

Statistical tests for identification of differentially expressed genes in cDNA microarray experiments

J Sreekumar* and K K Jose¹

Central Tuber Crops Research Institute, Sreekariyam, Thiruvananthapuram 695 017, India

¹Department of Statistics, St Thomas College, Pala, Kottayam 686 574, India

Received 20 July 2006; revised 15 January 2008; accepted 16 March 2008

Microarrays experiments are becoming a common laboratory tool for monitoring expression level in cells for thousand of genes simultaneously. The new data promise to enhance fundamental understanding of life on a molecular level and may prove useful in medical diagnosis, treatment and drug design. The greatest challenge to array technology lies in the analysis of gene expression data to identify which genes are differentially expressed across tissue samples or experimental conditions. A simple fold change was used to test the differential expression of genes. Ordinary t-test and t-test approaches with minor variations are usually used in finding differentially expressed genes under two conditions. Analysis of variance (ANOVA) and mixed model ANOVA proved to be powerful under multiple conditions or several sources of variation. Since thousands of hypotheses are tested simultaneously there is increased chance of false positives and it becomes necessary to adjust for multiple testing when assessing statistical significance of findings. Bayesian variable selection and empirical Bayesian approaches offer yet another avenue.

Keywords: ANOVA, Bayesian inference, bioinformatics, differential gene expression, DNA microarrays, t-test

Introduction:

Microarray technology provides biologists the ability to measure the expression levels of thousands of genes simultaneously¹. Currently, two platforms have emerged to dominate the microarray field, *oligonucleotide* and *spotted cDNA* arrays. The first approach developed by the Affymetrix group² involves the *in situ* synthesis of oligonucleotides (less than 30 bp long) onto solid substrates using photolithographic techniques. The second platform developed at Stanford University³ involves robotically printing cDNA clone inserts (200 to 2000 bp long) onto glass microscope slides. The supplement to *Nature Genetics*, *The Chipping Forecast*; the books, *DNA Microarrays: A Practical Approach*⁴, and *Microarray Biochip Technology*⁵; and the article, *Microarrays-Chipping in Genomics*⁶, provide general overviews of microarray technologies and different areas of application of microarrays. Image data from the arrays leads to gene specific numerical intensities representing the relative expression levels and these in turn form the input to computational analysis designed to assess the significance and relationship

across biological conditions. Some applications of this technology are: deducing functions of unknown genes, identifying disease profiles, deciphering regulatory mechanisms, classification of biological conditions, genotyping and drug development.

Novel computation tools and reliable data processing procedures are essential for the meaningful and accurate interpretation of microarray data⁷. The challenge faced by bioinformaticians in analysis is that the data contains overwhelming number of genes compared to the number of samples. Analysis of these data requires statistical tools that are acceptable to high dimensional data to make them more comprehensive. Microarray experiments raise numerous statistical questions in diverse fields, such as, image processing, experimental designs, multiple hypothesis testing and discriminant analysis. The common strategy to reduce data variability and data dimensionality is to perform two preprocessing operations before undertaking any analysis of the data: *normalization* and *filtering*. Normalization removes systematic distortions across microarrays and filtering reduce variability and decrease the dimensionality of the data by removing genes that are not sufficiently differentiated.

One of the important problems to be addressed in analysis of microarray data is the identification of

*Author for correspondence:

Tel: 91-471-2598551; Fax: 91-471-2590063,

E-mail: sreejyothi_in@yahoo.com

differentially expressed gene for further investigation. Fold change is the simplest method for identifying differentially expressed genes^{8,9}. It is based on the observed ratio (or average of ratios) between two conditions. An arbitrary cut-off value (for example, 2-fold) is used to identify differentially expressed genes. Fold change gives no assessment of statistical significance. Identifying differentially expressed genes involves the selection of test statistic which will rank the genes in order of strength of evidence of differential expression and to choose a critical value for ranking of statistic, above which any value is considered to be significant. The standard t-test is usually used to identify significantly differentiated genes across two conditions, which assumes normally distributed data and equal variance within classes. Welch statistics for unequal variances, Wilcoxon nonparametric test, permutation test, penalized *t*-statistics¹⁰, significance analysis of microarrays (SAM)¹¹, regularized *t*-test (Cyber-T)¹², moderated *t*-statistics¹³ are also applied for identifying differentially expressed genes in two conditions. Cui and Churchill¹⁴ gave a review of test statistics for differential expression for microarray experiments. Under multiple conditions and different sources of variation ANOVA method and mixed model method are applied^{15,16}. Other approaches to linear models for microarray data analysis have been described by Chu *et al*¹⁷ and Yang and Speed¹⁸⁻¹⁹.

The biological questions of differential expression can be considered as a problem in multiple hypothesis testing. A number of articles have addressed the question of multiple hypotheses testing in the context of microarray experiments and see Shaffer²⁰ and Dudoit *et al*²¹ for a review of multiple testing procedures and complete references²²⁻²⁴. Bayesian methods were used in gene selection where the criterion for identifying the differentially expressed gene is based on the posterior probability of differential expression²⁵⁻²⁸. Efron *et al*²⁷ used a nonparametric Bayes procedure for gene profiling.

Several sophisticated statistical methods are available and the method we choose to analyze the data can have profound influence on the interpretation of the results. A basic understanding of the computational tools is, therefore, required for meaningful data analysis. The paper is organized as follows. A summary of data pre-processing methods, background correction and variance stabilization are discussed in section "Data Preprocessing". Different

methods of analysis of comparative experiments for identifying the differential expression of the genes along with the problem of multiple testing and the methods of correction for multiple testing are discussed in section "Statistical methods for comparative analysis". Various computing facilities and software currently available for differential analysis of gene expression data are tabulated and an example is illustrated in section on "Computing tools".

Data Preprocessing

A common strategy to reduce data variability and data dimensionality of microarray gene expression data is to perform two preprocessing operations *normalization* and *filtering* before undertaking any analysis. The goal of the first operation is to remove systematic distortions across microarrays in order to render comparable experiments conducted under different conditions. The aim of the filtering operation is two-fold: to reduce variability by removing those genes whose measurements are not sufficiently accurate and to decrease the dimensionality of the data by removing genes those are not sufficiently differentiated.

Normalization of Microarray Data

One well known problem of cDNA technology is the consistent unbalance of the fluorescent intensities of the two dyes Cy3 (G) and Cy5 (R). There are several normalization methods and little consensus about which one to use. The choice of the method is coupled with the experimental design and these methods to reduce experimental errors can be carried out within each array and between arrays. Kerr *et al*¹⁵ and Wolfinger *et al*¹⁶ used ANOVA models for normalization where the "normalized" microarray data provide estimates of changes in gene expression that are corrected for potential confounding effects. Chen *et al*²⁶ used the ratio based method of normalization. They assume that although individual genes might be up or down regulated, in closely related cells, the total quantity of RNA produced is approximately the same for essential genes. Using this assumption they developed an approximate

probability density ratio $T_k = \frac{R_k}{G_k}$ (where R_k and G_k are the measured red and green intensities for the k^{th} array). Another method suggested is to use a LOWESS (locally weighted scatter plot smoothing)

approach for each print-tip group and its original application is for smoothing scatterplots in a weighted, least-squares fashion²⁹. Because the LOWESS smoother is available in many statistical packages³⁰, the ideas of Dudoit *et al*²¹ are easily implemented, and the two-step procedure for separately taking normalization and identification has been commonly accepted without question. Another approach is to build a regression curve to fit the intensity versus variance relationship³¹.

