Biostatistics (2005), **6**, 1, *pp.* 59–75
doi: 10.1093/biostatistics/kxh018

Improved statistical tests for differential gene expression by shrinking variance components estimates

XIANGQIN CUI
The Jackson Laboratory, Bar Harbor, Maine 04609, USA

J. T. GENE HWANG
Department of Statistical Science and Department of Mathematics, Cornell University, Ithaca, NY 14853, USA

JING QIU
Department of Statistical Science, Cornell University, Ithaca, NY 14853, USA

NATALIE J. BLADES, GARY A. CHURCHILL[†]
The Jackson Laboratory, Bar Harbor, Maine 04609, USA
garyc@jax.org

SUMMARY

Combining information across genes in the statistical analysis of microarray data is desirable because of the relatively small number of data points obtained for each individual gene. Here we develop an estimator of the error variance that can borrow information across genes using the James–Stein shrinkage concept. A new test statistic (F_S) is constructed using this estimator. The new statistic is compared with other statistics used to test for differential expression: the gene-specific F test (F_1), the pooled-variance F statistic (F_3), a hybrid statistic (F_2) that uses the average of the individual and pooled variances, the regularized t-statistic, the posterior odds statistic B, and the SAM t-test. The F_S-test shows best or nearly best power for detecting differentially expressed genes over a wide range of simulated data in which the variance components associated with individual genes are either homogeneous or heterogeneous. Thus F_S provides a powerful and robust approach to test differential expression of genes that utilizes information not available in individual gene testing approaches and does not suffer from biases of the pooled variance approach.

Keywords: ANOVA model; F statistic; Linear mixed model; Permutation; Shrinkage estimator; Variance microarray.

1. INTRODUCTION

Microarray technology has become an important tool for simultaneously screening thousands of genes for changes in their patterns of expression. In a two-color microarray experiment, a mixture of two cDNA samples (targets) that are differentially labeled with fluorescent dyes is hybridized to thousands of DNA sequences (probes) immobilized on a glass slide (Schena *et al.*, 1995). Sequences from the two targets can hybridize to complementary probe sequences. The observed fluorescent signals at each spot are, therefore,

[†]To whom correspondence should be addressed.

correlated with the mRNA concentrations in the RNA samples from which the cDNA targets were reverse-transcribed. The ratio of the two fluorescent signals at each spot is commonly used to estimate the ratio of the mRNA concentrations in the two RNA samples. In a one-color system, such as the Affymetrix arrays, all samples are labeled with the same color and each cDNA sample is hybridized to a separated array (Affymetrix, 1999). In this paper, we use examples from two-color arrays, but the results are applicable to one-color arrays.

The large amount of data generated by microarray technology is due mainly to the large number of genes represented on the array. For each gene the number of RNA samples assayed is typically small. Therefore, the commonly used approach of testing for differential expression one gene at a time often has low power (Callow *et al.*, 2000). Assuming that all of the variances are equal and using a common variance estimator for testing can substantially increase the power to detect differential expression (Kerr *et al.*, 2000) but at the risk of generating false positive and negative results when the common variance assumption is not true.

Cui and Churchill (2003b) reviewed some methods for testing differential expression of genes in microarray experiments. In addition, they defined three test statistics based on an analysis of variance (ANOVA) model. The usual ANOVA F test compares an estimate of variation across conditions to an estimate of error variance. The t-test is a special case when the number of conditions is two. One test statistic (F_1) uses only data from individual genes and is in fact the classical F statistic. Another test statistic (F_3) assumes a common error variance across genes and uses a pooled estimator of the common variance. The third test statistic (F_2) achieves a compromise by using an average of gene-specific and pooled variance estimates. When applied to real or simulated data, the F_2 test seems to work well; however, we found it hard to justify taking the simple average of variance estimates.

The idea of modifying estimators of variance has been presented by others in similar contexts. The SAM t-test (Storey and Tibshirani, 2003) adds a small constant to the gene-specific variance estimate in order to stabilize the small variances. The regularized t-test proposed by Baldi and Long (2001) replaces the usual variance estimate with a Bayesian estimator based on a hierarchical prior distribution. Lönnstedt and Speed (2002) proposed an Empirical Bayes approach that combines information across genes. Kendziorski *et al.* (2003) and Newton *et al.* (2003) considered a hierarchical gamma-gamma model to combine information across genes. Other information sharing methods have also been provided using similar strategies (Wright and Simon, 2003; Smyth, 2004).

In this paper we propose a shrinkage estimator for gene-specific variance components based on the James–Stein estimator (Lindley, 1962) and use it to construct a test statistic called F_S. The shrinkage estimator makes no prior assumptions about the distribution of variances across genes. We show that the test based on F_S has the highest or nearly the highest power among various F-like statistics and that it compares well with other 'information-sharing' statistics. The F_S test is robust, performing well under a wide range of assumptions about variance heterogeneity. It behaves well when the variances are truly constant as well as when they vary extensively from gene to gene. Furthermore, the F_S test is quite general. It can be applied in the context of general experimental designs for microarray studies (Churchill, 2002) and is not limited to the pairwise comparison of treatments. F_S can be used to construct tests that account for multiple sources of variation, both biological and technical, in microarray experiments.

In Section 2, we describe how to obtain a shrinkage estimator of variance components that provides gene-specific variances but also uses information across all of the genes in the data to improve estimation. In Section 3, we show how to use shrinkage estimators of variances to construct F-like statistics for differential expression of genes in the context of the mixed model analysis of variance. In Section 4, we validate the properties of the tests based on these statistics using simulations and real data. We simulate a canonical case to consider the problem in its most general and abstracted form. We then look at simulations of a simple microarray experiment comparing five samples and a more complex microarray experiment with biological replicates.

2. SHRINKING VARIANCE ESTIMATORS

In this section, we construct improved estimators of variance from an ensemble of individual variance estimators by shrinking them toward their common corrected geometric mean. The amount of shrinkage depends on the variability of the individual variance estimators. When individual variance estimates are similar, indicating homogeneity, the shrinkage estimator effectively pools these estimates. When individual variance estimates are widely dispersed, indicating heterogeneity, the shrinkage estimator gives greater weight to the gene specific contributions. The key result of this section is the expression in equation (2.3) below.

Let X_g be the residual sum of squared errors (SSE) and σ_g^2 be the true variance of gene g. For $g = 1, \ldots, G$ genes, it is assumed that X_g/σ_g^2 are independent, each having a Chi-squared distribution with ν degrees of freedom. Such random variable will be denoted as χ_ν^2. Therefore, we have

$$X_g \sim \sigma_g^2 \chi_\nu^2.$$

We take a natural logarithmic transformation on X_g to obtain a common location problem as shown below. We then have

$$\ln \frac{X_g}{\nu} \sim \ln \sigma_g^2 + \ln \frac{\chi_\nu^2}{\nu}. \tag{2.1}$$

Hence, if we denote the mean of $\ln \frac{\chi_\nu^2}{\nu}$ as m, by subtracting m from both sides, we could write equation (2.1) as

$$X_g' \sim \ln \sigma_g^2 + \epsilon_g'$$

where $X_g' = \ln \frac{X_g}{\nu} - m$ and $\epsilon_g' = \ln \frac{\chi_\nu^2}{\nu} - m$. Let V be the variance of ϵ_g'. By using a first-order Taylor expansion of the last term in equation (2.1), $\mathrm{Var}(\ln \frac{\chi_\nu^2}{\nu}) \approx \mathrm{Var}(\frac{\chi_\nu^2}{\nu}) = \frac{2}{\nu}$. In Table 1, we give the ratio of V to $2/\nu$, which eventually converges to one. When applied to X_g' $(1 \leqslant g \leqslant G)$ in estimating $\ln \sigma_g^2$, the positive part James–Stein estimator that shrinks toward the common mean $\bar{X}' = \sum X_g'/G$ is

$$\bar{X}' + \left(1 - \frac{(G-3)V}{\sum (X_g' - \bar{X}')^2}\right)_+ \times (X_g' - \bar{X}') \tag{2.2}$$

where for any number a, a_+ denotes $\max(a, 0)$. The truncation enacted by the '+' is necessary to avoid overshrinking.

