Applying mixed models to assess gene significance with microarray data.

By

LOVENESS NYARADZO DZIKITI

Submitted in partial fulfillment of the requirements for the Master of Science degree in Mathematical Statistics in the Faculty of Natural & Agricultural Science

University of Pretoria
Pretoria

October 2005

INTRODUCTION 3
BACKGROUND 3
MICROARRAY EXPERIMENTS 3
NOTATION 5
INTRODUCTION 7
EXPERIMENTAL DESIGN 7
DATA NORMALISATION 8
FOLDCHANGE 9
T-TEST 10
LSMEAN 10
BONFERRONI ADJUSTMENT 10
VOLCANO PLOTS 12
MIXED MODELS 13
Assumptions of the Mixed Model 14
ESTIMATION OF PARAMETERS 15
Under Normality Conditions 16
Variance Components 19
Method of Moments 19
Maximum Likelihood estimation 23
COMMENTS 28
INTRODUCTION 29
EXPLORATORY ANALYSIS 29
Example 30
THE MODEL 33
CONCLUSION 41
Appendix P1: Dataupdate 42
Appendix P2: ARRAYCALC 46
Appendix P3: LOGICALULATION 48
Appendix P4: FREQUECYPROG 49
Appendix P5: GLMPROG 52
APPENDIX A1: Frequency procedure 55
APPENDIX A2: Use of BY option and Class statement in calculating gene models 58
REFERENCE 64

CHAPTER 1

INTRODUCTION

In chapter1, we give a brief introduction to micro-array experiments and the analysis of micro-array data. The four main steps usually followed when analyzing micro array data are:
1. Extracting spot intensities using a scanner.
2. Filtering out poor quality data.
3. Performing normalization.
4. Statistical analysis.

We also introduce the notation used in this paper. In Chapter two, we give an overview of the methods traditionally used to analyze micro array data. We then discuss some aspects of the mixed model which is the method suggested by Wolfinger et al. The third chapter deals with an application of the mixed model to analyze micro array gene expression data. Analysis of gene expression data seeks to identify genes that express themselves, (behave) differently under different conditions or treatments. Some output from the analysis and programs used to prepare and analyze the data are given in the appendix.

BACKGROUND

Micro-array analysis can reveal gene activities associated with biological processes, notably, large-scale expression analysis has revealed genes associated with disease states like cancer, informed the design of new methods of diagnosis and provided molecular targets for drug development. It also provides the opportunity to screen for a very broad spectrum of pathogens in a relatively short time. However, this is a very expensive exercise and the need to quantify and minimize error cannot be over stressed.

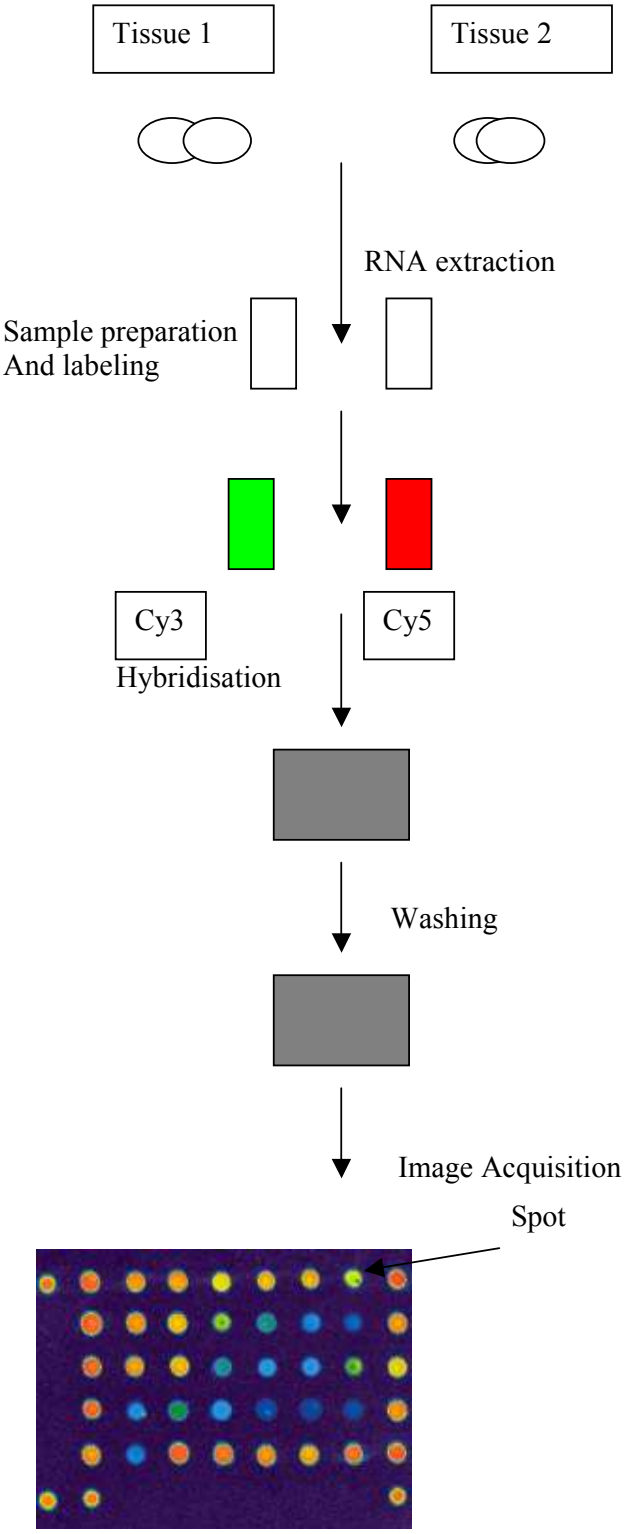
A DNA micro-array is a large number of genes or expressed sequence tags in a condensed array, on a solid face. The solid face can be glass slides, nylon slides or affymetrix chips. The latter two are very reliable but very expensive. Glass slides are usually used. The primary objective of a micro-array experiment is to look for differential expression, that is, changes in gene expression across a single factor of interest. For example;

- Different time points of a biological process.
- Different types of tissues.
- Different drug or stress treatments.

In this study, interest lies in the changes in gene expression across different drug treatments.

MICROARRAY EXPERIMENTS

A micro array experiment usually follows the sequence in the schematic diagram below:



Diagnostic microarray for drug susceptibility testing (RIF) in M. tuberculosis. Specimen courtesy of Drs. V. Mikhailovich and A. Mirzabekov, Centre for Biological Microchips, Englehardt Institute of Molecular Biology, Moscow, Russia.

The first step is to extract ribonucleic acid (RNA) from the tissue(s) of interest. The samples are then prepared and labeled. In a reference experimental design, the reference sample is labeled with the cy5 dye and the treatments of interest are labeled

with the cy3 dye. (See section on experimental design). Hybridization then follows. This is the process by which the RNA samples are placed on the micro array slides. The slides are then washed to remove excess RNA and to prevent cross hybridization. A scanner is then used to extract the image. The scanner has laser(s) that are focused onto the array inducing fluorescence. The intensity of fluorescence on each spot is measured and from this, the amount of DNA bound to each spot on the array is then inferred. A spot is more green if the gene is better expressed in the treatment labeled by the Cy3 dye and more red, if the gene is better expressed in the treatment labeled by the Cy5 dye. A yellow spot indicates equal expression or no differential expression.

Because of the large volume and intrinsic variation of the data obtained in a micro-array experiment, statistical methods have been applied to systematically extract information and to assess the associated uncertainty. The sources of variation need to be either removed or minimized. Some of these can be minimized through normalization. The aim of normalization is to remove the biases within each array involved in an experiment so that data can be comparable.

R.D. Wolfinger *et al*, 2001, used mixed models to detect differentially expressed genes in the *Saccharomyces cerevisiae swi/snf* mutation study of Sudarsanam et al. (2000). The data are available at *http://genome-www.stanford.edu/swisnf.*
C.Blake Gilks et al, 2005, studied the Distinction between serous tumors of low malignant potential and serous carcinomas based on global mRNA profiling. They used unsupervised hierarchical clustering to determine differential expression. Clustering is a technique, which aims to create groups of observations that are internally homogeneous and heterogeneous from group to group. The level of significance cannot be directly controlled as in the mixed model approach. The Fold Change method is one of the oldest methods used to identify differentially expressed gene. A value for Fold Change is calculated as will be discussed in the literature review. A cut-off value is decided upon. Genes whose fold change value falls below the cut off value will be classified as not differentially expressed. Though it has many critics, it is still widely used. The Mann-Whitney U test is sometimes preferred because of its non-parametric nature. It is only recommended with a large number of replicates, n>7. [B. Meunier *et al*. 2005]. Chen Y et al, (1997) described a maximum likelihood estimation approach. In this study, a recalculation of the work of Wolfinger and associates was done to get an understanding of the approach.

NOTATION

The following notation will be used.
Let x_{gtar} be the intensity of fluorescence on a spot for treatment t,
Let, g = 1,2…g, t=1,2 …t, a=1,2…a and r=1,2…r.

$\overline{x_{gt..}}$ Be the average intensity of fluorescence for gene g under treatment t, averaged over arrays and replicates.
Y_{gtar} Be the $\log_2(x_{gtar})$ of replicate r for gene g, on array a,

$\overline{Y}_{gt..} = \frac{1}{a}\sum_{a}\overline{Y}_{gat.}$ Be the average over arrays and replicates of $\log_2(x_{gtar})$ for gene g, (the ls-mean) and $\mu_{gt..} = E[\overline{Y}_{gt..}]$ Be the mean $\log_2(x_{gtar})$ for a gene g.

CHAPTER 2

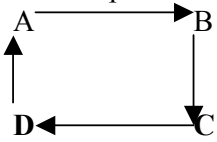
LITERATURE REVIEW

INTRODUCTION

In this chapter, we introduced the typical experimental designs used in micro-array analysis. It is standard procedure to normalize data obtained from a micro-array experiment before the data is analyzed. We introduce Data normalization in the micro-array context. It is customary to normalize all micro-array data before it can be analyzed. We also introduce techniques that are traditionally used in the analysis of micro-array data and justify the use of the Bonferroni adjustment for multiple tests. Though there are many possible adjustments for multiple testing, we only look at the Bonferroni adjustment as suggested by *Wolfinger et al*. An overview of some aspects of the mixed model is also given for example, its formulation and estimation of parameters. We also discuss the estimation of variance components using the method of moments and maximum likelihood estimation.

EXPERIMENTAL DESIGN

There are mainly two types of experimental designs used in analysis of differentially expressed data. That is, the reference or indirect design and the loop design. In a loop design, there is a direct comparison of the treatments. Two treatments are hybridised onto each array and compared to the other treatments in a loop fashion as indicated in the example below. For treatments A, B, C and D, the comparison would be as shown in the schematic diagram below.

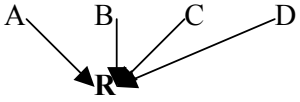


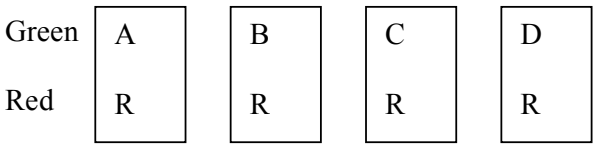
Green	A	B	C	D
Red	B	C	D	A

That is, we first compare A to B, then B to C, C to D and then D to A on subsequent arrays. Each treatment is labelled twice, once with red dye and once with green dye, which means dye effects is not confounded with treatment effects. Using the same number of arrays as the reference design, the loop collects twice as much data on the treatments of interest. Loop designs are more efficient in

estimating parameters of interest when comparing a small to moderate number of conditions or treatments. However, the results are not immediately intuitive and if one sample fails, the others are also affected. Some results suggest that the loop design stops being optimal for more than eight conditions. In this case it would be preferable to use the reference design.

In a reference design, each sample or treatment to be analysed is compared to a reference sample as shown below.





Each array contains a treatment of interest and a reference treatment.
Here, if one sample fails, we only lose information on that sample but the other samples are not affected. However, when using the reference design, more information is collected on the reference sample though it is usually not of much interest to the researcher.

DATA NORMALISATION

It is standard procedure to normalise the data before any micro-array analysis. Normalisation in Micro-array analysis is a general term for a collection of methods that are directed at resolving the systematic errors introduced by the micro array experimental platform. These include Data cleaning and transformation, within array normalization and between array normalization, background subtraction and taking logarithms. Data cleaning and transformation entails removal of flagged values, (flagged values are values for which the image processing software has detected some type of abnormality.) and making log transformation of the data. Log transformation is generally recommended for the following reasons:

- It is an easily invertible transformation. There is no loss of information about the original scale.
- Normalization is best done on the log scale; the variance stabilizing effect makes normalizations more robust to outliers.
- Log transformation to base two is the most natural scale for fold changes. On the original scale a twofold increase and a twofold decrease do not

correspond to the same absolute change, as is the case with the log scale. This is illustrated in the table below:

	Original scale	Log2 scale
Twofold increase	2	1
Unchanged	1	0
Twofold decrease	0.5	-1

- Log transformation ensures symmetric treatment of the numerator and denominator in ratios.

Within-array normalization is essential because the slide may not be perfectly flat, hence focus may be different on different parts of the array, when measuring the fluorescence intensities. The use of a reference sample is also a form of within-array normalization. Between-array normalization is essential because each hybridization, (preparation of the array), reaction is different and the slides are not exactly the same. Normalization, makes the data from the different slides comparable.

