A random variance model for detection of differential gene expression in small microarray experiments

George W. Wright* and Richard M. Simon

National Cancer Institute Biometric Research Branch, 6130 Executive Blvd., MSC 7434, Bethesda, MD 20892-7434, USA

ABSTRACT

Motivation: Microarray techniques provide a valuable way of characterizing the molecular nature of disease. Unfortunately expense and limited specimen availability often lead to studies with small sample sizes. This makes accurate estimation of variability difficult, since variance estimates made on a gene by gene basis will have few degrees of freedom, and the assumption that all genes share equal variance is unlikely to be true.

Results: We propose a model by which the within gene variances are drawn from an inverse gamma distribution, whose parameters are estimated across all genes. This results in a test statistic that is a minor variation of those used in standard linear models. We demonstrate that the model assumptions are valid on experimental data, and that the model has more power than standard tests to pick up large changes in expression, while not increasing the rate of false positives.

Availability: This method is incorporated into BRB-ArrayTools version 3.0 (http://linus.nci.nih.gov/BRB-ArrayTools.html).

Supplementary material: ftp://linus.nci.nih.gov/pub/techreport/RVM_supplement.pdf

Contact: wrightge@mail.nih.gov

1 INTRODUCTION

Microarray technology allows a scientist to view the expression of thousands of genes from an experimental sample simultaneously. By observing changes in expression levels across multiple samples, it is possible to generate and test a multitude of hypotheses relating gene expression to other characteristics of the samples. One of the primary goals of microarray analysis is to identify genes whose expression level varies between different classes of samples. Testing whether a single observed quantity varies across different classes of observations is a problem that is well understood, but the volume of information available on each sample creates new analysis challenges.

In modeling gene expression we use a linear model framework. This model is very versatile and can be applied to a large number of different experimental designs. Further, other familiar tests such as the t-test, paired t-test and F-test, and Analysis of Variance (ANOVA) are actually special cases of the more general linear model. Kerr and Churchill (2001) suggested the application of ANOVA to the red and green log intensity values as an alternative to the use of log ratios. But even at the level of log ratio data, use of an ANOVA-type model can provide a very powerful framework for evaluating multiple competing factors that might influence gene expression.

One question that arises when fitting a linear model to microarray data is how to estimate residual error. One approach is to form a single linear model that pools the residual sum of squares across all genes into a single variance estimate. This method makes the assumption that once all of the factors included in the model have been taken into account, all genes are equally variable. In practice we find that this is not an accurate assumption. It is impossible to take into account all sources of biological variation in a single linear model. Those factors that have not been taken into account will vary from gene to gene and be incorporated into the residual, leading to large differences in the residual variance across genes. A second approach is to form a separate linear model for each gene and estimate the gene-specific residual variance using only the data from that gene. This method has the disadvantage that there is no information shared between genes. If the sample size is small, there may be very few degrees of freedom available to estimate the residual variance, leading to statistical tests with low power.

We propose a hybrid approach in which it is assumed that the variance of the residuals change from gene to gene, but represent random selections from a single distribution. By observing the residual sum of squares within each gene, we estimate the form of this distribution. Then for an individual gene we adjust the observed residual sum of squares in light of the distribution. By sharing the variance estimate across multiple genes, we can form a better estimate for the true

---

*To whom correspondence should be addressed.

residual variance of a given gene and effectively boost the residual degrees of freedom. The test statistics produced by this shared variance model are very similar in form to those for standard linear models, meaning that this model can be easily implemented using standard statistical packages.

2 MODEL FORMULATION

We will denote by y_{ij} the normalized expression values for sample i and gene j. How the expression values are actually formulated will depend on the application. If we wish to follow the methodology of Kerr and Churchill, then the y_{ij} will represent normalized log intensity values of a single channel. Alternatively, the y_{ij} can represent the normalized log-ratios of a 2-color array, or normalized log signal for Affymetrix GeneChip^{TM} Arrays.

The type of model we wish to consider is the following:

$$y_{ij} = x_i'\beta_j + \varepsilon_{ij}. \qquad (1)$$

The x_is are vectors of design variables specific to the samples, the β_js are vectors of unknown coefficients that are specific to particular genes, and the ε_{ij}s are unobserved residuals with mean 0 and unknown variance. The x_is represent the characteristics of the samples that we wish to correlate with gene expression. The residuals ε_{ij} encompass both the array measurement error as well as biological variability of the gene across samples.

