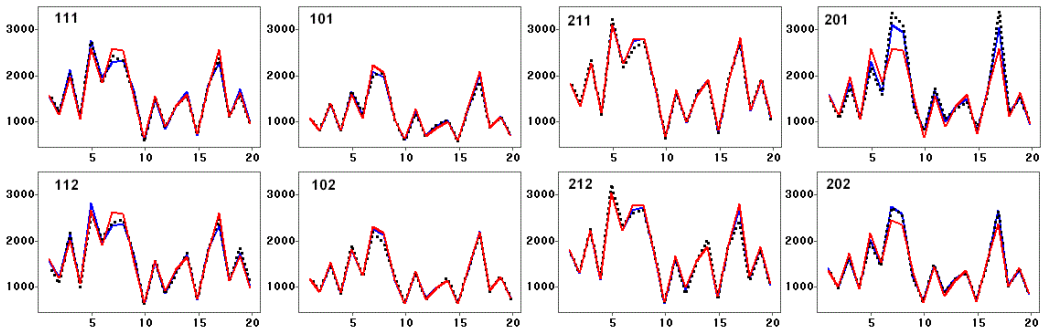



**Figure 3.4:** **A.** Histograms of the logarithm transformed estimates of the random components from [MM] and [LW]. "sigma2", "sigma2_a", and "omega2" indicate $\hat{\sigma}^2, \hat{\sigma}_a^2,$ and $\hat{\omega}^2$, respectively. **B.** Expression profiles from gene 2863 (accession number U18300).