Transformation back to the original scale gives the shrinkage estimator for σ_g^2,

$$\tilde{\sigma}_g^2 = \left(\prod_{g=1}^{G} (X_g/\nu)^{1/G}\right) B \times \exp\left[\left(1 - \frac{(G-3)V}{\sum (\ln X_g - \overline{\ln X_g})^2}\right)_+ \times (\ln X_g - \overline{\ln X_g})\right], \tag{2.3}$$

where $\overline{\ln X_g} = \frac{1}{G}\sum \ln(X_g)$, and $B = \exp(-m)$ is a bias correction. Note that multiplying the geometric mean $(\prod_{g=1}^{G}(X_g/\nu))^{1/G}$ by B gives an unbiased estimator of σ^2 when $\sigma_g^2 = \sigma^2$ for all g.

The values of B (and also V) depend on ν. They can be simulated easily and values are given in Table 1. Note that B is always larger than one, hence, the geometric mean without B underestimates σ^2 when all σ_g^2 are equal to σ^2.

Taylor expansion applied to the inverse log-transformed estimator in equation (2.3) demonstrates that it is similar to Ghosh *et al.*'s estimator (Ghosh *et al.*, 1984) (derivation not shown). If the collection of

Table 1. *Values of B (bias correction) and $V/(2/\nu)$ as a function of ν. These values are used in equation (2.3) to construct the estimates that shrink the unbiased estimators of variances to their corrected geometric mean. When ν is greater than 50, B and $V/(2/\nu)$ are effectively 1*

ν	B	$V/(2/\nu)$	ν	B	$V/(2/\nu)$
1	3.53	2.45	13	1.08	1.08
2	1.77	1.64	14	1.08	1.08
3	1.44	1.39	15	1.07	1.07
4	1.31	1.27	16	1.07	1.06
5	1.24	1.22	17	1.06	1.06
6	1.19	1.18	18	1.06	1.06
7	1.16	1.15	19	1.06	1.05
8	1.14	1.13	20	1.05	1.05
9	1.12	1.12	25	1.04	1.04
10	1.11	1.11	30	1.04	1.03
11	1.10	1.10	40	1.03	1.03
12	1.09	1.09	50	1.02	1.02

all X_g $(g = 1, \ldots, G)$ is represented by $\boldsymbol{X}$, it has been shown that Ghosh *et al.*'s estimator dominates $\boldsymbol{X}/(\nu + 2)$, which is better than $\boldsymbol{X}/\nu$ from the collection of individual variance estimators, according to the sum of squared invariant losses (Ghosh *et al.*, 1984). This provides a theoretical foundation that the estimator in equation (2.3) may work well as an estimator of variance. Extensive comparisons among several variations on this estimator show that the version (2.3) presented here behaves best in construction of test statistics as described in Section 3. In particular, the estimators in (2.3) provide a test statistic with better performance than similar statistics based on the Ghosh *et al.* (1984) estimator.

3. CONSTRUCTING F-LIKE STATISTICS

To illustrate how to construct F-like statistics using different variance estimators, we start with the general F statistic for a general linear mixed model and then introduce the statistics based on shrinkage estimators.

A general linear mixed model (Searle *et al.*, 1992) can be written as

$$Y = X\beta + Zu + \epsilon \tag{3.1}$$

where Y is the vector of observations, X is the design matrix of fixed effects β, Z is the design matrix of random effects u, and ϵ is the vector of the residuals.

The variances of the random effects u and residuals ϵ in equation (3.1) can be estimated using the restricted maximum likelihood method (REML) (Searle *et al.*, 1992). Estimation of the corresponding fixed effects ($\hat{\beta}$) and the prediction of the random effects ($\hat{u}$) can be obtained through generalized least squares using the estimated variance components (Searle *et al.*, 1992; Witkovsky, 2002).

The variance covariance matrix of $\hat{\beta}$ and $\hat{u}$ can be estimated as

$$\hat{C} = \begin{bmatrix} X'\hat{R}^{-1}X & X'\hat{R}^{-1}Z \\ Z'\hat{R}^{-1}X & Z'\hat{R}^{-1}Z + \hat{G}^{-1} \end{bmatrix}^{-}, \tag{3.2}$$

where $\hat{R}$ is a matrix with the estimates of residual variances on the diagonal and 0 elsewhere, and $\hat{G}$ is a matrix with the variance components estimated for random effects u on the diagonal and 0 elsewhere. The

'$-$' sign represent the generalized inverse of the matrix. Linear combinations of the fixed effects (denoted by L) in equation (3.1) can then be tested using an F statistic (Littell *et al.*, 1996) constructed as

$$F = \frac{\hat{\beta}' L' (L' \hat{C} L)^{-1} L \hat{\beta}}{\text{rank}(L)}. \tag{3.3}$$

When a linear mixed model is fit to microarray data one gene at a time, the design matrices of X and Z are the same for all genes. Therefore, the general linear mixed model for gene g can be expressed as

$$Y_g = X\beta_g + Zu_g + \epsilon_g \tag{3.4}$$

The statistic defined in equation (3.3) can then be used to test the fixed effect β_g directly for each gene. We refer to this as the gene-specific F test (F_1) (Wolfinger *et al.*, 2001). The variance components in this test are estimated using data from only one gene and the power of this test is likely to be low in experiments with only a few RNA samples. Other F-like statistics, F_2 and F_3, defined by Cui and Churchill (2003b) can borrow information across genes when estimating the variance components. F_3 uses the pooled variance estimator $\hat{\sigma}^2_{\text{pool}}$ for each variance component. For balanced designs, $\hat{\sigma}^2_{\text{pool}}$ is an average across genes of the individual variance estimates. F_2 uses the average of $\hat{\sigma}^2_g$ and $\hat{\sigma}^2_{\text{pool}}$ for each component. In this paper we define a new F-like statistic, F_S, which uses $\tilde{\sigma}^2_g$ from the shrinkage estimator in equation (2.3) as the variance component estimator for each gene. The variance component estimators are then used in equations (3.2) and (3.3) to compute the corresponding F statistics.

Consider a fixed effects ANOVA model in which Z and u are empty. If we denote the sum of squares of relative expression across samples for gene g as Δ_g, then the four F tests can be written as

$$\begin{aligned} F_1 &= \Delta_g / \hat{\sigma}^2_g, \\ F_2 &= \Delta_g / \tfrac{1}{2}(\hat{\sigma}^2_g + \hat{\sigma}^2_{\text{pool}}), \\ F_3 &= \Delta_g / \hat{\sigma}^2_{\text{pool}}, \\ F_S &= \Delta_g / \tilde{\sigma}^2_g. \end{aligned} \tag{3.5}$$

This form highlights the intuition behind the construction of these statistics.