FOLDCHANGE

After the data has been normalized, one can look at the fold-change to determine differential expression. The Fold-change method is a simple method for identifying differentially expressed genes. It is based on the ratio between the mean observed intensities of the two treatments being compared. Fold-change for a gene g is calculated as below:

$$FC_g = \frac{\overline{x}_{g1..}}{\overline{x}_{g2..}}$$

Where FC_g is the fold change for a particular gene g, $\overline{x_{g1..}}$ is the mean observed intensity for a particular gene under treatment 1, and $\overline{x}_{g2..}$ is the mean observed intensity for the same gene, under treatment 2. In this method, fold change is calculated as above for each gene. An arbitrary cut-off value (for example, 2 fold) is chosen. Genes whose Fold-change values fall below the cut-off values are classified as being not differentially expressed and those above as being differentially expressed. The fold-change method has been criticized because the cut off value is chosen arbitrarily. For instance, if one is selecting genes with at least a two-fold change and the condition under study does not affect any genes to the point of inducing a two-fold change, no genes will be selected, resulting in zero sensitivity. Equally, if the condition is such that many genes change dramatically, the method will select too many gene and we will have low

specificity. Fold change is not a statistical test and there is no associated level of confidence. We cannot talk of significance levels when using the fold change method.

T-TEST

The normalized data can be analyzed using the t-test. A standard t-test is conducted for each gene. The t-test statistic for paired data is calculated as follows:

$$t = \left(\frac{\overline{Y_{g1..}} - \overline{Y_{g2..}}}{\sqrt{\frac{\sigma^2_{g1}}{n_{g1}} + \frac{\sigma^2_{g2}}{n_{g2}}}} \right)$$

where σ^2_{g1} is the standard deviation of observations for gene g, under treatment one, and n_{g1} is the number of spots under treatment one, for the particular gene. σ^2_{g2} is the standard deviation of observations for gene g under treatment two, and n_{g2} is the number of spots under treatment two, for the particular gene.

It is more convenient to work with p-values than the t-statistics. Genes with p-values falling below a prescribed level may be regarded as being statistically significant. However, there is low statistical power because of the small sample size. T-tests alone are not appropriate for micro-array analysis because some genes may be statistically significant but have little or no biological importance.

LSMEAN

Ls-mean is short for Least squares mean. As indicated in the SAS User's Guide: Statistics, in an unbalanced design, the Ls-mean for each level of an effect is the arithmetic mean one would expect for that particular level if the design were balanced and all covariates were held at their mean values. A design is balanced if there is equal replication in each treatment level and unbalanced otherwise. In a balanced design, it is equal to the mean. In an unbalanced design, the two are not equal.

BONFERRONI ADJUSTMENT

A micro-array experiment usually involves a large number of genes hence giving rise to the problem of multiple testing. We test the hypothesis:

H_0: Gene g is not differentially expressed against H_1: Gene g is differentially expressed for each one of the genes.
If one significance test is performed with α being the level of significance, then P (type 1 error) = α. That is, α is the probability of rejecting the individual null hypothesis when it is in fact true. In this case, α is the comparison-wise error rate (CER) or individual error rate. Hence the probability of not rejecting a true null hypothesis is 1- α. If k independent tests are performed, each at α level of significance, then the probability of rejecting all k null hypotheses when in fact all are true is $k\alpha$ as shown below;
Let A_i be the event that a true null hypothesis is rejected for test i, such that $P(A_i) = \alpha$.
The probability of rejecting the null hypothis for at least one test,

$$\begin{aligned} P(\bigcup_i A_i) &= 1 - P(\bigcap_i A_i^c) \\ &= 1 - P(\bigcup_{i=1}^{k} A_i) \\ &\geq 1 - \sum_{i=1}^{k} P(A_i) \\ &\geq 1 - k\alpha \end{aligned}$$

Bonferroni proposed that, if it is desired to have an EER, which is at most equal to α, then for each individual experiment, we must have $CER = \frac{\alpha}{k}$.

That is $EER = 1 - (1 - k\alpha)$
$= k\alpha$

Hence CER, say $\tilde{\alpha} = \frac{\alpha}{k}$.

Dividing the individual levels of significance results in a conservative measure of the overall significance probability. The overall significance is better than or equal to the desired level. The Bonferroni adjustment has been criticized for being too conservative. Because of the large number of genes k, involved, CER becomes very small. Other methods of adjusting for multiple testing have been put forward. Examples are the Holm procedure, which is a modification of the Bonferroni adjustment, and the procedures of Duncan, Student-Newman-Keuls and Ryan-Einot-Gabriel-Welsch. (R. Bender, S. Lange / Journal of Clinical Epidemiology, 2001).

VOLCANO PLOTS

The volcano plot presents both fold change and t-test criteria. It is a scatter plot of $-\log_{10}(p)$, (where p represents p-values from the t-tests), against the $\log_2(FC)$. To compare treatments 1 and 2, on the x-axis, we plot

$$\log_2(FC_g) = \log_2 \frac{\overline{x_{g1..}}}{\overline{x_{g2..}}}$$
$$= \log_2 \overline{x_{g1..}} - \log_2 \overline{x_{g2..}} .$$
$$= \overline{Y}_{g1..} - \overline{Y}_{g2..}$$

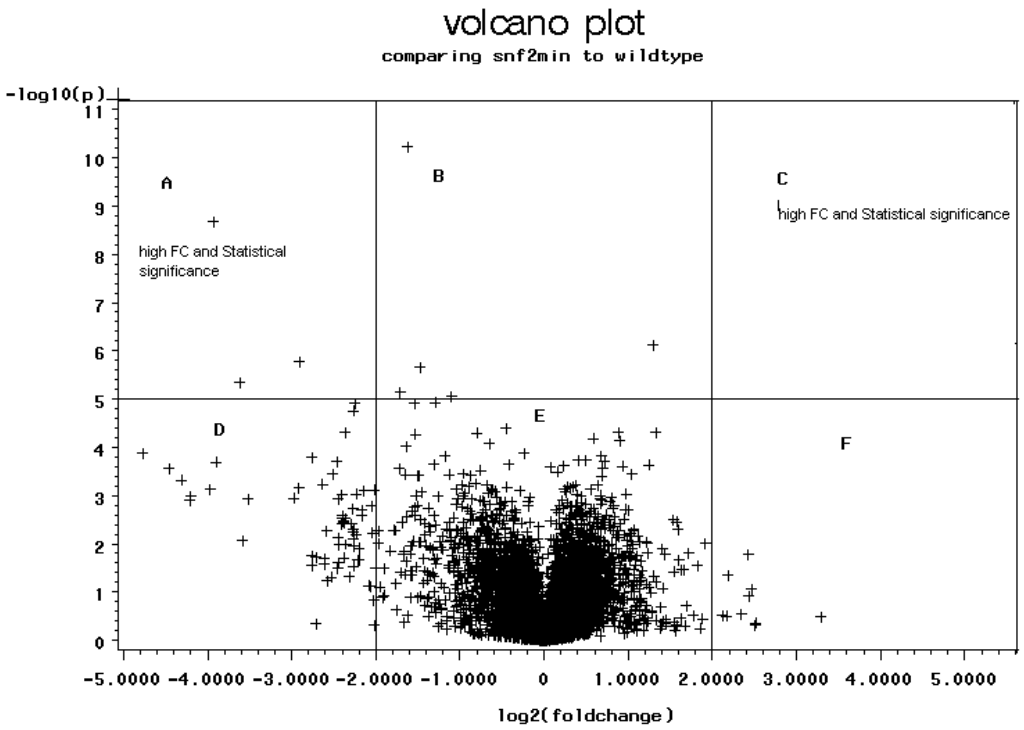
In the data used for illustration by Wolfinger et al., five treatments were tested, four treatments being compared to one treatment, wild type. Hence we need four pair-wise differences of $\log_2$ means. The p-values are obtained by testing the following hypothesis on each gene :

$H_o : \mu_{g1..} = \mu_{g2..}$ vs.

$H_1 : \mu_{g1..} \neq \mu_{g2..}$

Genes that are statistically significant lie above horizontal threshold line usually, -log10(p) = 5 where p is the p-value obtained from the t-tests. Genes that are biologically significant lie outside a pair of vertical threshold lines, usually $\log_2(FC)=-2$ and $\log_2(FC)=+2$. The significant genes tend to locate in the upper left or upper right of the graph.

An example of a Volcano plot is shown below;



The sectors, marked A and C have both high statistical significance and practical significance. These are the genes of interest that would be ear marked for further research depending on the application. Sector E has low statistical significance and fold change. The corresponding change is insignificant. Sectors D and F represent likely false positives when using the fold change method. The genes have high fold change but low statistical significance.

MIXED MODELS

Wolfinger et al suggested the use of mixed models to determine differential expression instead of the traditionally used methods.
In general, we have three broad categories of linear models, that is, fixed models, random models and mixed models. A mixed model takes the form:
$Y = Xb + Zu + \varepsilon$,
Where Y is the vector of response variables, b is the vector of fixed effects and u is the vector of random effects. The vectors b and u are the model parameter vectors. X and Z are the corresponding incidence matrices.

Assumptions of the Mixed Model

1. There is a linear relationship between the dependent variable and independent variables.

2. $E\begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

3. $Var\begin{pmatrix} u \\ \varepsilon \end{pmatrix} = \begin{bmatrix} D & 0 \\ 0 & R \end{bmatrix}$

If there are no random effects, the model becomes a fixed effects model. If there are no fixed effects, the model becomes a random effects model.
In a mixed model, there is at least one fixed effect and at least one random effect. An effect is fixed if research interest lies only in the specific selected levels of the effect used in the experiment. Inferences drawn from the data apply only to these levels. In our study, treatments are considered to be a fixed effect. Each treatment is of particular importance and interest.
When inferences will be made about a population of levels of an effect, from which those used in the experiment were sampled, then the effect is considered to be random. Levels of the effect used in the coding of the effect are a random sample of the total number of levels in the population of the effect. Nothing important conditions the choice of one level over the other. Arrays used in a micro-array study are considered to be a random sample of a conceptually large number of arrays that could have been constructed in the study. [*R.J. Templeman*, Veterinary immunology and immunopathology, 2005.]. Hence array is a random effect.
The normalization model used in the study is given below:

$$Y_{gtar} = \mu + T_t + A_a + (TA)_{ta} + \varepsilon_{gtar,}$$

Y_{tag} is the $\log_2(x)$ of gene g, treatment t and array a,

μ is the overall mean effect

A_a is the array effect, a=1,2,3.

T_t is the treatment effect, t=1,2.

ε_{tag} is the random error associated with y_{tag} .

This is a mixed model since there is a fixed effect (treatment) and a random effect (array). The effects A_a, $(TA)_{ta}$ and ε_{tag} were all assumed to be normally distributed with zero means and variance components $\sigma_A^2, \sigma_{TA}^2$ and σ_ε^2 respectively. These random effects are assumed to be independent both across their indices and with each other.

ESTIMATION OF PARAMETERS

In this section, we have followed the illustrations in Searle, filling in a few gaps.

The assumption that $E\begin{bmatrix} u \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ implies that $Var(u) = E[uu']$ and $Var(\varepsilon) = E[\varepsilon\varepsilon']$.

Let

$$V = Var(Y)$$

$$V = E[YY'] - E[Y]E[Y']$$

$$= E[Xb(Xb)' + Zu(Zu)' + \varepsilon\varepsilon'] - Xb(Xb)'$$

$$= Xbb'X' + ZE[uu']Z' + E[\varepsilon\varepsilon'] - Xbb'X'$$

$$= ZDZ' + R$$

Note Cov(u,ε)=0.

The parameters cannot be estimated using Ordinary Least Squares because V is not of the form $\sigma_\varepsilon^2 I$. It is of the form $V = \sum_{i=1}^{a}{}^{+}(\sigma_\varepsilon^2 I + \sigma_\alpha^2 J)$. Generalized Least Squares is used instead. Hence Y must be standardized first.

Let $W = V^{-\frac{1}{2}}Y$ where $V = V^{\frac{1}{2}}V^{\frac{1}{2}}$. $V^{\frac{1}{2}}$ is symmetric, and its inverse is also symmetric. Then $W = V^{\frac{-1}{2}}Xb + V^{\frac{-1}{2}}Zu + V^{\frac{-1}{2}}\varepsilon$.

$$Var(W) = Var(V^{\frac{-1}{2}}Zu) + Var(V^{\frac{-1}{2}}\varepsilon)$$

$$= E[V^{\frac{-1}{2}}Zu - E(V^{\frac{-1}{2}}Zu)][V^{\frac{-1}{2}}Zu - E(V^{\frac{-1}{2}}Zu]' + E[V^{\frac{-1}{2}}\varepsilon - E(V^{\frac{-1}{2}}\varepsilon)][V^{\frac{-1}{2}}\varepsilon - E(V^{\frac{-1}{2}}\varepsilon)]'$$

$$= E[V^{\frac{-1}{2}}Zuu'Z'V^{\frac{-1}{2}}] + E[V^{\frac{-1}{2}}\varepsilon\varepsilon'V^{\frac{-1}{2}}]$$

$$= V^{\frac{-1}{2}}ZE[uu']Z'V^{\frac{-1}{2}} + V^{\frac{-1}{2}}E[\varepsilon\varepsilon']V^{\frac{-1}{2}}$$

$$= V^{\frac{-1}{2}}ZDZ'V^{\frac{-1}{2}} + V^{\frac{-1}{2}}RV^{\frac{-1}{2}}$$

$$= V^{\frac{-1}{2}}[ZDZ' + R]V^{\frac{-1}{2}}$$

$$= V^{\frac{-1}{2}}VV^{\frac{-1}{2}}$$

$$= I$$

Then the sum of squares of residuals,

$$\varepsilon'\varepsilon = [W - E(W)]'[W - E(W)]$$
$$= [W - V^{\frac{-1}{2}} Xb]'[W - V^{\frac{-1}{2}} Xb]$$
$$= W'W - W'V^{\frac{-1}{2}} Xb - b'X'V^{\frac{-1}{2}} W + b'X'V^{-1} Xb$$
$$= W'W - 2b'X'V^{\frac{-1}{2}} W + b'X'V^{-1} Xb$$

To get estimates for b, we differentiate the sum of squares of residuals with respect to b and equate the differential to zero.

$$\frac{\partial \varepsilon'\varepsilon}{\partial b} = -2X'V^{\frac{-1}{2}} W + 2X'V^{-1} Xb$$

Equating to zero leads to

$$2X'V^{-1}X\hat{b} = 2X'V^{\frac{-1}{2}} W$$
$$X'V^{-1}X\hat{b} = X'V^{\frac{-1}{2}} W$$
$$\hat{b} = (X'V^{-1}X)^{-1} X'V^{\frac{-1}{2}} W$$
$$= (X'V^{-1}X)^{-1} X'V^{\frac{-1}{2}} V^{\frac{-1}{2}} Y$$
$$= (X'V^{-1}X)^{-1} XV^{-1} Y$$

If V is singular, then V^{-1} is replaced by the generalized inverse V^{-}. Under normality conditions, $\hat{b}$ is also the maximum likelihood estimate for the parameter vector b.

