Spring 5-31-2006

Comparative analysis of parametric, nonparametric and permutation methods for differential expression

Rahul Patil
New Jersey Institute of Technology

ABSTRACT

COMPARATIVE ANALYSIS OF PARAMETRIC, NONPARAMETRIC AND PERMUTATION METHODS FOR DIFFERENTIAL EXPRESSION

by

Rahul Patil

DNA microarrays permit us to study the expression of thousands of genes simultaneously. They are now used in many different contexts to compare mRNA levels between two or more samples of cells. Microarray experiments typically give us expression measurements on a large number of genes. Increasing popularity of microarray technology has resulted in a number of tests being proposed to detect differentials expression.

The purpose of study is to compare the parametric, non parametric and permutation tests when applied to microarray data for differential expression analysis. t test (parametric), Mann Whitney test (nonparametric) and Significance of analysis (permutation) test are compared. The study focused on comparison of tests based on the ranking of genes by different tests. Biological and simulation data was used to test compare the performance of statistical tests. The result shows that the SAM test outperform the other two tests, under Normal as well as Lognormal data simulation in case of both low and high number of replicates. Application to simulated data also brings out the fact that with increase in number of replicates the performance all the tests improves.

COMPARATIVE ANALYSIS OF PARAMETRIC, NONPARAMETRIC AND PERMUTATION METHODS FOR DIFFERENTIAL EXPRESSION

by

Rahul Patil

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
In Partial Fulfillment of Requirement for the Degree of
Master of Science in Computational Biology

Department of Computer Science

May 2006

APPROVAL PAGE

COMPARATIVE ANALYSIS OF PARAMERTRIC, NONPARAMETRIC AND PERMUTATION METHODS FOR DIFFERENTIAL EXPRESSION

Rahul Patil

Dr. Michael Recce, Thesis Advisor — Date
Associate Professor of Information Systems, NJIT

Dr. Marc Qun Ma, Committee Member — Date
Assistant Professor of Computer Science, NJIT

Dr. Usman W. Roshan, Committee Member — Date
Assistant Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Rahul Patil

Degree: Master of Science

Date: May 2006

Undergraduate and Graduate Education:

- Master of Science in Computational Biology
  New Jersey Institute of Technology, Newark, NJ, 2006

- Bachelor of Engineering
  Thadomal Shahani Engineering College, Mumbai, India, 1999

Major: Computational Biology

To

Shri Satyanarayan Goenka, my Parents and Kavita

ACKNOWLEDGEMENT

I would like to thank my Thesis advisor Dr. Michael Recce. This work would not have been possible without his support and encouragement. I have learned so much from him. I would also like to thank Dr. Marc Ma and Dr. Usman Roshan for reviewing my work and offering the constructive comments.

Finally, I would like to thank my friends Alex Patterson, Haibo Zhang and Ajay Diwakran for their support and encouragement in this endeavor.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES
(Continued)

CHAPTER 1

INTRODUCTION

Historically, the gene expression studies in molecular biology have focused on a very limited number of genes at one time. But with publication of a historical paper in 1995 [1] in science magazine changed the way gene expression studies are conducted and more importantly the scale on which they are conducted. This paper described a new technology allowing the quantitative, simultaneous monitoring of the expression of thousands of genes using a new tool termed a DNA microarray. Since then there has been a deluge of microarray papers in last decades focusing on different aspects of microarray technology.

The closest analogy one can find to compare the microarray technology is the revolution in computer industry over last couple of decades with more powerful computers and falling prices, it has changes the way people work and live, one of the most remarkable aspect of this revolution has been the tremendous improvement in the productivity of the individual and the society as whole. Similarly, with researchers putting more spots per chip the cost of microarray is reducing making it possible for even the modest budget institutes to perform microarray experiments and in the process enables a large number of scientist to perform novel experiments, in the process this has made it possible not only to do things better but also to generate new hypothesis and test it which would have been impossible with out the scale of microfarad. Clearest example of this is the expression profiles for various kinds of cancers generated in last couple of years. This will be even more evident in coming years when microarray makes forays into the clinical diagnostics.

Although microarray technology has origins into monitoring DNA expressions, it has since spanned into other areas of molecular biology like DNA copy number, DNA protein interactions and DNA sequencing, SNP detection etc.

In the differential expression experiments, the expression of a gene is measured by ability to hybridize to a target sequence localized to a specific region on a chip. To measure this hybridization, RNA, extracted from a biological sample of interest, is reverse-transcribed into cDNA that ideally represents a quantitative copy of genes expressed at the time of sample collection. This cDNA is labeled with a molecule such as fluorescent nucleotide. The labeled cDNA is then hybridized to the DNA chip that contains thousands of gene targets. Ideally, each molecule in the labeled cDNA will only bind to its appropriate complementary target sequence on the array. The measurement of hybridization allows measurement of the amount of labeled cDNA that hybridized to each target sequence, resulting in the identification and relative quantification of the genes expressed in the original biological sample. The overall microarray experimental procure is depicted in Figure 1.1

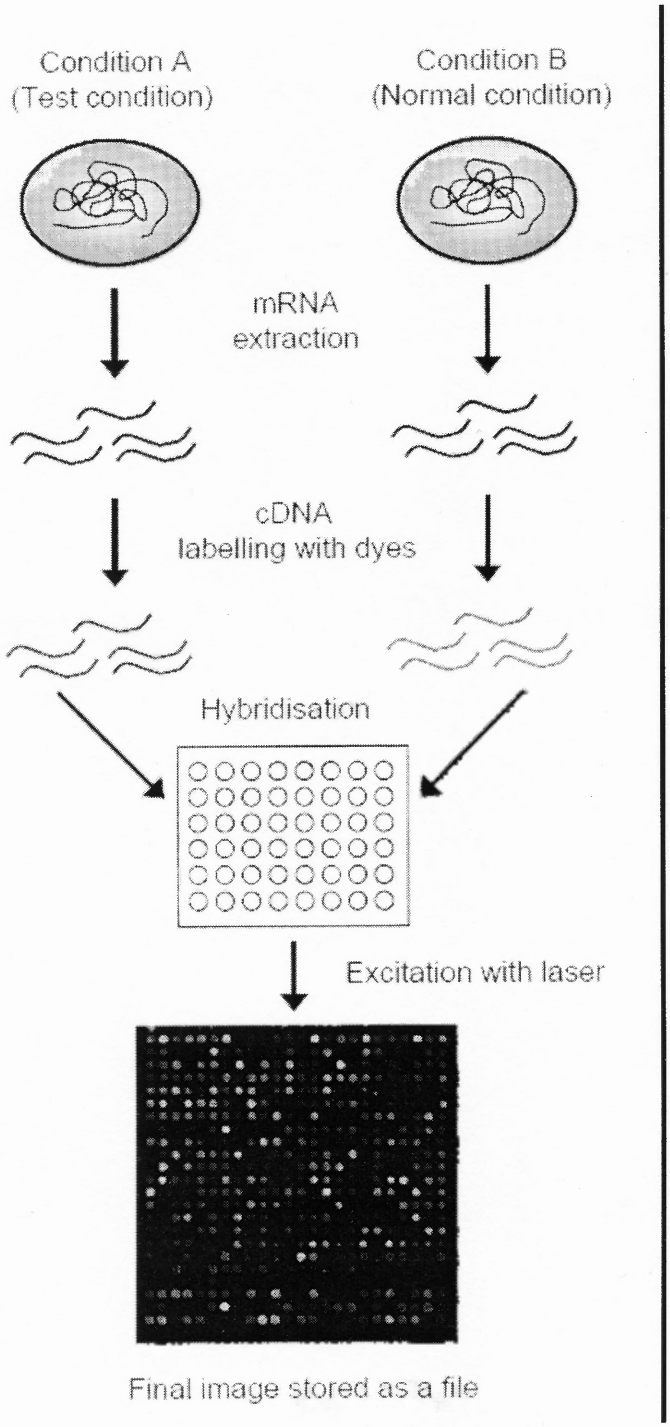

Figure 1.1 Microarray experimental process.

There is couple of variations of microarray technology. These include deposition or spotted cDNA, spotted oligomer, or synthesized oligomer chips. The cDNA microarray consists of a collection of partial gene sequences that are spotted individually

into precise locations within the DNA chip. These sequences usually range in size from 500–2000 bp and may be chosen from different regions of the gene depending on the goal of the project. When using a microarray to monitor gene expression, ideally the investigator must detect the precise gene that is affected without cross-reactivity with other family members. Advantage of these cDNA chips is their ease of use and attainability. Complete instructions on their manufacture within a laboratory are freely available. The ability to manufacture chips within a research laboratory or an institute provides the advantages of flexibility and customization of design as needed by the scientific goal of the project.

Another form of DNA microarray is based on deposition or on-chip synthesis of oligonucleotides for generation of targets. Several approaches to the manufacturing of these chips lead to the same end result, a chip that contains short oligomers ranging from 25–80 bases as the target sequences. One disadvantage of these oligomer-based chips is their availability only from commercial manufacturers. The cost of custom development of an oligomer-based chip is beyond the budget of a majority of researchers, but more affordable, oligomer-based chips are available and may be sufficiently informative for researchers.

Despite rapid technological developments, the statistical tools required to analyze these fundamentally different types of DNA microarray data are not in place. Data often consist of expression measures for thousands of genes, but experimental replication at the level of single genes is often low. This creates problems of statistical inferences because many genes will show fairly large changes in gene expression purely by chance alone.

Therefore, to interpret data from DNA microarray it is necessary to employ statistical methods capable of distinguishing chance occurrences from biologically meaningful data.

The next chapter presents the ways to visually inspect the microarray data and inference one can draw from it. Chapter 3 presents some of the popular statistical tests used for detection of differential expression. Since the focus of this study is the t test, Mann Whitney test and Significance of analysis test, detailed explanation of these tests is given, while others are described briefly. Finally, results and conclusion are presented.

CHAPTER 2

NATURE OF MICROARRAY DATA

The statistical tests applied to the data are influenced by the nature of data, and so it's worth exploring the kind of data microarray experiments generate. The best way to inspect the data is through visual inspection. The visual inspection of data makes it easy to understand and interpret the data, also one can get the idea about the outliners and shape of data.