Variance Stabilization

In a pioneering paper Chen *et al*²⁶ derived a model based on the assumption of a constant coefficient of variation and derived the distribution of the ratio intensities. The distribution has one parameter, the coefficient of variation and according to the model it is same for all the probes on the array. To fit their model to the intensity data from a two colour cDNA array, they used a multiplicative calibration. These concepts have been widely used in microarray data analysis. However, it has become clear that for many data sets that are encountered in practice they are insufficient and the limitation mostly affects the data from weakly expressed genes^{28,32-34}.

Huber *et al*³⁵ introduced a statistical model for microarray gene expression data that comprises data calibration, the quantification of differential expression and the quantification of measurement error. They derive a transformation h for intensity measurements and a difference statistic Δh whose variance is approximately constant along whole intensity range. For transformation h , the parametric form $h(x) = \text{arcsinh}(a + bx)$ is derived from a model of the variance-versus-mean dependant for the microarray intensity data, using the method of variance stabilizing transformations. For large intensities, h coincides with the log-ratio. Rocke and Durbin³² modeled the measured expression levels from microarray data as,

$$y = \alpha + \mu + e^\eta + \varepsilon \quad \dots (1)$$

where y is the measured raw expression level for a single colour, α is the mean background noise, μ is the true expression level, and η and ε are normally distributed error terms with mean 0 and variance σ_η^2 and σ_ε^2 , respectively. The variance of y under this model is,

$$\text{Var}(y) = \eta_2 S_\eta^2 + \sigma_\varepsilon^2 \quad \dots (2)$$

where $S_\eta^2 = e^{\sigma_\eta^2 - 1}$. It is shown that for a random variable z satisfying $\text{Var}(y) = a^2 + b^2 \mu^2$ with $E(y) = \mu$, there is a transformation that stabilizes the variance to the first order, meaning that the variance is almost constant no matter what the mean might be^{35,36}.

There are several equivalent ways of writing this transformation for example,

$$f_c(z) = \ln \frac{(z + z^2 + c^2)}{2} \quad \dots (3)$$

where $c = \frac{a}{b}$. This transformation converges to $\ln(z)$ for large z , and is approximately linear at 0 ³⁶. Since this is exactly the natural logarithm when $c = 0$, it was called the generalized logarithm or glog transformation. The inverse transformation is

$$f_c^{-1}(w) = e^w - c^2 \frac{e^{-w}}{4} \quad \dots (4)$$

Both f_c and its inverse are monotonic functions, defined for all values of z and w , with derivatives of all orders. For array data, the random variable satisfies (exactly or approximately)

$$\text{Var}(z) = a^2 + b^2 E(z)^2 \quad \dots (5)$$

Overview on Statistical Methods for Comparative Analysis

Fold Change

A simple microarray experiment is carried out to detect the differences in expression between two conditions. Each condition may be represented by one or more RNA samples. In earlier days simple fold changes was used to test the differential expression of genes. The procedure was to evaluate the log ratio between expression levels in two conditions and to consider all genes for which the ratio is higher than an arbitrary cutoff value to be differentially expressed. It is known to be unreliable because statistical variability was not taken into account²⁶. Fold change method is subject to bias if the data have not been properly normalized. Tanaka *et al*³⁷ illustrated the danger of false positive and false negative when looking strictly to fold change.

t-Test

The introduction of various statistical methods relating gene's expression level under different

conditions or time periods and ranking of genes accordingly allows for more accurate analysis and for most exact feature selection. The *t*-test is a simple, statistically based method for detecting differentially expressed genes. When there are replicated samples under each condition the straight forward method is to adopt the traditional two sample *t*-test³⁸. Here we summarize the procedure as follows: Let X_1 and X_2 be two independent gene expression data under two conditions, for example normal and diseased, for a particular gene, the two sample *t*-statistic is computed as in Eq 6.

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \dots (6)$$

where s^2 , n_1 and n_2 are the pooled sample variance and the number of observations in each condition, respectively and s is computed as Eq 7.

$$s = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} \quad \dots (7)$$

where s_1^2 and s_2^2 are the variances of replicated observations in each condition. The computed value of the *t* is compared with table value of *t* with $n_1 + n_2 - 2$ degrees of freedom. Because *t*-test utilizes the variance of the samples, it has potential of addressing some of the shortcomings of fold change approach. However the small sample sizes in microarray studies can affect the variance estimates.

***t*-Test (Unequal Variance)**

Student *t*-test assumes that features within each group have similar variances. But this is rarely seen in microarray experiments. The heterogeneous variance in microarray experiments can occur due to heterogeneous sample of subjects, cell samples or expression change of transcription factor can change correlation structure. Hence, the Welch (Satterwaite's) *t*-test method developed for independent samples with unequal variances would be much more appropriate. This test is based on the difference between sample means and assumes normal distribution of the samples but allows different variances of samples. Welch method proposes an elaborate correction for degrees of freedom under unequal variances of samples as in Eq 8³⁹.

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{(n_2 - 1)}} \quad \dots (8)$$

where n_1 and n_2 are the sample variances and the number of observations in each category.

Modifications of *t*-Test

To address the shortcomings of the *t*-test in microarray experiments, several modifications of *t*-test have been proposed, which borrows information across genes and manipulates the denominator of the *t*-test statistic that combines the gene-specific variance estimate and a predicted variance (often the microarray-wide average) in different ways. They can be roughly generalized in the form

$$t^* = \frac{\bar{M}}{(s + a) / \sqrt{n}} \quad \dots (9)$$

With this modification, genes with small fold changes will not be selected as significant. Efron *et al*⁴⁰ used a percentile of the distribution of sample standard deviations as a . The SAM *t*-test (S-test) estimated a from all the individual gene variances¹¹. Broberg⁴¹ considered the two sample problem and proposed a computationally intensive method of determining the a by minimizing a combination of estimated false positive and false negative rates over a grid of significance levels and variance.

The regularized *t*-test combines information from gene-specific and global average variance estimates by using a weighted average of the two as the denominator for a gene specific *t*-test. The idea of using a *t*-statistic with a Bayesian adjusted denominator was also proposed by Baldi and Long¹² who developed the useful cyber T program. The statistic is of the form in Eq 10.

$$t^* = \frac{\bar{M}}{(s) / \sqrt{n}} \quad \dots (10)$$

where

$$s^2 = \frac{v_0 \sigma_0^2 + (n - 1) s^2}{v_0 + n - 2}$$

is the regularized standard deviation. ν_0 is the strength of the prior and σ_0^2 is the background variance estimated from all genes or a set of subset genes.

All these methods are based on using the two sample t test or its minor variation that differ in how associate a significance level to the corresponding statistic, leading to possibly large difference in resulting significance levels and number of genes selected. The p -value or observed significance level p is the probability of getting a test statistic as or more extreme than the observed one, under the null hypothesis H_0 of no differential expression.