Under Normality Conditions

1. $\varepsilon \sim N(0, R)$
2. $u \sim N(0, D)$

The joint density of Y and u, which we will call the likelihood of Y and u for the remainder of the discussion,
$f(Y,u) = g(Y/u)h(u)$.

Because of the normality assumption, $h(u) = (2\pi)^{\frac{-n}{2}} |D|^{\frac{-1}{2}} \exp\{-\frac{1}{2}(u'D^{-1}u)\}$ and

$g(Y/u) = (2\pi)^{\frac{-n}{2}} |R|^{\frac{-1}{2}} \exp\{-\frac{1}{2}(Y - Xb - Zu)'R^{-1}(Y - Xb - Zu)\}$.

Note: In the expression for g(Y/u), u is treated as a fixed parameter vector because it is given.
Hence

$$f(Y,u) = (2\pi)^{-n} |R|^{\frac{-1}{2}} |D|^{\frac{-1}{2}} \exp\{-\frac{1}{2}(Y - Xb - Zu)' R^{-1}(Y - Xb - Zu) \exp\{-\frac{1}{2} u' D^{-1} u\}$$

.

$$L = \log_e f(Y,u) = -n \log_e (2\pi) - \frac{1}{2}\log_e |R| - \frac{1}{2}\log_e |D| - \frac{1}{2}(Y - Xb - Zu)' R^{-1}(Y - Xb - Zu) - \frac{1}{2} u' D^{-1} u$$

$$= -n \log_e (2\pi) - \frac{1}{2}\log_e |R| - \frac{1}{2}\log_e |D| - \frac{1}{2}[Y' R^{-1} Y - Y' R^{-1} Xb - Y' R^{-1} Zu - b' X' R^{-1} Y + b' X' R^{-1} Xb$$

$$+ b' X' R^{-1} Zu - u' Z' R^{-1} Y + u' Z' R^{-1} Xb + u' Z' R^{-1} Zu + u' D^{-1} u]$$

$$= -n \log_e (2\pi) - \frac{1}{2}\log_e |R| - \frac{1}{2}\log_e |D| - \frac{1}{2}[Y' R^{-1} Y - 2b' X' R^{-1} Y - 2u' Z' R^{-1} Y$$

$$+ b' X' R^{-1} Xb + b' X' R^{-1} Zu + u' Z' R^{-1} Zu + u' Z' R^{-1} Xb + u' D^{-1} u]$$

$$\frac{\partial L}{\partial b} = X' R^{-1} Y - X' R^{-1} Xb + X' R^{-1} Zu$$

Equating to zero yields:

$$X' R^{-1} Y = X' R^{-1} X\tilde{b} + X' R^{-1} Z\tilde{u}$$

$$\frac{\partial L}{\partial u} = Z' R^{-1} Y - Z' R^{-1} Xb - Z' R^{-1} Zu + D^{-1} u$$

Equating to zero yields:

$$Z' R^{-1} X\tilde{b} + Z' R^{-1} Z\tilde{u} + D^{-1}\tilde{u} = Z' R^{-1} Y$$

$$Z' R^{-1} X\tilde{b} + (Z' R^{-1} Z + D^{-1})\tilde{u} = Z' R^{-1} Y$$

This solution can be written more compactly as:

$$\begin{bmatrix} X' R^{-1} X & X' R^{-1} Z \\ Z' R^{-1} X & Z' R^{-1} Z + D^{-1} \end{bmatrix} \begin{bmatrix} \tilde{b} \\ \tilde{u} \end{bmatrix} = \begin{bmatrix} X' R^{-1} Y \\ Z' R^{-1} Y \end{bmatrix}$$

Solving the two equations simultaneously,

$$X' R^{-1} X\tilde{b} - Z' R^{-1} X\tilde{b} + X' R^{-1} Z\tilde{u} - (Z' R^{-1} Z + D^{-1})\tilde{u} = X' R^{-1} Y - Z' R^{-1} Y$$

$$(X' - Z') R^{-1} X\tilde{b} + [X' R^{-1} Z - (Z' R^{-1} Z + D^{-1})]\tilde{u} = (X' - Z') R^{-1} Y$$

$$[X' R^{-1} Z - (Z' R^{-1} Z + D^{-1})]\tilde{u} = (X' - Z')(R^{-1} Y - R^{-1} X\tilde{b})$$

$$X' R^{-1} Z\tilde{u} - (Z' R^{-1} Z + D^{-1})\tilde{u} = X' R^{-1} Y - X' R^{-1} X\tilde{b} - Z' R^{-1} Y + Z' R^{-1} X\tilde{b}$$

$$(Z' R^{-1} Z + D^{-1})\tilde{u} = -X' R^{-1} Z\tilde{u} - X' R^{-1}(Y - X\tilde{b}) + Z' R^{-1} Y - Z' R^{-1} X\tilde{b}$$

$$\tilde{u} = (Z' R^{-1} Z + D^{-1})^{-1}(Z' R^{-1} Y - Z' R^{-1} X\tilde{b})$$

Substituting the value of

$\tilde{u}$,

$$X'R^{-1}X\tilde{b} + X'R^{-1}Z[(Z'R^{-1}Z + D^{-1})^{-1}(Z'R^{-1}Y - Z'R^{-1}X\tilde{b})] = X'R^{-1}Y$$

$$X'R^{-1}X\tilde{b} + X'R^{-1}Z[(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}Y - (Z'R^{-1}Z + D^{-1})^{-1}(Z'R^{-1}X\tilde{b}] = X'R^{-1}Y$$

$$X'R^{-1}X\tilde{b} + X'R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}Y - X'R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}X\tilde{b} = X'R^{-1}Y$$

$$X'[R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}]X\tilde{b} = X'[R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}]Y$$

Let

$$W = R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}]$$

$$X'WX\tilde{b} = X'WY$$

But

$$WV = [R^{-1} - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}][ZDZ' + R]$$

$$= R^{-1}ZDZ' - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}ZDZ' + R^{-1}R - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}Z'R^{-1}R$$

$$= R^{-1}ZDZ' + I - R^{-1}Z(Z'R^{-1}Z + D^{-1})^{-1}(Z'R^{-1}Z + D^{-1})DZ'$$

$$= I$$

Therefore

$$W = V^{-1}$$

and

$$X'V^{-1}X\tilde{b} = X'V^{-1}Y$$

$$\tilde{b} = (X'V^{-1}X)^{-1}X'V^{-1}Y = \hat{b}$$

Therefore, under normality conditions, the generalized least squares estimate for b is also the maximum likelihood estimator for b.

In general, V=ZDZ'+R is not diagonal and V^{-1} is not always easy to calculate. A set of equations for estimating the parameters, which does not involve V^{-1} can be established.

Suppose u were in fact a fixed parameter vector. Then

$$\begin{bmatrix} X & Z \end{bmatrix}' R^{-1} \begin{bmatrix} X & Z \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X & Z \end{bmatrix}' R^{-1} Y$$

$$\begin{bmatrix} X' \\ Z' \end{bmatrix} R^{-1} \begin{bmatrix} X & Z \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X' \\ Z' \end{bmatrix} R^{-1} Y$$

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}$$

If we add D^{-1} to the lower right hand sub matrix, we get

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + D^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}Y \\ Z'R^{-1}Y \end{bmatrix}$$

Which is the same set of equations obtained using maximum likelihood estimation under normality conditions.

$\tilde{b} = \hat{b} = \hat{b}$. Thus inversion of V can be avoided.

Variance Components

Estimation techniques of the variance components for the random effects in a model depend on whether the data are balanced or unbalanced. For categorical data, data are balanced, if there are an equal number of replicates in all the subclasses and unbalanced if the number of replicates is unequal. We are going to discuss only two methods of estimation, that is:

1. Method of moments.
2. Maximum Likelihood estimation, MLE.

Method of Moments

The Method of Moments is as follows:

1. Calculate the ANOVA table as if the model were a fixed effects model.
2. Derive expected mean squares values under the random (or mixed) model.
3. Equate expected mean to their observed values.
4. Solve for estimators of the variance components.

The analysis of variance table for a two-way classification, mixed model, as defined above, with balanced data is shown below. It also shows expected mean square values.

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	Expected Mean Squares
Mean	1	$SSM = N\bar{y}^2...$	$MSM = SSM$	$abn(\mu+\bar{\alpha}.)^2 + an\sigma_\beta^2 + \sigma_\varepsilon^2$
A-factor, α	a-1	$SSA = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2$	$MSA = SSA/(a-1)$	$\frac{bn}{a-1}\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}.)^2 + \sigma_\varepsilon^2$
B-factor, β	b-1	$SSB = an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$	$MSB = SSB/(b-1)$	$an\sigma_\beta^2 + \sigma_\varepsilon^2$
Residual error	ab(n-1)	$SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$	$MSE = SSE/ab(n-1$	σ_ε^2
Total	N = abn	$SST = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}y_{ijk}^2$		

Let

$$\bar{\alpha}. = \frac{\sum_{i=1}^{a}\alpha_i}{a}, \bar{\beta}. = \frac{\sum_{j=1}^{b}\beta_j}{b}, \bar{\varepsilon}... = \frac{1}{N=abn}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{ijk}, \bar{\varepsilon}_{i..} = \frac{1}{bn}\sum_{j=1}^{b}\sum_{k=1}^{n}\varepsilon_{ijk},$$

$$\bar{\varepsilon}_{.j.} = \frac{1}{an}\sum_{i=1}^{a}\sum_{k=1}^{n}\varepsilon_{ijk}$$

$$E[\bar{\varepsilon}...]^2 = E[\frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\varepsilon^2{}_{ijk}}{N^2}]$$

$$= \frac{1}{N^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}E(\varepsilon^2{}_{ijk})$$

$$= \frac{N}{N^2}\sigma_\varepsilon^2$$

$$= \frac{\sigma_\varepsilon^2}{N}$$

$$E[\bar{\beta}.^2] = E[\frac{1}{b^2}\sum_{j=1}^{b}\beta_j^2$$

$$= \frac{1}{b^2}\sum_{j=1}^{b}E[\beta_j^2]$$

$$= \frac{1}{b}\sigma_\beta^2$$

Note that: $E[\varepsilon_{ijk}] = 0$ and $E[\beta_j] = 0$.

$$E[MSM] = E[N\bar{y}^2...]$$
$$= NE[\mu + \bar{\alpha}. + \bar{\beta}. + \bar{\varepsilon}...]^2$$
$$= N\{(\mu + \bar{\alpha}.)^2 + E[\bar{\beta}.^2] + E[\bar{\varepsilon}^2...]\}$$
$$= N(\mu + \bar{\alpha}.)^2 + N(\frac{\sigma_\beta^2}{b} + \frac{\sigma_\varepsilon^2}{N})$$
$$= abn(\mu + \bar{\alpha}.)^2 + an\sigma_\beta^2 + \sigma_\varepsilon^2$$

Note that: Cross product terms are equal to zero since:

(i) Random terms are independent of each other by assumptions of the mixed model.

(ii) $E[(\mu + \alpha)\bar{\beta}.] = (\mu + \alpha)E[\bar{\beta}.] = 0$

(iii) $E[(\mu + \alpha)\bar{\beta}.] = (\mu + \alpha)E[\bar{\varepsilon}...] = 0$

(iv) $E[\bar{\beta}.^2] = Var(\bar{\beta}.) + E[\bar{\beta}.]^2 = \sigma_{\bar{\beta}.}^2 = \frac{\sigma_\beta^2}{b}$

(v) $E[\bar{\varepsilon}^2...] = Var(\bar{\varepsilon}...) + E[\bar{\varepsilon}...]^2 = \sigma_{\bar{\varepsilon}...}^2 = \frac{\sigma_\varepsilon^2}{N}$

$$E[MSA] = E[\frac{\text{bn}}{a-1}\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}...)^2$$
$$= \frac{\text{bn}}{a-1}\sum_{i=1}^{a}E(\bar{y}_{i..} - \bar{y}...)^2$$
$$= \frac{\text{bn}}{a-1}\sum_{i=1}^{a}E[\mu + \alpha_i + \bar{\beta}. + \bar{\varepsilon}_{i..} - (\mu + \bar{\alpha}. + \bar{\beta}. + \bar{\varepsilon}...)]^2$$
$$= \frac{\text{bn}}{a-1}\sum_{i=1}^{a}E[(\alpha_i - \bar{\alpha}.) + (\bar{\varepsilon}_{i..} - \bar{\varepsilon}...)]^2$$
$$= \frac{\text{bn}}{a-1}\sum_{i=1}^{a}[(\alpha_i - \bar{\alpha}.)^2 + \frac{a-1}{N}\sigma_\varepsilon^2]$$
$$= \frac{\text{bn}}{a-1}\sum_{i=1}^{a}(\alpha_i - \bar{\alpha}.)^2 + \sigma_\varepsilon^2$$

Note that: i) Cross product terms are equal to zero because $E[(\alpha_i - \bar{\alpha}.)(\varepsilon_{i..} - \bar{\varepsilon}...)] = (\alpha_i - \bar{\alpha}.)E[\varepsilon_{i..} - \bar{\varepsilon}...] = 0$.