The microarray data is about the expression of genes under consideration, abundance of each gene is quantified by the intensity of measurement. Figure 2.1 shows the simplest of graphs the intensities of two channels (in case of cDNA array), green vs. red

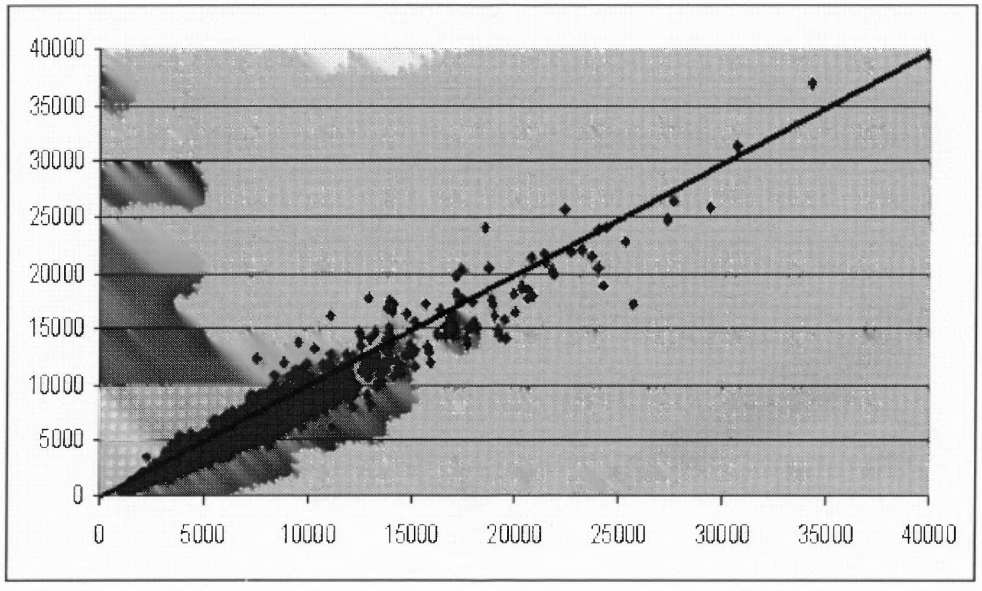

Figure 2.1 Raw intensity plot of two channel microarray data.

Two things stand out from above graph, first thing, the range of intensity values generated in microarray experiments is large and second thing to note is that the graph has funnel shape, genes with lower expression having bigger variation compared to higher intensity genes. Another observation that can be made from the graph is that, since it is the graph of intensities from control vs. the sample, under the condition of no gene being expressed, ideally the intensities of both green and red channel for each gene would be same and hence, the graph would look like a straight line passing through center at 45. Which means that, farther the gene from this 45 line, more likely that it is differentially expressed. Also, the genes away from the 45 lines in upper half are likely to be over expressed and on the other hand those away in the lower half are likely to be under expressed as seen in Figure 2.2

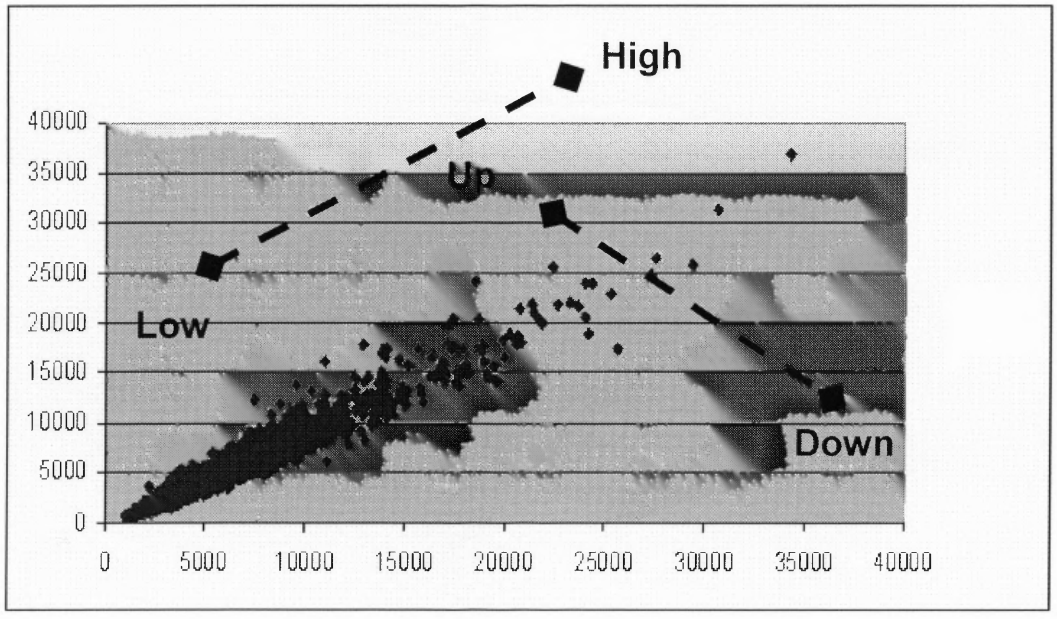

Figure 2.2 Up and down regulated genes.

Data transformation techniques are frequently used in data analysis. Transformations are a remedy for outliers, failures of normality, linearity, and homoscedasticity[3]. Data transformation usually involves applying some mathematical function to the given data. In case of microarray data log transformation is frequently used, it serves two purposes.

As seen in Figure 2.1, the range of intensity values in the microarray experiments is indeed large, the log function compresses the larger values to reduce the range with out affecting the underlying relationship. But more importantly, the log treats the number and the it's reciprocal symmetrically, log(2) = 1 and log(1/2) = -1, this tends to make the expression data normal. Since many of the tests in the statistics assume normality of the data this log transformation has important benefits.

During the early days of microarray, the frequently used method for expression change was fold change [2]. In simplest form a change of 2 fold in intensity would qualify the gene to be declared differentially expressed. In the graph 1 this would involve drawing a line at two fold, and the genes above the line would be over regulated and genes below would be under regulated, as shown in Figure 2.3,

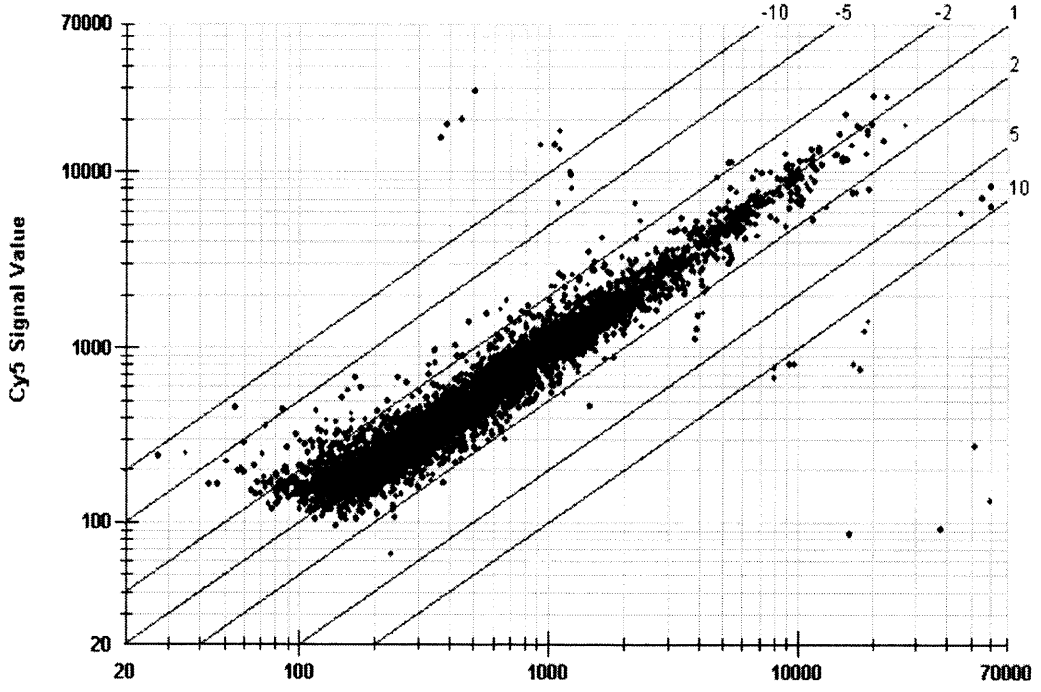

Figure 2.3 Fold change.

One of the first questions that arise from this kind of analysis is, how dose one know that the observed difference is not due to chance? Also bigger the fold change more likely that the change not due to chance, but how much of fold change is required to be differentially expressed? Clearly this form of analysis lacks statistical rigor But it may still serve the purpose if the objective of investigator is to obtain the list of probable differentially expressed genes for further investigation.

CHAPTER 3

STATISTICAL TESTS FOR DIFFERENTIAL EXPRESSION

3.1 Tests for Differential Expression

A parametric test makes assumptions about the population distribution. The most common assumption is that the data is normally distributed. Though this may sound as a restriction, but many of the datasets in biological experiments amazingly enough do follow the normal distribution and when they do fit the distribution the parametric methods are indeed the methods of choice.

Non parametric tests are class of statistical test that do not make any assumptions about the data, they allow for hypothesis testing that do not involve statements about population parameters. Also non parametric test are much easier to implement and less computationally intensive compared to parametric test. Another advantage of non parametric data is that they can be applied to all kind of data including the ranked data, analysis of which cannot be done using parametric methods. Another advantage of non parametric data is that they are more immune to outliners, the reason being non parametric test tend to take median as measure of central tendency in contrast to mean as done in parametric test. What would happen if one were to apply the parametric test to the non normal data? It would result in higher type I errors meaning there would be more false positives. Non parametric test would reduce these false positives.

But such simplicity comes at a price, non parametric test are less powerful that the parametric tests and for such reason in case of data that can be handled by parametric test it would be prudent to use parametric tests. If the sample size is big(>30) and the data not normal both the parametric and non parametric test would give similar kind of result, if

if the data were normal then also result would be almost identical except non parametric test would be slightly less powerful.

For most hypothesis tests, one starts with the assumptions and work forward to derive the sampling distribution of the test statistic under the null hypothesis. For permutation tests the procedure is reversed, since the sampling distribution involves the permutations the method is called permutation test.

In order for the permutation test to work a large number of permutations are required, which are limited by the number of samples available. Also because of large number of permutations to be performed these test tend to be computationally expensive.

The permutation test normally involves selecting the test statistics, resembling and recomputing test statistics, rejecting or accepting null hypothesis.

3.2 t TEST

The differential expression is instance of class comparison problem that has been studied extensively in statistics. Under the normal distribution assumption how dose one determine that there is difference in expression between two samples? One choice could be to compare the means of two samples, as seen in fold change. But it would not give enough information, for example if the means of two samples differ by X, what dose it mean? How dose one know if X is significant enough to say that there is no differential expression or there is differential expression? Clearly what is needed is some frame of reference against which difference in mean can be compared. That's were t distribution comes into picture. If one were to draw repeated pair of samples from the normally distributed population and for each pair calculate the mean difference between pair of

samples, then that mean difference follows a distribution known as t distribution. Logically this distribution will have the mean of zero, because the average difference of means of two repeated samples would be equal and hence the difference would be zero, the samples are drawn form same population. It also follows that this t distribution, of difference in mean has standard deviation of,

$$\sigma^2_{M-M} = \sigma^2_{population} / N_1 + \sigma^2_{population} / N_2 \quad (3.1)$$

Were N_1 and N_2 are the sample sizes of two samples.

The interesting thing to notice in above equation is that, though the t distribution for the different sample sizes will have same mean of 0, the variance will depend on the sample size and will be more for smaller sample sizes. This is expected because larger samples will have mean difference closer to the true mean of 0 and hence the distribution would be more tightly clustered around mean.

Above explanation has important assumption that there is prior knowledge about the population parameters, which is hardly ever the case in real scenario. Fortunately there is a way around it, use the information in the given samples to estimate the population parameters and use these parameters along with degree of freedom to draw the sampling distribution of mean difference.

Having established the theory the following paragraphs would explain the way t test is actually used to test differential expression. Analysis starts with the available data that is expression level from the replicates for two samples. And ultimate aim is to test how significant the difference in the mean of two samples.