Regression Modeling Approach

Thomas *et al*⁴² proposed a regression modeling approach which follows basic idea of Efron *et al*⁴⁰. The idea is to model the expression level y_{gi} of gene g ($g = 1, 2, \dots, p$) in the array i ($i = 1, 2, \dots, n_1, n_1 + 1, \dots, n_2$) as follows. Suppose the first n_1 and last n_2 arrays are obtained under the two conditions respectively, the general statistical model is represented as Eq 11.

$$y_{gi} = \alpha_i + b_g x_i + \epsilon_{gi} \quad \dots (11)$$

where $x_i = 1$, for $i = 1, 2, \dots, n_1$ and $x_i = 0$, for $i = n_1 + 1, \dots, n_2$. The ϵ_{gi} are the random errors with mean 0. Hence, b_g represents the difference of expression levels of gene g across two conditions. The test for differential expression, thus, becomes testing for the null hypothesis in $H_0 : b_g = 0$ against $H_1 : b_g \neq 0$. The α_i and b_g estimated using weighted least square approach and estimated the variance of b_g using the robust or sandwich variance estimator. The test statistic in Eq 12 is used with reference to the asymptotically normal distribution.

$$z_g = \frac{\hat{b}_g}{\sqrt{\text{Var}(\hat{b}_g)}} \quad \dots (12)$$

The final inference to this modeling approach is based on large sample approximations and may be a serious limitation of this approach in microarray studies where the sample sizes are too small.

Alternative Approaches

The Wilcoxon rank sum test (equivalent to Mann-Whitney test) for two groups or Kruskal-Wallis test for two or more groups has also been used as

alternative method in testing differential expression, which can be applied especially when the data are not following a normal distribution. Critical importance to possibly identify a small subset of genes that best discriminate between tissues under different conditions. The B statistic in Eq 13 proposed by Lonnstedt and Speed⁴³ is log posterior odds ratio of differential expression versus non-differential expression.

$$B = \log \frac{\text{Pr}\{DE\}}{\text{Pr}\{\text{not } DE\}} \quad \dots (13)$$

It allows for gene specific variances but it also combines information across many genes using an empirical Bayes approach and thus should be more stable than the t -statistic and it is equivalent to moderated t -statistics in terms of ranking of genes. The moderated t -statistic is shown to follow a t -distribution with augmented degrees of freedom. The moderated t inferential approach extends to accommodate tests involving two or more contrasts through the use of moderated F -statistics, *i.e.*, when the expression data is from more than two conditions.

ANOVA

The experimental design issues involved in two-colour cDNA microarray experiments have recently been receiving more attention. Kerr *et al*^{15,44} were first to propose the study of gene expression data using analysis of variance (ANOVA) models. These models perform both normalization and identification of differentially expressed genes. They applied ideas from optimal experimental designs to suggest efficient designs for the some of the common microarray experiments¹⁵. They explored the connection between the microarray designs and classical block design and gave family of ANOVA models as a guide to choose a design. The works of Churchill *et al*⁴⁵ and Yang *et al*^{18,19} have provided comprehensive insights into two-colour cDNA microarray experimental design. Yang *et al* proposed an average variance approach to choose an experimental design, while Churchill *et al* raised several principles involved in two-colour cDNA microarray experimental design. A typical ANOVA model is in Eq. 14.

$$y_{ijk} = \mu + A_i + D_j + AD_{ij} + G_g + AG_{ig} + VG_{kg} + DG_{jg} + \epsilon_{ijk} \quad \dots (14)$$

where Y_{ijk} is the measured intensity from array i , dye j , variety k and gene g on appropriate scale (typically log scale). They called “variety” to refer the mRNA samples under the study. The “varieties” may be treatment and control samples, cancer and normal cells or different time points of a biological process. In this model μ refers to the overall mean and the terms A, D and AD account for all aspects that are not gene specific. The gene effect G_g capture the average levels of expression of genes and the array-by-gene interaction (AG_{ig}) accounts for the difference due to varying sizes of spots on arrays. The dye-by-gene interaction (DG_{jg}) represents gene specific dye effects. None of these effects are of biological interest and helps in normalization of the data for ancillary source of variation. The effect of interest is the interaction between the gene and varieties (VG_{kg}), differences among this variety by gene interaction accounts for the relative gene expression. For example to estimate the relative gene expression of gene g in varieties 1 and 2, we should estimate $(VG_{1g}) - (VG_{2g})$. The error terms ϵ_{ijk} are assumed to be independent with mean 0 and variance σ^2 .

The ANOVA has several advantages, first it provides explicit quantitative term for each factor considered in the noise, thus helps in assessing the array as well as dye effects and normalization becomes unnecessary. Wolfinger *et al*¹⁶ considered two interconnected ANOVA model where the residual from the usual ANOVA model, which represents the normalized values, were fitted separately to a gene model to make inferences of variability. The normalization model is as in Eq 15.

$$Y_{gij} = \mu + T_i + A_j + TA_{ij} + \epsilon_{gij} \quad \dots (15)$$

where μ represents overall mean value, T is the main effect for treatments, A is the main effect for arrays, TA is the interaction effect of arrays and treatments, and ϵ is stochastic error. Let Y_{gij} denotes the residual from this model, then the gene model under consideration is as in Eq 16.

$$Y_{gij} = G_g + GT_{gi} + GA_{gj} + Y_{gij} \quad \dots (16)$$

where all effects are indexed by g and are assumed to serve the similar role as of normalization model but at the gene level. The effects A_j , TA_{ij} , ϵ_{gij} , GA_{gj} and Y_{gij} are all assumed to be normally distributed random variable with zero mean and variance

components σ_A^2 , σ_{TA}^2 , σ_ϵ^2 , σ_{GA}^2 and σ_Y^2 , respectively. Hypothesis testing in ANOVA model involves comparison of null model of no differential expression and an alternative model with differential expression. F-statistics are computed for individual genes from the residual sum of squares from fitting of the two models. Thus F-statistic is computed as in Eq. 17.

$$F_1 = \frac{\frac{rss_0 - rss_1}{df_0 - df_1}}{\frac{rss_1}{df_1}} \quad \dots (17)$$

Other F-like statistics (F_2 and F_3) defined by Cui and Churchill¹⁴ borrow information from other genes for estimating the variance components. F_3 uses the pooled variance estimator σ_{pool}^2 for each variance component. For balanced designs, σ_{pool}^2 is an average across genes of the individual variance estimates. F_2 uses the average of individual gene variance and pooled variance for each component. Thus F_2 and F_3 are estimated as:

$$F_2 = \frac{\frac{rss_0 - rss_1}{df_0 - df_1}}{\frac{\frac{rss_1}{df_1} + \sigma_{pool}^2}{2}} \quad \dots (18)$$

$$F_3 = \frac{\frac{rss_0 - rss_1}{df_0 - df_1}}{\sigma_{pool}^2} \quad \dots (19)$$

where rss_0 , rss_1 , and df_0 , df_1 are the residual sum of squares and degrees of freedom for the null and alternative models, respectively.

The mixed model ANOVA treats some of the factors in the experimental design as random samples from a population and there are multiple levels of variance (biological, array, spot and residual). Constructing an appropriate F-statistics using the mixed model is tricky⁴⁶. We can also apply random effects models which use BLUP (best linear unbiased prediction) for the estimation of the gene expression effects⁴⁷.

Multiple Hypothesis Testing

The biological question of differential expression can be considered as a problem in multiple hypothesis testing in which m null hypotheses were simultaneously tested, where m can be considerably

large. In such situations, false discoveries (true null hypothesis declared significant or genes identified as differentially expressed which are truly not) are inevitable. Thus, it is important in any multiple testing problems to control the error rate of false discoveries. Multiple testing procedures consist of choosing a vector of cut offs for the test statistics such that a suitably defined false positive rate is controlled at an a priori specified level α . A standard approach to the multiple testing problems consists of two aspects:

1. Computing a test statistic T_j for each gene j
2. Applying multiple testing procedures to determine which hypotheses to reject while controlling a suitably defined Type I error rate^{11,21,48}.