$$\sum_{i=1}^{a} E[(\bar{\varepsilon}_{i..} - \bar{\varepsilon}...)^2] = \sum_{i=1}^{a} E(\bar{\varepsilon}_{i..}^2 - 2\bar{\varepsilon}_{i..}\bar{\varepsilon}... + \bar{\varepsilon}^2...)$$

ii)
$$= \sum_{i=1}^{a} \frac{\sigma_\varepsilon^2}{bn} - 2a\frac{\sigma_\varepsilon^2}{N} + \sum_{i=1}^{a} \frac{\sigma_\varepsilon^2}{N}$$

$$= a\frac{\sigma_\varepsilon^2}{bn} - a\frac{\sigma_\varepsilon^2}{N}$$

$$= (a-1)\frac{\sigma_\varepsilon^2}{bn}$$

$$E[MSB] = E[\frac{an}{b-1}\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}...)^2$$

$$= \frac{an}{b-1}\sum_{j=1}^{b} E[(\bar{y}_{.j.} - \bar{y}...)^2]$$

$$= \frac{an}{b-1}\sum_{j=1}^{b} E[\mu + \bar{\alpha}. + \beta_j + \varepsilon_{.j.} - (\mu + \bar{\alpha}. + \bar{\beta}. + \bar{\varepsilon}...)]^2$$

$$= \frac{an}{b-1}\sum_{j=1}^{b} E[(\beta_j - \bar{\beta}.) + (\varepsilon_{.j.} - \bar{\varepsilon}...)]^2$$

$$= \frac{an}{b-1}\sum_{j=1}^{b} \{E[(\beta_j - \bar{\beta}.)]^2 + E[(\varepsilon_{.j.} - \bar{\varepsilon}...)]^2\}$$

$$= \frac{an}{b-1}\sum_{j=1}^{b} [\frac{b-1}{b}\sigma_\beta^2 + \frac{b-1}{abn}\sigma_\varepsilon^2]$$

$$= an\sigma_\beta^2 + \sigma_\varepsilon^2$$

Note that: i) Random terms are independent of each other by assumptions of the mixed model therefore cross product terms are equal to zero.

$$\sum_{j=1}^{b} E(\beta_j - \bar{\beta}.)^2 = \sum_{j=1}^{b} E(\beta^2{}_j - 2\beta_j\bar{\beta}. + \bar{\beta}.^2)$$

ii) $$= b\sigma_\beta^2 - b\frac{\sigma_\beta^2}{b}$$

$$= (b-1)\sigma_\beta^2$$

$$\sum_{j=1}^{b}(\bar{\varepsilon}_{.j.}-\bar{\varepsilon}_{...})^2 = E\sum_{j=1}^{b}(\bar{\varepsilon}_{.j.}^2 - 2\bar{\varepsilon}_{.j.}\bar{\varepsilon}_{...}+\bar{\varepsilon}^2$$

iii) $$= b\frac{\sigma_\varepsilon^2}{an} - b\frac{\sigma_\varepsilon^2}{abn}$$

$$= \frac{b-1}{an}\sigma_\varepsilon^2$$

$$E[MSE] = E[\frac{1}{ab(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2$$

$$= \frac{1}{ab(n-1)}E[\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[\mu+\alpha_i+\beta_j+\varepsilon_{ijk}-(\mu+\alpha_i+\beta_j+\bar{\varepsilon}_{ij.})]^2$$

$$= \frac{1}{ab(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}E[\varepsilon_{ijk}-\bar{\varepsilon}_{ij.}]^2$$

$$= \frac{1}{ab(n-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}n-1\frac{\sigma_\varepsilon^2}{N=abn}$$

$$= \sigma_\varepsilon^2$$

Equating the expected mean square values to their observed values yields the respective estimates of the variance components.

1. $\tilde{\sigma}_\varepsilon^2 = MSE$

   $MSB = an\tilde{\sigma}_\beta^2 + \tilde{\sigma}_\varepsilon^2$

2. $\Rightarrow \tilde{\sigma}_\beta^2 = \frac{MSB - MSE}{an}$

When the data is unbalanced, the method of moments for mixed models yields biased estimates. Methods for correcting this bias have been suggested though they are not without problems. (Searle 1997: Linear Models.) Other methods of estimation can be used instead.

Maximum Likelihood estimation

The One-way analysis of variance model will be used to illustrate maximum likelihood estimation of variance components using guidelines from Searle [1]. The mixed model can be expressed as a one-way analysis of variance model. The normal distribution is going to be assumed.

$$y_{ir} = \mu + \alpha_i + \varepsilon_{ir}$$

y_{ij} is the r-th replicate for the i-th factor, μ is the overall mean, α_i is the effect of the random factor i and ε_{ij} is the error associated with y_{ij}.
Let l be the likelihood function.

$$l = (2\pi)^{ar} |V|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y-\mu)'V^{-1}(y-\mu)\}$$

The variance V can be written as $V = \sum_{i=1}^{a}{}^{+}(\sigma_\varepsilon^2 I + \sigma_\alpha^2 J)$ and $|V| = \Pi_{i=1}^{a+}|\sigma_\varepsilon^2 I + \sigma_\alpha^2 J|$ is the determinant of matrix V. Consider a 2*2 matrix,

$$V = \begin{pmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{pmatrix}$$

$$|V| = \begin{vmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{vmatrix}$$

$$= (\sigma_\varepsilon^2 + \sigma_\alpha^2)^2 - \sigma_\alpha^4$$

$$= 2\sigma_\alpha^2\sigma_\varepsilon^2 + \sigma_\varepsilon^4$$

$$V^{-1} = \frac{1}{2\sigma_\alpha^2\sigma_\varepsilon^2 + \sigma_\varepsilon^4}\begin{pmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & -\sigma_\alpha^2 \\ -\sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{pmatrix}$$

Consider a 3*3 matrix,

$$V = \begin{pmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{pmatrix}$$

$$|V| = \begin{vmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{vmatrix}$$

$$= (\sigma_\alpha^2 + \sigma_\varepsilon^2)^3 + 2\sigma_\alpha^6 - 3\sigma_\alpha^4(\sigma_\alpha^2 + \sigma_\varepsilon^2)$$

$$= (\sigma_\alpha^2 + \sigma_\varepsilon^2)(\sigma_\alpha^4 + \sigma_\varepsilon^4 + 2\sigma_\alpha^2\sigma_\varepsilon^2) + 2\sigma_\alpha^6 - 3\sigma_\alpha^6 - 3\sigma_\alpha^4\sigma_\varepsilon^2$$

$$= 2\sigma_\alpha^6 + 3\sigma_\alpha^2\sigma_\varepsilon^4 + 3\sigma_\alpha^4\sigma_\varepsilon^2 + 2\sigma_\alpha^6 - 3\sigma_\alpha^6 - 3\sigma_\alpha^4\sigma_\varepsilon^2$$

$$= 3\sigma_\alpha^2\sigma_\varepsilon^4 + \sigma_\alpha^6$$

In general,

$$|V| = \prod_{i=1}^{a}[r\sigma_\varepsilon^{2r-2}\sigma_\alpha^2 + \sigma_\varepsilon^{2r}]$$

$$= [\sigma_\varepsilon^{2(r-1)}(\sigma_\varepsilon^2 + r\sigma_\alpha^{2r})]^a$$

For the 2*2 matrix, the inverse,

$$V^{-1} = \frac{1}{2\sigma_\alpha^2\sigma_\varepsilon^2 + \sigma_\varepsilon^4}\begin{pmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & -\sigma_\alpha^2 \\ -\sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{pmatrix}$$

$$= \frac{1}{\sigma_\varepsilon^2(\sigma_\alpha^2 + 2\sigma_\varepsilon^2)}\begin{pmatrix} \sigma_\varepsilon^2 + \sigma_\alpha^2 & -\sigma_\alpha^2 \\ -\sigma_\alpha^2 & \sigma_\varepsilon^2 + \sigma_\alpha^2 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2}{\sigma_\varepsilon^2(\sigma_\alpha^2 + 2\sigma_\varepsilon^2)} & -\sigma_\alpha^2 \\ -\sigma_\alpha^2 & \frac{\sigma_\alpha^2 + \sigma_\varepsilon^2}{\sigma_\varepsilon^2(\sigma_\alpha^2 + 2\sigma_\varepsilon^2)} \end{pmatrix}$$

$$= \frac{1}{\sigma_\varepsilon^2(2\sigma_\alpha^2 + \sigma_\varepsilon^2)}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} + \frac{1}{(2\sigma_\alpha^2 + \sigma_\varepsilon^2)}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \frac{1}{\sigma_\varepsilon^2}\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2(2\sigma_\alpha^2 + \sigma_\varepsilon^2)}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

For the 3*3 matrix,

$$V^{-1} = \frac{1}{\sigma_\varepsilon^2(3\sigma_\varepsilon^2\sigma_\alpha^2 + \sigma_\alpha^4)}\begin{pmatrix} 2\sigma_\varepsilon^2\sigma_\alpha^2 + \sigma_\alpha^4 & -\sigma_\varepsilon^2\sigma_\alpha^2 & -\sigma_\varepsilon^2\sigma_\alpha^2 \\ -\sigma_\varepsilon^2\sigma_\alpha^2 & 2\sigma_\varepsilon^2\sigma_\alpha^2 + \sigma_\alpha^4 & -\sigma_\varepsilon^2\sigma_\alpha^2 \\ -\sigma_\varepsilon^2\sigma_\alpha^2 & -\sigma_\varepsilon^2\sigma_\alpha^2 & 2\sigma_\varepsilon^2\sigma_\alpha^2 + \sigma_\alpha^4 \end{pmatrix}$$

Using the same approach as above it can be shown that,

$$V^{-1} = \frac{1}{\sigma_\varepsilon^2} I - \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2(\sigma_\varepsilon^2 + 3\sigma_\alpha^2)} J$$

In general,

$$V^{-1} = \sum_{i=1}^{a}[\frac{1}{\sigma_\varepsilon^2} I - \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2(\sigma_\varepsilon^2 + r\sigma_\alpha^2)} J]$$

$$V^{-1} = \sum_{i=1}^{a}[\frac{1}{\sigma_\varepsilon^2} I + \frac{1}{r}(\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2(\sigma_\varepsilon^2 + r\sigma_\alpha^2)} - \frac{1}{\sigma_\varepsilon^2})J]$$

Substituting for $|V|$ and V^{-1} in the likelihood equation,

$$l=\frac{\exp[-\frac{1}{2}\{(y-\mu)'(\sum_{i=1}^{a}[\frac{1}{\sigma_\varepsilon^2}I+\frac{1}{r}(\frac{\sigma_\alpha^2}{\sigma_\varepsilon^2(\sigma_\varepsilon^2+r\sigma_\alpha^2)}-\frac{1}{\sigma_\varepsilon^2})J])(y-\mu)\}]}{[(2\pi)^{\frac{ar}{2}}(\sigma_\varepsilon^2)^{\frac{a(r-1)}{2}}(\sigma_\varepsilon^2+r\sigma_\alpha^2)^{\frac{a}{2}}]}$$

$$=\frac{\exp[-\frac{1}{2}(\sum_{i=1}^{a}\sum_{j=1}^{r}\frac{1}{\sigma_\varepsilon^2}(y_{ij}-\bar{y}_{i.})^2+\sum_{i=1}^{a}r\frac{(\bar{y}_{i.}-\bar{y}_{..})^2}{(\sigma_\varepsilon^2+r\sigma_\alpha^2)}+\frac{ar(\bar{y}_{..}-\mu)^2}{(\sigma_\varepsilon^2+r\sigma_\alpha^2)})]}{[(2\pi)^{\frac{ar}{2}}(\sigma_\varepsilon^2)^{\frac{a(r-1)}{2}}(\sigma_\varepsilon^2+r\sigma_\alpha^2)^{\frac{a}{2}}]}$$

$$=\frac{\exp[-\frac{1}{2}(\frac{SSE}{\sigma_\varepsilon^2}+\frac{SSA}{\sigma_\varepsilon^2+r\sigma_\alpha^2}+\frac{ar(\bar{y}_{..}-\mu)^2}{\sigma_\varepsilon^2+r\sigma_\alpha^2}}{[(2\pi)^{\frac{ar}{2}}(\sigma_\varepsilon^2)^{\frac{a(r-1)}{2}}(\sigma_\varepsilon^2+r\sigma_\alpha^2)^{\frac{a}{2}}]}$$

$$L=\log_e l=[-\frac{1}{2}(\frac{SSE}{\sigma_\varepsilon^2}+\frac{SSA}{\sigma_\varepsilon^2+r\sigma_\alpha^2}+\frac{ar(\bar{y}_{..}-\mu)^2}{\sigma_\varepsilon^2+r\sigma_\alpha^2})]$$

$$-\frac{1}{2}[ar\log_e(2\pi)+a(r-1)\log_e(\sigma_\varepsilon^2)+a\log_e(\sigma_\varepsilon^2+r\sigma_\alpha^2)]$$

$$\frac{\partial L}{\partial\mu}=-\frac{1}{2}\frac{(-2ar(\bar{y}_{..}-\mu))}{\sigma_\varepsilon^2+r\sigma_\alpha^2}$$

$$=\frac{ar(\bar{y}_{..}-\mu)}{\sigma_\varepsilon^2+r\sigma_\alpha^2}$$

Equating to zero yields:

$$ar(\bar{y}_{..}-\mu)=0$$

$$\tilde{\mu}=\bar{y}_{..}$$

$$\frac{\partial L}{\partial\sigma_\alpha^2}=-\frac{1}{2}\left[-\frac{rSSA}{(\sigma_\varepsilon^2+r\sigma_\alpha^2)^2}-\frac{ar^2(\bar{y}..-\mu)^2}{(\sigma_\varepsilon^2+r\sigma_\alpha^2)^2}\right]-\frac{ar}{2(\sigma_\varepsilon^2+r\sigma_\alpha^2)}$$