In practice, the investigator is often interested in identifying differences in gene expression between two or more classes, in which case the x_i' will be vectors of class indicators. However, this model is not restricted just to class comparison problems. For example, it may be reasonable to consider fitting a regression of gene expression against time, in which case x_i could be the continuous variable representing elapsed time.

Let n denote the number of samples, m denote the number of genes, and k denote the dimension of each x_i and β_j. We use X to denote the $n \times k$ dimensional design matrix whose columns are x_i, and make the additional simplifying assumption that X is of full rank. For X not of full rank it is always possible to reparametrize the model into a lower dimensional model with X of full rank.

The crux of our model is in the handling of the ε_{ij}s. We assume that for each gene j,

$$\varepsilon_{ij} \sim \text{NID}(0, \sigma_j^2) \qquad (2)$$

with the σ_j^2s being random variables themselves with an inverse Gamma distribution. That is

$$\sigma^{-2} \sim G(x; a, b) \equiv \frac{x^{a-1}\exp(-x/b)}{\Gamma(a)b^a} \qquad (3)$$

for some unknown parameters a and b. This choice of the Inverse-Gamma distribution as a prior distribution of variance is a standard choice in Bayesian analysis due to its computational convenience, but as we will show later, it models

true variance structure of microarray data surprisingly well. We will refer to the linear model described above as the Randomized Variance Model, or RVM.

This use of a distribution for σ is similar to a model suggested by Baldi and Long (2001) with some exceptions. First, our analysis is frequentist in perspective rather than Bayesian. Second, we consider the more general case of linear models while Baldi and Long concentrated on a 2-class distinction. Third, we make the assumption that the a and b parameters in the distribution of σ_j^2, are the same for all genes and show how they can be estimated from the data. Finally, we make no assumptions about a prior distribution for β_j. This was a conscious decision. The distribution of β depends on how the genes on the array were chosen. A significant proportion of the genes on the array are likely to have constant expression across samples. Any prior distribution on the mean structure would have to include a singular component at zero, and any hypothesis tests on β would be heavily dependent on the choice of prior.

3 HYPOTHESIS TESTING FOR β

Hypothesis tests for β_j will be performed on a gene by gene basis. Therefore in this section and the next, we will suppress the j subscript and concern ourselves with tests of a single gene. In these sections we will also make the temporary assumption that the parameters a and b are known constants. In a later section we will show how to use the full set of data to estimate values for a and b and argue that the substitution of these estimates for the true values will not invalidate the tests. For the sake of brevity, mathematical proofs of all claims have been omitted, but are available in the supplementary materials.

Following the standard framework of hypotheses on linear models, we wish to test the hypothesis $H_0 : \beta \in \omega$, where ω is a linear subspace of $\mathbb{R}^k$. We will use $r \equiv k - \dim(\omega)$ to denote the number of linear constraints imposed by ω. In class comparison problems, the subspace is likely to be that all of the components of β are equal (indicating that all classes have the same average expression). In that case r will be equal to the number of classes minus 1. In regression problems, the constraint will be that some of the components of β are equal to zero, meaning that change in the respective component of x_i has no effect on gene expression.

In standard linear models in which there is no distribution for σ, the maximum likelihood estimate for β over $\mathbb{R}^k$ is given by

$$\widehat{\beta} \equiv (X'X)^{-1}X'y \qquad (4)$$

and the maximum-likelihood estimate under ω, is

$$\widehat{\widehat{\beta}} \equiv (X_\omega'X_\omega)^{-1}X_\omega'y, \qquad (5)$$

where X_ω represents the design matrix X projected into the subspace ω.

To test the hypothesis $H_0 : \beta \in \omega$ against the alternative $H_1 : \beta \in \mathbb{R}^k$, we would consider the respective sum of square

residuals,

$$\widehat{SS} = \|y - X'\widehat{\beta}\|^2 \quad \text{and} \quad \widehat{\widehat{SS}} = \|y - X\widehat{\widehat{\beta}}\|^2 \quad (6)$$

and then the likelihood ratio test statistic would be

$$\mathcal{F} = \frac{n-k}{r} \frac{\widehat{\widehat{SS}} - \widehat{SS}}{\widehat{SS}}, \quad (7)$$

which under H_0 has an F distribution with r and $n-k$ degrees of freedom.