The justification for choosing one of these four statistics depends on our assumptions about the variability of the variances across genes. If all variance components are constant across genes, then F_3 is the right statistic. If the variance components are gene specific, then F_1 is the right statistic. However, a statistic like F_S should be more efficient when there is limited information to estimate the gene specific variance components. Comparisons of these tests in different situations are described in Section 4.

For simple microarray experiments, fixed effects ANOVA models, a special case of the general linear mixed model with empty Z and u in equation (3.1), can be used for modeling and computational convenience. The error variance for each gene can be estimated using the residual mean square error (MSE), which is the SSE divided by its degrees of freedom (ν). Thus, the denominators of F_1, F_2, F_3, and F_S can be estimated based on these MSEs across the genes in equation (3.5).

The null distributions of the modified F statistics are not readily available. The F_1 test for a fixed-effect ANOVA model, which is used for small or simple experiments, has a standard F distribution and critical values could be obtained from the F tables under typical distributional assumptions; however, when mixed-effects ANOVA models are used for large and complicated experiments, the F_1 in equation (3.5) does not strictly follow the F distribution, although a conservative approximation can be obtained (Littell *et al.*, 1996). Since F_2, F_3, and F_S are not standard F statistics, their null distributions can be approximated by permutation analysis (Wu *et al.*, 2003). It may be prudent to establish all critical values by permutation analysis because distributional assumptions are often questionable for microarray data.

Permutation analysis is a nonparametric approach to establish the null distribution of a test statistic. The key to developing a permutation strategy is to identify units in the experiment that are exchangeable under the null hypothesis. In microarray experiments, if we allow for gene-specific variance heterogeneity, then the unit must be whole arrays. Furthermore, the arrays that are to be shuffled will depend on the design of the experiment and the factor(s) being tested. Two-color arrays are slightly more complex than single-color systems as the pairing between the two channels of the array must be maintained in the permuted units. To execute the permutation analysis we generate random shuffles ($p = 1, \ldots, P$) of whole array units and compute a new set of statistics $F_g^{(p)}$ ($g = 1, \ldots, G$). Due to the large computational demand, we can typically perform only 100 permutations. For example, a 2000-gene experiment with 30 arrays requires about an hour on our 32-node Beowulf cluster. To reduce the granularity of the gene-specific null distribution, a common null distribution for each test statistic is established using the entire collection of $F_g^{(p)}$ values over indices p and g based on the assumption that the F statistics have common null distributions across genes (Storey and Tibshirani, 2003).

4. SIMULATION STUDIES

In order to compare the tests based on each of the four F statistics in their ability to identify differentially expressed genes, we first simulated an abstracted canonical form and then simulated data based on real microarray experiments. For the latter we simulated data based on models using estimated parameters from real data sets. We also used resampling methods based on real data. The first microarray experiment that we considered is based on a five-sample comparison with no biological replicates and the second is based on a three-sample comparison with biological replicates.

4.1 *Canonical simulation*

To evaluate the tests based on the four F statistics in a general setting, we simulated data in a canonical form and studied the successful detection rate (the percentage of true positives identified), which is analogous but not identical to the average power in Dudoit *et al.* (2003), of each test at several levels of variance heterogeneity, represented by coefficient of variation (CV) of the variances and degrees of freedom (ν).

We define the canonical form of this problem as $y_{g,t} = \theta_{g,t} + \epsilon_{g,t}$ for gene $g = 1, \ldots, G$ and treatment $t = 1, \ldots, T$, where $\theta_{g,t}$ represents the relative expression level of gene g under treatment condition t, and $\epsilon_{g,t}$ is the gene-specific residual error ($\epsilon_{g,t} \sim N(0, \sigma_g^2)$) associated with estimating $\theta_{g,t}$.

In this simulation, the residual variances, σ_g^2, were drawn randomly from the 15 600 residual variance estimates from the tumor data set described in Section 4.3. To vary the CV of these residual variances while keeping their geometric means constant, we rescaled them using a tuning parameter τ:

$$Z_g = \frac{\sigma_g^{2\tau}}{\mathrm{gm}(\sigma_g^{2\tau})} * \mathrm{gm}(\sigma_g^2), \tag{4.6}$$

where *gm* stands for geometric mean. When $\tau = 0$, $CV = 0$, corresponding to the homogeneous variance case. We study four cases where $\tau = 0$, 0.78, 1.5 and 2.3, which correspond to $CV = 0$, 1, 4 and 20. The two middle cases are typical of real microarray data.

The treatment effect for each gene can be estimated as

$$\hat{\Delta}_g = \frac{1}{t-1} \sum_{t=1}^{T} (y_{g,t} - \overline{y_{g.}})^2. \tag{4.7}$$

This is also the common numerator for all four F statistics in equation (3.5). In this case, the denominators of all F statistics are obtained using residual MSE_g in the place of $\hat{\sigma}_g^2$ in equation (3.5). The residual MSE_g for each gene was generated from a chi-square distribution and scaled by gene-specific residual variance Z_g, $\text{MSE}_g \sim Z_g \chi_\nu^2/\nu$, where ν are the degrees of freedom associated with MSE_g. We studied many degrees of freedom but only report $\nu = 2, 6$, and 50 here to represent small, moderate and large microarray experiments.

To establish the null distribution for the F tests, we set $\theta_{g,t} = 0$ for all $g = 1, \ldots, 5000$, $t = 1, \ldots, 5$. We calculated F_1, F_2, F_3, and F_S for each gene and then use the 95% quantiles as the critical values.

To calculate the successful detection rate for each F test, we generated a number of non-zero θ_g. Because the successful detection rate of a test depends on the magnitude of the effect ($\Delta_g = \frac{1}{t-1}\sum(\theta_{g,t} - \overline{\theta}_{g.})^2$), we study it as a function of Δ_g. Specifically, we let $Q_{g,t} \sim N(0, 1^2)$ and $\theta_{g,t} = K Q_{g,t}/\sqrt{\sum_{t=1}^5 Q_{g,t}^2}$, consequently, $K = \sqrt{\Delta_g(t-1)}$. By varying K, we can vary the treatment effect. For each K value we studied, we generated 5000 genes and recorded the percentage that were identified by each test. Figure 1 shows the successful detection rate of the four tests as a function of $\sqrt{\Delta_g(t-1)}$ for degrees of freedom, $\nu = 2, 6, 50$, and heterogeneity, $CV = 0, \approx 1, \approx 5$, and ≈ 20. When all the treatments are identical, $\sqrt{\Delta_g(t-1)} = 0$, the null hypothesis H_0 holds. In general, F_1 shows good power only when ν is large ($\nu > 6$). F_3 only has good power when variance heterogeneity is low ($CV < 1$). F_2 is similar to F_3 but more robust. It still has good power when CV is about 4. The power of the F_2 and F_3 tests decrease when the CV increases. When the CV is larger than 10, F_3 loses power completely and F_2 loses most of its power. Compared with the other tests, F_S is the most robust and is usually most powerful or nearly so. F_S is more powerful than or as powerful as F_1 and F_3 in all the situations. The improvement over F_1 is quite substantial when ν is small. It also has a substantial advantage over F_2 and F_3 when the CV is large. When the CV is small, the power of F_S is still comparable to that of F_3.