Equating to zero yields:

$$\frac{rSSA}{2(\tilde{\sigma}_\varepsilon^2+r\tilde{\sigma}_\alpha^2)^2}-\frac{ar}{2(\tilde{\sigma}_\varepsilon^2+r\tilde{\sigma}_\alpha^2)^2}=0$$

$$rSSA-ar(\tilde{\sigma}_\varepsilon^2+r\tilde{\sigma}_\alpha^2)=0$$

$$SSA=a(\tilde{\sigma}_\varepsilon^2+r\tilde{\sigma}_\alpha^2)$$

$$\tilde{\sigma}_\alpha^2=(\frac{SSA}{a}-\tilde{\sigma}_\varepsilon^2)/r$$

$$\frac{\partial L}{\partial \sigma_\varepsilon^2} = -\frac{1}{2}\left[\frac{-SSE}{\sigma_\varepsilon^4} - \frac{SSA}{(\sigma_\varepsilon^2 + r\sigma_\alpha^2)^2} - \frac{ar(\bar{y}.. - \mu)^2}{(\sigma_\varepsilon^2 + r\sigma_\alpha^2)^2}\right] - \left[\frac{a(r-1)}{2\sigma_\varepsilon^2} + \frac{a}{2(\sigma_\varepsilon^2 + r\sigma_\alpha^2)}\right]$$

Equating to zero:

$$\frac{SSE}{\tilde{\sigma}_\varepsilon^4} + \frac{SSA}{(\tilde{\sigma}_\varepsilon^2 + r\tilde{\sigma}_\alpha^2)^2} - \frac{a(r-1)}{\tilde{\sigma}_\varepsilon^2} - \frac{a}{(\tilde{\sigma}_\varepsilon^2 + r\tilde{\sigma}_\alpha^2)} = 0$$

$$\frac{SSE}{\tilde{\sigma}_\varepsilon^4} + \frac{a(\tilde{\sigma}_\varepsilon^2 + r\tilde{\sigma}_\alpha^2)}{(\tilde{\sigma}_\varepsilon^2 + r\tilde{\sigma}_\alpha^2)^2} - \frac{a(r-1)}{\tilde{\sigma}_\varepsilon^2} - \frac{a}{(\tilde{\sigma}_\varepsilon^2 + r\tilde{\sigma}_\alpha^2)} = 0$$

$$\frac{SSE}{\tilde{\sigma}_\varepsilon^4} + \frac{a}{(\tilde{\sigma}_\varepsilon^2 + r\tilde{\sigma}_\alpha^2)} - \frac{a(r-1)}{\tilde{\sigma}_\varepsilon^2} - \frac{a}{(\tilde{\sigma}_\varepsilon^2 + r\tilde{\sigma}_\alpha^2)} = 0$$

$$\frac{SSE}{\tilde{\sigma}_\varepsilon^4} - \frac{a(r-1)}{\tilde{\sigma}_\varepsilon^2} = 0$$

$$SSE = a(r-1)\tilde{\sigma}_\varepsilon^2$$

$$\tilde{\sigma}_\varepsilon^2 = \frac{SSE}{a(r-1)} = MSE$$

Maximum likelihood equations for estimating variance components from unbalanced data cannot be solved explicitly. This will be illustrated again using the one-way analysis of variance model:
When the data are unbalanced,

$V = \sigma_\varepsilon^2 I_N + \sigma_\alpha^2 \sum_{i=1}^{a} {}^{+} J_{n_i}$ n_i is the number of replicates in factor i and N is the total number of observations.

$$|V| = \sigma_\varepsilon^{2(N-a)} \prod_{i=1}^{a} (\sigma_\varepsilon^2 + n_i\sigma_\alpha^2)$$

$$V^{-1} = \frac{1}{\sigma_\varepsilon^2} I_N + \sum_{i=1}^{a} \frac{1}{n_i}\left(\frac{1}{\sigma_\varepsilon^2 + n_i\sigma_\alpha^2} - \frac{1}{\sigma_\varepsilon^2}\right) J_{n_i}$$

Again, assuming normality,

$$l = (2\pi)^{ar} |V|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y-\mu)'V^{-1}(y-\mu)\}$$

Substituting for $|V|$ and V^{-1} in the likelihood equation and taking natural logarithms,

$$L = -\frac{1}{2}[N\log_e(2\pi) + (N-a)\log_e\sigma_\varepsilon^2 + \sum_{i=1}^{a} l\,\mathrm{og}_e(\sigma_\varepsilon^2 + n_i\sigma_\alpha^2)$$

$$+\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2 + \sum_{i=1}^{a}\frac{n_i(\bar{y}_{i.}-\mu)^2}{\sigma_\varepsilon^2 + n_i\sigma_\alpha^2}$$

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{a}\frac{{n_i}^2(\bar{y}_{i.}-\mu)}{(\sigma_\varepsilon^2 + n_i\sigma_\alpha^2)^2}$$

Equating to zero and solving for μ gives:

$$\tilde{\mu} = \frac{\sum_{i=1}^{a}\frac{n_i\bar{y}_{i.}}{\tilde{\sigma}_\varepsilon^2 + n_i\tilde{\sigma}_\alpha^2}}{\sum_{i=1}^{a}\frac{n_i}{\tilde{\sigma}_\varepsilon^2 + n_i\tilde{\sigma}_\alpha^2}} = \bar{y}_{i.}$$ When n=n_i (balanced data).

$$\frac{\partial L}{\partial \sigma_\varepsilon^2} = -\frac{1}{2}\left[\frac{N-a}{\sigma_\varepsilon^2} + \sum_{i=1}^{a}\frac{1}{\sigma_\varepsilon^2 + n_i\sigma_\alpha^2} - \frac{1}{\sigma_\varepsilon^4}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2 - \sum_{i=1}^{a}\frac{n_i(\bar{y}_{i.}-\mu)^2}{(\sigma_\varepsilon^2 + n_i\sigma_\alpha^2)^2}\right]$$

$$\frac{\partial L}{\partial \sigma_\alpha^2} = -\frac{1}{2}\left[\sum_{i=1}^{a}\frac{n_i}{\sigma_\varepsilon^2 + n_i\sigma_\alpha^2} - \sum_{i=1}^{a}\frac{{n^2}_i(\bar{y}_{i.}-\mu)^2}{(\sigma_\varepsilon^2 + n_i\sigma_\alpha^2)^2}\right]$$

Clearly, there are no explicit solutions for μ, σ_α^2 and σ_ε^2. Even if solutions can be found, when data are unbalanced, the estimator for σ_α^2 can be negative.

COMMENTS

1. Analyses for differentially expressed genes are best carried out via statistical tests rather than using the fold-change method.
2. Further studies could extend the use of mixed models to analyze data from affymetrix chips. This study applied mixed models to cDNA micro arrays.

CHAPTER 3

APPLICATION

INTRODUCTION

In this chapter, we apply the mixed model to determine differential expression as suggested by *R.D Wolfinger et al*, 2001. The data used in this study was obtained from the Internet site *http://genome-www.stanford.edu/swisnf.* The experiment investigates the behavior of yeast genes in five treatments. The mixed model was used to identify genes, which are differentially expressed, that is genes that behave differently depending on the treatment applied.
In this data, a reference design was used. Interpretation is simple, less RNA is used and there is less sensitivity to bad array elements when a reference design is used. All the treatments are compared via the reference treatment. Wild type was used as the reference. We have five treatments, that is wild type, snf2ypd, snf2min, swi1ypd and swi1min. Wild type was mixed with the other four. Wild type was labelled red and all the other four green. It is being used as the reference. Twelve arrays were made, that is, three replicates of each of the four arrays bellow:

Wild Type versus Snf2ypd	Wild Type Versus Snf2min	Wild Type Versus Swi1ypd	Wild Type Versus Swi1min

Measurements of fluorescence intensity x, and other variables were then taken from these twelve arrays using a scanner.

EXPLORATORY ANALYSIS

The SAS program **Data update (**Appendix P1) was used for cleaning the data. The Name column was updated using the Type column. If a name were missing, then the type would be entered. The Name column was then used to update the Gene column. If a gene were missing, then the name would be used. If the Flag value associated with a spot was not zero, then there is a problem with the spot, which can be faint signal, too much background noise or a host of many other problems. That is, the researcher decides what value to associate with a particular problem. In this case, a conservative approach has been taken and all spots with a non-zero flag value have been excluded from the analysis.

The SAS program **Arraycalc,** Appendix P2, was used to allocate arrays to the observations and **logicalculation** , Appendix P3, to calculate the base two logarithms, Y_{gtar} of the channel intensities; y and allocate treatment combinations. Transforming expression data to a log scale removes much of the proportional relationship between random error and signal intensity. Also, distributions of replicated logged expression values tend to be normal. [R. Nadon and J. Shoemaker, Trends in Genetics, 2002].

The variables of interest in this study are spot, name, type, gene, flag, Ch1d, Ch2d, x_{gtar} r and logi. See Chapter 1 for definition of variables. x_{gtar} and logi, (Y_{gtar}) were calculated as follows:
x_{gtar} = ch2d if treatment is wild type and Ch1d if treatment is swi1ypd, swi1min, snf2ypd or snf2min.
Logi = (Y_{gtar})= $\log_2$ (x_{gtar}).
It is convenient to use log base two because each unit, after transformation corresponds to a two-fold difference.
Example
If $\overline{Y}_{ga2}$ =14 and $\overline{Y}_{ga1}$ = 11, then intensity for treatment 2 is $2^{(14-11)= 3}$ = 8 fold greater than intensity for treatment 1.

A frequency procedure, (**frequencyprog**, Appendix P4**)** was performed on the data and a univariate procedure was run on the variables x_{gtar} and logi. The results of the univariate procedure are in appendix A1. They show that negative and zero values of logi = Y_{gtar} were removed. All remaining values appear to be reasonable. We could not reject the null hypothesis that the mean, $\mu_0 = 0$ at 0.0001 level of significance for both variables.

A frequency analysis of the Flag variable yielded the following;

Treatment combination	FLAG	0	1	2	3
1	ARRAY 1	12444	2	4	166
1	2	15734			256
1	3	15054		2	934
2	4	12432			196
2	5	14874			1116
2	6	10640			5350
3	7	12471			145
3	8	15550			440
3	9	15376			614
4	10	12434	62	4	128
4	11	15216			774
4	12	14340			1650

All the observations with a non-zero flag value were discarded. Spots with a negative x_{gtar} value were also excluded from the analysis. Background noise was greater than channel intensity on these particular spots. Zero x_{gtar} values were also discarded; no information can be gained from these spots. Also we cannot calculate logarithm of a number less or equal to zero. Hence no such data was included in the data set analysed.

Treatment	1	1	1	2	2	2	3	3	3	4	4	4
Array	1	2	3	4	5	6	7	8	9	10	11	12
No. Of Obs. $x_{gtar} \leq 0$	384	77	69	1019	235	258	375	80	96	201	120	78

The frequency procedure showed that some genes were spotted more than once on the same array, so the data are unbalanced. A summary of the frequency table is follows:

Array	Number of spots for a specific gene	Number of genes.	Gene Name	Total number of Genes
1	1	6000		
	2	23		
	4	4		
	24	1	GENOMIC	
	41	1	CONTROL	
	95	1	NORF	
				6030
2	1	4857		
	2	1259		
	3	82		
	4	20		
	6	1		
	8	1		
	43	1		
	109	1		
				6222
3	1	4757		
	2	1160		
	3	77		
	4	18		
	5	1	PWP2	
	6	1	GENOMIC 0.	
	35	1	3XSSC	
	101	1	NORF	
				6016
4	1	5959		
	2	56		
	4	4		
	12	1	GENOMIC 1.	
	24	1	GENOMIC 0.	
	93	1	NORF.	
				6022
5	1	4725		
	2	1136		
	3	71		
	4	22		
	5	1	PWP2	
	40	1	3XSSC	
	94	1	NORF	
				5957
6	1	3692		
	2	671		
	3	48		
	4	12		
	26	1	3XSSC	
	68	1	NORF	
				4425
7	1	6012		
		23		
	4	4		
	24	1	GENOMIC	
	41	1	CONTROL	
	96	1	NORF	
				6042
8	1	4851		

	2	1216		
	3	86		
	4	19		
	6	1	PWP2	
	8	1	GENOMIC 0.	
	39	1	3XSSC	
	105	1	NORF	
				6176
9	1	4806		
	2	1207		
	3	80		
	4	18		
	6	1	PWP2	
	7	1	GENOMIC 0.	
	42	1	3XSSC	
	101	1	NORF	
				6115
10	1	5956		
	2	57		
	4	4		
	12	1	GENOMIC 1.	
	24	1	GENOMIC 0.	
	95	1	NORF	
				6020
11	1	4755		
	2	1208		
	3	66		
	4	21		
	6	1	PWP2	
	7	1	GENOMIC 0.	
	42	1	3XSSC	
	100	1	NORF	
				6054
12	1	4651		
	2	1050		
	3	80		
	4	9		
	5	2	PWP2, GENOMIC 0.	
	36	1	3XSSC	
	97	1	NORF	
				5794

3XSSC is used for quality control purposes.

THE MODEL

The program **glmprog,** Appendix P5, was used to calculate the normalization and gene models. The data were normalised to get rid of systematic errors. The residuals from the normalisation model are considered to be the normalised values and are used to calculate the gene models. The normalisation model used is:

$Y_{gtar} = \mu + T_t + A_a + (TA)_{ta} + \varepsilon_{gtar}$,

Where μ represents the overall mean value, T is the main effect for treatments, i=1…5, A is the main effect for arrays, j=1, 2… 12, TA is the interaction effect

between treatments and arrays and ε is the random error. Wolfinger *et al* assumed no replication of spots within arrays. An extra subscript r has been included to cater for the replicates.