Type I and Type II Error Rates

In any testing situation in microarray experiments, two types of errors can be committed: a *false positive* or *Type I error*, is committed by rejecting a true null hypothesis that the gene is not differentially expressed, and a *false negative* or *Type II error* is committed when the test procedure fails to reject a false null hypothesis. The situation can be summarized by Table 1 below, where the number of Type I errors is V_n and the number of Type II errors is U_n . The numbers h_0 and h_1 of true and false null hypotheses, respectively are *unknown parameters*. The number of rejected hypotheses R_n by applying a particular test is an *observable random variable*, and the entries in the body of the table, U_n , h_1-U_n , V_n , and h_0-V_n are *unobservable random variables*. Ideally, one would like to simultaneously minimize both the chances of committing a Type I error and a Type II error. Unfortunately, that is not feasible and one seeks a trade-off between the two types of errors. A standard approach is to specify an acceptable level α for the Type I error rate and derive testing procedures, *i.e.*, rejection regions, that aim to minimize the Type II error rate, *i.e.*, maximize *power*, within the class of tests with Type I error rate at most α .

The commonly used Type I error rates in multiple hypotheses testing are:

- Per-comparison error rate (PCER): the expected value of the number of Type I errors over the number of hypotheses, $PCER = \frac{E(V_n)}{m}$
- Per-family error rate (PFER): the expected number of Type I errors, $PCER = E(V_n)$

Table 1—Type I and Type II errors in multiple hypothesis testing

		Null hypothesis		
		Not rejected	Rejected	
Null hypothesis	True (non different genes)	$h_0 - V_n$	V_n	h_0
	False (different genes)	U_n	$h_1 - U_n$	h_1
		$m - R_n$	R_n	m

- Family-wise error rate: the probability of at least one Type I error $FWER = Pr(V_n = 1)$
- False discovery rate (FDR) is the expected proportion of Type I errors among the rejected hypotheses $FDR = E\left(\frac{V_n}{R_n}; R_n > 0\right) = E\left(\frac{V_n}{R_n} \mid R_n > 0\right) \times Pr(R_n > 0)$
- Positive false discovery rate (pFDR): the rate that discoveries are false $pFDR = E\left(\frac{V_n}{R_n} \mid R_n > 0\right)$

Family-wise Error Rate

The family-wise error rate (FWER) is the probability of rejecting at least one true null hypothesis, *i.e.*, the probability of making at least one Type I error. For strong control of the FWER at level α , the Bonferroni procedure rejects any hypothesis H_i with p -value less than or equal to $\frac{\alpha}{m}$. The corresponding Bonferroni single-step adjusted p -values (\tilde{p}_i) are thus given by

$$\tilde{p}_i = \min[(mp_i], 1) \quad \dots (20)$$

Closely related to the Bonferroni procedure is the Sidak procedure, which is exact for protecting the FWER when the raw p -values are independently and uniformly distributed over $[0,1]$ ⁴⁹. The Sidak single-step adjusted p -values are given by

$$\tilde{p}_i = 1 - (1 - p)^m \quad \dots (21)$$

A more general and less conservative definition of adjusted p -values, which takes into account the dependence structure between variables, is proposed by Westfall and Young⁵⁰. The Westfall and Young single-step minP adjusted p -values are defined by

$$\tilde{p}_i = Pr(\min_{k=1, \dots, m} p_k \leq p_i \mid H_0) \quad \dots (22)$$

where H_0 denotes the complete null hypothesis.

While single-step procedures are simple to implement, they tend to be conservative for control of the FWER. Improvement in power, while preserving strong control of the FWER, may be achieved by step-down procedures. Let $p_{r1}, p_{r2}, \dots, p_{rm}$ denote the ordered raw p -values. For control of the FWER at level α , the Holm's step down procedure is given by⁵¹

$$\bar{p}_i = \max_{k=1, \dots, i} \left\{ \min \left((m - k + 1) p_{rk}, \alpha \right) \right\} \quad \dots (23)$$

Holm's procedure is less conservative than the standard Bonferroni procedure, which would multiply the p -values by m at each step. The Westfall & Young step-down minP adjusted p -values are defined by

$$\bar{p}_i = \max_{k=1, \dots, i} \left\{ \Pr \left(\min_{i=k, \dots, m} p_{ri} \leq p_{rk} | H_0 \right) \right\} \quad \dots (24)$$

Similar method (maxT) under the assumption that the statistics Tg are equally distributed under the null hypothesis and computed as above by replacing \bar{p}_i by $|Tg|$ and min by max and it is computationally less intensive⁵⁰.

False Discovery Rate

FDR is defined to be the expected value of the ratio of the number of incorrectly rejected hypotheses and the total of number of rejected hypotheses. Controlling FDR proves more powerful than FWER and has become increasingly adopted for genomic studies. Assume a usual p -value is available for each hypothesis corresponding to individual genes, based on the p -values of the hypotheses, Benjamini and Hochberg⁵² provided a multiple testing procedure that guarantees the FDR to be less than or equal to a prefixed value q . The procedure rejects H_0 for all genes i for which $p_i \leq p_k$, where

$$k = \max_{i=1, 2, \dots, n} \left\{ i: p_i \leq q \times \frac{i}{n} \right\} \quad \dots (25)$$

where $p_i, i = 1, 2, \dots, n$, are the p values sorted in ascending order with $p_0 = 0$.

Benjamini and Yekutieli⁵³ proposed a simple conservative modification of the original procedure which controls FDR under general dependence. Many related measures, such as, the positive FDR (p FDR) made popular by Storey⁵⁴⁻⁵⁷, local FDR^{27,58} and conditional FDR⁵⁹ have been suggested in the literature. Storey proposed a mixture model setup for evaluating or estimating p FDR that seems convenient

and appropriate in many multiple testing situations⁵⁶. In his framework, the test statistics are supposed to be independent and identically distributed. Each null hypothesis has a fixed probability, p_0 , of being true. Thus, the number of true null hypotheses, m_0 , is taken to be a random variable distributed as binomial ($m; p_0$). Also, marginal p -value distribution, F , is then a mixture of the uniform distribution (distribution when the null hypothesis is true) and an alternative distribution F_1 (distribution when the alternative hypothesis is true). Several papers connect empirical Bayes methods with false discovery rates. One of the key issues in estimating FDR is the assumption regarding the underlying null distribution. The Significance Analysis of Microarrays (SAM) method uses a full permutation strategy, sampling across all genes and conditions to generate such a null distribution. In Permutation test some elements of the data are permuted (shuffled) to create multiple new pseudo-data sets. One then evaluates whether a statistic quantifying departure from the null hypothesis is greater in the observed data than a large proportion of the corresponding statistics calculated on the multiple pseudo-data sets. Also, the popular software SAM computes *false discovery rates* from a frequentist viewpoint. Furthermore, the use of a mixture of distributions under the null hypothesis of no differential expression and under the alternative hypothesis of differential expression accounts, for certain, within-gene dependence. It is also possible to account for multiple testing issues by using such models. Recently methods are developed for improving estimation of false discovery rate by assuming empirical distribution of p -values^{60, 61}.