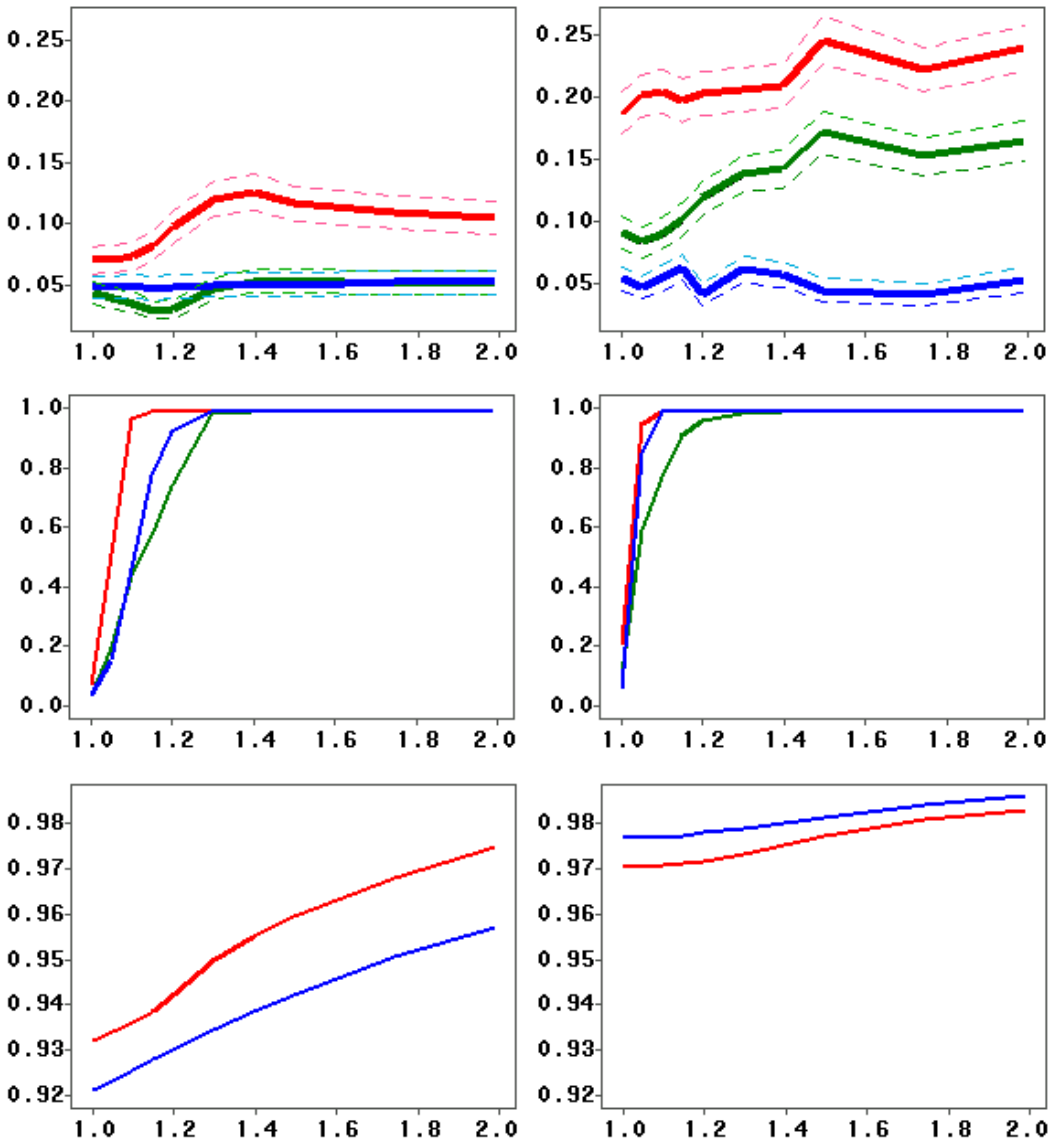



**Figure 3.5:** Comparison of simulation results for scenarios 1 and 2. The top two plots compare the Type 1 Errror (false positive) rate for testing line effect. The dashed curves are 95% confidence bands. The middle two plots compare power for testing the treatment effect. The bottom two plots compare average $R^2$ values. The three plots on the left are results from the simulation using [LW] as the true model, and the three plots on the right are those using [MM] as the true model. Blue curves represent the results from [MM], red curves from [LW], and green curves from the proposed [LW] Wald Z. X-axes in all six plots represent for $\delta$ value (treatment fold change). Y-axes in the top 4 plots represent for testing rejection rate. Y-axes in the bottom 2 plots represent for $R^2$ value.

# References


Bishop, J. F. (1997). The treatment of adult acute myeloid leukemia. *Semin. Oncol.* **24**(1), 57-69.

Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomedical Optics* **2**(4), 364-374.

Chu, T., Weir, B., and Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math. Biosci.* **176**, 35-51.

Cripe, L. D. (1997). Adult acute leukemia. *Curr. Probl. Cancer* **21**(1), 1-64.

DNA-Chip Analyzer (dChip). http://www.biostat.harvard.edu/complab/dchip/

Dudoit, S., Fridlyand, J., and Speed, T. P. (2000a). Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report #576, University of California, Berkeley.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2000b). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical Paper, University of California, Berkeley.

Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000). Microarrays and their use in a comparative experiment, Technical Report, Stanford University.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher V. (2001). Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Asso.* **96**, 1151-1160.

Golub, T. R., Slonim, D. K., Tamauo, P., Huard, C., Hassenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999). Molecular classification of cancer: discovery of class prediction by gene expression monitoring. *Science* **286**, 531-537.

Irizarry, R. A., Hobbs, B., and Speed, T. (2001). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. (http://oz.berkeyley.edu/users/terry/zarray/Affy/GL_Worksop/genelogic2001.html)

Jin, W., Riley, R., Wolfinger R. D., White K. P., Passador-Gurgel, G., and Gibson G. (2001). Contributionss of sex, genotype and age to transcriptional variance in Drosophila melanogaster, *Nature Genetics*, in press.

Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis*, 3rd edition. Prentice Hall.

Kerr, M. K., Martin, M., and Churchill G. A. (2000). Analysis of variance for gene expression microarray data, *J. Com. Bio.* **7**(96), 819-837.

Lazaridis, E. N., Sinibaldi, D., Bloom, G., Mane, S., and Jove, R. (2001). A simple method to improve probe set estimates from oligonucleotide arrays, University of South Florida (2001).

Lemon, W. J., Palatini, J. J. T., Krahe,R., and Wright , F. A. (2001). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. (http://thinker.med.ohio-state.edu/projects/fbss/index.html)

Li, C. and Wong, W. H. (2001a). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Nat. Acad. Sci. USA* **98**(1), 31-36.

Li, C. and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol* **2**(8), research0032.1-0032.11.

Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature Genetics Sup.* **21**, 20-24.

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models.* SAS Institute Inc., Cary, NC.

Lockhart, D., Dong, H., Byrne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotech.* **14**, 1675-1680.

McConkey, E. H. (1993). *Human Genetics the Molecular Revolution*. Jones and Barlett Publishers, Boston.

McCulloch, C. E. and Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York, NY.

McGall, G., Labadie, J., Brock, P., Wallraff, G., Nguyen, T., and Hinsberg, W. (1996). Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc. Nat. Acad. Sci. USA* **93**, 13555-13560.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Bio.*, **8**, 37-52.

Sarle, W. S. (1983). Cubic Clustering Criterion. SAS Technical Report, A-108.

SAS Institute Inc. (1999a). *SAS/OR Software Version 8*. SAS Institute, Inc., Cary, NC.

SAS Institute Inc. (1999b). *SAS/STAT Software Version 8*. SAS Institute, Inc., Cary, NC.

Schadt, E. E., Li, C., Su, C., and Wong, W. H. (2000). Analyzing high-density oligonucleotide gene expression array data. *J. Cell Bioche.* **80**(2), 192-202.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**. 467-470.