If we repeat this type of analysis under the RVM assumption of a distribution for σ^2, expressed in Equation (3) we obtain the same maximum likelihood estimates, $\widehat{\beta}$ and $\widehat{\widehat{\beta}}$, for β within $\mathbb{R}^k$ and ω, respectively, but there is a change in the likelihood ratio test statistic. The residual sum of squares in the denominator of the $\mathcal{F}$ is replaced with

$$\widetilde{SS} = \widehat{SS} + 2b^{-1} \quad (8)$$

and the number of degrees of freedom in the denominator changes from $n-k$ to $n-k+2a$. Thus the adjusted statistic becomes

$$\widetilde{\mathcal{F}} = \frac{n-k+2a}{r} \left(\frac{\widehat{\widehat{SS}} - \widehat{SS}}{\widetilde{SS}} \right), \quad (9)$$

which under H_0 will have F distribution with r and $n-k+2a$ degrees of freedom (see supplementary material).

To interpret this change, it is useful to consider

$$\widehat{\sigma}^2 = \frac{\widehat{SS}}{n-k} \quad \text{and} \quad \widetilde{\sigma}^2 = \frac{\widetilde{SS}}{n-k+2a}, \quad (10)$$

which are the maximum likelihood estimates for σ^2 under the standard linear model and under RVM. After some algebra we find that

$$\widetilde{\sigma}^2 = \frac{(n-k)\widehat{\sigma}^2 + 2a(ab)^{-1}}{(n-k)+2a}. \quad (11)$$

ab is the mean value for σ^{-2} under its distribution, so $(ab)^{-1}$ could be thought of as an estimate of σ^2 based on the shared variance distribution, and $\widetilde{\sigma}^2$ as a weighted average of this and the sample variance for the specific gene. The degree to which each of these is weighted depends on the number of samples and the parameter a. The larger the sample size, the more confident we will be in $\widehat{\sigma}^2$, our sample estimate of σ^2. On the other hand, large values for a will indicate a highly peaked inverse gamma distribution, making it more likely that the true σ^2 is close to $(ab)^{-1}$. The increase in the degrees of freedom from $n-k$ to $n-k+2a$ indicates the additional information about σ^2 provided by the distribution. Again, a larger value for a represents a more informative distribution and so more certainty about $\widetilde{\sigma}^2$ as an estimate of σ^2, that in turn translates to more degrees of freedom.

4 APPLICATION TO t AND F TESTS AND ANOVA

A very common problem in microarray analysis is that of determining genes that are differentially expressed between two or more tissue varieties. To represent this in our framework, we set k equal to the number of groups, and set the pth component X_j to an indicator of membership of sample j in variety p. Just as our hypothesis tests for the linear model under the RVM assumption was similar to the standard tests for linear models, so also are our tests for differences across varieties of similar form to the t- and F-tests that are often used.

For testing between two varieties, with sample means $\widehat{\mu}_1$, $\widehat{\mu}_2$; sample variances $\widehat{\sigma}_1^2$, $\widehat{\sigma}_2^2$; and sample sizes n_1, n_2 the standard test statistic is

$$t = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\widehat{\sigma}_{\text{pooled}}\sqrt{1/n_1 + 1/n_2}}, \quad (12)$$

where

$$\widehat{\sigma}_{\text{pooled}}^2 = \frac{(n_1 - 1)\widehat{\sigma}_1^2 + (n_2 - 1)\widehat{\sigma}_2^2}{n_1 + n_2 - 2} \quad (13)$$

and this has a t-distribution with $n-2$ degrees of freedom.

In the case of the RVM model, we would consider

$$\widetilde{t} = \frac{\widehat{\mu}_1 - \widehat{\mu}_2}{\widetilde{\sigma}\sqrt{1/n_1 + 1/n_2}}, \quad (14)$$

where

$$\widetilde{\sigma}^2 = \frac{(n-2)\widehat{\sigma}_{\text{pooled}}^2 + 2b^{-1}}{(n-2)+2a}. \quad (15)$$

This would still have a t-distribution, but the degress of freedom increase to $n-2+2a$.