The gene models are: $r_{gia} = G_g + (GT)_{gi} + (GA)_{ga} + \gamma_{gia}$

Where r_{gia} is the residual for gene g, in array a with treatment i, from the normalisation model, G_g is the mean response associated with gene g, $(GT)_{gi}$ is the gene-treatment interaction, $(GA)_{gj}$ is the gene-array interaction and γ_{gij} is the random error associated with r_{gij}.

These could be obtained in one step by including gene in the class statement of **Proc Mixed**. However, this is heavy on computer resources. A similar analysis can be done for each gene, using the **BY** option in **Proc Mixed.** The **BY** statement is used to obtain separate analyses on observations in groups defined by the variables used in the **BY** statement. Separate analysis yields the following model for a particular gene;

$r_{ij} = \mu + T_i + A_j + \gamma_{ij}$.

Where r_{ij} is the residual for a gene in treatment i and array j, μ is the mean response associated with a particular gene, T_i is the effect of treatment i on the particular gene, A_j is the effect of array j on the gene and γ_{ij} is the random error associated with r_{ij}.

An artificial data set was used to show that the two approaches are equivalent. See appendix A2 for results.

Consider gene1, treatment1 and array1, Defining gene as a class variable yielded equation 1and equation 2. was obtained using the **BY** statement;

1. $r_{gij} = 10.5354 - 2.9887 + 0.8496 - 1.1376 = 7.2587$
2. $r_{ij} = 7.5467 + 0.8496 - 1.1376 = 7.2587$

Note: The value of the treatment effect obtained using the BY gene option is exactly the same as the gene-treatment effect obtained using gene in the class statement. The same goes for the array effect and the gene-array effect. However, we get a better fit, R^2=0.307 using gene as a class than when we do separate analyses for the genes, $R^2 \leq 0.123$ for each of the three genes. We also have fewer degrees of freedom when we do separate analyses.

The differences in $\log_2$ means used in the volcano plots were obtained by using the **ls-means / diff control** option in **Proc Mixed.** The procedure constructs an approximate t-test to test the null hypothesis that the associated population parameter equals zero. In this case, it tests the hypothesis that the difference in

treatment means is equal to zero. That is, the particular gene for which the hypothesis is being tested, behaves in the same way in the treatments under consideration.

RESULTS AND DISCUSSION

The p-values obtained from the t-tests used to compare each of the four experimental treatments to the reference treatment and the fold changes are presented in the form of volcano plots. Thus we can see at a glance which genes are differentially expressed. Genes in the top left and right hand corners of the volcano plots are the genes that are both statistically significant and biologically important. The lower left and right hand corners are false positives. If we had only considered fold change, they would have been declared differentially expressed. But they have little or no statistical significance which means that most of the variation is due to chance. The top middle sction of the volcano plots has high statistical significance but has little biological importance.

(i) When comparing snf2ypd to wildtype, the following genes were found to statistically significant and biologically important; YDR033W and CY30.1X

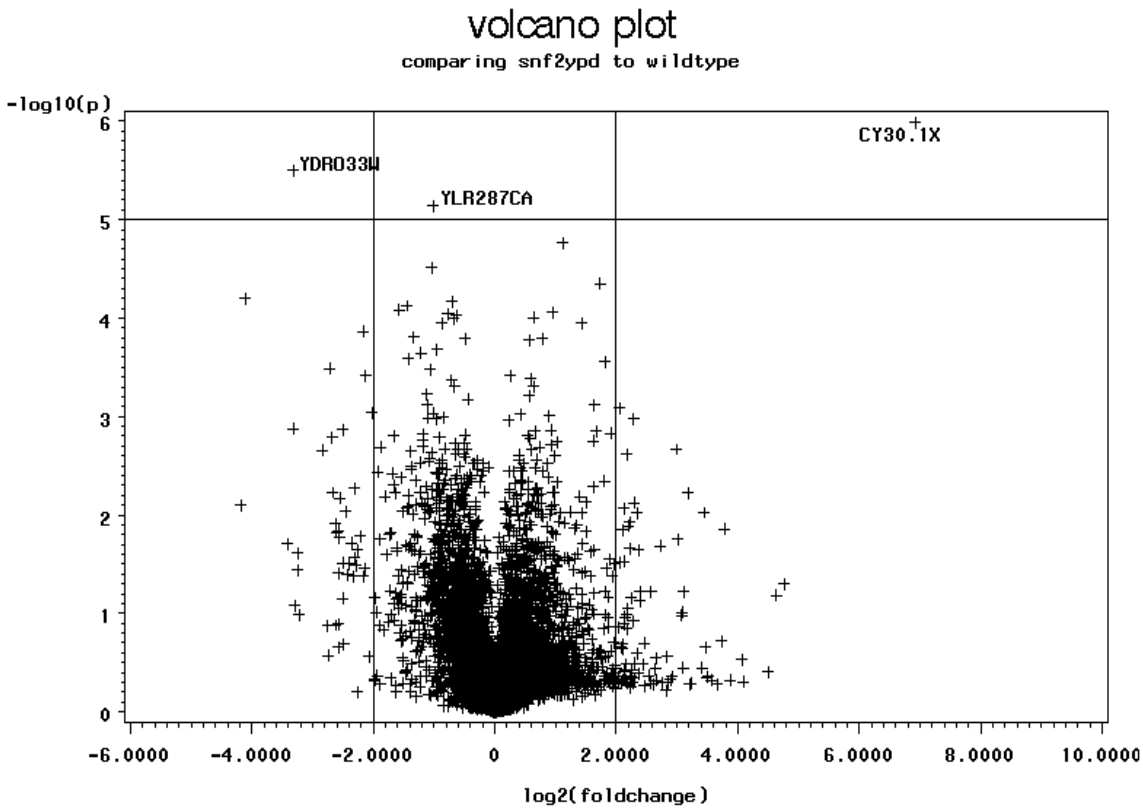
(ii) When comparing snf2ymin to wildtype, the following genes were found to statistically significant and biologically important; ASP3, YDR033W, HSP30, CY30.1X

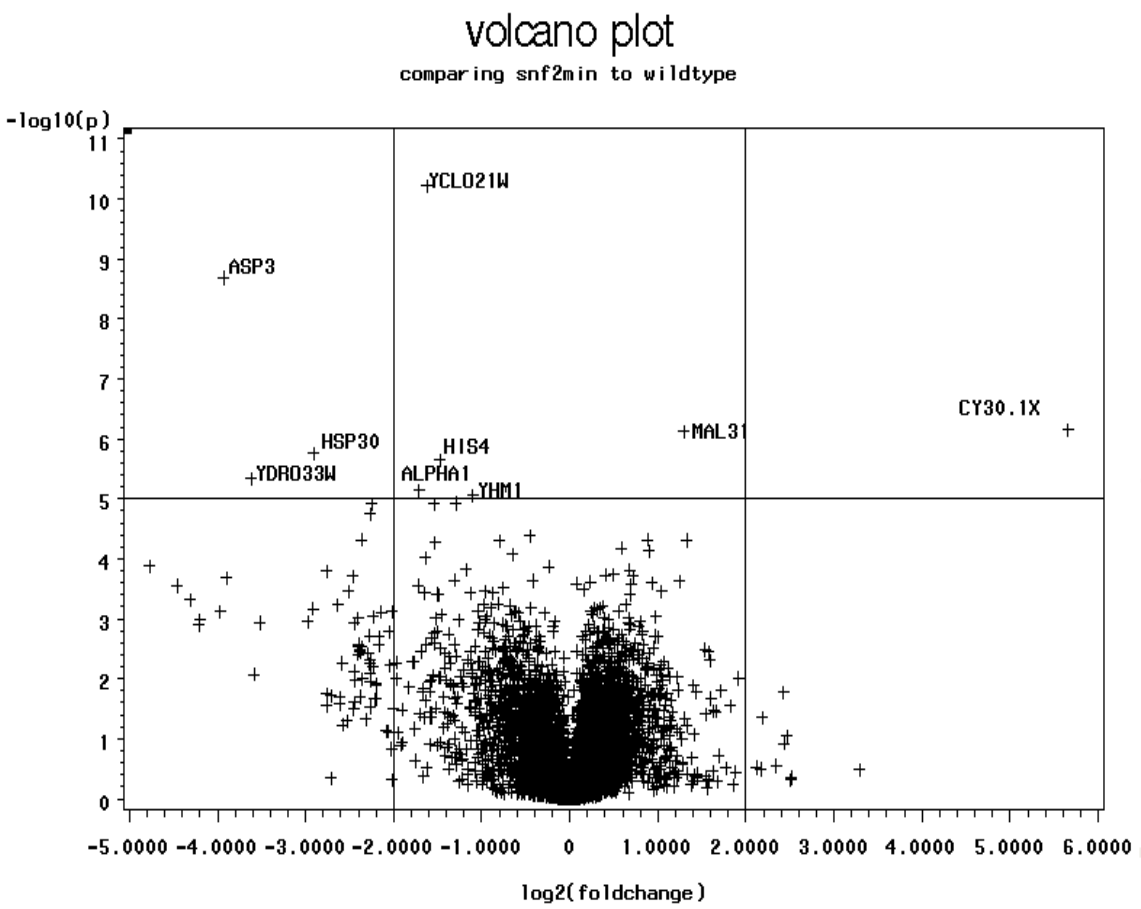
(iii) When comparing sw1ypd to wildtype, the following genes were found to statistically significant and biologically important; YBR244W, ZE01

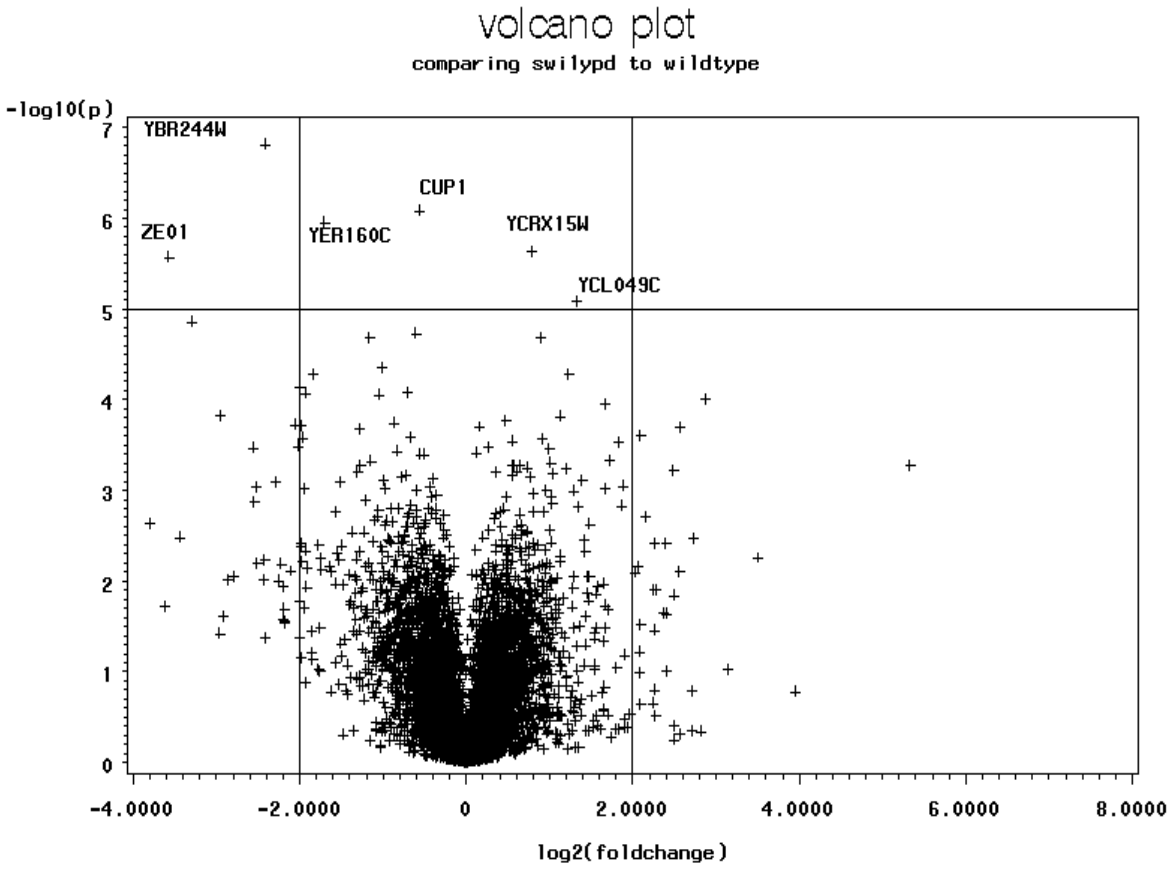
(iv) When comparing swi1min to wildtype, the following genes were found to statistically significant and biologically important; ASP3, YDR033W, YER160C