The parameters (mean and variance) of sampling distribution of mean differences need to be estimated from available data. The mean of this distribution is zero; the

standard deviation of distribution is given in Equation 3.1. From the Equation 3.1 it is clear that, first estimate the variances of population from the sample data is required, which is estimated as,

The variance of population from the sample can be estimated as,

$$S^2 = SS / N - 1 \tag{3.2}$$

But there are two samples in the present case and that would result in two estimates of the population variances. So question is which one to use? Better answer would be to use the information from both the samples, and a composite variance is estimated as, which is also know as pooled variance.

$$\overset{2}{S}_P = SS_1 + SS_2 / (N_1 - 1) + (N_2 - 1) \tag{3.3}$$

Were SS_1 and SS_2 are sum of the squares for two samples and s^2_p is called pooled sampled variance.

The value obtained from above equation can be plugged in to equation 1 to get the standard deviation of sampling distribution of sample mean.

Once all the parameters of given distribution are available, next step would be to calculate the t test statistics given as,

$$t = M_1 - M_2 / est \underset{M-M}{\sigma} \tag{3.4}$$

Were M_1 and M_2 are means of two samples.

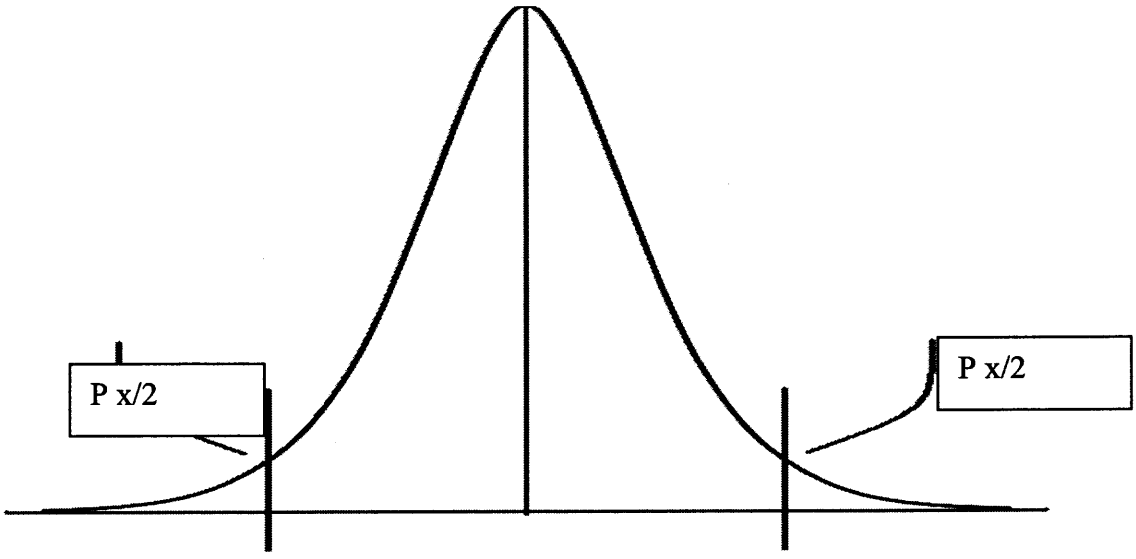

Figure 3.1 t –test.

What this equation simply does is to transform the raw mean difference between two samples as difference per standard deviation. The t statistic on it's own is a mere number and needs to be converted to probability to make meaningful inference. This probability is known as p value, in simple terms p values is the probability of getting t value as large or larger than the one obtained from the observations under the null hypothesis. So the p value of .05 would be observed less than 5% of times.

Next is to compare the value obtained above to the table of critical values of t for the given significance level, for degree of freedom $(N_1—1)+(N_2—1)$. The df has one value less per sample because while calculating the mean one df per sample is already used. The significance of df lies in the fact that the t distribution curve would have the same shape and mean but different spread depending on the df as shown in Figure 3.2

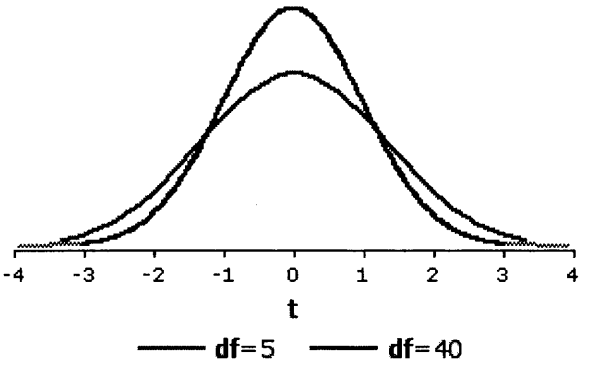

Figure 3.2 Degrees of freedom and spread of data.

There are different ways to interpret the t statistic; one is to say that it is ratio of between class to within class variability. Other interpretation of it would be to say that the denominator, in the t statistic is the estimator of the variability of M1-M2 one would observe if one repeatedly draws two samples form large population.

$$t = \frac{\text{(Random Variation + Differential expression)}}{\text{Difference expected by chance (Random Variation)}} \qquad (3.5)$$

In a away numerator (difference), is the outcome of microarray experiment and denominator is the estimator. Greater is the difference between two, more likely the gene being differentially expressed, meaning higher the absolute t value more likely that gene is differentially expressed.

3.2.1 Paired t test

The t test for dependent samples helps us to take advantage of one specific type of design in which an important source of within-group variation can be easily identified and excluded from the analysis. Specifically, if two groups of observations (that are to be compared) are based on the same sample of subjects who were tested twice (e.g., say a drug is tested on a group on individuals), then a considerable part of the within-group variation in both groups of scores can be attributed to the initial individual differences between subjects. In case of independent samples, one cannot do any thing, about this with in group variation because we cannot identify the variation due to individual differences in subjects. However, if the same sample was tested twice, then we can easily identify this variation and subtract it. Instead of treating each group separately, and analyzing raw scores, we can look only at the differences between the two measures in each subject. By subtracting the first score from the second (e.g., pre trial and post trial), for each subject and then analyzing only those differences, the entire part of the variation attributed to intrinsic variation in data set can be eliminated. The remaining variation can be explained by factor under consideration. Paired test always more sensitive than unpaired t test, due to this reason only.

The paired t test takes advantage of specific experiment design, question is how dose this fit into microarray experiments?. Especially because in microarray the same sample(mRNA) is not treated twice as would be in case of say drug resting in patients. The answer is to treat both control and sample in such way that both can be subjected to similar noise.

Consider a microarray experiment design in which the both the control and the treatment are put on the same spot on the microarray chip. Under these conditions it is reasonable to assume that what ever differ observed can be contributed solely to the experimental factor under study since the experimental conditions for both the samples are same. Under such circumstances it would be prudent to take in to account the measured difference in the expression than the raw expression values.

The t statistics for unpaired t test is given as

$$t = M_1 - M_2 / est \underset{M-M}{\sigma} \tag{3.6}$$

Paired t test tries to modifies above statistic by taking into account only the individual differential expression in the observations of two samples. The statistics for the pair t test is given as ,

$$t = \frac{\bar{d}}{SE\,(\bar{d})} \tag{3.7}$$

The numerator id the average difference in expression and the denominator is the standard deviation estimated for the mean difference distribution. Also the degree of freedom in this case would change to (N-1), since there are only N observations.

It is important to note that the paired t test is result of particular experimental design in this case the test and control being printed on the same spot. The advantage of such design is that both the test and control are subjected to same kind of environment and hence similar random noises, which cancels out the net effect when the difference is taken.

3.2.2 Unequal Variances

Procedure outlined above assumes that the two samples are presumed to come from distributions with equal variances. Due to small number of replicates in microarray analysis this assumption may not be met. If this cannot be assumed (as a result of say F test) then the procedure would still be same except that the degrees of freedom would change to

$$df = (s_1^2 / N_1 + s_1^2 / N_2)^2 / (s_1^2 / N_1)^2 / (N_1 - 1) + (s_2^2 / N_2)^2 / (N_2 - 1) \qquad (3.8)$$

3.3 ANOVA

The t test can be used to test the differential expression between two conditions. But in case purpose is to study the gene under more than two conditions simultaneously, how would one know, whether there is differential expression and if there is, under which conditions? One simple way could be to perform the pair wise t test for all the conditions. But this has interesting fallout. Say, one were to perform ten pair wise t test, at 5% significance, which means that for each test there is probability of 5% for Type I error, but for the group of ten test the Type I error add up and is much more than 5%

Each observation in given sample has variation around its sample mean. But observation also varies around the global mean. The former is known as within group variation and the later as between group variations. ANOVA studies the relationship between these two variations (variances) hence the name Analysis of Variance.

The variations can be measured by sum of squared deviates, thus the total Sums of Square (SS) would be sum of components,

$$SS_{Total} = SS_{within\ group} + SS_{between\ Groups} \qquad (3.9)$$

As in the case of t test, within group and inter group variability can be used to estimate the population variances for appropriate degree of freedom. It follows that under the condition of no Differential expression these estimates would be more or less equal, which can be tested by F test.

As in case of t test the F test can be viewed as

$$F = \frac{\text{Observed variance between sample means (Between Group Variance)}}{\text{Variance expected by chance (Within Group Variance)}} \quad (3.10)$$

As shown in the following figure, the samples were the with in group variation is much smaller than the between group variation we can reject the hypothesis that the two samples are from same distribution.

Larger between group variation

Between group variation comparable to within group variation

Figure 3.3 Relationship between within group and between group variation.

But the real utility of ANOVA is not really in the above F test like application. ANOVA can be used to build an explicit model about the sources of variance that affect

the measurements, and then use the data to estimate the variance of each individual variable in the model. One of the papers [3] proposes the following model to account for the multiple sources of variation in a microarray experiment:

$$\log yijkg = \mu + Ai + Dj + Vk + Gg + (AG)ig + (VG)kg + \varepsilon ijkg \qquad (3.11)$$

In this model, μ is the overall mean signal of the array, *Ai* is the effect of the *ith* array, *Dj* is the effect of the *jth* dye, *Vk* is the effect of the *kth* condition, *Gg* is the variation of the *gth* gene, *(AG)ig* is the effect of a particular spot on a given array, *(VG)kg* represents the interaction between the *kth* variety and the *gth* gene, and *εijkg* represents the error term for array *i*, dye *j*, condition *k* and gene *g*. The error is assumed to be independent and have a mean of zero. Finally, log(*yijkg*) is the measured log-ratio for gene *g* of variety *k* measured on array *i* using dye *j*.

The advantage of ANOVA is that each source of variance is taken into consideration. Because of this, it is easy to distinguish between interesting variations, such as differential expression, and side effects, such as differences caused by different dyes or arrays. The problem is that ANOVA requires very careful experimental design it cannot be after thought.

3.4 Wilcoxon Rank Sum Test

This is a non parametric equivalent of paired t test. Instead of using the actual expression values directly to detect differential expression this test use the rank of expression values. The steps involved in the Wilcoxon test are as follows,

Calculate the difference between the two samples for each corresponding observation.

Get the absolute value for each observation.

Rank the absolute values.

For each rank add a sign it had before taking the absolute value.