4.2 *Analysis and simulation of a microarray experiment: Case I. Technical replication*

To compare the four F-like tests in a simple microarray experiment, we applied them to experimental data and performed simulations based on the results of this experiment. The experiment compared two human colon cancer cell lines, CACO2 and HCT116, and three human ovarian cancer cell lines, ES2, MDAH2774 and OV1063, using a design in which the samples were arranged in a loop and no reference sample was used (Figure 2A). Fluorescent dye labeled cDNA targets were hybridized to cDNA microarrays containing 9600 human cDNA clones from the Research Genetics sequence verified human cDNA collection (Invitrogen, Carlsbad, CA) spotted in duplicate. Slides were scanned using the GenePix4000 microarray scanner and the median intensities of each spot were calculated using an image processing software (Axon Instruments, Inc., Foster City, CA).

To simplify the analysis, the two spots for the same gene on each array were averaged at the original signal level. The data were then intensity LOWESS transformed (Cui *et al.*, 2003) and normalized before fitting the following ANOVA model to each gene:

$$y_{ij} = \mu + A_i + D_j + S_{k(i,j)} + \epsilon_{ij}. \tag{4.8}$$

In this model, μ is the gene mean; A_i ($i = 1, \ldots, 10$) is the array effect; D_j ($j = 1, 2$) is the dye effect; $S_{k(i,j)}$ ($k = 1, \ldots, 5$) is the sample effect. The sample index k is determined by the array and channel indices i and j. Here ϵ_{ij} is the residual, terms μ, D_j and $S_{k(i,j)}$ are treated as fixed while term A_i is treated as random. To put this model in the context of the general linear mixed model (equation 3.4), μ, D_j and $S_{k(i,j)}$ belong to β and A_i belongs to u. The dimension of the X matrix is 20×8 with rank of 6

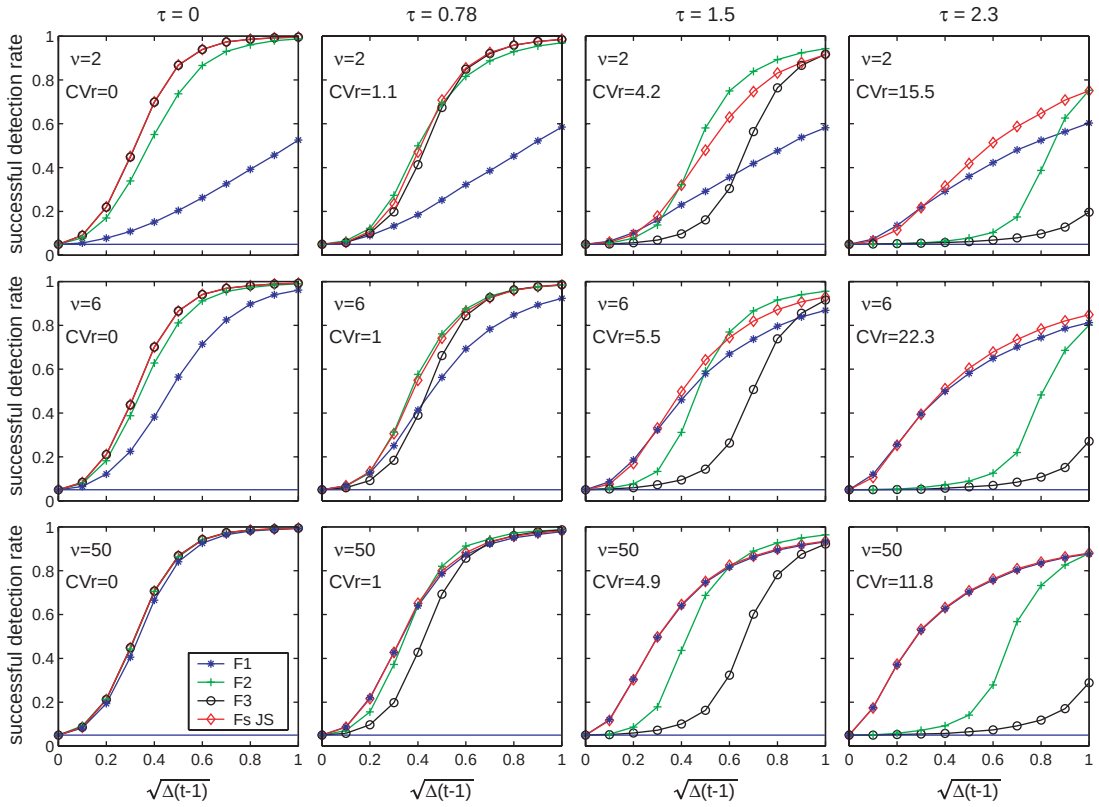

Fig. 1. Successful detection rate comparison among the four F tests using the canonical simulations. In each panel, the successful detection rate of each F test is plotted against the treatment effect, $\sqrt{\Delta(t-1)}$. The variability of the individual variances is controlled by τ shown on the top and is reflected by the coefficient of variation (CV) shown at the upper left corner of each panel. The degrees of freedom ($\nu = 2$, 6, and 50) are noted in each panel at the upper left corner. The nominal type I error rate of 0.05 is indicated by a solid blue line.

and the dimension of the Z matrix is 20×10 with rank of 9. The variance components of A_i and ϵ_{ij} were estimated (Searle *et al.*, 1992) for each gene and their distributions were compared (Figure 3A). The array variance is substantially larger than the residual variance but it has similar heterogeneity ($CV = 1.34$) to the residual variance ($CV = 1.79$). We note that array variance has little impact on the F tests because of the experimental design (Cui and Churchill, 2003a); thus in simple experiments like this one treating array as a fixed effect simplifies the computation with little impact on the results.

The four F test statistics were constructed under model (4.8) and their null distributions were established by permutation analysis (Kerr *et al.*, 2000; Wu *et al.*, 2003; Cui and Churchill, 2003b). The permutation unit in this case is one array. At a nominal significance level of 0.01, F_1, F_2, F_S and F_3 detected 1588, 2012, 1896 and 981 significant genes, respectively. The volcano plot (Figure 3B) illustrates the differences among the four F tests. The significant genes for F_1 are located above the horizontal line and those for F_3 are located right of the vertical line. The significant genes identified by F_S and F_2 are indicated by yellow and red coloring respectively and are generally in the upper right corner.

To study the false positive and the successful detection rates of each F test, we simulated 10 data sets based on this design, each with 1000 constant genes and 1000 differentially expressed genes. The

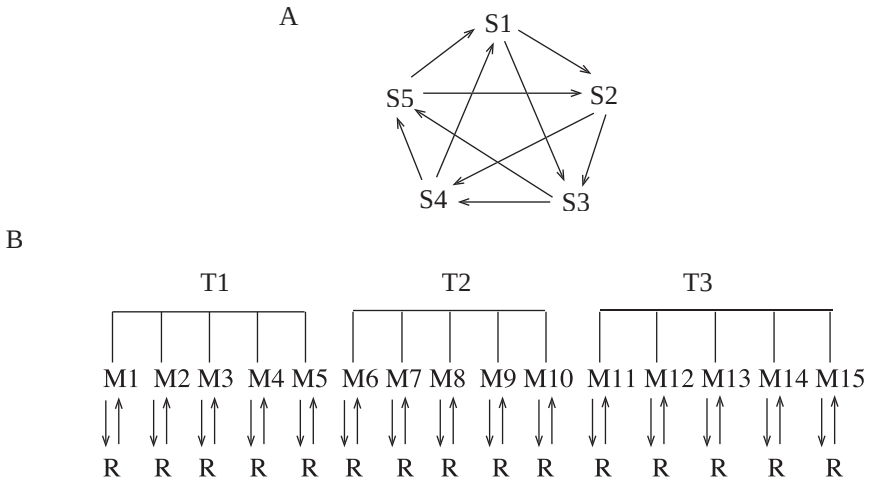

Fig. 2. Illustration of the microarray designs used in the paper. Panel A is a double-loop design comparing five samples (S1–S5). Panel B is a reference design used in the tumor experiment to compare three treatments (T1, T2, and T3) with five mice in each treatment group and one pair of arrays for each mouse. Arrows represent arrays with head pointing to the sample with Cy3 labeling and tail pointing to the sample with Cy5 labeling. R denotes reference sample.