For testing $k > 2$ varieties, it is common to consider the following F-test statistic involving mean sums of squares (MSS):

$$\mathcal{F} = \frac{\text{MSS(between varieties)}}{\text{MSS(within varieties)}}, \quad (16)$$

which under the null hypothesis has an F-distribution with $k-1$ and $n-k$ degrees of freedom. The RVM model in this case will be identical with the exception that

$$\widetilde{\text{MSS}}(\text{within varieties})$$
$$\equiv \frac{(n-k)\text{MSS(within varieties)} + 2b^{-1}}{(n-k)+2a} \quad (17)$$

replaces MSS(within varieties), and the degrees of freedom for the denominator of the F-statistic becomes $n-k+2a$, instead of $n-k$.

For more complicated ANOVA models, such as Latin square designs, interaction effects, or other fixed effect models, the tests in the case of RVM are identical to those in the classical ANOVA case, with the exceptions that the residual mean sum of squares is increased by the amount $2b^{-1}$ and the residual degrees of freedom are increased by $2a$.

5 ESTIMATION OF a, b

In the previous sections we assumed complete knowledge of a and b. In practice a and b must be estimated. However, unlike σ_j and β_j, the values a and b are assumed constant across genes, so it is possible to use the data from all genes to form an estimate. This means that our estimates of these parameters can be based on thousands of data points and so can in principle be extremely accurate. For this reason we can substitute estimated values for a and b in the above tests without invalidating them.

To estimate a and b we consider $\widehat{\sigma}_j^2 = \widehat{SS}_j/(n-k)$ the empirical estimates of σ_j^2. It can be shown (see supplemental material) that under the assumptions of RVM,

$$ab(\widehat{\sigma}_j^2) \sim F_{(n-k),2a}. \tag{18}$$

Therefore, we can estimate the parameters a and b by fitting an F-distribution to the observed $\widehat{\sigma}_j^2$.

Although computationally simpler, we recommend against using a method of moments estimator. The higher moments of the F-distribution are infinite, and so such an estimator will be unstable. Instead we recommend numerically maximizing the likelihood under the F-distribution with respect to a and b. This provides more accurate estimates for a and b, and while computationally intensive is not overly so, since this estimation needs to be performed only once for the entire data set.

In Sections 3 and 4, all analyses are performed on a gene by gene basis, so any correlations between genes does not affect the results. However, the estimation of a and b are performed across genes and so correlations in individual variance estimates will influence estimates of a and b. Still, provided that the correlation is not extreme, maximum-likelihood estimates are consistent with the correlation acting only to increase the variability of the estimate.

6 SIMULATION RESULTS

To evaluate the tests presented, we generated 2000 simulated data sets, of the expressions of 6000 genes from 2 groups of 5 samples each. For each gene, a random σ_j^2 value was chosen from an inverse gamma distribution with $a = 3$, $b = 1$. The σ_j^2 values were chosen separately for each data set. Independent, normally distributed random numbers with mean 0 and variances σ_j^2 were then assigned as gene expression values. We chose the values $a = 3$, $b = 1$ because estimates of a and b in actual microarray data were found to be in the vicinity of these values. For 3000 of these genes, an amount between 0.1 and 2.0 was added to the expression of samples from the first group (150 genes at every 0.1 units between 0.1 and 2.0). These 3000 genes would represent genes that were truly differentially expressed between the two groups. The remaining 3000 genes would represent non-differentially expressed genes. This procedure was then repeated for an additional 2000 data sets, with 10 samples per group.

Table 1. Parameter estimates in simulation

		Method of moments	Maximum likelihood
5 samples per group	a	3.12 ± 0.57	3.00 ± 0.086
	b	1.01 ± 0.33	1.00 ± 0.034
	$(ab)^{-1}$	0.333 ± 0.004	0.333 ± 0.004
10 samples per group	a	3.01 ± 0.141	3.00 ± 0.068
	b	1.00 ± 0.051	1.00 ± 0.026
	$(ab)^{-1}$	0.333 ± 0.003	0.333 ± 0.003

Table 2. Observed type 1 error in simulation

	Variance model	$P < 0.01$	$P < 0.005$	$P < 0.001$
5 samples per group	Individual	0.00998	0.00498	0.00098
	Pooled	0.01963	0.01377	0.00687
	Random	0.00999	0.00499	0.00100
10 samples per group	Individual	0.01003	0.00496	0.00099
	Pooled	0.0336	0.0255	0.0145
	Random	0.00996	0.00496	0.00099

For each gene in each data set, we tested for mean differences between sample groups according to three different one-sided t-tests. For the first t-test, a pooled variance estimate across the 6000 genes included in the data set was used, and the t-statistic was assumed to have infinite degrees of freedom. For the second t-test, the variance estimate was made separately within each gene, and $n - 2$ degrees of freedom were used. For the final t-test, we estimated a and b from the observed residual variances for the data set and then used the modified t-test presented in equation (14).