Bayesian Variable Selection

Many hypothesis testing problems in microarray rely on the strong distributional assumptions on the data intensities^{25,26,28}. The literature on microarray data is mainly based on two distributions: the log-normal and the gamma distributions, that often appear to be effective when used in a Bayesian hierarchical framework^{62,63}. In the case of empirical Bayes studies, inference is usually made on some quantities related to the posterior distribution of the parameter of interest, or of a certain type of hypothesis. West *et al*²⁵ considered an approach based on probit model and linear regression for characterizing differential gene expression. Another approach is to model the gene expression data through binary probit model for classification and use Bayesian variable selection methodology to select important

genes⁶⁴. Lee *et al*⁶⁵ proposed a hierarchical bayesian model and employed latent variables to specialize the model to a regression setting and applied variable selection to select differentially expressed genes. Bae and Mallick⁶⁶ proposed an alternative method to perform Bayesian gene selection using a two-level hierarchical Bayesian model. Another method which can simultaneously perform gene selection and clustering in high dimensional data has been proposed in microarray data analysis⁶⁷.

Bayesian models with correlated priors were proposed to model gene expression and to classify between normal and tumor tissues⁶⁸. First, empirical Bayes methods have proven to be very efficient, particularly in the situation where the number of observations available for each test is small, which is often the case in cDNA microarray studies. The main advantage of this type of approach over the traditional p -value based methods is that it allows a gene-specific inference, through the use of posterior probabilities, without the need of estimating a set of parameters for each gene. Regardless the method used to choose the parameters of the model (full Bayes or empirical Bayes), posterior probabilities allow a re-adjustment of the prior information assumed using the actual observations. As a consequence, posterior inference is in essence less sensitive to extreme values and is less model-dependent, compared to frequentist inference. The parameters of the model are typically estimated using the data for all the genes (in an empirical Bayes manner), allowing for a certain sort of dependence between genes.

Computing Tools

There are plenty of statistical softwares available for identification of differentially expressed gene from microarray data, which utilizes the statistical methods explained earlier. A list of packages and algorithms are given in Table 2. There are many packages developed for microarray data analysis in the freely available R computing environment³⁰. Also the commercial statistical packages such as SAS, MATLAB and GenStat have included microarray analysis modules, for example MATLAB bioinformatics toolbox, SAS microarray solutions and the sub-menu for microarray analysis in GENSTAT provides algorithms for microarray analysis.

Examples

SAM (Significance Analysis of Microarrays)

SAM is a statistical technique for finding significant genes in a set of microarray experiments. It

was proposed by Tusher *et al*¹¹. The software was written by Balasubramanian Narasimhan and Robert Tibshirani. The academic version of the program can be downloaded freely from <http://www-stat.Stanford.edu/~tibs/SAM/>, which after installation works as Excel add-in. This needs the pre-installation of the statistical computing package R. The data should be put in an Excel spreadsheet. The first row of the spreadsheet has information about the response measurement; all remaining rows have gene expression data, one row per gene. The columns represent the different experimental samples. The response variable may be a grouping like untreated, treated (either unpaired or paired), a multi class grouping (like breast cancer, lymphoma, and colon cancer), a quantitative variable (like blood pressure) or a possibly censored survival time. SAM imputes missing values via K-Nearest Neighbour algorithm normalization. SAM computes a statistic d_i for each gene i , measuring the strength of the relationship between gene expression and the response variable. It uses repeated permutations of the data to determine if the expression of any genes is significantly related to the response. The cut off for significance is determined by a tuning parameter Δ , chosen by the user based on the false positive rate. One can also choose a fold change parameter to ensure that called genes are differentially expressed at least a pre-specified amount.

To begin, highlight an area of the spreadsheet that represents the data and click on the SAM button in the toolbar. A dialog form as shown in Fig. 1 pops up. Select the type of response variable, and if desired, change any of values of the default parameters. For two class and paired data, one has to specify whether the data is in the logged (base 2) scale or not. Click the OK button to do the analysis. The software adds three more worksheets to the workbook. There is one which is hidden called SAM Plot data and should be left alone. The sheet named SAM Plot contains the plot that the user can interact with. The sheet named SAM Output is used for writing any output. Initially a slider pops up along with the plot shown in Fig. 2 that allows one to change the parameter and examine the effect on the false positive rate.

Positive significant genes are labeled in red on the SAM Plot, negative significant genes are green. When you have settled on a value for p , click on the List Significant Genes button, for a list of significant genes. The List Delta Table button lists the number of

Table 2—List of packages and tools for identification of differentially expressed genes from microarray data

Purpose	Tools	Author(s)/Affiliation
MA plots, t-test, lowess curve fitting, Bayesian methods	Statistics for Microarray Analysis (SMA)	Speed Berkely Research Group
Analysis cDNA microarray experiments	MAANOVA (R,MATLAB* & SAS*)	Gary Churchill's Statistical Genetics Group
Regression modeling, mixed model, multiple testing	SAS*,	SAS Institute Inc, Cary, NC, USA
R packages for the analysis of cDNA microarray data	www.bioconductor.org	Dudoit and Yang
R package for oligonucleotide array statistical analysis	affy	Irizarry, Gautier and Cope
Regularised t-test	Cyber-T	Tony Long & Harry Mangalam
FDR controlling procedures	R & S-PLUS*	Reiner A, Yekutieli D, Bejamini Y
Computing q values	Qvalue (R)	Storey J D
Variance stabilizing	vsN (R)	Wolfgang Huber
Comparing expression level between two groups	permax (R)	Robert J Gray
Gene filtering and annotation	www.bioconductor.org	Gentleman and Carey
DNA-chip analyzer	(d-Chip) DCHIP	Li and Wong
Expression profiler	Expression Profiler	Vilo,Kapusshesky,Kemmeren
S-plus library for microarray data	S-PLUS 6* GENOMICS LIBRARY	Sarkans and Brazma
Differential expression , FDR	(Significance Analysis of Microarrays)	Lee and O'Connell www-stat.stanford.edu/ tibs/SAM/index.html