individual treatment effects $S_{k(i,j)}$ were drawn randomly from distribution $N(0, 0.3^2)$. The μ and D_j were generated from normal distributions $N(0, 0.65^2)$ and $N(0, 0.35^2)$, respectively. These fixed effects parameter values were held constant across all simulations. For each simulation, A_i was generated randomly from a normal distribution $N(0, 0.6^2)$ and the residuals (ϵ_{ij}) were drawn randomly from normal distribution $N(0, \sigma_g^2)$, where the gene specific variance σ_g^2 was sampled randomly without replacement from the 9600 estimates of residual variance of the loop data set. The variability of the residual variances was controlled by τ in the same fashion as for the canonical simulation, but the value of τ was set to be 0.8, 1, and 1.5 to only cover the ranges of variability that we have seen in real data sets. Corresponding CVs are about 1.2, 1.8, and 3.7.

The averaged results of the 10 simulations at nominal significance level of 0.05 are shown in Table 2. Among the 1000 null model genes, fewer than 50 false positives were detected by each F test, which indicates that the actual average type I error rate is somewhat lower than the expectation in each case. Among the 1000 differentially expressed genes, the majority were identified by all four F tests, but the number of identified true positives decrease as CV increases. F_1 and F_S are less affected by heterogeneity than F_2 and F_3. F_S identifies fewer true positives than F_2 when the CV is around 1.2 and 1.8, but it identifies more than F_2 when the CV is around 3.7. F_S identifies more true positives than F_1 and F_3 regardless of the degree of heterogeneity. The successful detection rate of these F tests plotted against the sample effect is shown in Figure 4 (A–C). The relationship among all four F tests are similar to those obtained in the canonical simulation. The volcano plots from one simulation (Figure 5) illustrate the false positives and false negatives from each F test. The results of all individual simulations are shown in supplemental Tables 1A–C.

4.3 *Analysis and simulation of a microarray experiment: Case II. Biological replication*

A promising trend in microarray experiments is to include biological replicates of samples in order to account for inherent biological variation. To accommodate this trend, mixed linear models with biological

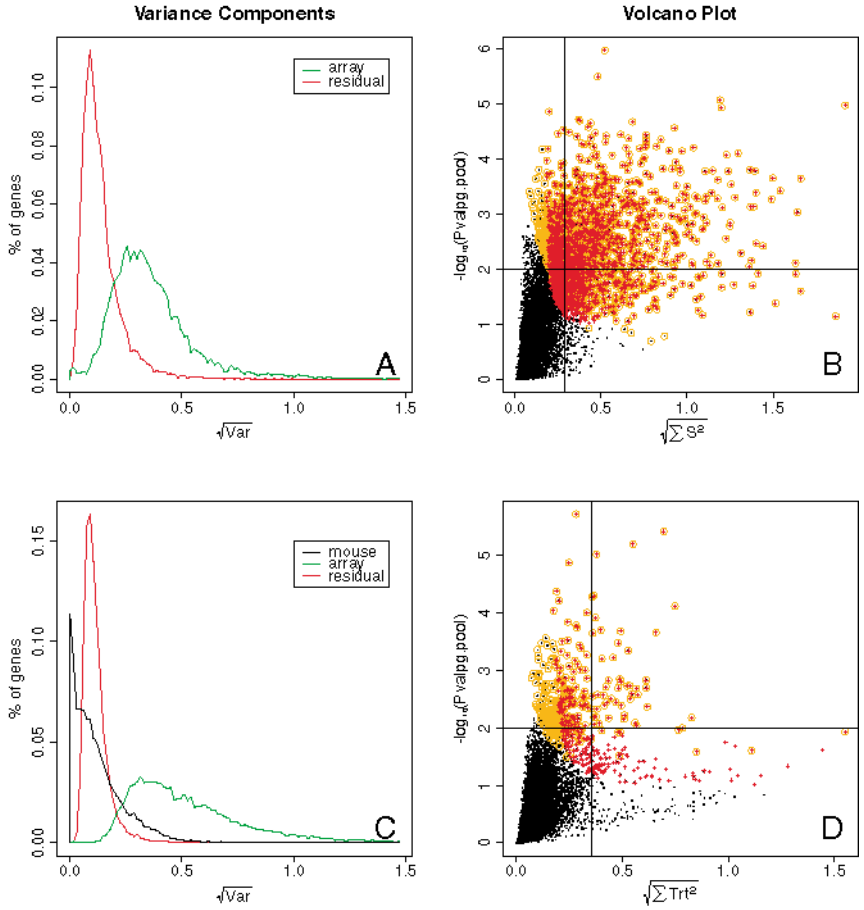

Fig. 3. Variance component plots and volcano plots of the loop and tumor data sets. Panel A shows smoothed histograms of the variance components of the loop data with A_i treated as random (equation 4.8). Panel B shows the volcano plot from the loop experiment. Panel C shows the smoothed histograms of the variance components from the tumor data set. Panel D shows the volcano plot of the tumor data. In the volcano plots, the -$\log_{10}$ p values based on permutation analysis of the F_1 statistic are plotted against the treatment effect. Horizontal and vertical lines represent the 0.01 nominal significance level for F_1 and F_3 respectively. Yellow square, F_S significant; Red '+', F_2 significant.

replicates treated as random effects are required. Here we analyze a representative data set and perform simulations based on these data to compare the properties of the four F-like tests in this experimental setting.

The granulosa cell tumor microarray experiment was performed using eight week old SWXJ-9 mice. The effects of dietary androgenic supplementation (DHEA, testosterone and control) were assessed. RNA samples from each mouse were compared to the Stratagene reference RNA using two microarrays with dye labeling reversed (Figure 2B). Fluorescent dye labeled cDNA targets were hybridized to DNA microarrays printed with the 15 000 NIA clone set spotted in duplicate. Slides were scanned and the mean intensities of each spot calculated using the GenePix4400 microarray scanner and image processing software.

Table 2. *Average number of true and false positives identified by each F test in 10 simulations of model (4.8). Significance level is nominal 0.05. The total number of genes is 2000, with 1000 constant genes and 1000 differentially expressed genes. CV_r, average CV of the residual variance; TP, true positives; FP, false positives. The results from individual simulations are shown in Supplemental Tables 1A–C*

	$CV_r = 1.2$		$CV_r = 1.8$		$CV_r = 3.7$	
	TP	FP	TP	FP	TP	FP
F_1	766	44	746	44	713	41
F_2	866	33	829	31	688	30
F_S	857	35	824	39	762	40
F_3	806	37	752	46	434	50