Searle, S., Casella, G., and McCulloch, C. (1993). ) *Variance Components*. Wiley, New York.

Steel, R. G. D, Torrie, J. H., and Dickey, D. A. (1997). *Principles and Procedures of Statistics: A Biometrical Approach, Third Edition*. McGraw-Hill Inc., New York.

Teng, Chi-Hse, Nestorowicz, A., and Reifel-Miller A. (1999). Experimental designs using Affymetrix GeneChips. *Nature Genetics* **23**, 78, **DOI: 10.1038/14415** Poster Abstracts.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Nat. Acad. Sci. USA* **98**, 5116-5121.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. Springer, New York, NY.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* **54**, 426-482.

Westfall, P.H., Zaykin, D.V., and Young, S.S. (2001). Multiple tests for genetic effects in association studies. *Methods in Molecular Biology*, vol. 184: Biostatistical Methods, pp. 143-168. Stephen Looney, Ed., Humana Press, Toloway, NJ.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed model. *J. Comp. Bio.*, **8**, 625-637.

Wrzesien-Kus, A. and Krykowski, E. (1997). Treatment of acute lymphoblastic leukemia in adults. *Przegl Lek* **54**(9), 639-646.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cDNA microarray data. Technical paper. University of California, Berkeley.

# Appendix

Li and Wong (2001b) suggest a confidence interval for fold change based on the Q statistic being distributed as chi-square. This can be extended for testing treatment effects. Suppose $\tilde{\theta}_i \sim N(\theta_i, \delta_i^2)$, $i=1,2,\ldots,I$. Here $\tilde{\theta}_i$, $i=1,2,\ldots,m$ are the MBEI for treatment group1 and the remaining $\tilde{\theta}_i$'s are the MBEI for treatment group 2. Write

$$\tilde{\eta}_1 = (\frac{1}{m})\sum_1^m \tilde{\theta}_i \quad .$$
$$\tilde{\eta}_2 = (\frac{1}{I-m})\sum_{m+1}^I \tilde{\theta}_i \quad .$$

Since $\tilde{\theta}_i$'s are independent:

$$\tilde{\eta}_i \sim N(\eta_i, \tau_i^2), \; i=1,2,$$

where

$$\eta_1 = \frac{1}{m}\sum_1^m \theta_i,$$
$$\eta_2 = \frac{1}{I-m}\sum_{m+1}^I \theta_i,$$

$$\tau_1^2 = \frac{1}{m^2}\sum_1^m \delta_i^2,$$
$$\tau_2^2 = \frac{1}{(I-m)^2}\sum_{m+1}^I \delta_i^2.$$

Define $r = \eta_1/\eta_2$. Then the 1 df chi-square statistic is

$$Q = \frac{(\tilde{\eta}_1 - r\tilde{\eta}_2)^2}{\tau_1^2 + r^2\tau_2^2}.$$

For a fixed significance level, a confidence interval for *r* can be obtained and be used for testing treatment effects. When the confidence interval for *r* does not include 1, the treatment effect is significant. This implies that the corresponding gene has significantly different expressions between treatment groups. A Wald statistic can also be used to test whether $\eta_1/\eta_2 = 1$. However, it is more common to test whether $\eta_1-\eta_2 = 0$. According to [LW], the log-likelihood function

$$L(\nu,\phi,\theta,\omega^2) = IK \log(\frac{1}{\sqrt{2\pi\omega^2}}) - \frac{1}{2\omega^2}\sum_i\sum_k (PM_{ik} - \nu_k - \theta_i\phi_k)^2 .$$

For illustration purpose, only one subscript, i, is used for chip index. The MLEs can be solved by minimizing numerically the least square function, the second term of above equation. The variance of the MLEs can be approximated from the inverse of the observed information matrix. Let $\boldsymbol{\beta} = (\nu,\phi,\theta,\omega^2)^t$. Then, an explicit form of the approximated covariance matrix can be derived.

$$Cov(\hat{\beta}) \approx \hat{C} = \left[\frac{-\partial^2}{\partial\beta\,\partial\beta^t} L(\hat{\nu},\hat{\phi},\hat{\theta},\hat{\omega}^2)\right]^{-1},$$