Table 1 shows the mean and standard deviations of the estimates for a and b over the 4000 simulations, first by method of moments and then by maximum likelihood. We observe that both estimates were unbiased, but the maximum likelihood estimate had less variability than the method of moments estimate. We also observed that there was a strong inverse correlation between our estimates of a and of b, such that $(ab)^{-1}$ was estimated with extreme accuracy even within the method of moments estimation. In light of equation (11), the estimation of b is important only insofar as ab^{-1} is estimated accurately. Therefore, any error in our statistic through inaccurate estimation of a and b will really be through the mis-estimation of a.

Table 2 shows the proportion of false positives (i.e. non-differentially expressed genes that the test declared to be differentially expressed) for each of the tests at different levels. We observe that both the RVM test, and the t-test with the
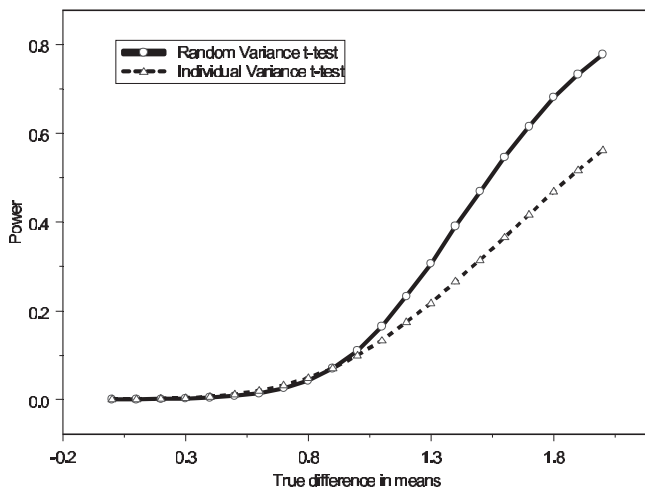
Fig. 1. Detection power for $P < 0.001$ level test by true mean difference for 5 samples/group in simulation.
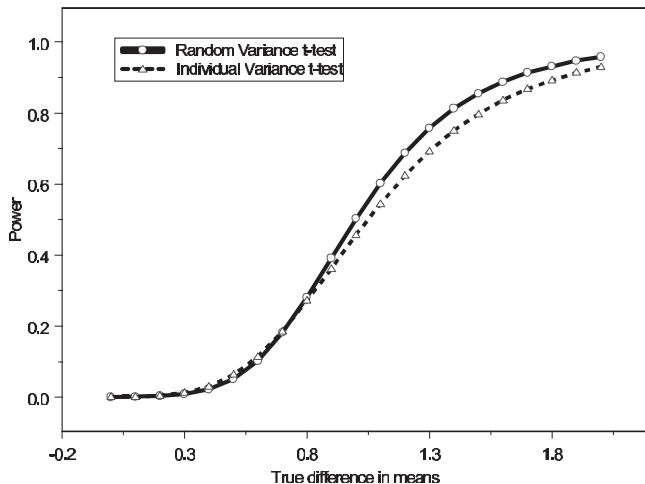

Fig. 2. Detection power for $P < 0.001$ level test by true mean difference for 10 samples/group in simulation.

variances estimated within individual genes, appear to have the proper false-positive rates. However, using a single value for all variances results in an excessive false-positive rate. This is due to the fact that highly variable genes may exhibit large fold-differences even when they are not actually differentially expressed. When these fold differences are then divided by the relatively small pooled residual variance estimate, they will appear significant.

Figures 1 and 2 show the power for detecting true differences between groups according to the RVM t-test and the standard t-test in which the variances were computed individually for each gene. The X-axis represents the true difference between the means of the two groups, and the Y-axis shows the proportion of such genes that were found to be significant with

$P < 0.001$. Power estimates for the test in which the variance was pooled across genes were not included since this test failed to control type one error and so was not comparable to the other tests. We observe that the RVM t-test generally performed better than the standard t-test, particularly when there was a large difference between the two groups and a very small p-value was required for significance. The difference was less prominent when there were 10 samples in each group than when there were 5 samples in each group. This is due to the fact that as sample size increases, the RVM test statistic approaches the standard t-statistic.