Sum the ranks obtained ($\sum R$)

It can be seen that if there is no difference between two means then the rank sum should be zero and that is the Null hypothesis that will be tested. But if the rank sum turns out to be say X, how significant is this X? What is the probability that this X can be due to mere chance? Like the t test some frame of reference is required to test the hypothesis.

For a sample size of three, the possible number of rank combinations would be $2^N = 8$ as shown in Table 3.1

Table 3.1 Rank Combination

1	2	3	Rank
+	+	+	+6
—	+	+	+4
+	—	+	+2
+	+	—	0
—	—	+	0
—	+	—	-2
+	—	—	-4
—	—	—	-6

The range of the rank sum would be +- N(N+1)/2, which is +-6 in this case. All the other possible values would lie between these two extremes and plot of the frequency distribution would be as follows

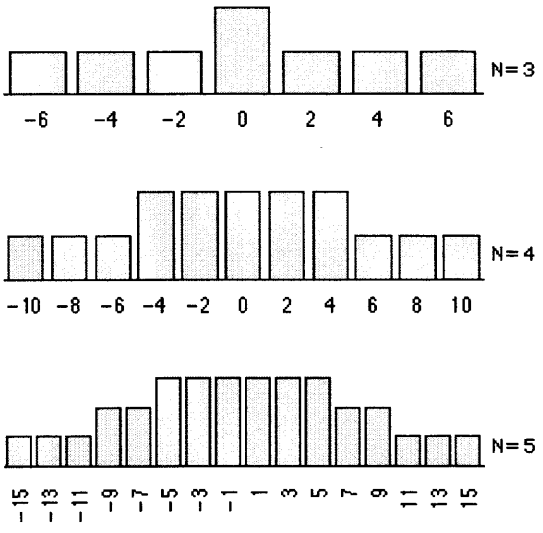

Figure 3.4 Rank distribution.

As the number of samples increase, the distribution will resemble to Normal distribution with standard deviation of

$$\sigma = (N(N+1)(2N+1)/6)^{1/2} \tag{3.12}$$

This is the reference, with which one can test the hypothesis of zero mean for given samples, with Z ration as follows,

$$Z = \sum R + (-).5/\sigma_R \tag{3.13}$$

.5 in above equation is a correction factor and is - .5 when $\sum R$ is positive and +.5 when $\sum R$ is negative. Once Z ratio is obtained, the hypothesis testing is simply the matter of comparing against the critical value.

Consider a simple case of rank a theoretical rank S for four samples. An *S* of 6 might be 1+2+3, or might be 2+4, or might be 1+5, or might represent a single rank of 6. p is the probability of finding a value of *S* smaller than or equal *to* the one observed. Thus the fact that four different patterns in this example all yield *S*-values of 6, means that all these patterns are counted in the p computed when any one of the patterns is observed. That raises the values of p and thus lowers the power of the test.

3.5 Mann Whitney Test

This is a non parametric equivalent of un paired t test. Like in case of Wilcoxon test the Mann Whitney takes median as a measure of central tendency rather than the mean, which makes these tests more robust compared to parametric tests. There are two ways of calculating the statistics for Mann Whitney test. If the sample size is more than 20, following procedure is adopted.

The steps involved in the Mann Whitney test are,

Let X1, X2, X3,……..Xn and Y1,Y2,…Ym, be the measure of samples from control and test . Note that these are observations form one gene and the test would be repeated for each gene.

The next step would be to rank these observations according to the values,

X1 < Y5 < Y2 < X5……..<Y8

1 2 3 4 (n+m)

Combine observations from two samples and rank them.

In case of tied ranks each such observation gets the average rank, for e.g., if Y2, X5,Y5 in above case then each would receive the ranking 3, and the next ranking will

start at 5 and not 4. Which makes sense because the rank is representation of relative distance in the observations.

Separate the ranks in two groups according to samples they came from and calculate the rank sum for each group.

If there were no differential expression the average rank sum of both the groups will be expected to be equal. The reason for using the average rank sum instead of raw rank sum is that number of observations in both the samples might not be same.

For combined observation N the average rank sum would be N(N+1)/2. This gives us the expected rank sum for two samples as,

$$N_1*N(N+1)/2 \text{and} N_2*N(N+1)/2 \tag{3.14}$$

For the samples size greater than 20 the sampling distribution for the samples tend to be normal distribution with mean as above and standard deviation of ,

$$\sigma_s = Sqrt(N_1N_2(N+1)/12) \tag{3.15}$$

Once mean standard deviation and distribution are known, which in this case is normal, the null hypothesis can easily be evaluated using z test as,

$$U = (\sum R - U) \pm .5) / \sigma_s \tag{3.16}$$

Interestingly the Z value obtained above would be same for both samples except for the opposite sign (mirror images).

The above discussion is only valid if the sample sizes are at least 5 for both samples, which might not be true in case microarray. Under such circumstances a variation of above test is used. The basis of this variation is to calculate the statistic U, which is defined as the Difference between maximum possible rank sum for given sample minus the observed rank sum.

$$U = \sum_{Max} R - \sum_{Obs} R \tag{3.17}$$

This is the observed value of U, but under the Null hypothesis of no differential expression expected value of U is,

$$U = \sum_{Max} R - \sum_{Exp} R \tag{3.18}$$

This null hypothesis value of U can be calculated as n*m/2

The term $\sum_{Exp} R$ can be calculated as explained in the procedure 1.

Suppose there are two samples of

Let X=X1 , X2, X3,……..Xn and Y=Y1,Y2,…Ym,

Then the rank sum of X is maximum when all the samples of X are greater then the sample Y,

Y3 < Y2 < Y5 <…Ym < X2 < X1 <….Xn

1 2 3 m m+1 m+n

And this maximum rank sum of X, can be calculated as

Rx[max] = (m+1) + (m+2) + (m+3) …(m+n)

Rx[max]= m*n + (1+2+3…n)

Rx[max] = m*n+(n *(n+1))/2,

The thing to note is that when rank sum of X is max, the rank sum of Y is minimum, since all its observations are less than sample X and this minimum rank sum would be

Ry[min] = 1+2+3…..m

Ry[min]= m*(m+1)/2

Similarly the maximum rank sum of Y would be

Ry[max] = m*n+(m *(m+1))/2

Once the expected value of U_{null} under null hypothesis and the observed value of U_{obs}, are known, how likely is it to obtain U_{obs} by mere chance. The total number of possible combination for given ranks are

$$(m+n)!/n! *m! \tag{3.19}$$

So for the sample size of m=5 n=5 there are , the possible combinations are 252. Then the task would be to find out the possible number of combinations that would produce value as large as $U_{obs.}$ This number divided by the total number of combination would give probability of obtaining value as large as $U_{obs.}$ Important thing to note is that one could have very well define the statistics U as

$$U = \sum_{\min} R - \sum_{Obs} R \tag{3.20}$$

Instead of

$$U = \sum_{Max} R - \sum_{Obs} R \tag{3.21}$$

And still the result would be same.

Generally speaking, the t-test varies with the ratio of the average difference between the two groups to the standard error of the average difference. A large numerator or small denominator will increase the value of the t-statistic. On the other hand, for the Mann-Whitney U test, if for example expression of a gene on four arrays in class A is always greater the expression for the same genes in four arrays of class B, then the ranks will be 5,6,7,8 in class A and 1,2,3,4 in class B, resulting in some standard normal *Z* statistic. For other genes having the same relationship, the ranks will be the same, i.e., ranks 5,6,7,8 for group A vs. ranks 1,2,3,4 for group B and the *Z* statistic will

be identical. While the t statistic is usually never the same for different genes, the Mann-Whitney U value can be identical for many genes.

3.6 Significance of Analysis (SAM)

Consider a simple experiment involving 10000 genes; let's say t-test is applied to detect the differential expression. Even if not a single of 10000 genes is differentially expressed, still at .5 significance (which is normal in biological experiments), the t-test would result in 500 significant genes! This is known as problem of multiple hypothesis testing. Another issue with t test like test is that they treat all the genes in same way applying the same statistic, but as shown in Figure 1.1 the distribution of microarray data intensities is such that non linear models are required to better explain the variations as shown below [4]. The permutation models discussed in this and next section are such non linear models.

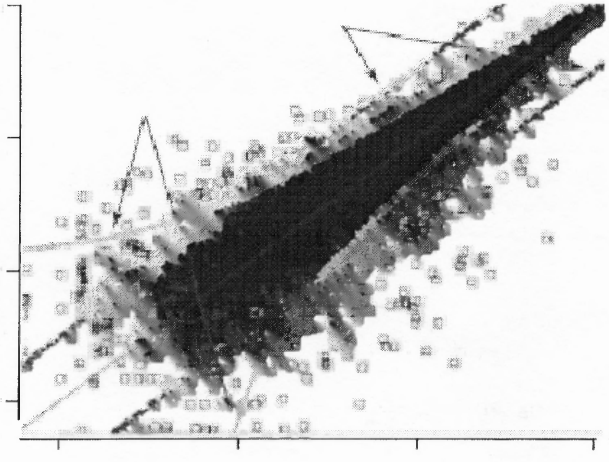

Figure 3.5 Nonlinear and linear cutoff for differential expression.

For small number of replicates, which is indeed the case in microarray experiments because of cost factor, t statistics tend to be co related to the error term in

denominator. The consequence is that t test would tend to find more significant genes form the genes with low sample variance than genes with higher sample variance. The reason this is more of issue in case of microarray experiments is because there are thousands of genes tested simultaneously, and there would always be some genes with low variance by chance. The net effect of above problem in case of t test is increase of false positives in case of genes with low variance and increase of false negative in case of genes with high variance.

SAM [5] tries to address the issues raised above by adding a constant in the denominator of the t statistic known as 'fudge factor', so the statistic for the SAM is,

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0} \tag{3.22}$$

The challenge is to find the value of *So* such that dependent of *d(i)* on *s(i)* is as small as possible. Later in the section, the procedure for determining this is outlined. In case of normal t statistic the after calculating *d(i),* it is tested against the critical value that separates significant from non significant using the t table. But in case of SAM *d(i)* is not a t distribution, because of addition of fudge factor hence the critical value table for t test no longer valid. The significant of observed *d(i)* is accessed through permutation.

Next SAM generates the random datasets from the given data through permutation. For each such permutation, d values are calculated. Interestingly, these permutated data are observations when the null hypothesis is true, because the two samples are exchanged and there is no differential expression indeed. Thus the values of d for given genes obtained from permutation are the values expected under Null hypothesis.

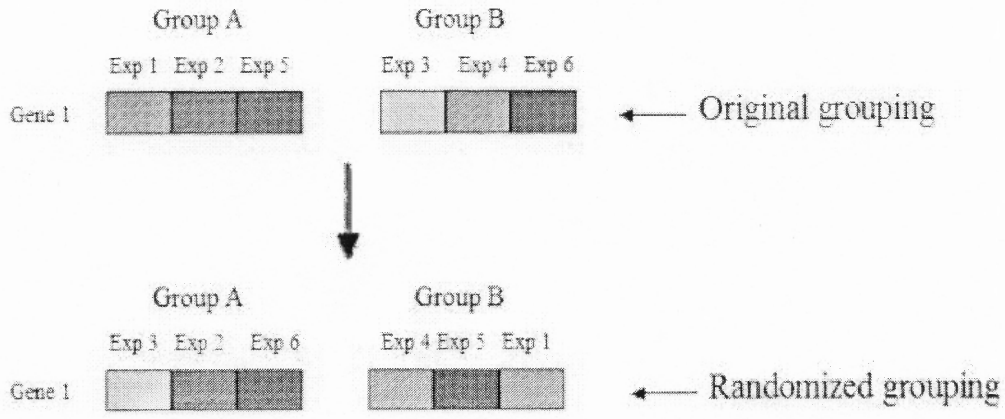

Figure 3.6 Permutation .

