A Conditional *t* Suite of Tests for Identifying Differentially Expressed Genes in a DNA Microarray Experiment with Little Replication

Dhammika AMARATUNGA and Javier CABRERA

Abstract

A fundamental experiment in functional genomics research is the comparison of two groups of microarray data to determine which genes are expressed differentially between the two (e.g., diseased versus normal tissue). Although these data could be simply analyzed gene by gene with a series of individual *t* tests, it should be possible to increase the power of the procedure substantially by borrowing strength across genes. We show how this can be realized via a model, which posits minimal distributional assumptions, and a conditional *t* suite of tests.

KEY WORDS: Borrowing strength; Ct; Critical curve; Resampling; SAM; *t* test; Target estimation; Wedge effect.

1. INTRODUCTION

An important objective of modern biological research is to understand how an organism's genome's behavior is altered in response to certain specific situations, such as when the organism is affected by a disease or is administered a drug. In particular, it is of interest to determine which of the organism's genes express differently in such situations.

DNA microarray technology (Schena 1999) has emerged as the primary screening tool with which the expression patterns of the thousands of genes that constitute an organism's genome could be explored in high throughput with this goal in mind. This is undertaken with the understanding that, as this is merely a screening stage, it is adequate to generally uncover evidence of differential expression without necessarily

Dhammika Amaratunga is Senior Research Fellow, Nonclinical Biostatistics, Johnson & Johnson Pharmaceutical Research & Development LLC, 1000 Route 202, Raritan, NJ 08869 (E-mail: damaratu@prdus.jnj.com). Javier Cabrera is Professor, Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08855 (E-mail: cabrera@stat.rutgers.edu).

© *2007 American Statistical Association, www.amstat.org*
Statistics in Biopharmaceutical Research

having to be sharply accurate or precise, meaning that the false positive rate and false negative rate of this screen would not need to be as small as is expected of a standard confirmatory study. The intention is to then follow up on just a handful of select genes that the screen, together with auxiliary information, flags as being the most interesting [see Amaratunga, Göhlmann, and Peeters (2007) for a pertinent case study related to drug target discovery].

Thus, for instance, to study the genomic effect of a test compound on the gastric mucosa of rats, one small group of n_1 rats would be treated with control and another small group of n_2 rats would be treated with the test compound. After some time, all $N = n_1 + n_2$ rats would be sacrificed and mRNA from their gastric mucosa extracted, amplified, fluorescently labeled, and exposed to N microarrays, one microarray per rat, with G genes arrayed on the microarrays. Typically G would be in the thousands and N would be between 5 and 50. Differences between the corresponding fluorescence intensities of the test microarrays versus the control microarrays should indicate which genes are differentially expressed across the two situations.

Example GM-1: We shall use, for illustrative purposes, data from an experiment called GM-1 whose objective was to study gene expression changes in the gastric mucosa of rats following treatment with a certain test compound as described above (Rose et al. 2003). There were $n_1 = n_2 = 4$ rats in each of the control and treatment groups. Gene expression levels were measured using Affymetrix RG-U34A microarrays which contain probe sets for $G = 8{,}799$ genes.

The data for such experiments typically consist of log transformed and suitably normalized intensities: X_{gij}, where $g (g = 1, \ldots, G)$ indexes the genes on the microarray, $j (j = 1, 2)$ indexes the groups, and $i (i = 1, \ldots, n_j)$ indexes the rats. Details of the preprocessing is itself a research topic and falls outside the scope of this article (see, e.g., Affymetrix 2002; Amaratunga and Cabrera 2001, 2004; Yang et al. 2002; Irizarry, Wu, and Jaffee 2006). Ideally, the goal of the experiment would be to characterize Γ, the subset of genes, among the G in the experiment, that are expressing differentially across the two groups. However, in the spirit of screening, a more modest goal is more realistic: produce a short list Γ^* of genes that is likely to contain a reasonably large proportion of Γ and only a small proportion of Γ^c. Naturally, in practice, biological considerations, as well as statistical considerations, would drive the choice of Γ^*, but, for the purpose of this article, we will only focus on the latter.

An assortment of methods for generating Γ^* are in use. Early researchers (e.g., Schena et al. 1995) used fold change, but this fails to take variability into account. Subsequent statistically motivated methods include t test statistics, a number of different regularized t test statistics [suggested by several authors with perhaps a Bayes or empirical Bayes justification: Lee et al. (2000); Efron et al. (2001); Tusher, Tibshirani, and Chu (2001); Baldi and Long (2001); Newton et al. (2001, 2004); Lonnstedt and Speed (2002); Ishwaran and Rao (2003); Broberg (2003); Wright and Simon (2003); Smyth (2004)] and median-based methods (Amaratunga and Cabrera 2001). Some sort of control for the massive amount of multiple testing involved, such as Bonferroni or, preferably, the positive False Discovery Rate (pFDR) (Storey and

Tibshirani 2001) may also be applied.