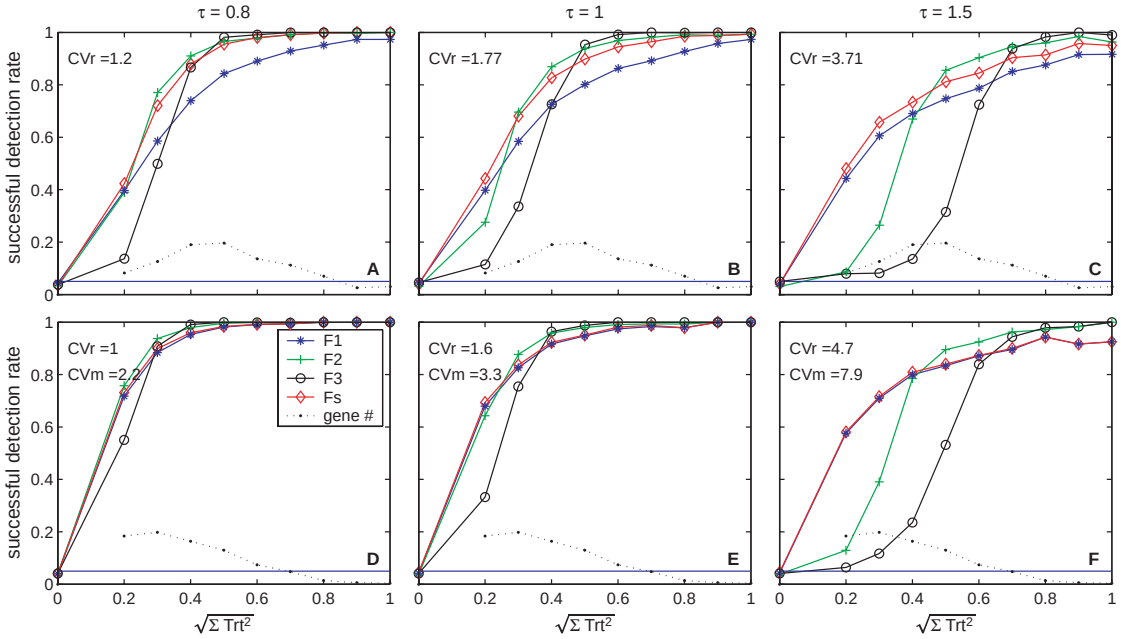

Fig. 4. The average successful detection rate of four F tests from 10 microarray simulations. The data were simulated with the fixed-effect ANOVA model (A–C) or mixed-effects ANOVA model with two variance components (D–F). The values of variance components in A–C and D–E were randomly drawn from the loop and tumor data, respectively. The variability of the variances across genes are controlled by τ (0.8, 1, 1.5) and reflected by CV_r and CV_m. CV_r, CV of the residual variance component; CV_m, CV of the mouse variance component.

The raw data were preprocessed as described above before fitting the following mixed ANOVA model (Wolfinger *et al.*, 2001) for each gene,

$$y_{ij} = \mu + A_i + D_j + T_{k(i,j)} + M_{l(i,j)} + R_{h(i,j)} + \epsilon_{ij}, \tag{4.9}$$

with μ for the gene mean, A_i for array effect ($i = 1, \ldots, 30$), D_j ($j = 1, 2$) for the dye effect, $T_{k(i,j)}$ ($k = 1, 2, 3$) for the treatment effect, and $M_{l(i,j)}$ ($l = 1, \ldots, 15$) for mouse effect. $R_{h(i,j)}$ is an indicator of reference ($h = 1$) versus tissue sample ($h = 2$). The indices of treatment, mouse and reference are determined by the combination of array and dye. We treat μ, D_j, $T_{k(i,j)}$ and $R_{h(i,j)}$ as fixed effects. The biological replicate, *mouse* ($M_{l(i,j)}$), effect is treated as a random effect. Therefore, the mouse variance is

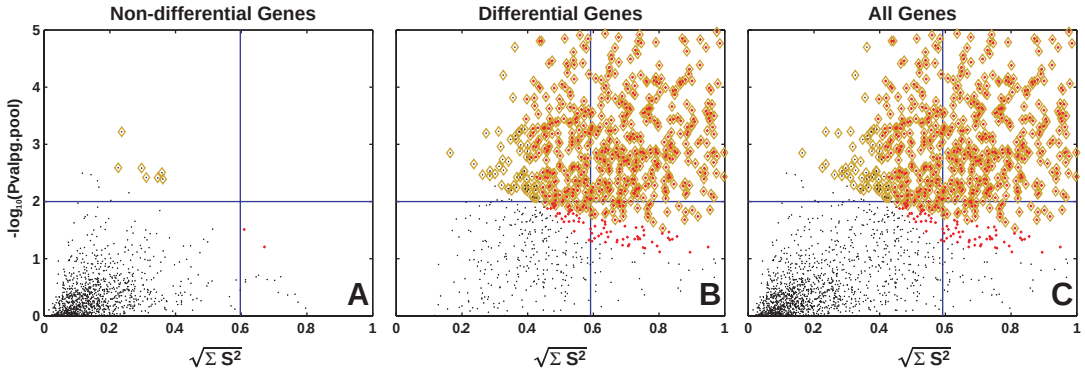

Fig. 5. Representative volcano plots of the simulations based on the loop data set using a fixed-effect ANOVA model (equation (4.8)) with $CV = 2.2$. Panels A, B and C correspond to non-differential, differential, and all genes, respectively. The selected genes in panel A are false positive and the unselected genes in panel B are false negative results. Horizontal and vertical lines indicate the 0.01 nominal significance level for F_1 and F_3, respectively. Red '+', F_2 significant; orange diamond, F_S significant.

included with the error variance in tests that compare treatments ($T_{k(i,j)}$) (McLean *et al.*, 1991; Churchill, 2002; Cui and Churchill, 2003b). The array effect (A_i) is treated as random effect but it has little impact on the F statistics.

The variance components $mouse$, $array$, and $residual$ in this model were estimated using REML (Searle *et al.*, 1992). Their distributions are shown in Figure 3C. The array variance is the largest component and has only moderate heterogeneity ($CV = 1.5$). The mouse variance is the smallest, but it has greatest heterogeneity ($CV = 3.4$). Most of the genes have small mouse variance, but a small proportion of genes show large variation across individual mice. The residual variances are intermediate, between array and mouse components, in size and have only moderate heterogeneity ($CV = 1.7$).

The four F statistics were computed for each gene and their null distributions were established using permutation analysis. The permutation unit in this case is mouse, which consists of a pair of arrays measuring the same RNA samples with the dye labels reversed, because mouse is a nested factor within treatment (Figure 2B). At a nominal significance level of 0.01, the F_1, F_2, F_S and F_3 tests detect 295, 348, 333, and 252 genes respectively. The volcano plot of these F tests is shown in Figure 3D.

To study the false positive rate and successful detection rate of the four tests in this experimental setting, we performed 10 simulations each having 1000 constant genes and 1000 differentially expressed genes based on the design and variance components estimates of this experiment. The simulations were similar to those in the previous section. The settings for the fixed effects μ, T_k and D_j were the same as the corresponding fixed effects of model (4.8). The settings of A_i and ϵ_{ij} were the same as before except that σ_g^2 was drawn randomly from the 15 600 estimates of residual variance of this data set. The settings for the random effect mouse, M_l, were sampled from mouse variance component estimates. The reference $R_{h(i,j)}$ effect was drawn randomly from distribution $N(0, 0.3^2)$.