For each permutation, the false positives are calculated as the number of genes d values is more than the value considered as significant and the False discovery rate is calculated as number of false positives divided by the number of genes in original dataset. The whole SAM procedure can be depicted as shown in Figure 3.7

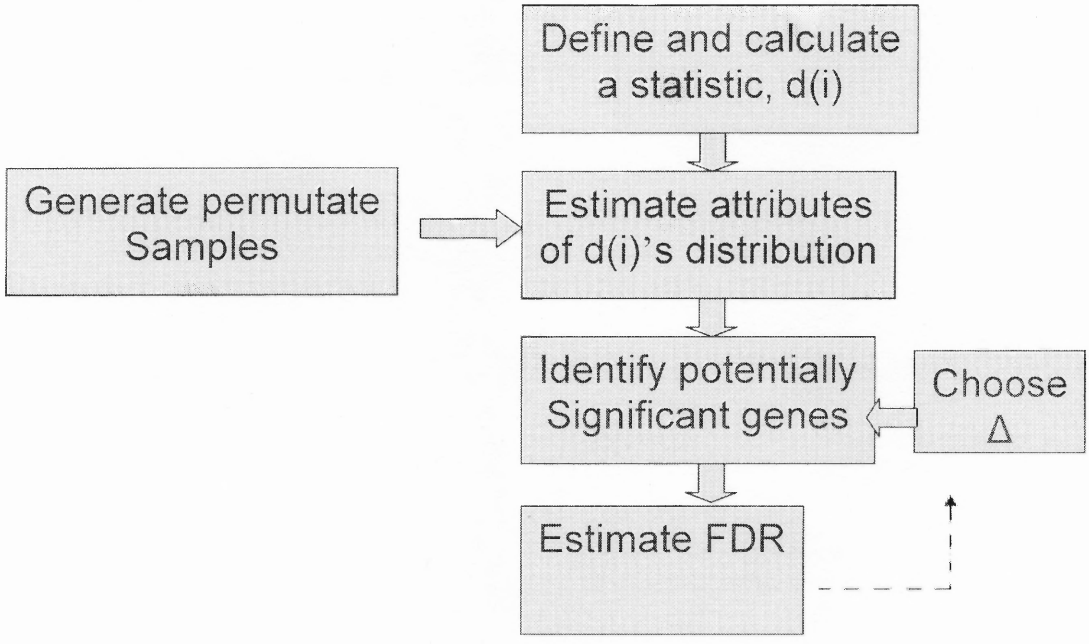

Figure 3.7 SAM overview.

3.6.1 SAM Details

The details of SAM algorithm are as follows [6]

SAM summarizes the observations using the statistic

$$t_i = \frac{\overline{x_{i1}} - \overline{x_{i2}}}{s_i} \tag{3.23}$$

Where i is the gene number, which ranges from 1 to n (total number of genes),j is the sample number, which ranges from 1 to m (total number of samples) t_i is the pooled variance test statistic. s_i is the denominator known as the error term,

$$s_i = \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sqrt{\frac{\sum_{j=1}^{k}(x_{ij} - \overline{x_{i1}})^2 + \sum_{j=1}^{k}(x_{ij} - \overline{x_{i2}})^2}{n_1 + n_2 - 2}} \tag{3.24}$$

Given two classes of sample size n_1 and n_2 with group means of $\overline{x_1}$ and $\overline{x_2}$, The pooled variance t-test for each gene is calculated as:

So, the overall equation for pooled variance t-test appears as

$$t_i = \frac{\overline{x_{i2}} - \overline{x_{i1}}}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\sqrt{\frac{\sum_{j=1}^{k}(x_{ij} - \overline{x_{i1}})^2 + \sum_{j=1}^{k}(x_{ij} - \overline{x_{i2}})^2}{n_1 + n_2 - 2}}} \tag{3.25}$$

with $n_1 + n_2 - 2$ degrees of freedom.

3.6.2 Calculation of Fudge Term

1. Calculate the pooled variance test statistic for each gene in the normal way. Store the numerator r_i and the error term s_i (the denominator)

2. Permute the data set arbitrary number of times to create P number of permuted datasets, $D_1^*, D_2^*, D_3^*, \ldots\ldots\ldots\ldots D_P^*$.

3. For each gene, calculate the pooled variance test statistic for each permuted data set (D*). Store the numerator and the denominator for each permuted data set. These are used to find differentially expressed genes.

4. For each gene individually, rank the error terms (denominators) over all permutations (D*). Divide each of these ranked lists into 100 quantiles.

5. For every fifth quantile(α), re-calculate the test statistic for each gene in the unpermuted data set using a modified error term which is the original error term s_i (without permutation) plus the error term at that quantile (s^{α}).

$$t_i^{\alpha} = \frac{\overline{x_{i1}} - \overline{x_{i2}}}{s_i + s^{\alpha}} \tag{3.26}$$

6. Construct a vector of these test statistics comprised of that or those for that quantile (α) and that or those of its next highest neighbor (α+1). The number of test statistics in this list will be twice the number of permutations/number of quantiles *2P/100*

7. For each gene, calculate the median absolute deviation of this collection of test statistics.

8. Find the quantile (α) where the coefficient of variation of the median absolute deviations is minimized. The error term at that quantile becomes the gene-specific 'fudge factor's_0.
9. For each gene, in the original data set, calculate a revised test statistic (the SAM test statistic) where the error term for each is the original error term plus the gene-specific fudge factor s_0.

3.6.3 How Genes are Deemed Differentially Expressed

10 Rank the genes according to the SAM test statistic in descending order.

11. Re-calculate the test statistics for all of the permuted data sets using the values stored in step 3 and the gene-specific s_0 (t_i^*).

12. Rank these revised test statistics (for the permuted data sets) and calculate the average test statistic value for each rank position ($\overline{t_i^*}$). This is the 'expected' value of the SAM test statistic at that rank position under the null hypothesis.

13. The difference between the SAM test statistic for a gene in the real data set and the expected SAM test statistic for that genes' rank (Δ) is calculated

$$\Delta = t_i - \overline{t_i^*} \tag{3.27}$$

and compared to a user-defined threshold value Δ_T, to determine whether the observed value of the test statistic for the gene in the real data set is higher or lower than the expected value of the SAM test statistic associated with Δ_T .

The SAM test is asymmetrical, i.e. the absolute value of the upper critical value of t_i associated with Δ_T is not necessarily the same as that for the lower critical value of t_i associated with Δ_T. Genes with test statistic values that are higher than the upper critical

value of t or less than the lower critical value of t associated with Δ_T are considered to be statistically significant.

3.6.4 Estimating FDR

Count the number of genes (J) determined to be significant (above the value t_i associated with Δ_T)

16 Determine the average number of genes found to be significant ($\overline{J^*}$) at values of t over all permuted data sets (D^*).

$$\hat{FDR}_{initial} = \frac{\overline{J}^*}{J} \tag{3.28}$$

Because there are differentially expressed genes in most experiments, this estimate is biased upwards. To correct for the upwards bias, the initial estimate is multiplied by a correction factor π_0.

17 To calculate π_0, calculate the ratio of the number of nonsignificant genes (R) in the real data set to the average number of nonsignificant genes ($\overline{R}^*$) over all permuted data sets (D^*). Because false positives are presumed to be rare at high values of t_i, this is performed at an arbitrarily small value of Δ_T (called Δ'). The value of π_0 is either itself or unity, whichever is lower

$$\pi_0 = \frac{n - J(\Delta')}{n - \overline{J}^*(\Delta')} \tag{3.29}$$

The revised estimate of FDR is obtained by multiplying the initial estimate of FDR and π_0

$$\hat{FDR} = \pi_0 \frac{\overline{J}^*}{J} \tag{3.30}$$

3.6.5 Significance of Delta

Clearly delta is the parameter that will determine how many genes are declared differentially expressed. So important question is how to choose this delta values? Smaller delta value would generate more differentially expressed genes, but as seen from graph below smaller delta would also result in higher FDR. The delta chosen would depend on the answers the experimenter is looking for from microarray study.

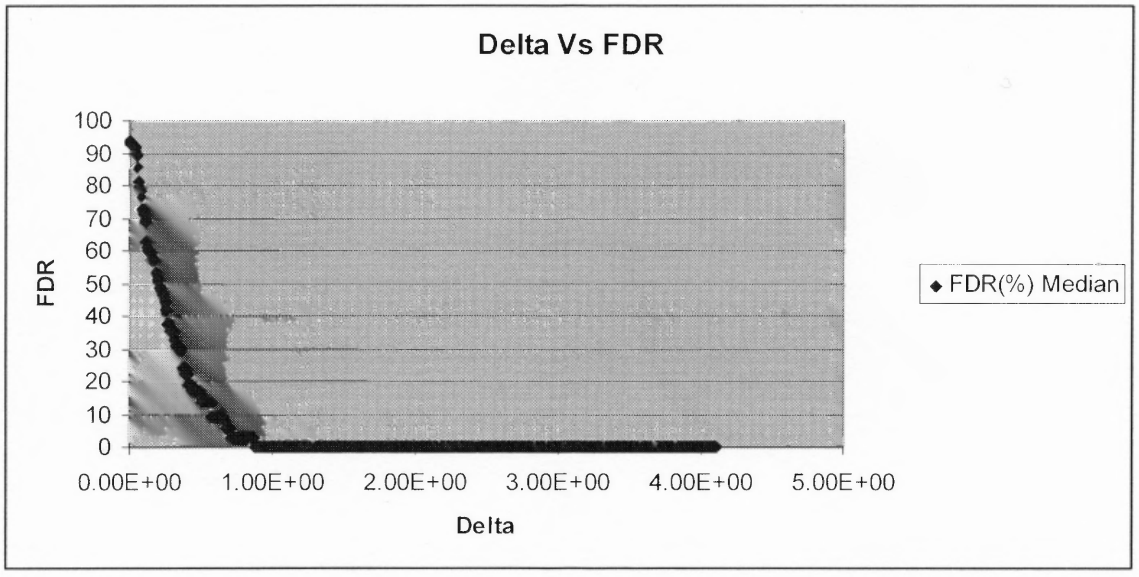

Figure 3.8 Plot of delta vs FDR .

3.7 Empirical Bayes

Let us say there are two samples X and Y obtained under two conditions. The null hypothesis is that there is no Differential expression hence the data in X and Y arises from joint probability density function $f_0(X, Y)$. On the other hand the data with differential expression come form the PDF $f_1(X, Y)$.

Apriori one do not know which distribution gene g comes from, so a parameter p is introduced. P is the probability that the gene is differentially expressed. (1-P) is the

probability that the gene is not differentially expressed. Thus the marginal distribution of data would be,

$pf_1(X,Y)+(1-p)f_0(X,Y)$ If p, f_1 and f_0 are known then the posterior probability of differential expression would be,

$$pf_1(X,Y)/\,pf_1(X,Y)+(1-p)f_0(X,Y) \tag{3.31}$$

There is couple of variations of Empirical bayes as to how these parameters are estimated. Some assume the distribution hence parametric while others are distribution free.