The simple t test is the most obvious approach. However, it has very low power as N is small. This is exacerbated by the fact that, within a gene, the intensities, though symmetric, are longer tailed than a normal and outliers are common [as can be seen with simple diagnostic tools such as qq-plots; see Amaratunga and Cabrera (2004)].

Since there are a multitude of genes, it can be conjectured that power could be greatly improved by borrowing strength across genes. This is correct, but since different genes express at different levels and with different variabilities and since the distribution of gene means across genes is heavily skewed, careful modeling is necessary to do this properly. This is what many of the regularization methods attempt, although the validity of some of the assumptions underlying some approaches may be debatable.

All in all, we decided to take a different approach as we found many users were quite comfortable with the t test itself; many nonstatisticians in particular had difficulty understanding the concepts behind regularization and the pros and cons of the different ways of doing it. Thus, we decided to retain the t test statistic as is but, instead of using the classical rejection rule, to generate a rejection region by borrowing strength across genes with minimal distributional assumptions, thereby avoiding strong parametric assumptions that are difficult to justify but are necessary for applying some of the regularization procedures. This produces a novel approach which, when compared to the ordinary t and to a popular regularization method, SAM, outperforms both under a wide range of situations.

2. MOTIVATION

The above situation could be modeled by

$$X_{gij} = \mu_{gj} + \sigma_g \varepsilon_{gij},$$

where μ_{gj} is the mean of the gth gene in the jth group and σ_g^2 is the variance of the gth gene. The treatment effect for the gth gene is:

$$\tau_g = \mu_{g2} - \mu_{g1}.$$

The random errors, $\varepsilon_{gij} \sim F_g$, an unspecified distribution with zero mean and unit variance. The t test statistic for testing H_0: $\tau_g = 0$ for gene g is

$$T_g = (\overline{X}_{g2} - \overline{X}_{g1})/(s_g(1/n_1 + 1/n_2)^{1/2}),$$

where s_g, the pooled standard error, is the positive square root of

$$s_g^2 = ((n_1 - 1)s_{1g}^2 + (n_2 - 1)s_{2g}^2)/(n_1 + n_2 - 2).$$

The conventional approach is to set $F_g = N(0, 1)$ and designate any gene g in the rejection region $RR(T) = \{|T_g| > t_{\alpha/2}\}$, for some specified value of $\alpha (0 < \alpha < 0.5)$, as statistically significant; $RR(T)$ is constructed such that $\text{Prob}_{\text{H}_0}[|T_g| > t_{\alpha/2}] = \alpha$.

Since n_1 and n_2 are both very small, each individual test has very low power and it seems worthwhile to consider borrowing strength across genes to improve the efficiency of the procedure.

In order to evaluate the merit of doing so, it is instructive to study the simple situation in which, for all g, $\sigma_g^2 = \sigma^2$ (i.e., the data are homoscedastic) and $F_g = N(0, 1)$. In this case, borrowing strength across all the genes yields $\hat{\sigma}^2 = \text{ave}(s_g^2)$ as an unbiased estimator for σ^2 with higher efficiency than s_g^2. In fact, since Gp is huge, $\hat{\sigma}^2 \cong \sigma^2$, the z-statistic

$$Z_g = (\overline{X}_{g2} - \overline{X}_{g1})/(\hat{\sigma}\,(1/n_1 + 1/n_2)^{1/2})$$

has a $N(0, 1)$ null distribution, and the corresponding z-test is asymptotically UMP.

Small-sample situations may be explored via simulation. We simulated a set of data: $X_{gij} \sim \text{NID}(\mu_{gi}, 1)$, with $n_1 = n_2 = n$, $\mu_{g1} = 0$ for all g, $\mu_{g2} = \pm\Delta$, with the sign assigned randomly, for $g = 1, \ldots, G_{\text{sig}}$, and $\mu_{g2} = 0$ for $g = (G_{\text{sig}} + 1), \ldots, G$, where G_{sig} denotes the number of differentially expressed genes. We compared the t and z tests for several different values of n, G_{sig}, G, and Δ by picking the m most significant genes produced by each method and, if d of these are among the G_{sig} differentially expressed genes, calculating the True Discovery Rate as TDR $= d/m$. In situations where the maximum possible value of TDR, TDR_{max}, is less than 1 (such as when $m > G_{\text{sig}}$), the raw TDR value is normalized by TDR_{max} (i.e, TDR $\leftarrow$ TDR/TDR_{max}) The simulation results (a subset of which is shown in Table 1) demonstrate that a substantial improvement in performance is indeed afforded by borrowing strength this way as TDR is consistently higher for the z test compared to the t test under homoscedasticity.