*Commercial packages

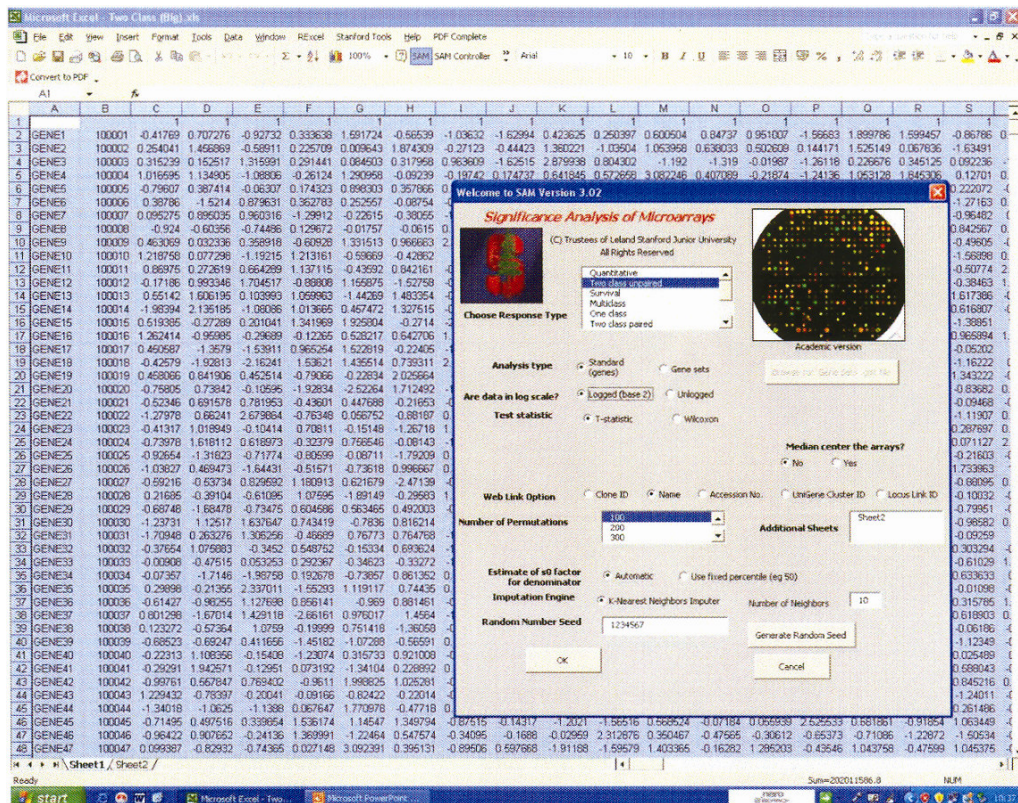


Fig. 1—SAM window in Excel with dataset and pop up menu for selection of parameters

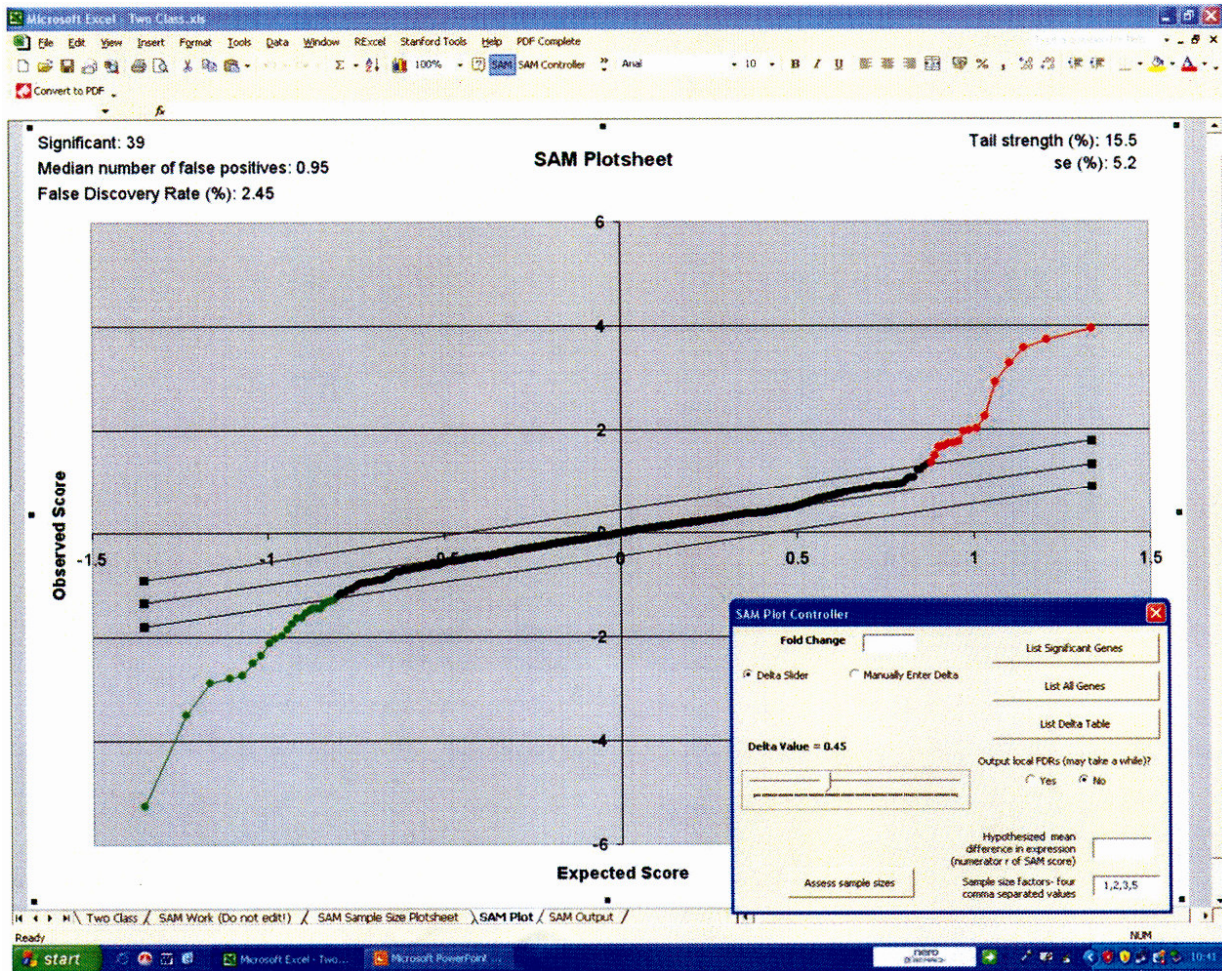


Fig. 2—SAM plot in Excel with SAM plot control menu

significant genes and the false positive rate for a number of values of p . Note that all output tables are sent to the worksheet named SAM Output Fig. 3. The List All Genes prints out all genes in the dataset. The Assess sample sizes button gives information on FDR, power etc. for various sample sizes.

Bioconductors Multtest Package

The multtest package was developed by Dudoit *et al*²¹, which estimates the joint null distribution of the test statistic T by a permutation algorithm, which reshuffle the labels of the experiments B times (permute all genes simultaneously). B is usually kept as 1000. Next, the vector of t -statistics for the permuted data: T_1, \dots, T_B were calculated. Then the package uses the empirical distribution formed by these permuted t -vectors as estimate for the null-distribution of T . Based on this permutation distribution, multtest offers several multiple testing

procedures which includes Westfall and Young minP and maxT procedures to control the FWER (family wise error rate), Benjamini/Hochberg or Benjamini/Yekutieli step-up procedures to control the FDR (false discovery rate).

Conclusion

Microarrays and gene expression data present researchers with a powerful tool to examine the function of known genes at the cellular level. A very common task in microarray data analysis is gene-by-gene selection. In the last few years scientists have been increasingly using microarray experiments to examine the behaviour of organism genes under different experimental conditions. Microarray experiments have the advantage of looking at many thousands of genes simultaneously rather than examining genes one by one, which is common practice in alternative areas of genetic research. Many