In one of the non parametric mixture models [7] the data is reduced to expression scores Z for each gene. If $f_1(z)$ is the density of differential genes and $f_0(z)$ the density of un differentiated genes then the mixture density can be written as,

$f(z)=p_0f_0(z)+p_1f_1(z)$ Were p_0 is the probability that gene is unaffected and $p_1(1\text{-}p_0)$ is probability that the gene is differentially expressed.

The posterior probability that the gene with score Z shows no differential expression is, $p_0(z)=p_0f_0(z)/f(z)$, $f(z)$ which can be directly obtained from the data. To estimate $f_0(z)$ data is permuted number of times and expression score z is obtained, from which $f_0(z)$ is estimated.

CHAPTER 4

RESULTS

4.1 Biological and Simulation Data

The biggest impediment in the comparison different tests for microarray differential expression is that there is no standard data to compare results against. In the absence of such "gold standard" how can one comment upon the efficacy of particular method in detecting differential expression. Increasingly researchers have turned to simulated data as a alternate. Though the simulated data dose not reflect the complexity inherent in the microarray data, nevertheless in dose give insight about the tests being compared.

4.1.1 Data Simulation

Two biological conditions (control and treatment) are simulated, m, n are numbers of samples under each of two conditions, the total number of genes simulated are 1000 and 5% of these are produced as differentially expressed. Simulation is done for two distributions normal and lognormal, so that the analysis is not biased towards particular test due to distribution of data, as t test do assume normal distribution and real microarray data may not be normally distributed. For both the distribution the parameters are, m1, m2 the means under two conditions and s1, s2 are standard deviation respectively. For the genes deemed as non differentiated, the means as well as standard deviation are same. For the differentially expressed genes mean of treatment is more than control (m2>m1) and standard deviation is same. For the purpose of simulation the parameters value chosen are m1 = 0 , m2=0 , s1 =.5, s2 =.5 under non differential

expression condition. For Differential expression the parameter values are m1 = 0 , m2=1 , s1 =.5, s2 =.5 .

Number of replicate in microarray is an issue, due to cost and other factors so to test the effectiveness of method under low and high replicates each distribution (normal and lognormal) were simulated with three and ten replicates. So in all four datasets of 1000 genes each were generated, two which are normal, two lognormal with three and ten replicates. The code for simulation was written in statistical language R.

4.1.2 Biological Data

cDNA data published as a with of publication "Liver gene expression in rats in response to the peroxisome proliferator-activated receptor-_ agonist ciprofibrate" by Yadetie etc [8] in journal physiolgenomics was applied to different tests under consideration. The theme of study was to examined the effects of ciprofibrate(a drug) on liver gene expression in rats using cDNA microarray. In all four hybridization were carried out involving more than 6000 genes between normal and treated samples. The study also carried out northern blot analysis of very few selected genes. The raw data was obtained from Gene expression omnibus (GEO) [9] repository at the NCBI website.

The raw data was preprocessed as follows. First spots with undetected fluorescence signals in either channel in at least one array were excluded. Further filtering to remove unreliable measurements based on fluorescence intensity values was performed by excluding spots with fluorescence signal intensity (mean of the four arrays) less than 300 in both channels, about 3000 genes were left after filtering. After back ground subtraction each array was globally normalized to balance intensity differences

between the two channels using the online normalization tool MIDAW [10]. The graph in 4.3 illustrates the effect of normalization for a single one sample.

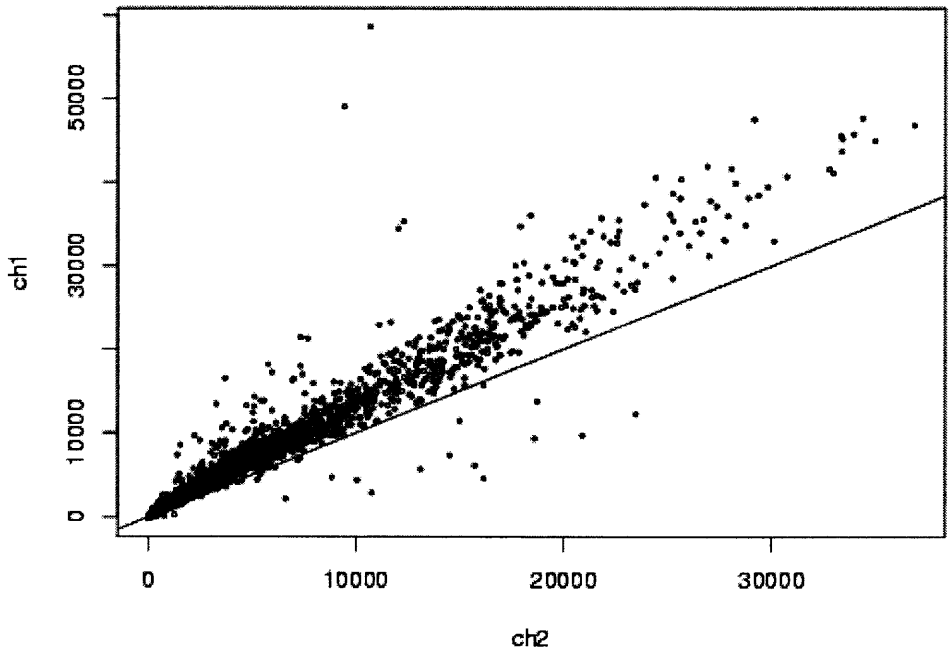

Figure 4.1 Raw Intensity plot.

Above graph is the result of raw intensity plot between two channels for a single sample, before any preprocessing, After background correction same graph would look like as follows

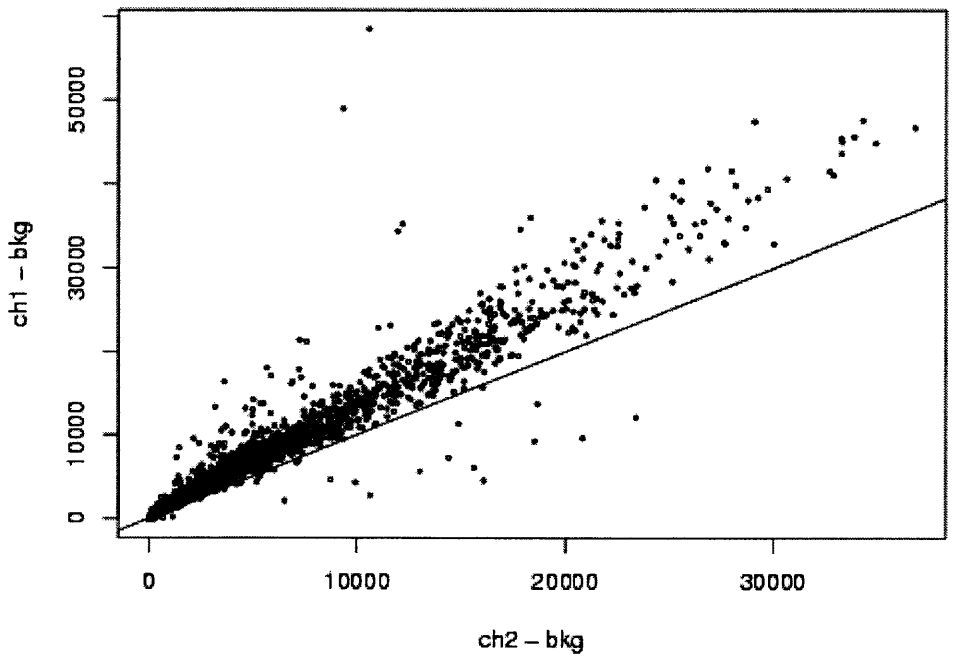

Figure 4.2 Intensity plot after background subtraction.

The global normalization was applied to each sample after background correction, and can been seen in Figure 4.3, the global normalization indeed balances the intensity difference between two channels.

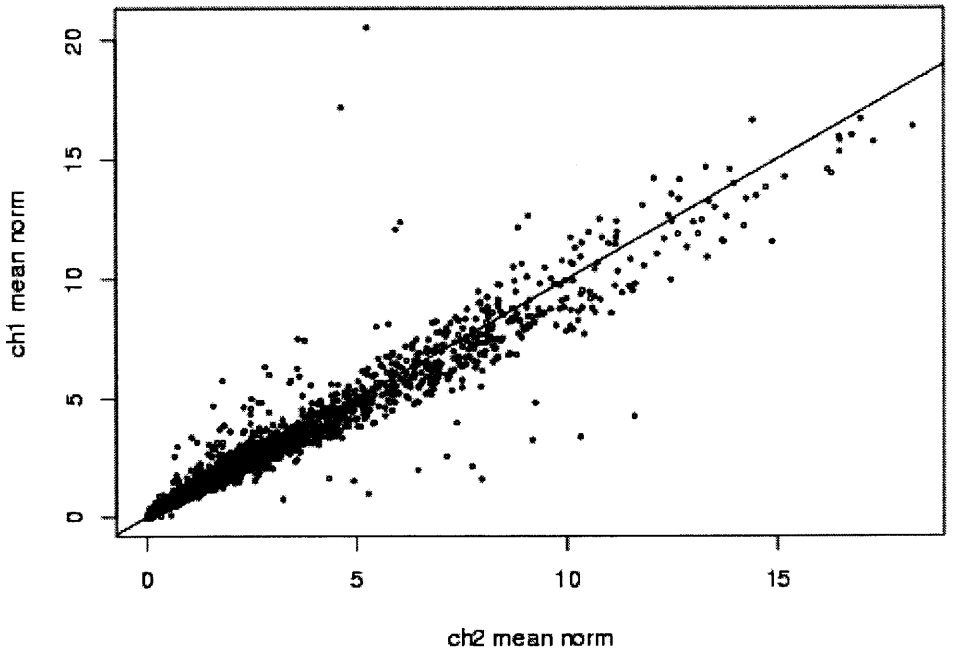

Figure 4.3 Intensity Plot after normalization.

4.2 Methodology of Comparison

Gene ranking is used as the metrics for comparison of the tests. Gene ranks are generated by descending order of the test statistic for each gene. For example if a gene x has t test score of 4 and gene y has a score of 5 then gene y would rank higher than gene x. Similarly for SAM genes are ranked according to d statistic and for Mann Whitney, U statistic is used. One can get fair bit idea about the efficacy of test by the way each test ranks the differentially expressed genes. Logically, the differentially expressed genes should receive higher test score compare to non differentially expressed genes and hence rank higher. Also ranking is important from practical point of view as in many of the

differential expression experiments, normally researcher would generate a list of important genes from expression data for further biological scrutiny.

For t test and Mann Whitney test, the functions available in software chipST2C were used while the SAM implementation software, available as excel plug-in released by the authors of SAM was used.