Remark: We use TDR for comparing the methods so as to be able to later include SAM in a performance assessment. The conclusion would be identical had we specified a significance level α and used False Positive Rate as comparator.

A key distinction between the z test and the t test is in the form of the rejection region. The rejection region of the t test is $RR(T) = \{|T_g| > t_{\alpha/2}\}$, whereas the rejection region of the z test is $RR(Z) = \{|Z_g| > z_{\alpha/2}\} = \{|T_g| > (\hat{\sigma}/s_g)z_{\alpha/2}\}$ since $T_g = Z_g(\hat{\sigma}/s_g)$. Thus, in terms of T, the rejection region of the t test is delimited by a constant, the critical value $t_{\alpha/2}$, while that of the z test is delimited by a function, a "critical curve", of the form $T = \eta_\alpha(s)$ with $\eta_\alpha(s) = k/s$, where k is a constant. In addition, a pertinent characteristic of $RR(Z)$ is depicted by the following lemma:

Lemma 1 *When $\sigma_g = \sigma$ for all g,* $\text{Prob}_{\text{H}_0}\,(|T_g| > (\hat{\sigma}/s_g)z_{\alpha/2}|s_g) \simeq \alpha$.

Because $\hat{\sigma} \simeq \sigma$ as it is based on a huge number of degrees of freedom, the approximation in Lemma 1 is almost an identity. This then implies that were we to compute the $(1 - \alpha/2)$ quantile curve for the T versus s graph in a null situation (i.e., one in which no genes were truly differentially expressed, $G_{\text{sig}} = 0$), we would obtain a close approximation to the $T = \eta_\alpha(s)$ critical curve. To illustrate this, consider the scatterplot of $|T|$ versus s from one of the null simulations above (see Figure 1(i)). It has a wedge shape because $T_g|s_g \sim N(0, \sigma^2/s_g^2)$. Observe the two curves running through the plot that almost coincide. One is the critical curve $\eta_\alpha(s)$ and the other is the nonparametric $(1 - \alpha/2)$ quantile regression

Table 1. Simulation results: Table of True Discovery Rates (TDR) for different values of n, G_{sig}, Δ and m and different choices for F and F_σ. TDR values that are within 0.010 of the corresponding TDR value for Z_{null} (or larger than Z_{null} due to simulation variability) are highlighted in bold; when there are none, the TDR value closest to the TDR value for Z_{null} is highlighted in bold.

					TDR for $\Delta = 1$					TDR for $\Delta = 2$				
F	F_σ	G_{sig}	m	n	Z_{null}	Z	T	Ct	SAM	Z_{null}	Z	T	Ct	SAM
$N(0,1)$	const	100	100	4	0.350	**0.350**	0.298	**0.348**	0.300	0.742	**0.742**	0.621	**0.738**	0.725
$N(0,1)$	$\chi^2_{10}/10$	100	100	4	0.399	0.338	0.342	**0.358**	0.311	0.762	0.729	0.652	**0.738**	**0.736**
$N(0,1)$	$\chi^2_3/3$	100	100	4	0.444	0.322	0.385	**0.400**	0.372	0.758	0.715	0.663	**0.735**	0.718
$N(0,1)$	χ^2_1	100	100	4	0.541	0.309	0.483	**0.504**	0.486	0.796	0.717	0.736	**0.769**	0.744
t_5	Const	100	100	4	0.353	**0.353**	0.300	**0.351**	0.302	0.740	**0.740**	0.620	**0.737**	0.724
t_5	$\chi^2_{10}/10$	100	100	4	0.400	0.330	0.340	**0.353**	0.305	0.748	**0.726**	0.640	**0.731**	**0.728**
t_5	$\chi^2_3/3$	100	100	4	0.449	0.320	0.387	**0.402**	0.369	0.768	0.723	0.687	**0.752**	**0.750**
t_5	χ^2_1	100	100	4	0.549	0.291	0.496	**0.509**	**0.499**	0.769	**0.750**	0.703	**0.752**	0.715
$N(0,1)$	Const	20	100	4	0.370	**0.370**	0.325	**0.370**	0.275	0.860	**0.860**	0.770	**0.860**	0.770
$N(0,1)$	$\chi^2_3/3$	20	100	4	0.475	0.400	0.415	**0.455**	0.345	0.865	**0.890**	0.810	**0.870**	0.815
$N(0,1)$	const	500	250	4	0.807	**0.807**	0.768	**0.807**	**0.797**	0.991	**0.991**	0.955	**0.990**	**0.988**
$N(0,1)$	$\chi^2_3/3$	500	250	4	0.897	0.792	**0.854**	**0.864**	0.816	0.998	0.971	0.977	**0.988**	**0.984**