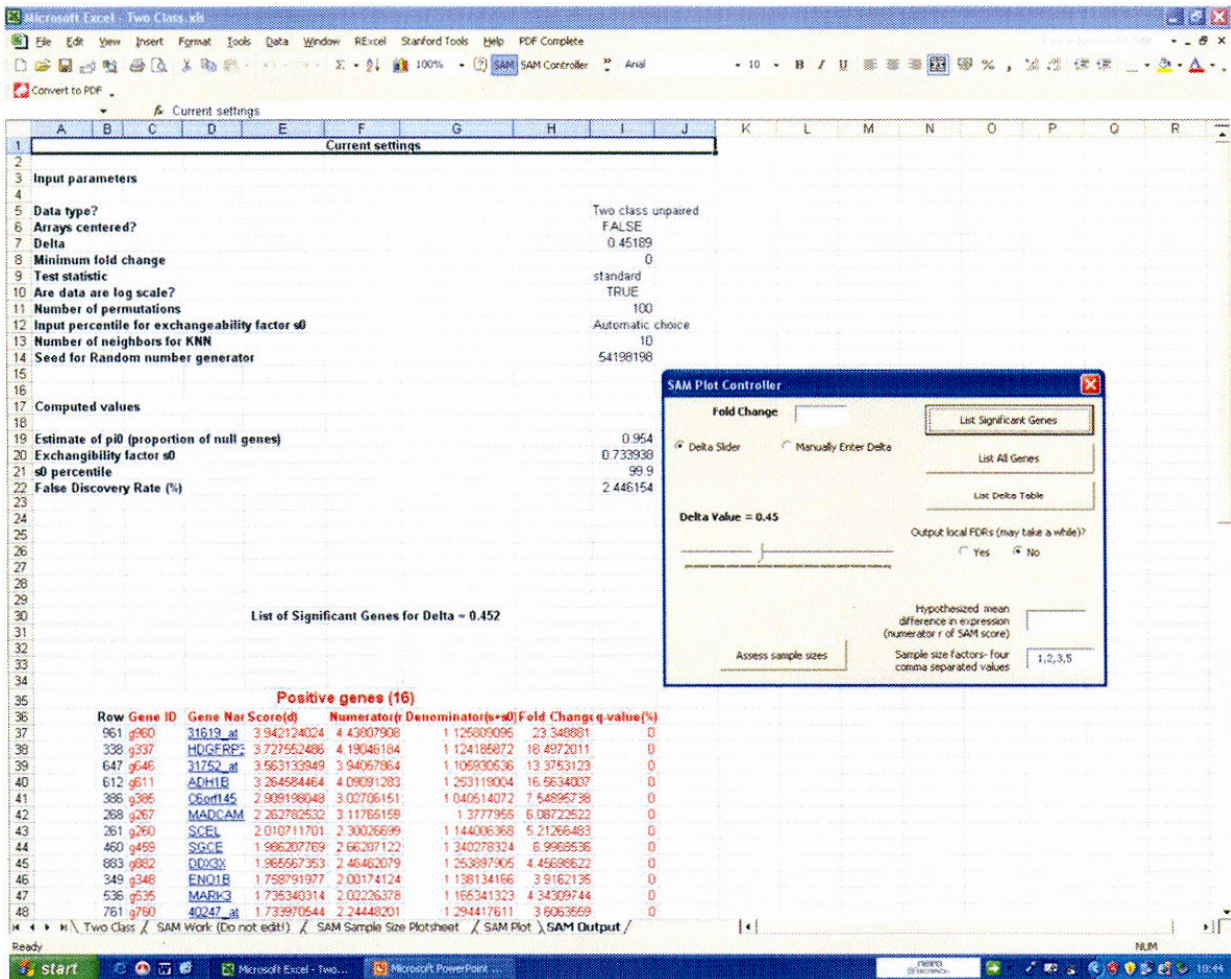


Fig. 3—SAM Output sheet in Excel with output parameters

interesting statistical questions arise when considering how to analyze data from such experiments, and in attempting to answer these it is important to be aware of the experimental processes involved in the preparation and production of microarray slides and their subsequent use. One of the important problems to be addressed in microarray analysis is to identify a set of genes, which are differentially expressed across biological condition or time periods. Usually t-test and its modification by borrowing information across genes are used for identification of genes. This paper reviews the different statistical methods for gene selection and the problem of multiple testing in the context of microarray data analysis. None of the method is considered as the best, one has to apply different methods and choose the method suitable a particular dataset. Selecting many genes as differentially expressed increases the risk of finding

genes that are not differentially expressed and selecting only a few genes increases the risk of missing differentially expressed genes. In the framework of hypothesis testing, one can reduce the type-I error (the number of false positives) at the cost of increasing the type-II error (the number of false negatives).

References

- 1 Lander E S, Array of hope, *Nat Genet*, 21 Suppl (1999) 3-4.
- 2 Lipschultz R J, Fodor S P, Gingeras T R & Lockhart D J, High density synthetic oligonucleotide arrays, *Nat Genet*, 21 Suppl (1999) 20-24.
- 3 Brown P O & Botstein D, Exploring the new world of the genome with DNA microarrays, *Nat Genet*, 21 Suppl (1999) 33-37.
- 4 Schena M, *DNA microarrays: A practical approach* (Oxford University Press, New York) 1999.
- 5 Schena M, *Microarray biochip technology* (Eaton Publishing, Natick MA) 2000.