4.3 Application to Biological Data

Often with real microarray data the absolute value of the *t*-statistic is a function of the standard error SE, and there is an erratic behavior of the statistic for small values of SE with an increased risk of false positives. SAM is a nonlinear test and it tries to overcome this shortcoming of t test. Figure 4.8 is the proof of this. It's a plot of numerator, which is, SE in case of t test and SE+S0 in case of SAM vs. t statistic and d statistic respectively. Clearly because of fudge factor S0 d statistic dampens the d statistic value at lower SE, which why SAM produces less false positives, which is evident from the simulation data in above section.

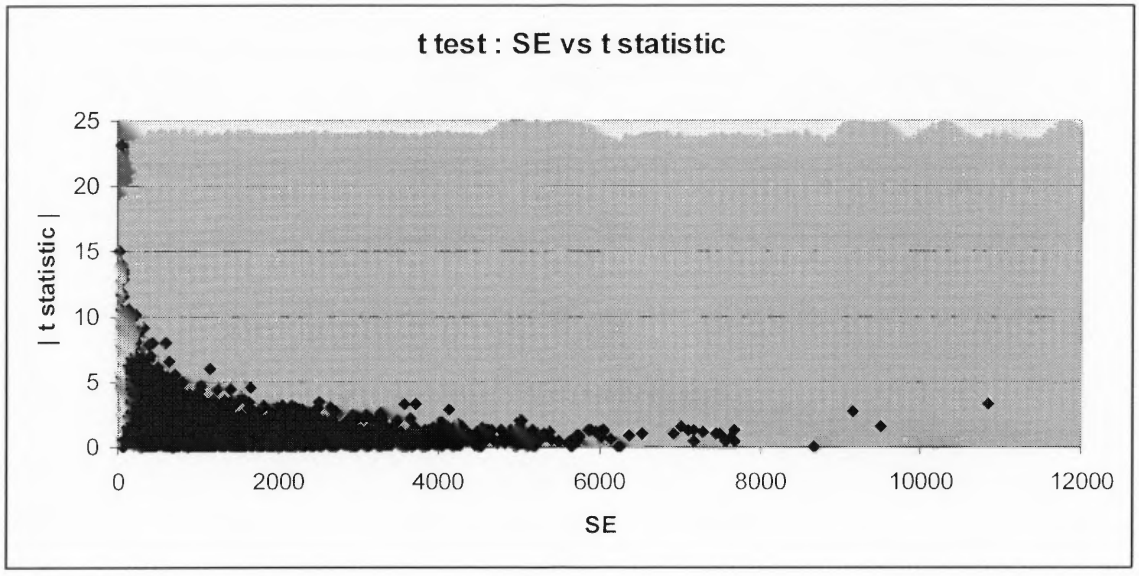

Figure 4.4 t test, variation of t statistic relative to SE.

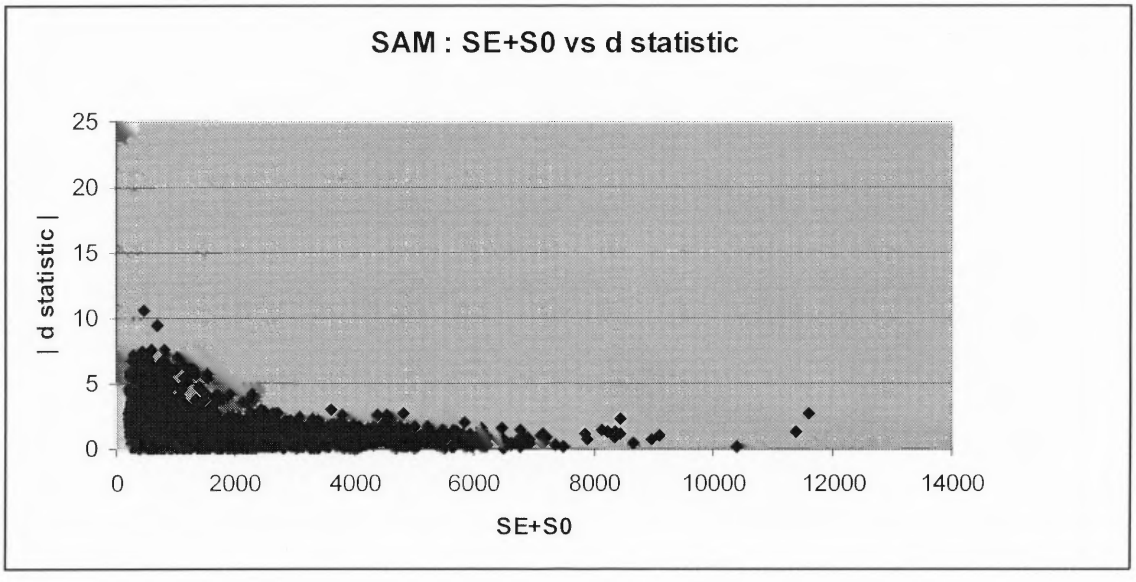

Figure 4.5 SAM, variation of d statistic relative to SE+ fudge factor.

Figure 4.9 shows the SAM graphical output generated during data analysis.

Following table summarizes as how these genes are ranked by each of the test under consideration,

Table 4.2 Genes Ranks for Different Tests

Ref ID	T test	Mann Whitney Test	SAM Test
3144	12	16	12
4662	406	637	153
2239	253	391	185
813	39	86	112
104	441	233	465
1185	44	106	21
1864	102	119	82
4872	115	567	439

The original study identifies 194 genes to be differentially expressed in the experiment, if one were to take that as standard, as shown above SAM not only identifies more genes but also ranks them higher compared to other two test

4.4 Application to Simulated Data

To understand how each test ranks the genes graphs are plotted as number of true genes detected by each test for the top 50 genes at the interval of 10 genes, for the simulated data under different distributions and replicates. The test with the uppermost graph on the plot, is the best one for given dataset, because that test would consistently pick up more true genes per interval compared to other tests.

As seen in the graph below at low number of replicates, SAM is by far the best, as it picks up a lot more number of true genes compared to t test and Mann Whitney test.

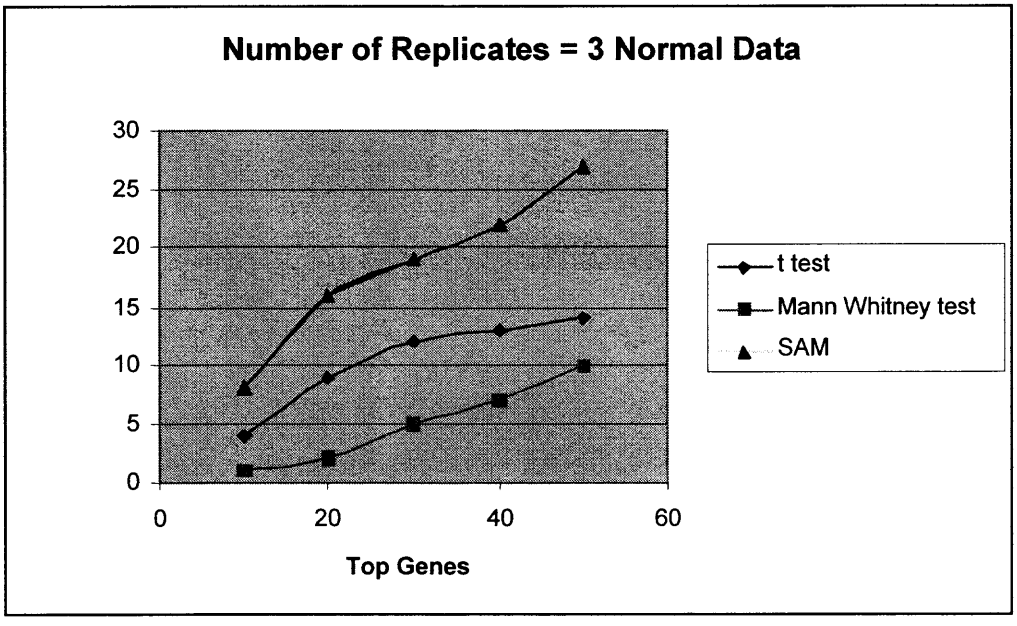

Figure 4.7 Gene ranking with normal data and few replicates.

Things look pretty much the same even with the lognormal data with three replicates as shown below, and t test and Mann Whitney test do roughly pick up same number of genes under normal and lognormal data. But the important thing to note is that even with lognormal data t test out performs Mann Whitney test. Another thing to note is the fact that SAM actually shows improvement in the total number of true genes detected in top 50 genes.

4.4 Application to Simulated Data

To understand how each test ranks the genes graphs are plotted as number of true genes detected by each test for the top 50 genes at the interval of 10 genes, for the simulated data under different distributions and replicates. The test with the uppermost graph on the plot, is the best one for given dataset, because that test would consistently pick up more true genes per interval compared to other tests.

As seen in the graph below at low number of replicates, SAM is by far the best, as it picks up a lot more number of true genes compared to t test and Mann Whitney test.

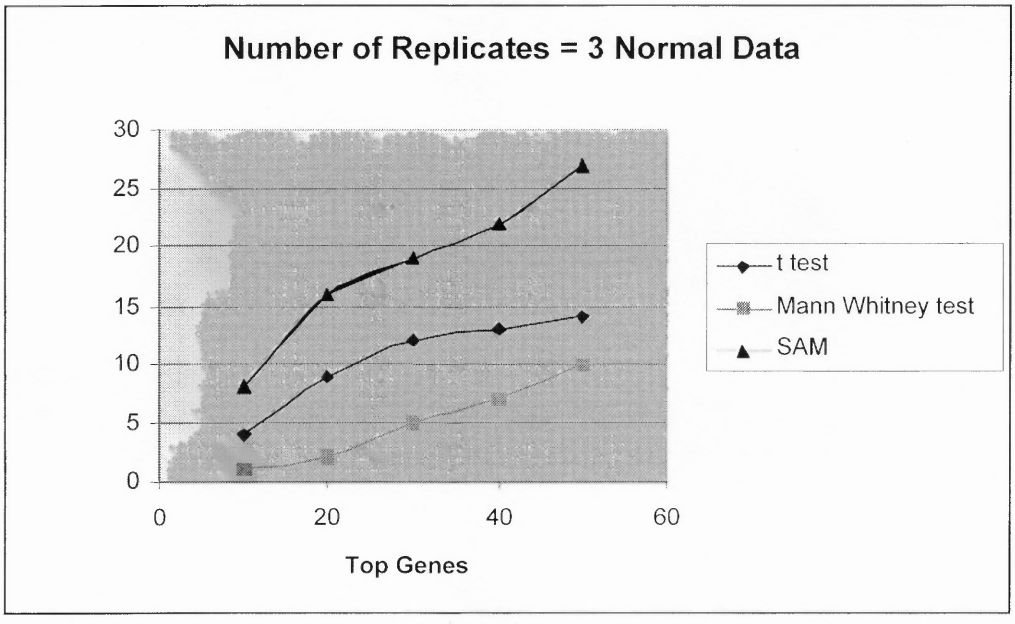

Figure 4.7 Gene ranking with normal data and few replicates.

Things look pretty much the same even with the lognormal data with three replicates as shown below, and t test and Mann Whitney test do roughly pick up same number of genes under normal and lognormal data. But the important thing to note is that even with lognormal data t test out performs Mann Whitney test. Another thing to note is the fact that SAM actually shows improvement in the total number of true genes detected in top 50 genes.