7 EXPERIMENTAL RESULTS

Although the RVM test performed well in simulations in which the model assumptions held exactly, we needed to determine how well the model assumptions hold in actual data. To check this, we looked at two different sets of data. The first set (referred to as the DLBCL data set) came from Rosenwald *et al*. (2002) and included 7399 genes measured on 274 Lymphoma samples which were divided into three groups (ABC, GCB, Type 3) according to their gene expression values. The second set (referred to as the BRCA data set) from Hedenfalk *et al*. (2001) included 3226 genes measured on 22 breast cancer samples, which were also divided into three classes (Sporadic, BRCA-1 and BRCA-2) according to their mutational status. These data sets can be downloaded from (http://llmpp.nih.gov/DLBCL) and (http://research.nhgri.nih.gov/microarray/ NEJM_Supplement).

The primary assumption of the RVM model is that the within group variance is distributed according to an inverse gamma distribution. This choice of distribution was made purely for computational convenience, and there is no intrinsic reason why the gene variances should follow this distribution. Since we are unable to observe the true within group variances directly to determine whether they were inverse gamma distributed, we instead observed the sample variances to determine the extent to which they could be fitted to the F-distribution described in equation (18), as should be the case if the true variances followed the inverse gamma distribution. We found that in both data sets, the distribution of the observed residuals was virtually indistinguishable from the corresponding F-distributions with fitted a and b (Figs 3 and 4). Other data sets have also been investigated (data not shown) and in each case the F-distribution appears to be a very good fit to the observed residuals.

In order to check the relative type 1 and type 2 errors of the various tests, we restricted ourselves to the DLBCL data, and in particular the 83 ABC patients and 134 GCB patients. With such a large sample size we can determine with great accuracy which genes were truly differentially expressed between these two groups and the size of the difference. By selecting small subsets of these groups, we can determine the extent to which the various methods can detect
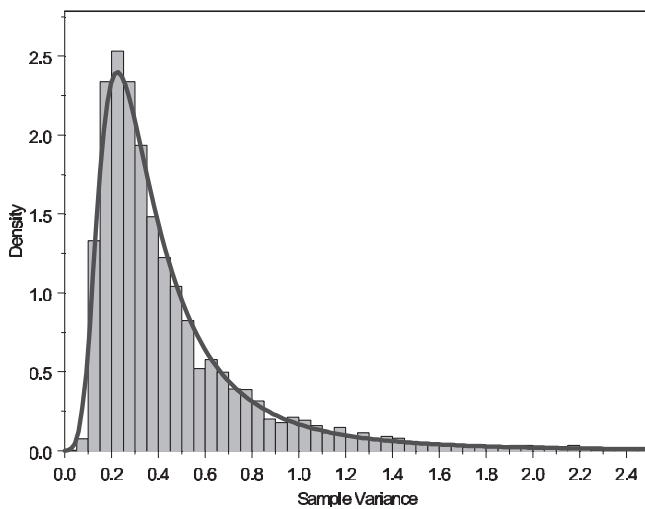
Fig. 3. Histogram of sample variances within DLBCL data set. Line represents theoretical sample variance distribution according to Equation (16).
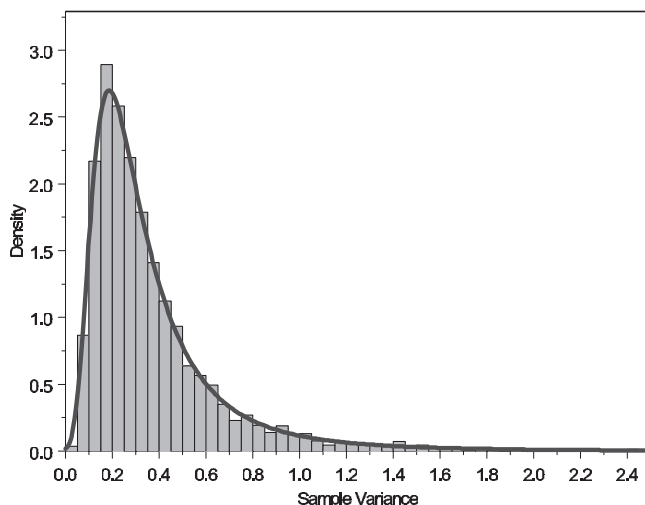

Fig. 4. Histogram of sample variances within BRCA data set. Line represents theoretical sample variance distribution according to Equation (16).