- 6 Biju J, Anupama S & Govindaswami K, Microarrays-chipping in genomics, *Indian J Biotechnol*, 1 (2002) 245-254.
- 7 Quackenbush J, Computational analysis of microarray data, *Nat Rev Genet*, 2 (2001) 418-427.
- 8 De Risi J L, Iyer V R & Brown P O, Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278 (1997) 680-685.
- 9 Eisen M B, Spellman P T, Brown P O & Botstein D, Cluster analysis and display of genome-wide expression patterns, *Proc Natl Acad Sci USA*, 95 (1998) 14863-14868.
- 10 Efron B, Tibshirani R, Goss V & Chu G, Microarrays and their use in a comparative experiment, *J Am Stat Assoc*, 96 (2001) 1151-1160.
- 11 Tusher V G, Tibshirani R & Chu G, Significance analysis of microarrays applied to the ionizing radiation response, *Proc Natl Acad Sci USA*, 98 (2001) 5116-5121.
- 12 Baldi P & Long A D, A Bayesian framework for the analysis of microarray expression data: Regularized t-test and statistical inferences of gene changes, *Bioinformatics*, 17 (2001) 509-519.
- 13 Smyth G K, Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Stat Appl Genet Mol Biol*, 3 (2004) 3.
- 14 Cui X & Churchill G A, Statistical tests for differential expression in cDNA microarray experiments, *Genome Biol*, 4 (2003) 210.
- 15 Kerr M K, Martin M & Churchill G A, Analysis of variance for gene expression microarray data, *J Comput Biol*, 7 (2000) 819-837.
- 16 Wolfinger R D, Gibson G, Wolfinger E D, Bennett L, Hamadeh H *et al*, Assessing gene significance from cDNA microarray expression data via mixed models, *J Comput Biol*, 8 (2001) 625-637.
- 17 Chu T M, Weir B & Wolfinger R, A systematic statistical linear modeling approach to oligonucleotide array experiments, *Math Biosci*, 176 (2002) 35-51.
- 18 Yang Y H & Speed T P, Design issues for cDNA microarray experiments, *Nat Rev Genet*, 3 (2002) 579-588.
- 19 Yang Y H & Speed T P, Design and analysis of comparative microarray experiments, in *Statistical analysis of gene expression microarray data*, edited by T P Speed (Chapman & Hall/CRC Press, Boca Raton, USA) 2003, 35-91.
- 20 Shaffer J P, Multiple hypothesis testing, *Annu Rev Psychol*, 46 (1995) 561-584.
- 21 Dudoit S, Yang Y H, Callow M J & Speed T P, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Stat Sin*, 12 (2002) 111-139.
- 22 Tibshirani R, Hastie T, Narashiman B & Chu G, Diagnosis of multiple cancer types by shrunken centroids of gene expression, *Proc Natl Acad Sci USA*, 99 (2002) 6567-6572.
- 23 Efron B, Tibshirani R, Goss V & Chu G, *Microarrays and their use in a comparative experiment* (Technical Report, Department of Statistics, Stanford University, USA) 2000.
- 24 Manduchi E, Grant G R, McKenzie S E, Overton G C, Surrey S *et al*, Generation of patterns from gene expression data by assigning confidence to differentially expressed genes, *Bioinformatics*, 16 (2000) 685-698.
- 25 West M, Nevin J R, Marks J R, Spang R & Zuzan H, Bayesian regression analysis in the large p, small n paradigm with application in DNA microarray studies, (Technical Report, Duke University, USA) 2000.
- 26 Chen Y, Dougherty E R & Bittner M L, Ratio-based decisions and the quantitative analysis of cDNA microarray images, *J Biomed Opt*, 2 (1997) 364-374.
- 27 Efron B, Tibshirani R, Storey J D & Tusher V, Empirical Bayes analysis of a microarray experiment, *J Am Stat Assoc*, 96 (2001) 1151-1160.
- 28 Newton M A, Kendziorski C M, Richmond C S, Blattner F R & Tsui K W, On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, *J Comp Biol*, 8 (2001) 37-52.
- 29 Cleveland W S, Robust locally weighted regression and smoothing scatterplots, *J Am Stat Assoc*, 74 (1979) 829-836.
- 30 R Development Core Team, *R: A language and environment for statistical computing* (R Foundation for Statistical Computing, Vienna, Austria) 2005.
- 31 Jain N, Thattai J, Braciale T, Ley K, O'Connell M *et al*, Local pooled error test for identifying differentially expressed genes with a small number of replicated microarrays, *Bioinformatics*, 19 (2003) 1945-1951.
- 32 Rocke D M & Durbin B, A model for measurement error for gene expression arrays, *J Comput Biol*, 8 (2001) 557-569.
- 33 Rocke D M & Durbin B, Approximate variance-stabilizing transformations for gene-expression microarray data, *Bioinformatics*, 19 (2003) 966-972.
- 34 Theilhaber J, Bushnell S, Jackson A, & Fuchs R, Bayesian estimation of fold changes in the analysis of gene expression: The FOLD algorithm, *J Comput Bio*, 8 (2001) 585-614.
- 35 Huber W, Von Heydebreck A, Sultmann H, Poustka A & Vingron M, Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, 18 Suppl (2002) S96-S104.
- 36 Durbin B P, Hardin J S, Hawkins D M & Rocke D M, A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics*, 18 Suppl (2002) S105-S110.
- 37 Tanaka T S, Jaradat S A, Lim M K, Kargul G J, Wang X *et al*, Genome-wide expression profiling of mid-gestation placenta and embryo using a 15,000 mouse developmental cDNA microarray, *Proc Natl Acad Sci USA*, 97 (2000) 9127-9132.
- 38 Devore J & Peck R, *Statistics: The exploration and analysis of data*, 3rd edn, (Duxbury Press, Pacific Grove, CA) 1997.
- 39 Welch B L, The significance of the difference between two means when the population variances are unequal, *Biometrika*, 29 (1938) 350-362.
- 40 Efron B, Tibshirani R, Goss V & Chu G, Microarrays and their use in a comparative experiment, *J Am Stat Assoc*, 96 (2001) 1151-1160.
- 41 Broberg P, Statistical methods for ranking differentially expressed genes, *Genome Biol*, 4 (2003) R41.
- 42 Thomas J G, Olson J M, Tapscott S J & Zhao L P, An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles, *Genome Res*, 11 (2001) 1227-1236.
- 43 Lonnstedt I & Speed T P, Replicated microarray data, *Stat Sin*, 12 (2002) 31-46.
- 44 Kerr M K & Churchill G, Experimental design for gene expression microarrays, *Biostatistics*, 2 (2001) 183-201.

- 45 Churchill G A, Fundamentals of experimental design for cDNA microarrays, *Nat Genet*, 32 Suppl (2002) 490-495.
- 46 Littell R C, Milliken G A, Stroup W W & Wolfinger R D, *SAS system for mixed models* (SAS Publishing, SAS Institute Inc, Cary, NC) 1996.
- 47 Henderson C R, *Applications of linear models in animal breeding* (University of Guelph, Canada) 1984.
- 48 Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M *et al*, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, 286 (1999) 531-537.
- 49 Sidak Z, Rectangular confidence regions for the means of multivariate normal distributions, *J Am Stat Assoc*, 62 (1967) 626-633.
- 50 Westfall P H & Young S S, *Resampling-based multiple testing: Examples and methods for p-value adjustment* (John Wiley & Sons, NY) 1993.
- 51 Holm S, A simple sequentially rejective multiple test procedure, *Scand J Stat*, 6 (1979) 65-70.
- 52 Benjamini Y & Hochberg Y, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *J R Stat Soc, Ser B*, 57 (1995) 289-300.
- 53 Benjamini Y & Yekutieli D, The control of the false discovery rate in multiple hypothesis testing under dependency, *Ann Stat*, 29 (2001) 1165-1188.
- 54 Storey J D & Tibshirani R, *Estimating false discovery rates under dependence, with applications to DNA microarrays* (Technical Report, Department of Statistics, Stanford University, USA) 2001.
- 55 Storey J D, *The positive false discovery rate: A Bayesian interpretation and the q-value* (Technical Report, Department of Statistics, Stanford University, USA) 2001.
- 56 Storey J D, A direct approach to false discovery rates, *J R Stat Soc, Ser B*, 64 (2002) 479-498.
- 57 Storey J D & Tibshirani R, Statistical methods for identifying differentially expressed genes in DNA microarrays, *Methods Mol Biol*, 224 (2003) 149-157.
- 58 Efron B & Tibshirani R, Empirical Bayes methods and false discovery rates for microarrays, *Genet Epidemiol*, 23 (2002) 70-86.
- 59 Tsai C A, Hsueh H M & Chen J J, Estimation of false discovery rates in multiple testing: Application to gene microarray data, *Biometrics*, 59 (2003) 1071-1081.
- 60 Pounds S & Morris S, Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values, *Bioinformatics*, 19 (2003) 1236-1242.
- 61 Pounds S & Cheng C, Improving false discovery rate estimation, *Bioinformatics*, 20 (2004) 1737-1745.
- 62 Newton M A, Noueiry A, Sarkar D & Ahlquist P, Detecting differential gene expression with a semiparametric hierarchical mixture method. *Technical Report*, 2003. <http://www.stat.wisc.edu/newton/papers/abstracts/tr1074a.html>.
- 63 Kendzierski C M, Newton M A, Lan H, & Gould M N, On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles, 2003. <http://www.stat.wisc.edu/newton/research/arrays.html>.
- 64 Brown P J, Vannucci M & Fearn T, Multivariate Bayesian variable selection and prediction, *J R Stat Soc, Ser B*, 60 (1998) 627-641.
- 65 Lee K E, Sha N, Dougherty E R, Vannucci M & Mallick B K, Gene selection: A Bayesian variable selection approach, *Bioinformatics*, 19 (2003) 90-97.
- 66 Bae K & Mallick B, Gene selection using a two-level hierarchical Bayesian model, *Bioinformatics*, 20 (2004) 3423-3430.
- 67 Tadesse M, Sha N & Vannucci M, Bayesian variable selection in clustering high dimensional data, *J Am Stat Assoc*, 100 (2005) 602-617.
- 68 Ibrahim J G, Chen M H & Gray R J, Bayesian models for gene expression with DNA microarray data, *J Am Stat Assoc*, 97 (2002) 88-99.