The average numbers of true and false positives over 10 simulations by each F test at nominal significance level of 0.05 are shown in Table 3. The numbers of false positives are all close to expectation (50), indicating that average type I error is controlled at the specified level. The successful detection rate of each F test decreases as CV increases, especially for F_2 and F_3. Again, F_S shows an advantage over F_1. More importantly, it shows more advantage over F_2 and F_3 at large CVs than observed from the microarray simulation without biological replication in Table 2, indicating that when biological variation

Table 3. *Average number of true and false positives identified by each F test in 10 simulations of model (4.9). Significance level is nominal 0.05. The total number of genes is 2000, with 1000 constant and 1000 differentially expressed. CV_r, average CV of the residual variance; CV_m, average CV of the mouse variances; TP, true positives; FP, false positives*

	$CV_r = 1.0$ $CV_m = 2.2$		$CV_r = 1.6$ $CV_m = 3.3$		$CV_r = 4.7$ $CV_m = 7.9$	
	TP	FP	TP	FP	TP	FP
F_1	789	41	758	43	666	38
F_2	806	40	751	37	493	42
F_S	796	41	764	43	671	38
F_3	742	39	658	41	282	48

Table 4. *Average number of true and false positives identified by each F_s, B, Regularized-t and SAM tests in 10 simulations as described in Section 4.4. Significance level is nominal 0.05. The total number of genes is 2000, with 1000 constant and 1000 differentially expressed. CV_r, average CV of the residual variance; CV_m, average CV of the mouse variances; TP, true positives; FP, false positives*

	$CV_r = 1.0$ $CV_m = 2.2$		$CV_r = 1.7$ $CV_m = 3.2$		$CV_r = 4.5$ $CV_m = 7.8$	
	TP	FP	TP	FP	TP	FP
F_S	591	42	563	41	500	41
B	619	51	574	54	527	92
R-t	583	46	553	45	493	44
SAM	212	0.2	157	0.7	203	2.5

is included in the computation of the error variance, F_S could be advantageous. The successful detection rate comparison among all four F tests against the treatment effect is shown in Figure 4 (D to F). The results from each of the 10 simulations are shown in Supplemental Table 1D–F.

4.4 *Comparison between F_S and some other information-sharing statistics*

In order to evaluate F_S against other information-sharing statistics, we compared F_S with SAM (Storey and Tibshirani, 2003; Lönnstedt and Speed, 2002), and regularized t-statistics (Baldi and Long, 2001) using the simulations described in the microarray simulation case II. Because B and regularized t-statistics are not applicable to multiple group comparisons, we restricted the simulations to two groups. The SAM, B, and regularized t-source codes were incorporated into our analysis from their current implementations in *siggenes* at `http://www.bioconductor.org/`, *SMA* at `http://cran.r-project.org/src/contrib/PACKAGES`, and *hdarray* at `http://visitor.ics.uci.edu/genex/cybert/hdarray`, respectively. Significance levels for each statistic were established by permutation analysis as discussed in Section 3. The average number of true and false positives identified by each test is shown in Table 4. The individual simulation results are shown in supplemental Table 2. Figure 6 shows the comparison of the successful detection rates of these four statistics as the size of the treatment effect changes. The successful detection rates of F_S, B and regularized t are similar in most of the cases. The B statistic shows a slightly higher false positive rate than expected when the heterogeneity of variances is high ($CV_r = 4.47$ and $CV_m = 7.76$). The SAM statistic is conservative compared to the other three in identifying differentially expressed genes regardless of the heterogeneity of the variance components.

In order to relax the normality assumptions in the simulation model (4.9), we conducted 10 additional simulations with all parameters drawn from the estimates of the tumor microarray experiment. For each

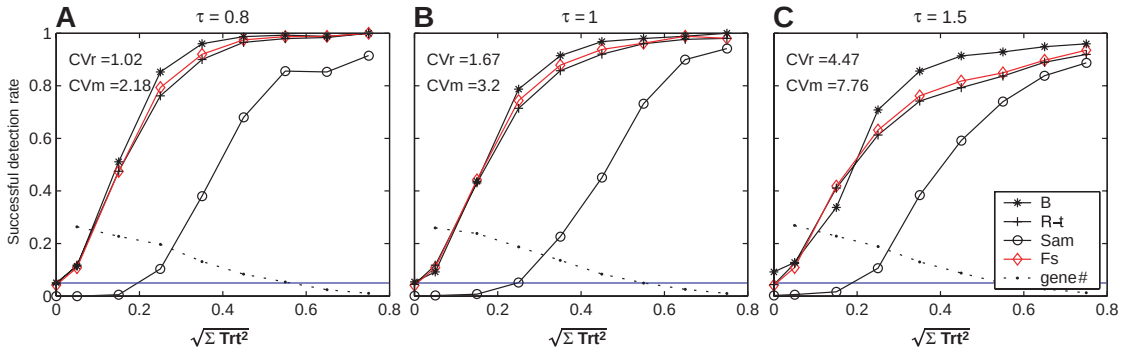

Fig. 6. The average successful detection rate of F_S, B, SAM, and regularized t (R-t) from 10 microarray simulations. Data were simulated according to model (4.9) with two variance components and normal effects as described in the text. Variance components were randomly drawn from the tumor data. Significance level is nominal 0.05 type I error rate. The variability of the variances across genes was controlled by τ (0.8, 1, 1.5) and is reflected by CV_r and CV_m. CV_r, CV of the residual variance component; CV_m, CV of the mouse variance component.

Table 5. *Average number of true and false positives identified by F_S, B, Regularized-t and SAM under per gene type I error or FDR control at 0.05. Ten simulations were conducted with all parameter drawn from the estimates of the tumor data. The total number of genes is 2000, with 1000 constant and 1000 differentially expressed. Average CV of the residual variance $CV_r = 2.3$; Average CV of the mouse variances $CV_m = 3.6$. TP, true positives; FP, false positives*

	Type I error 0.05		FDR 0.05	
	TP	FP	TP	FP
F_S	338	45	159	2.2
B	310	46	111	2.2
R-t	354	52	195	4.9
SAM	185	6.3	42	0.0

simulation, μ and R were randomly drawn from the estimates of these two parameters. The values of D at gene g were set to be the estimates of a randomly picked gene. The A, T, M, and ϵ were set in the same way as D except that only two of the three treatments were sampled. The sampled data depart dramatically from normal assumptions. The comparison of F_S with B, regularized t, and SAM is shown in Figure 7A, Table 5, and supplemental Table 3. Results from this simulation are similar to those obtained from previous simulations except that SAM appears to be slightly less conservative than before.

False discovery rate (FDR) is often used to address the multiple testing issues in microarray data analysis. FDR is the expected proportion of false positives among the rejected null hypotheses (Benjamini and Hochberg, 1995; Storey, 2002). When the average nominal is controlled at a specified level, a total number of false positives are controlled. A more powerful test will detect more true positive genes. Therefore, the FDR of the detected gene list is usually smaller for a more powerful test. On the other hand, if we control FDR at fixed level, a more powerful test will generally give a longer significant gene list. To compare F_S with B, SAM and regularized t-statistics we controlled FDR using Benjamini and Hochberg's adaptive control procedure (Benjamini and Hochberg, 2000), which is similar to Storey's positive FDR (Storey, 2002). Figure 7B and Table 5 show the comparison of F_S with B, regularized t, and SAM when FDR is controlled at 0.05. F_S and regularized t perform well. The B statistic is less powerful when the treatment effect is small but recovers quickly when the effect increases. SAM identifies