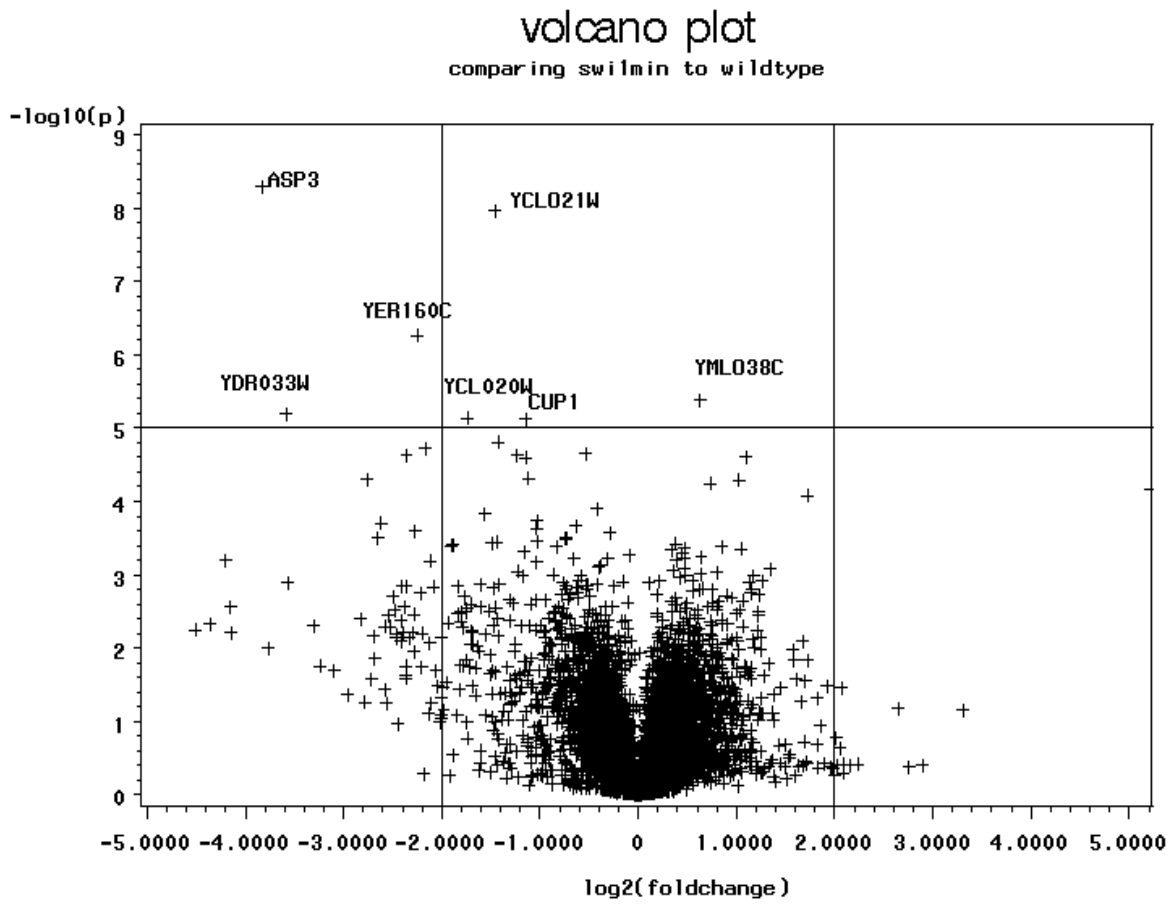
The genes that were found to be both statistically significant and biologically important are the genes that should be considered for further studies. The graph for ‘All pair-wise comparisons’ is similar to the graph obtained by Wolfinger *et al, Fig. 1. B.*

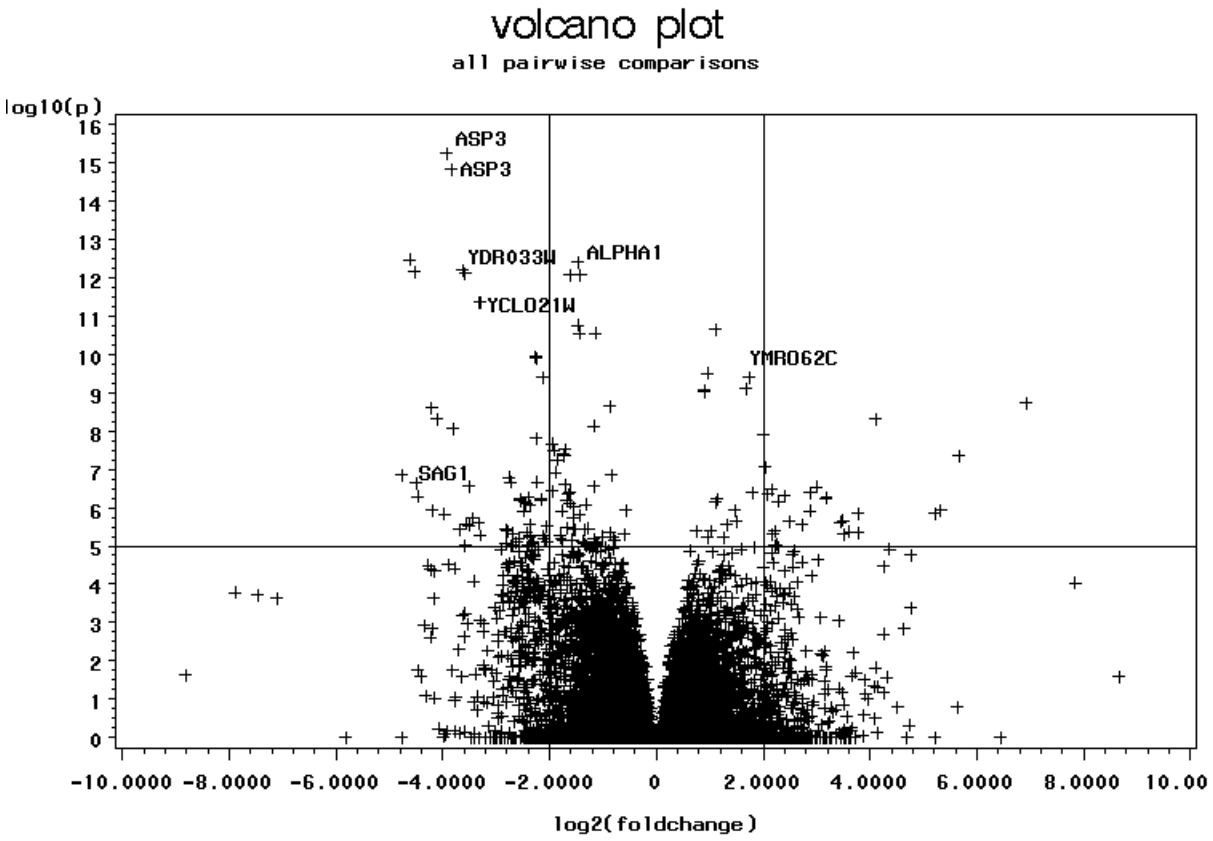
The volcano plots are follow:











CONCLUSION

The methodology used in this study enables us to statistically infer differential gene expression in way that controls both false positives and false negatives. False positives are genes that are declared to be differentially expressed but in fact are not differentially expressed. False negatives are genes that are declared to be not differentially expressed but in reality are differentially expressed. Some genes may have high fold change but exhibit a lot of variation across arrays and thus possess little statistical significance. Combining p-values obtained from the gene models with biological knowledge ensures that researchers do not chase noise or chance variation. The methodology can be used as a precursor to clustering to ensure that the inputs are statistically meaningful. It can also be used after clustering to explore and validate implied associations. The gene models were fit separately for each gene. This enables identification of genes that are separately differentially expressed. However, there may be pairs or groups of genes that are differentially expressed jointly but not so much individually. Further work is needed to explore how to identify such pairs or groups.

Appendix P1: Dataupdate

```
data wolfinge.testdata1;
set wolfinge.testdata1;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;


data wolfinge.testdata2;
set wolfinge.testdata2;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata3;
set wolfinge.testdata3;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
    If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata4;
```

```
set wolfinge.testdata4;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata5;
set wolfinge.testdata5;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata6;
set wolfinge.testdata6;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata7;
set wolfinge.testdata7;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
```

```
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata8;
set wolfinge.testdata8;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata9;
set wolfinge.testdata9;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata10;
set wolfinge.testdata10;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
```

```
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata11;
set wolfinge.testdata11;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
data wolfinge.testdata12;
set wolfinge.testdata12;
name = upcase (name);
      type = upcase (type);
      gene = upcase (gene);
      If (name=" ") then name=type;
      If (gene=" ") then gene =name;
      If (flag = "1") then delete;
      If (flag ="2") then delete;
      If (flag ="3") then delete;
      If (treatment="wildtype") then y=ch2d;
      else y=ch1d;
      If (y<="0") then delete;
      output;
      run;
```

Appendix P2: ARRAYCALC

```
data wolfinge.a1;
set wolfinge.testdata1;
a=1;
drop array;
run;
data wolfinge.a2;
set wolfinge.testdata2;
a=2;
drop array;
run;
data wolfinge.a3;
set wolfinge.testdata3;
a=3;
drop array;
run;
data wolfinge.a4;
set wolfinge.testdata4;
a=4;
drop array;
run;
data wolfinge.a5;
set wolfinge.testdata5;
a=5;
drop array;
run;
data wolfinge.a6;
set wolfinge.testdata6;
a=6;
drop array;
run;
data wolfinge.a7;
set wolfinge.testdata7;
a=7;
drop array;
run;
data wolfinge.a8;
set wolfinge.testdata8;
a=8;
drop ARRAY;
run;
data wolfinge.a9;
set wolfinge.testdata9;
a=9;
drop ARRAY;
run;
data wolfinge.a10;
set wolfinge.testdata10;
```

```
a=10;
drop array;
run;
data wolfinge.a11;
set wolfinge.testdata11;
a=11;
drop array;
run;
data wolfinge.a12;
set wolfinge.testdata12;
a=12;
drop array;
run;
data wolfinge.all;
set wolfinge.a1 wolfinge.a2 wolfinge.a3 wolfinge.a4 wolfinge.a5
wolfinge.a6
wolfinge.a7 wolfinge.a8 wolfinge.a9 wolfinge.a10 wolfinge.a11
wolfinge.a12;
run;
```

Appendix P3: LOGICALULATION

```
data wolfinge.a;
set wolfinge.all;
logi=log2(y);
output;
keep spot a treatment name gene y logi;
run;
quit;
data wolfinge.a;
set wolfinge.a;
If (a="1" or a="2" or a="3") then combination="1";
run;
quit;
data wolfinge.a;
set wolfinge.a;
If (a="4" or a="5" or a="6") then combination="2";
run;
quit;
data wolfinge.a;
set wolfinge.a;
If (a="7" or a="8" or a="9") then combination="3";
run;
quit;
data wolfinge.a;
set wolfinge.a;
If (a="10" or a="11" or a="12") then combination="4";
run;
quit;
```

Appendix P4: FREQUECYPROG

```
data s;
set wolfinge.all;
proc sort data =s out = t;
by gene ;
proc freq data=t;
tables gene / out=wolfinge.freq;
run;
proc freq data=wolfinge.freq;
tables gene;
run;


data s1;
set wolfinge.a1;
proc sort data =s1 out = t1;
by gene;
proc freq data=t1;
tables gene /out=wolfinge.freq1;
run;
proc freq data=wolfinge.freq1 ;
run;
data s2;
set wolfinge.a2;
proc sort data =s2 out = t2;
by gene;
proc freq data=t2;
tables gene /out=wolfinge.freq2;
proc freq data=wolfinge.freq2;
run;
data s3;
set wolfinge.a3;
proc sort data =s3 out = t3;
by gene;
proc freq data=t3;
tables gene /out=wolfinge.freq3 ;
run;
proc freq data=wolfinge.freq3;
run;

data s4;
set wolfinge.a4;
proc sort data =s4 out = t4;
by gene;
proc freq data=t4;
tables gene /out=wolfinge.freq4;
run;
proc freq data=wolfinge.freq4;
```

```
run;

data s5;
set wolfinge.a5;
proc sort data =s5 out = t5;
by gene;
proc freq data=t5;
tables gene / out=wolfinge.freq5;
run;
proc freq data=wolfinge.freq5;
run;

data s6;
set wolfinge.a6;
proc sort data =s6 out = t6;
by gene;
proc freq data=t6;
tables gene / out=wolfinge.freq6;
run;
proc freq data=wolfinge.freq6;
run;

data s7;
set wolfinge.a7;
proc sort data =s7 out = t7;
by gene;
proc freq data=t7;
tables gene / out=wolfinge.freq7;
run;
proc freq data=wolfinge.freq7;
run;

data s8;
set wolfinge.a8;
proc sort data =s8 out = t8;
by gene;
proc freq data=t8;
tables gene / out=wolfinge.freq8;
run;
proc freq data=wolfinge.freq8;
run;

data s9;
set wolfinge.a9;
proc sort data =s9 out = t9;
by gene;
proc freq data=t9;
tables gene / out=wolfinge.freq9;
run;
proc freq data=wolfinge.freq9;
```

```
run;

data s10;
set wolfinge.a10;
proc sort data =s10 out = t10;
by gene;
proc freq data=t10;
tables gene / out=wolfinge.freq10;
run;
proc freq data=wolfinge.freq10;
run;

data s11;
set wolfinge.a11;
proc sort data =s11 out = t11;
by gene;
proc freq data=t11;
tables gene / out=wolfinge.freq11;
run;
proc freq data=wolfinge.freq11;
run;

data s12;
set wolfinge.a12;
proc sort data =s12 out = t12;
by gene;
proc freq data=t12;
tables gene / out=wolfinge.freq12;
run;
proc freq data=wolfinge.freq12;
run;
quit;
proc univariate data=wolfinge.all;
var y;
run;

proc univariate data=wolfinge.a;
var logi;
run;
```

Appendix P5: GLMPROG

```
/* normalization of data*/
proc mixed data = wolfinge.a;
class a treatment;
model logi = a treatment a*treatment/outp
=wolfinge.residual(keep=a gene name spot treatment resid
combination) ;
random a a*treatment;
run;
quit;

/*calculation of gene models*/
proc sort data=wolfinge.residual;
by gene a spot;
run;
quit;
ods listing close;
run;
proc mixed data=wolfinge.residual ;
where (combination="1");
by gene;
class a treatment ;
model resid =a treatment;
lsmeans treatment/diff adjust=bon ;
ods output diffs=wolfinge.pval;
run;
quit;
data wolfinge.pval1;
set wolfinge.pval;
logp=log10(adjp);
neglogp=-logp;
run;
proc gplot data=wolfinge.pval1;
label estimate=log2(foldchange) neglogp=-log10(p);
plot neglogp*estimate/href=-2 2 vref=5;
title volcano plot;
title2 comparing snf2min to wildtype;
run;
quit;
ods listing close;
run;
proc mixed data=wolfinge.residual ;
where (combination="2");
by gene;
class a treatment ;
model resid =a treatment;
lsmeans treatment/diff adjust=bon ;
```

```
ods output diffs=wolfinge.pval2;
run;
quit;
data wolfinge.pval21;
set wolfinge.pval2;
logp=log10(adjp);
neglogp=-logp;
run;
proc gplot data=wolfinge.pval21;
label estimate=log2(foldchange) neglogp=-log10(p);
plot neglogp*estimate/href=-2 2 vref=5;
title volcano plot;
title2 comparing snf2ypd to wildtype;
run;
quit;
ods listing close;
run;
proc mixed data=wolfinge.residual ;
where (combination="3");
by gene;
class a treatment ;
model resid =a treatment;
lsmeans treatment/diff adjust=bon ;
ods output diffs=wolfinge.pval3;
run;
quit;
data wolfinge.pval31;
set wolfinge.pval3;
logp=log10(adjp);
neglogp=-logp;
run;
proc gplot data=wolfinge.pval31;
label estimate=log2(foldchange) neglogp=-log10(p);
plot neglogp*estimate/href=-2 2 vref=5;
title volcano plot;
title2 comparing swi1min to wildtype;
run;
quit;
ods listing close;
run;
proc mixed data=wolfinge.residual ;
where (combination="4");
by gene;
class a treatment ;
model resid =a treatment;
lsmeans treatment/diff adjust=bon ;
ods output diffs=wolfinge.pval4;
run;
quit;
data wolfinge.pval41;
```

```
set wolfinge.pval4;
logp=log10(adjp);
neglogp=-logp;
run;
proc gplot data=wolfinge.pval41;
label estimate=log2(foldchange) neglogp=-log10(p);
plot neglogp*estimate/href=-2 2 vref=5;
title volcano plot;
title2 comparing swi1ypd to wildtype;
run;
quit;
```

APPENDIX A1: Frequency procedure

A univariate procedure was run on the variables x_{gtar} and $\log i = Y_{gtar}$ r. The results of the univariate procedure are in appendix A1. They show that we managed to remove negative and zero values of $\log i = Y_{gtar}$. All remaining values appear to be reasonable.