Table 3. Parameter estimates in experiemental data

		Method of moments	Maximum likelihood
5 samples	a	2.67 ± 0.843	2.45 ± 0.355
per group	b	1.53 ± 0.681	1.45 ± 0.374
	$(ab)^{-1}$	0.282 ± 0.041	0.283 ± 0.040
10 samples	a	2.51 ± 0.318	2.45 ± 0.216
per group	b	1.43 ± 0.283	1.45 ± 0.248
	$(ab)^{-1}$	0.288 ± 0.029	0.288 ± 0.029

the differentially expressed genes in small samples, while controlling the number of non-differentially expressed genes found to be significant. We repeatedly took sub-samples of size 10 (5 GCB, 5 ABC) and 20 (10 GCB, 10 ABC) and calculated for each gene the p-value for the difference between the samples observed for the two sets. This was repeated 2000 times, and the proportion of times the gene was found to be significant at $p < \{0.001, 0.005, 0.01\}$ was recorded.

To avoid having subsets in which no data for a gene was available, we excluded all those that were missing values in more than 20% of the total data. This resulted in 6027 genes which were used to estimate parameters a and b for each

sub-sampling of the data. The results of this estimation are indicated in Table 3. We notice that the variability of the estimates was larger in the experimental data than it was in simulation. This is likely due to the fact that there is correlation between genes, which will act to increase the variability of the estimates. Additionally, the experimental data had some missing values, which may have also contributed to the increased variability. However, even with the increased variability, estimates of ab^{-1} remained very consistent, while errors in the estimates of a would generally be less than 0.5 and so would alter the number of degrees of freedom of the final statistic by less than 1.

A t-test was performed on the full data set (83 ABC, 134 GCB patients) to determine those genes that were truly differentially expressed between the two groups. The 1621 genes that had a p-value less than 0.001 were declared to be differentially expressed and used to compute power, while the 2916 genes that had a p-value greater than 0.05 were declared non-differentially expressed and used to compute the false-positive rates. The remaining 1490 genes were on the borderline between differentially expressed and non-differentially expressed, and so were not used in either the power or false positive rate calculations. In order to consider the effect of true fold-difference in our ability to detect genes, we calculated for each gene the mean log ratio within the ABC and GCB subtypes on the total data , then used the difference between these two means rounded to the nearest 0.1.

The results of this analysis are similar to what we found in the simulation. The false positive rates of the t-test and random variance tests are very close to the desired values in both the 10-sample and 20-sample subsets, while the pooled variance model had a much larger rate of false-positives than was desired, indicating that this test was unacceptable (Table 4). Again we also observe that the RVM model performs much better than the t-test in identifying those genes with large true fold differences between classes (Figs 5 and 6).

8 DISCUSSION

Analysis of microarray data clearly indicates that all genes are not equally variant within a sample. Therefore, some method