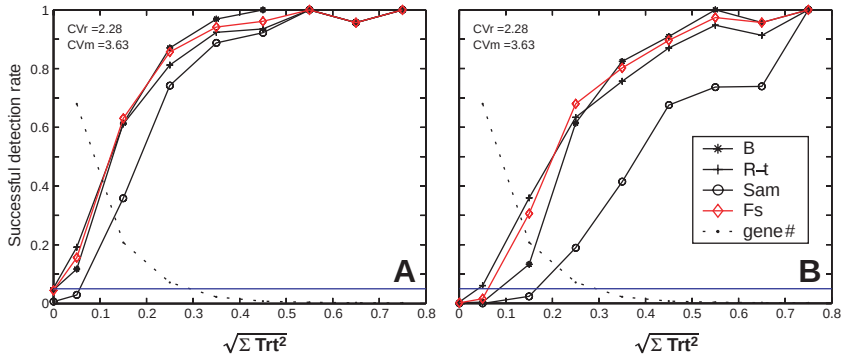

Fig. 7. Comparison between F_S, B, SAM, and regularized t (R-t) using data-based resampling simulations. Data were simulated according to model (4.9) where all parameters were drawn from the estimated values of the tumor data set. The average number of true positives of 10 simulations identified by each test is plotted. CV_r, CV of the residual variance component; CV_m, CV of the mouse variance component. A, nominal type I error is controlled at 0.05 for each gene. B, FDR is controlled at 0.05.

the smallest number of genes among the four statistics regardless of the size of the treatment effect. We also performed a larger simulation with 1000 differential genes and 9000 constant genes. The successful detection rates of the four tests are lower than those shown in Figure 7B but the relative comparison among methods is the same.

5. Discussion

Variance components in microarray experiments display varying degrees of heterogeneity, across experiments, across variance components, and across genes within a variance component (Cui and Churchill, 2003a). Assumptions of variance heterogeneity lead to the use of individual gene-specific tests, such as F_1, but these tests can suffer from low power due to small sample size per gene. On the other hand, the assumption of common variance leads to powerful tests, such as F_3, but at the risk of generating false positive and negative in the event that the common variance assumption is not true. A better approach is to use the tests based on variance estimates that are gene specific but combine information across many genes. We gain power by utilizing more information in the data but can also avoid bias.

In this paper, we apply James–Stein shrinkage to improve estimated variance components in a linear mixed model. We show that the resulting test statistic F_S performs better than the standard gene-specific test F_1 and that the improvement in successful detection rate can be substantial when the degrees of freedom are small, a common situation for microarray experiments. Compared with some other information-sharing statistics, such as B, SAM, and regularized t, F_S has comparable or better power in identifying differentially expressed genes. Moreover, F_S is more general as it can be applied to a wider range of experimental designs, i.e. not restricted to two-sample comparisons with single variance component.

By taking a shrinkage approach to improve variance estimation, we make only weak prior assumptions about the distribution of the variance components. Although the James–Stein shrinkage estimator was developed in the context of a normal model, it is the sampling distribution of the logarithm of the variance estimators, not the values of the variance themselves, that are assumed to be normal. Parametric Empirical Bayes methods require explicit distributional assumption on the true variances. In some simple settings, such as estimating a normal mean which has a normal prior, the Empirical Bayes approach and the

shrinkage approach lead to exactly the same estimator (Efron and Morris, 1973). But in this setting, the Empirical Bayes approach is complicated (Wright and Simon, 2003). Our proposed statistic, F_S, has an explicit expression and is computationally simple.

In summary, we have proposed a variation on the general mixed model testing strategy using shrinkage estimates of variance components to construct test statistics that are powerful and robust in respect to variance heterogeneity in gene expression data.

Supplemental tables, software and data sets cited in this paper are available at `http://www.jax.org/staff/churchill/labsite/pubs/index.html`.

ACKNOWLEDGMENTS

We would like to thank Hao Wu for software support, Qian Li for assistance with data handling, Ann Dorward at the Jackson Laboratory and John Quackenbush at TIGR for providing sample data sets. This research is supported by grants CA88327, HL66620, and HL55001 from the National Institute of Health.

REFERENCES

AFFYMETRIX (1999). *Affymetrix Microarray Suite User Guide*. Santa Clara, CA: Affymetrix.

BALDI, P. AND LONG, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519.

BENJAMINI, Y. AND HOCHBERG, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **85**, 289–300.

BENJAMINI, Y. AND HOCHBERG, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics* **25**, 60–83.

CALLOW, M. J., DUDOIT, S., GONG, E. L., SPEED, T. P. AND RUBIN, E. M. (2000). Microarray expression profiling identifies genes with altered expression in hdl-deficient mice. *Genome Research* **10**, 2022–2029.

CHURCHILL, G. A. (2002). Fundamentals of experimental design for cDNA microarrays. *Nature Genetics* **32**, 490–495.

CUI, X. AND CHURCHILL, G. A. (2003a). How many mice and how many arrays? Replication of cDNA microarray experiments. In Lin, S. M. and Allred, E. T. (eds), *Methods of Microarray Data Analysis III*, New York: Kluwer.

CUI, X. AND CHURCHILL, G. A. (2003b). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biology* **4**.

CUI, X., KERR, M. K. AND CHURCHILL, G. A. (2003). Transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology* **2**.

DUDOIT, S., SHAFFER, J. P. AND BOLDRICK, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science* **18**, 71–103.

EFRON, B. AND MORRIS, C. (1973). Stein's estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association* **68**, 117–130.

GHOSH, M., HWANG, J. AND TSUI, K. (1984). Construction of improved estimators in multiparameter estimation for continuous exponential familes. *Journal of Multivariate Analysis* **14**, 212–220.

KENDZIORSKI, C. M., NEWTON, M. A., LAN, H. AND GOULD, M. N. (2003). On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. `http://www.stat.wisc.edu/~newton/research/arrays.html`.

KERR, M. K., MARTIN, M. AND CHURCHILL, G. A. (2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

LINDLEY, D. V. (1962). Discussion of Professor Stein's paper, 'Confidence sets for the mean of a multivariate normal distribution'. *Journal of the Royal Statistical Society Series B* **24**, 265–296.

LITTELL, R. C., MILLIKEN, G., STROUP, W. W. AND WOLFINGER, R. D. (1996). *SAS System for Mixed Models*. Cary, NC: SAS Institute Inc.

LÖNNSTEDT, I. AND SPEED, T. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

MCLEAN, R. A., SANDERS, W. L. AND STROUP, W. (1991). A unified approach to mixed linear models. *The American Statistician* **45**, 54–64.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. AND AHLQUIST, P. (2003). Detecting differential gene expression with a semiparametric hierarchical mixture method. `http://www.stat.wisc.edu/~newton/papers/abstracts/tr1074a.html`.

SCHENA, M., SHALON, D., DAVIS, R. AND BROWN, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

SEARLE, S., CASELLA, G. AND MCCULLOCH, C. (1992). *Variance Components*. New York: Wiley.

SMYTH, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

STOREY, J. AND TIBSHIRANI, R. (2003). SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (eds), *The Analysis of Gene Expression Data: Methods and Software*, New York: Springer.

WITKOVSKY, V. (2002). Matlab algorithm mixed.m for solving Henderson's mixed model equations. `http://www.mathpreprints.com`.

WOLFINGER, R. D., GIBSON, G., WOLFINGER, E. D., BENNETT, L., HAMADEH, H., BUSHEL, P., AFSHARI, C. AND PAULES, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* **8**, 625–637.

WRIGHT, G. W. AND SIMON, R. M. (2003). A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* **19**, 2448–2455.

WU, H., KERR, M. K., CUI, X. AND CHURCHILL, G. A. (2003). Maanova: a software package for the analysis of spotted cDNA microarray experiments. In Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. L. (eds), *The Analysis of Gene Expression Data: Methods and Software*, New York: Springer.

[*Received October 16, 2003; first revision March 22, 2004; second revision May 26, 2004; accepted for publication June 1, 2004*]