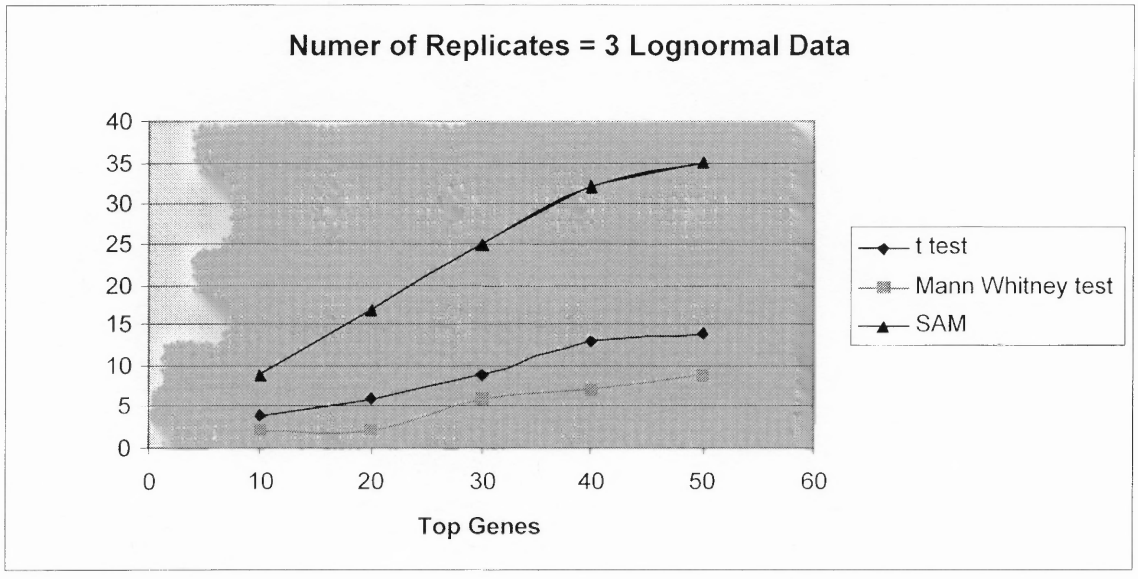

Figure 4.8 Gene ranking with lognormal data and few replicates.

More the number of replicates better the estimation of population parameters, hence with more replicates one would expect better performance from all the tests, which indeed is the case as shown below. Though SAM still ranks genes better that other two tests, the performance gap between SAM and other tests reduces dramatically with increase in number of replicates.

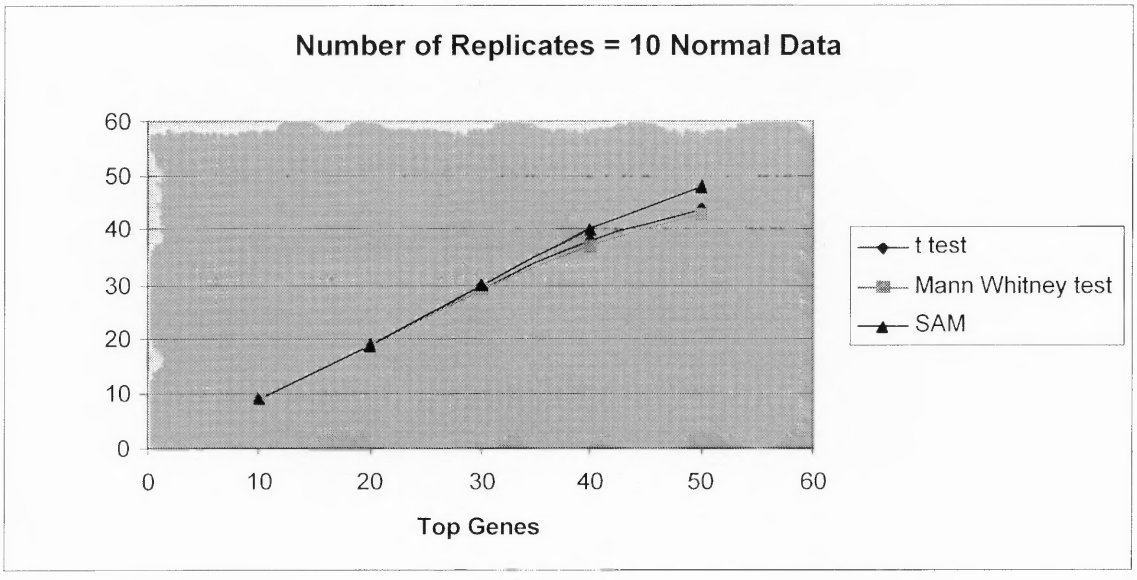

Figure 4.9 Gene ranking with normal data and large number of replicates.

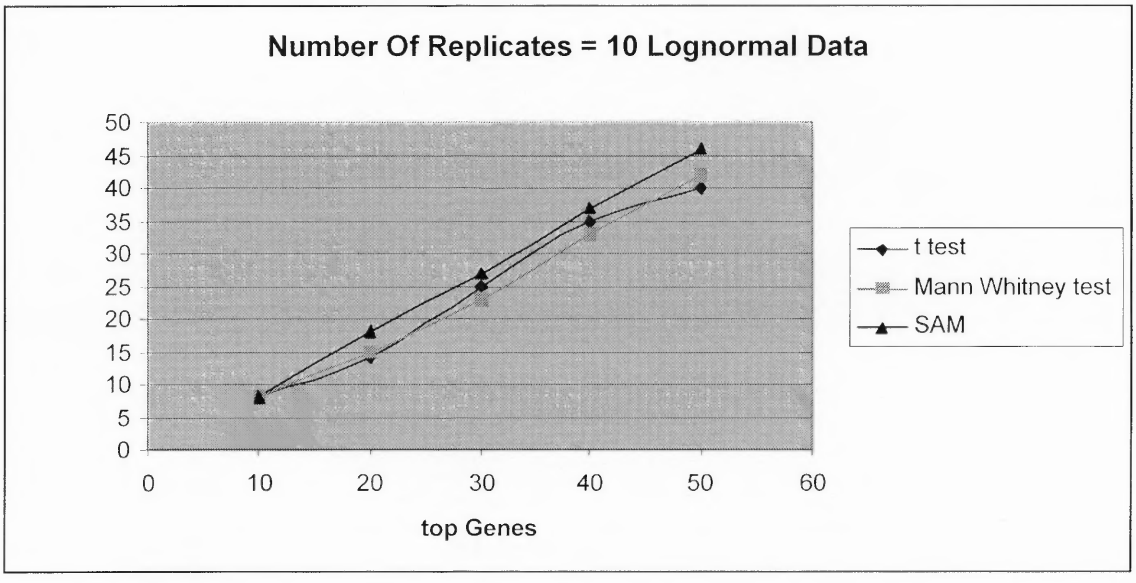

Figure 4.10 Gene ranking with lognormal data and large number of replicates.

Below is the table listing the total number of genes correctly identified by different test, in the top 50 genes for all the simulations, as seen from graphs above SAM detects more differentially expressed genes in top 50 compared to other two tests across

CHAPTER 5

CONCLUSION

The report discusses the popular parametric, on parametric and permutation tests to detect the differentially expressed genes in microarray experiment. Further comparison of the performance of t test, Mann Whitney test and SAM was carried out using ranking of genes by these tests as criteria. Since currently there is no gold standard to access the performance of the tests, data was simulated with Normal and Lognormal distributions with three and ten replicates. This enables us to access the gene ranking ability of different test under different distribution and with low and high number of replicates. The study also discusses the nature of microarray data, it's distribution and visual inspection techniques to get the feel of differentially expressed genes.

As shown in the result section, SAM has better performance in ranking the genes compared to t test and clearly stands out as the best. While the performance of both t test and Mann Whitney is found to be almost similar, with t test being marginally better.

The application of the t test and SAM to biological data also brings out the fact that SAM indeed dampens the effect of some lowly expressed genes turning out to be false positives in the as in case of t test.

One of the issues with simulated data is that it dose not account for outliners, which are reality in real microarray data also the fact that simulation treats the genes independent from each other in terms of expression, which is not the case in real scenario, where genes are part of some network and expression of one can be dependent on the other. Simulation dose not account for this kind of dependence. It would be

worthwhile in the future to explore this dependence using simulation, especially since t-test assumes independence between the gene expressions.

APPENDIX

SAMPLE SIMULATION CODE

```
library(stats)
# setting seed for random number generation
set.seed(553554)
gen_num <- 1000  # Total number of genes involved in simulation
m=10;n=10 # Number of Replicates
perc.changed <-0.05
utmat<-matrix(nrow=gen_num,ncol=m+n)
for (i in 1:gen_num){
temp<-sample(1:3,1)
mix <- runif(1)
if(mix < perc.changed) {
sample1<-rnorm(m,mean=0,sd=.5)
sample2<-rnorm(n,mean=1,sd=.5)
trueres[i,j]<-1
row<-cbind(t(sample1),t(sample2))
}
else {
sample1<-rnorm(m,mean=0,sd=.5)
sample2<-rnorm(n,mean=0,sd=.5)
trueres[i,j] <- 0
row<-cbind(t(sample1),t(sample2))
```

```
}
utmat[i,]<-row
}
utmat1=cbind(utmat,trueres)
ctr=0
for(i in 1:1000){
if(utmat1[i,21]>0){
ctr=ctr+1
}
}
for(i in 1:1000){
if(utmat1[i,21]>0){
utmat1[i,21]=w
w=w+1
}
else{
utmat1[i,21]=v
v=v+1
}
}
write.table(utmat1,"10_nor_test.txt",sep="\t")
```

REFERENCES

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

2. Schena M, Shalon D, Heller R, Chai A, Brown P, Davis W: **Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.** *Proc Natl Acad Sci USA* 1996, **93**: 10614-10619

3. Kerr , Martin M, Churchill A: **Analysis of variance for gene expression microarray data.** *Comput. Biol.* 2000, **7**: 819-837

4. Draghici S: **Statistical intelligence: effective analysis of high-density microarray data.** *Drug Discovery Today* 2002, **7**:S55-S63

5. Tushar G, Tibshirani, R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc. Natl. Acad. Sci. U.S.A* 2001, **98**:5116-5121

6. **Significance Analysis of Microarray in the GEDA Web Application** [http://bioinformatics.upmc.edu/Help/SAM/SAMINFO.htm]

7. Efron B, Tibshirani R, Storey D, Tusher V: **Empirical Bayes analysis of a microarray experiment.** *J. Amer. Statist. Assoc.* 2001, **96**: 1151-1160

8. Yadetie F, Laegreid A, Bakke I, Kusnierczyk W, Komorowski J, Waldum HL, Sandvik AK: **Liver gene expression in rats in response to the peroxisome proliferator-activated receptor-alpha agonist ciprofibrate.** *Physiol Genomics* 2003, **15**:9–19

9. Barrett T, Suzek O, Troup B, Wilhite E, Ngau C, Ledoux P, Rudnev D, Lash E, Fujibuchi W, Edgar R: **NCBI GEO: mining millions of expression profiles database and tools.** *Nucleic Acids Res.* 2005, **33**: D562-D566

10. Romualdi C, Vitulo N, Favero D, Lanfranchi G: **MIDAW: a web tool for statistical analysis of microarray data.** *Nucleic Acids Res.* 2005 **33**:W644-W649