The UNIVARIATE Procedure

Variable: Y= CH2D if treatment is wild type
Y= CH1D otherwise.

Moments

N	163572	Sum Weights	163572
Mean	1003.44056	Sum Observations	164134780
Std Deviation	2152.48125	Variance	4633175.55
Skewness	7.92215385	Kurtosis	100.229294
Uncorrected SS	9.22553E11	Corrected SS	7.57853E11
Coeff Variation	214.51009	Std Error Mean	5.32212286

Basic Statistical Measures

Location		Variability	
Mean	1003.441	Std Deviation	2152
Median	395.000	Variance	4633176
Mode	77.000	Range	55248
		Interquartile Range	821.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	188.5414	Pr > |t|	<.0001
Sign	M	81786	Pr >= |M|	<.0001
Signed Rank	S	6.689E9	Pr >= |S|	<.0001

Quantile	Estimate
100% Max	55249
99%	9884
95%	3774
90%	2273
75% Q3	980
50% Median	395
25% Q1	159
10%	61
5%	32
1%	7
0% Min	1

Extreme Observations

----Lowest----		----Highest----	
Value	Obs	Value	Obs
1	149342	50566	85311
1	149341	51260	12059
1	149340	51835	54115
1	149339	52474	12060
1	149338	55249	85310

The UNIVARIATE Procedure
Variable: logi = $\log_2(y)$

Moments

N	163572	Sum Weights	163572
Mean	8.56448793	Sum Observations	1400910.42
Std Deviation	2.10030277	Variance	4.41127172
Skewness	-0.3258411	Kurtosis	0.71782915
Uncorrected SS	12719636.5	Corrected SS	721556.127
Coeff Variation	24.5233899	Std Error Mean	0.00519311

Basic Statistical Measures

Location		Variability	
Mean	8.564488	Std Deviation	2.10030
Median	8.625709	Variance	4.41127
Mode	6.266787	Range	15.75366
		Interquartile Range	2.62375

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	1649.202	Pr > |t|	<.0001
Sign	M	81657.5	Pr >= |M|	<.0001
Signed Rank	S	6.668E9	Pr >= |S|	<.0001

Quantile	Estimate
100% Max	15.75366
99%	13.27088
95%	11.88188
90%	11.15038
75% Q3	9.93664
50% Median	8.62571
25% Q1	7.31288
10%	5.93074
5%	5.00000
1%	2.80735
0% Min	0.00000

Extreme Observations

----Lowest----		-----Highest-----	
Value	Obs	Value	Obs
0	149342	15.6259	85311
0	149341	15.6455	12059
0	149340	15.6616	54115
0	149339	15.6793	12060
0	149338	15.7537	85310

APPENDIX A2: Use of BY option and Class statement in calculating gene models

The gene models $r_{gia} = G_g + (GT)_{gi} + (GA)_{ga} + \gamma_{gia}$ could be obtained in one step by including gene in the class statement of **Proc Mixed**. However, this is heavy on computer resources. An analysis can be done for each gene, using the **BY** option in **Proc Mixed.** The **BY** statement is used to obtain separate analyses on observations in groups defined by the variables used in the **BY** statement. Separate analysis yields the following model for a particular gene:
$r_{ij} = \mu + T_i + A_j + \gamma_{ij}$.

An artificial data set was used to show that the two approaches are equivalent.

```
                                The GLM Procedure
                            Class Level Information
                          Class          Levels    Values
                          gene                3    1 2 3
                          array               2    1 2
                          treatment           3    1 2 3
                          Number of observations    165
Dependent Variable: logi

                                        Sum of
   Source                       DF          Squares      Mean Square    F Value    Pr
> F
   Model                        11       448.711318        40.791938       6.17
<.0001
   Error                       153      1011.147713         6.608809
   Corrected Total             164      1459.859030

                       R-Square     Coeff Var      Root MSE     logi Mean
                       0.307366      29.49555      2.570760      8.715758

   Source                       DF       Type I SS     Mean Square    F Value    Pr
> F
   gene                          2      369.3490964     184.6745482      27.94
<.0001
   gene*array                    3       25.5541176       8.5180392       1.29
0.2803
   gene*treatment                6       53.8081037       8.9680173       1.36
0.2355

   Source                       DF     Type III SS     Mean Square    F Value    Pr
> F
   gene                          2      269.3324214     134.6662107      20.38
<.0001
```

```
   gene*array                  3      27.5782986       9.1927662       1.39
0.2477
   gene*treatment              6      53.8081037       8.9680173       1.36
0.2355

                                                          Standard
        Parameter                    Estimate             Error    t Value    Pr >
|t|
        Intercept                10.53541530 B      0.68414518      15.40
<.0001
        gene           1         -2.98869312 B      1.16373827      -2.57
0.0112
        gene           2         -2.05270624 B      0.90219175      -2.28
0.0243
        gene           3          0.00000000 B       .                .         .
        gene*array     1 1        0.84962632 B      0.81908120       1.04
0.3012
        gene*array     1 2        0.00000000 B       .                .         .
        gene*array     2 1        1.15742053 B      0.66450621       1.74
0.0836
        gene*array     2 2        0.00000000 B       .                .         .
        gene*array     3 1        0.19849758 B      0.78955077       0.25
0.8018
        gene*array     3 2        0.00000000 B       .                .         .
        gene*treatment 1 1       -1.13761289 B      0.92795868      -1.23
0.2221
        gene*treatment 1 2       -1.90948946 B      0.87226896      -2.19
0.0301
        gene*treatment 1 3        0.00000000 B       .                .         .
        gene*treatment 2 1       -1.09719671 B      0.91526110      -1.20
0.2325
        gene*treatment 2 2       -1.16209391 B      0.72260487      -1.61
0.1099
        gene*treatment 2 3        0.00000000 B       .                .         .
        gene*treatment 3 1        0.33730106 B      1.09857148       0.31
0.7592
        gene*treatment 3 2        0.43212778 B      0.88288954       0.49
0.6252
        gene*treatment 3 3        0.00000000 B       .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was
used to solve the normal equations. Terms whose estimates are followed by the
letter 'B' are not uniquely estimable.
------------------------------------------- gene=1 -----------------------------
----
                                  The GLM Procedure

                              Class Level Information
                            Class          Levels    Values
                            array               2    1 2
                            treatment           3    1 2 3
```

Number of observations 54

Dependent Variable: logi

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	40.2080091	13.4026697	1.90	0.1421
Error	50	353.2995835	7.0659917		
Corrected Total	53	393.5075926			

R-Square	Coeff Var	Root MSE	logi Mean
0.102178	37.86400	2.658193	7.020370

Source	DF	Type I SS	Mean Square	F Value	Pr > F
array	1	8.50969391	8.50969391	1.20	0.2777
treatment	2	31.69831521	15.84915760	2.24	0.1167

Source	DF	Type III SS	Mean Square	F Value	Pr > F
array	1	7.11090976	7.11090976	1.01	0.3206
treatment	2	31.69831521	15.84915760	2.24	0.1167

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		7.546722178	B	0.97341653	7.75	<.0001
array	1	0.849626319	B	0.84693861	1.00	0.3206
array	2	0.000000000	B	.	.	.
treatment	1	-1.137612892	B	0.95951907	-1.19	0.2414
treatment	2	-1.909489457	B	0.90193531	-2.12	0.0392
treatment	3	0.000000000	B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

```
------------------------------------------- gene=2 ----------------------------
----

                                The GLM Procedure
                            Class Level Information
                          Class          Levels    Values
                          array               2    1 2
                          treatment           3    1 2 3
                          Number of observations    63
Dependent Variable: logi
                                       Sum of
   Source                      DF         Squares     Mean Square    F Value    Pr
> F
   Model                        3      36.8014912      12.2671637       2.78
0.0491
   Error                       59     260.6616834       4.4179946
   Corrected Total             62     297.4631746

                      R-Square     Coeff Var      Root MSE     logi Mean
                      0.123718      24.54038      2.101903      8.565079

   Source                      DF       Type I SS     Mean Square    F Value    Pr
> F
   array                        1     16.53964030     16.53964030       3.74
0.0578
   treatment                    2     20.26185086     10.13092543       2.29
0.1099

   Source                      DF      Type III SS     Mean Square    F Value    Pr
> F
   array                        1     20.04968019     20.04968019       4.54
0.0373
   treatment                    2     20.26185086     10.13092543       2.29
0.1099
                                                      Standard
             Parameter            Estimate             Error    t Value    Pr >
|t|
             Intercept         8.482709050 B      0.48086520      17.64
<.0001
             array     1       1.157420530 B      0.54331294       2.13
0.0373
             array     2       0.000000000 B       .                .         .
             treatment 1      -1.097196709 B      0.74833492      -1.47
0.1479
             treatment 2      -1.162093909 B      0.59081551      -1.97
0.0539
             treatment 3       0.000000000 B       .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was
used to solve the normal equations. Terms whose estimates are followed by the
letter 'B' are not uniquely estimable.
```

```
---------------------------------------------- gene=3 -----------------------------
----
                                   The GLM Procedure
                                Class Level Information
                              Class          Levels    Values
                              array               2    1 2
                              treatment           3    1 2 3
                              Number of observations    48

Dependent Variable: logi

                                          Sum of
   Source                       DF         Squares     Mean Square    F Value    Pr
> F
   Model                         3       2.3527210       0.7842403       0.09
0.9669
   Error                        44     397.1864457       9.0269647
   Corrected Total              47     399.5391667

                       R-Square     Coeff Var      Root MSE     logi Mean
                       0.005889      27.76580      3.004491      10.82083

   Source                       DF       Type I SS     Mean Square    F Value    Pr
> F
   array                         1      0.50478336      0.50478336       0.06
0.8142
   treatment                     2      1.84793761      0.92396880       0.10
0.9029

   Source                       DF     Type III SS     Mean Square    F Value    Pr
> F
   array                         1      0.41770866      0.41770866       0.05
0.8307
   treatment                     2      1.84793761      0.92396880       0.10
0.9029

                                                    Standard
              Parameter           Estimate             Error    t Value    Pr >
|t|
              Intercept        10.53541530 B      0.79957197      13.18
<.0001
              array     1       0.19849758 B      0.92276125       0.22
0.8307
              array     2       0.00000000 B       .                .         .
              treatment 1       0.33730106 B      1.28391894       0.26
0.7940
              treatment 2       0.43212778 B      1.03184783       0.42
0.6774
              treatment 3       0.00000000 B       .                .         .
```

```
NOTE: The X'X matrix has been found to be singular, and a generalized inverse was
used to solve the normal equations. Terms whose estimates are followed by the
letter 'B' are not uniquely estimable.
```

Consider gene1, treatment1 and array1.

$r_{gij} = 10.5354 - 2.9887 + 0.8496 - 1.1376 = 7.2587$

$r_{ij} = 7.5467 + 0.8496 - 1.1376 = 7.2587$

Note: The treatment and array effects are exactly the same as the gene-treatment and gene-array interactions respectively. Hence the two approaches are equivalent.

REFERENCE

1. R.S Searle (1997): Linear Models, Wiley Classics Library edition.
2. Dov Stekel (2003): Microarray Bioinformatics, Cambridge University Press.
3. Wolfinger et al (2001): Assessing Gene Significance from cDNA Microarray Expression Data via Mixed Models. In: Journal of Computational Biology.
4. Terry Speed (2003): Statistical Analysis of Gene Expression Microarray Data, Chapman and Hall/CRC Press.
5. Efron B, Tibishirani R, Goss V, Chu G, (2000): Microarray and their use in a comparative experiment. http://www-stat.stanford.edu/~tibs/research.html.
6. R. Nadon and J Shoemaker, (2002) : Statistical issues with microarrays: processing and analysis: Trends in Genetics volume 18.
7. R. J. Templeman (2005) : Assessing Statistical precision, power and robustness of alternative experimental designs for two colour microarray platforms based on mixed effects models. In: Veterinary Immunology and Immunopathology 105.