Table 4. Observed type 1 error in experimental data

	Variance model	$P < 0.01$	$P < 0.005$	$P < 0.001$
5 samples per group	Individual	0.0095	0.0047	0.00091
	Pooled	0.0212	0.0150	0.00745
	Random	0.0095	0.0046	0.00092
10 samples per group	Individual	0.0102	0.0051	0.000998
	Pooled	0.0213	0.0152	0.00749
	Random	0.0103	0.0051	0.00102
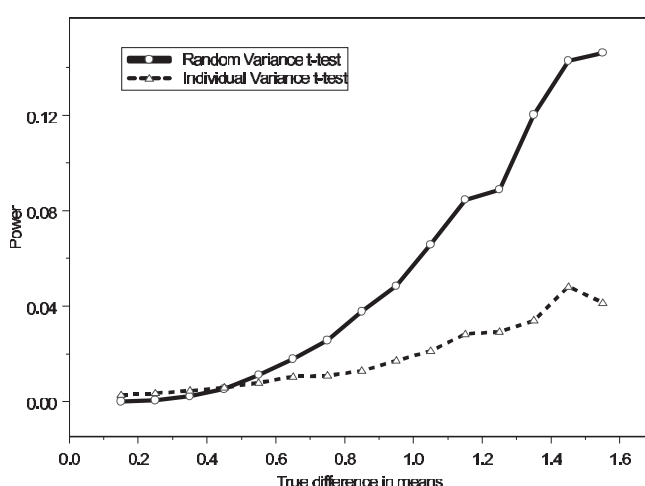
Fig. 5. Detection power for $P < 0.001$ level test by true mean difference for 5 samples/group in experimental data.
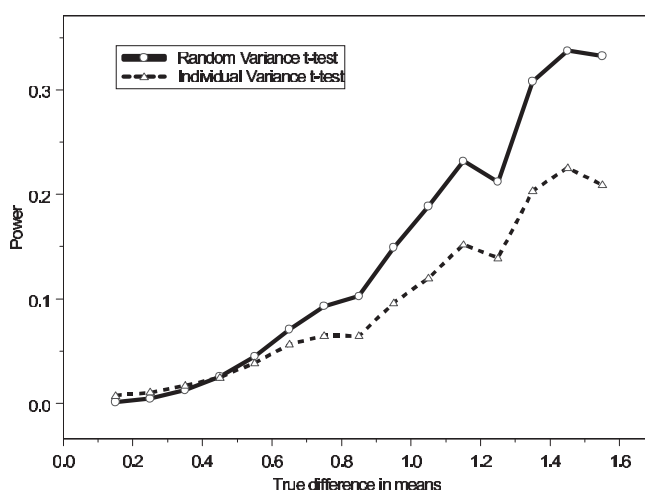
Fig. 6. Detection power for $P < 0.001$ level test by true mean difference for 10 samples/group in experimental data.

of estimating an individual gene's variability must be taken into account when determining statistical significance. While directly estimating the variance within each gene works well for large samples, when the sample size is small such an estimate can be imprecise, resulting in a test of low power. We have proposed a model through which information from the entire set of genes is used to influence the variance estimate of a single gene, while still allowing the differences in gene variances. This model can be applied to any linear regression framework and so is very versatile.

This model has the advantage that it is very simple to implement, requiring only slight modifications to tests that are already available in most statistical packages. We have demonstrated that the underlying assumption of our model, that the within gene variances are distributed according to an inverse gamma distribution, appears to be correct in actual experimental data. As a result, our model correctly controls type 1 error in these experimental situations. Our method has greater power than does the standard *t*-test in detecting genes with large fold differences. Since these are the genes that are most likely to be of biological interest, we feel that this new method provides an easily implemented improvement to standard techniques where genes are analyzed individually.

We have not addressed the issue of interpretation of the RVM *p*-values in light of the large number of comparisons possible. We recommend using the RVM *p*-values in the context of a larger framework that controls the number or proportion of false discoveries, such as the method of Benjamini and Hochberg (1995), or the multivariate permutation tests of Korn *et al.* (2003). The latter method is permutation based and the RVM *p*-values can be used while avoiding the assumption that the errors have Gaussian distributions. The RVM *t*-statistics can also be used in the SAM procedure (Tusher *et al.*, 2001) for controlling the false discovery rate as an alternative to the somewhat *ad hoc* relative difference statistic. The relative difference is the difference in class means for a gene divided by a constant plus the within-gene estimate of residual standard deviation. Our Equations (14) and (11) suggest an alternative with a stronger theoretical basis, but the permutational framework of SAM can still be used.

ACKNOWLEDGEMENTS

REFERENCES

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. *J. R. Stat. Soc.*, Ser. B, **57**, 289–300.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Kallioniemi,O. *et al*. (2001) Gene expression profiles of hereditary breast cancer. *N. Engl. J. Med.*, **344**, 539–548.

Kerr,K. and Churchill,G. (2001) Statistical design and the design of gene expression microarrays. *Biostatistics*, **2**, 183–201.

Korn,E.L., Troendle,J.F., McShane,L.M. and Simon,R. (2003) Controlling the number of false discoveries: application to high-dimensional genomic data. *J. Stat. Planning and Inference* (in press)

Menezes,R. (2003) Heirarchical modeling to handle heteroscedacisity in microarray data. *Presentation at 3-day Workshop on Statistical Analysis of Gene Expression Data*, Richardson,S. and Brown,P. Organizers, 11–14 July 2003, Wye, UK.

Rosenwald,A., Wright,G., Chan,W.C., Connors,J.M., Campo,E., Fisher,R.I., Gascoyne,R.D., Muller-Hermelink,H.K., Smeland,E.B., Giltnane,J.M. *et al*. (2002) The use ofmolecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS* **98**, 5116–5121.