$$= -\begin{bmatrix} \frac{\partial^2 L}{\partial\nu\,\partial\nu^t} & \frac{\partial^2 L}{\partial\phi\,\partial\nu^t} & \frac{\partial^2 L}{\partial\theta\,\partial\nu^t} & \frac{\partial^2 L}{\partial\omega^2\,\partial\nu^t} \\ \frac{\partial^2 L}{\partial\nu\,\partial\phi^t} & \frac{\partial^2 L}{\partial\phi\,\partial\phi^t} & \frac{\partial^2 L}{\partial\theta\,\partial\phi^t} & \frac{\partial^2 L}{\partial\omega^2\,\partial\phi^t} \\ \frac{\partial^2 L}{\partial\nu\,\partial\theta^t} & \frac{\partial^2 L}{\partial\phi\,\partial\theta^t} & \frac{\partial^2 L}{\partial\theta\,\partial\theta^t} & \frac{\partial^2 L}{\partial\omega^2\,\partial\theta^t} \\ \frac{\partial^2 L}{\partial\nu\,\partial\omega^2} & \frac{\partial^2 L}{\partial\phi\,\partial\omega^2} & \frac{\partial^2 L}{\partial\theta\,\partial\omega^2} & \frac{\partial^2 L}{\partial(\omega^2)^2} \end{bmatrix}^{-1}.$$

The diagonal terms of the matrix above are:

$$\frac{\partial^2 L}{\partial \nu \, \partial \nu^{\mathrm{t}}} = \frac{-I}{\hat{\varpi}^2} I_K \, .$$

$$\frac{\partial^2 L}{\partial \phi \, \partial \phi^{\mathrm{t}}} = \frac{-\sum_i \hat{\theta}_i^2}{\hat{\varpi}^2} I_K \, .$$

$$\frac{\partial^2 L}{\partial \theta \, \partial \theta^{\mathrm{t}}} = \frac{-\sum_k \hat{\phi}_k^2}{\hat{\varpi}^2} I_I \, .$$

$$\frac{\partial^2 L}{\partial (\omega^2)^2} = \frac{IK}{2(\hat{\varpi}^2)^2} - \frac{1}{(\hat{\varpi}^2)^3} \sum_{ik} (PM_{ik} - \hat{\nu}_k - \hat{\theta}_i \hat{\phi}_k)^2 \, .$$

Here, $I_K$ and $I_I$ are identity matrices with size $K$ and $I$. The terms below the diagonal of the covariance matrix are:

$$\frac{\partial^2 L}{\partial \nu \, \partial \phi^t} = \frac{-\sum \hat{\theta}_i}{\hat{\varpi}^2} I_K \, .$$

$$\frac{\partial^2 L}{\partial \nu \, \partial \theta^t} = \frac{-1}{\hat{\varpi}^2} \begin{bmatrix} \hat{\phi}_1 & \hat{\phi}_2 & \cdots & \hat{\phi}_K \\ \hat{\phi}_1 & \hat{\phi}_2 & \cdots & \hat{\phi}_K \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\phi}_1 & \hat{\phi}_2 & \cdots & \hat{\phi}_K \end{bmatrix}_{I*K} \, .$$

$$\frac{\partial^2 L}{\partial \nu \, \partial \varpi^2} = \frac{-1}{(\hat{\varpi}^2)^2} \left[ \sum_i (PM_{ik} - \hat{\nu}_k - \hat{\theta}_i \hat{\phi}_k) \right]_{1*K} \, .$$

$$\frac{\partial^2 L}{\partial \phi \, \partial \theta^t} = \frac{1}{(\hat{\varpi}^2)^2} \left[ PM_{ik} - \hat{\nu}_k - 2\hat{\theta}_i \hat{\phi}_k \right]_{I*K} \, .$$

$$\frac{\partial^2 L}{\partial \phi \, \partial \varpi^2} = \frac{-1}{(\hat{\varpi}^2)^2} \left[ \sum_i \hat{\theta}_i (PM_{ik} - \hat{\nu}_k - \hat{\theta}_i \hat{\phi}_k) \right]_{1*K} \, .$$

$$\frac{\partial^2 L}{\partial \theta \, \partial \varpi^2} = \frac{-1}{(\hat{\varpi}^2)^2} \left[ \sum_k \hat{\phi}_k (PM_{ik} - \hat{\nu}_k - \hat{\theta}_i \hat{\phi}_k) \right]_{1xI} \, .$$

The terms above the diagonal are the transposes of the corresponding terms below the diagonal. Then,

$$h(\beta) \equiv \frac{1}{m} \sum_{i=1}^{m} \theta_i - \frac{1}{I-m} \sum_{i=m+1}^{I} \theta_i = \eta_1 - \eta_2 \, .$$

$$H(\beta) \equiv \frac{\partial h(\beta)}{\partial \beta^t} = (0,0,\cdots,0,\frac{1}{m},\cdots,\frac{1}{m},\frac{-1}{I-m},\cdots,\frac{-1}{I-m},0) \, .$$

Then, the Wald $Z$ statistic

$$W = \frac{h(\hat{\beta})}{[H(\hat{\beta})\hat{C}H(\hat{\beta})^t]^{-1/2}} \sim N(0,1).$$