					TDR for $\Delta = 1$					TDR for $\Delta = 2$				
F	F_σ	G_{sig}	m	n	Z_{null}	Z	T	Ct	SAM	Z_{null}	Z	T	Ct	SAM
$N(0,1)$	const	100	100	10	0.590	**0.590**	0.556	**0.588**	0.576	0.954	**0.954**	0.921	**0.951**	**0.948**
$N(0,1)$	$\chi^2_{10}/10$	100	100	10	0.653	0.581	0.621	**0.628**	0.619	0.938	**0.935**	0.906	**0.931**	**0.929**
$N(0,1)$	$\chi^2_3/3$	100	100	10	0.632	0.562	0.603	0.617	**0.619**	0.918	**0.916**	0.892	**0.909**	0.900
$N(0,1)$	χ^2_1	100	100	10	0.687	0.542	0.667	**0.672**	0.666	0.893	**0.883**	0.863	0.875	0.869
t_5	const	100	100	10	0.591	**0.591**	0.558	**0.589**	0.577	0.955	**0.955**	0.922	**0.951**	**0.950**
t_5	$\chi^2_{10}/10$	100	100	10	0.616	0.581	0.585	**0.604**	0.598	0.939	**0.931**	0.903	**0.930**	**0.929**
t_5	$\chi^2_3/3$	100	100	10	0.662	0.588	0.635	**0.647**	0.646	0.920	**0.921**	0.897	**0.912**	0.908
t_5	χ^2_1	100	100	10	0.709	0.544	0.690	**0.697**	0.695	0.893	**0.883**	0.863	**0.875**	0.869
$N(0,1)$	const	20	100	10	0.695	**0. 695**	0.665	**0. 695**	0.555	0.995	**0. 995**	**0. 995**	**0. 995**	**0. 995**
$N(0,1)$	$\chi^2_3/3$	20	100	10	0.860	0.790	**0.845**	**0.845**	0.805	0.990	**0.970**	0.960	**0.970**	0.960
$N(0,1)$	const	500	250	10	0.957	**0.957**	0.940	**0.956**	**0.953**	1.000	**1.000**	**1.000**	**1.000**	**1.000**
$N(0,1)$	$\chi^2_3/3$	500	250	10	0.987	0.930	0.975	**0.976**	0.956	0.998	**1.000**	**1.000**	**1.000**	**1.000**

					TDR for $\Delta = 1$					TDR for $\Delta = 2$				
F	F_σ	G_{sig}	m	n	Z_{null}	Z	T	Ct	SAM	Z_{null}	Z	T	Ct	SAM
$N(0,1)$	const	100	100	20	0.807	**0.807**	0.787	**0.804**	**0.800**	0.998	**0.998**	**0.994**	**0.996**	**0.997**
$N(0,1)$	$\chi^2_{10}/10$	100	100	20	0.816	0.799	0.799	**0.809**	**0.814**	0.991	**0.989**	**0.983**	**0.987**	**0.989**
$N(0,1)$	$\chi^2_3/3$	100	100	20	0.794	0.778	0.781	**0.789**	**0.793**	0.983	**0.976**	**0.970**	**0.973**	**0.974**
$N(0,1)$	χ^2_1	100	100	20	0.807	0.773	**0.797**	**0.801**	**0.801**	0.968	**0.962**	**0.956**	**0.959**	**0.958**
t_5	const	100	100	20	0.806	**0.806**	0.785	**0.803**	**0.800**	0.998	**0.998**	**0.994**	**0.996**	**0.997**
t_5	$\chi^2_{10}/10$	100	100	20	0.804	**0.795**	0.787	**0.797**	**0.803**	0.992	**0.989**	**0.984**	**0.985**	**0.989**
t_5	$\chi^2_3/3$	100	100	20	0.803	0.785	0.792	**0.798**	**0.805**	0.980	**0.974**	0.967	**0.970**	**0.972**
t_5	χ^2_1	100	100	20	0.786	0.771	**0.776**	**0.781**	**0.780**	0.968	**0.962**	0.956	**0.959**	**0.958**
$N(0,1)$	const	20	100	20	0.920	**0.920**	**0.910**	**0.920**	0.885	1.000	**1.000**	**1.000**	**1.000**	**1.000**
$N(0,1)$	$\chi^2_3/10$	20	100	20	0.925	**0.950**	**0.920**	**0.925**	0.910	1.000	**1.000**	**1.000**	**1.000**	**1.000**
$N(0,1)$	const	500	250	20	0.997	**0.997**	**0.995**	**0.996**	**0.996**	1.000	**1.000**	**1.000**	**1.000**	**1.000**
$N(0,1)$	$\chi^2_3/3$	500	250	20	1.000	0.985	**0.999**	**0.999**	**0.995**	1.000	**1.000**	**1.000**	**1.000**	**1.000**

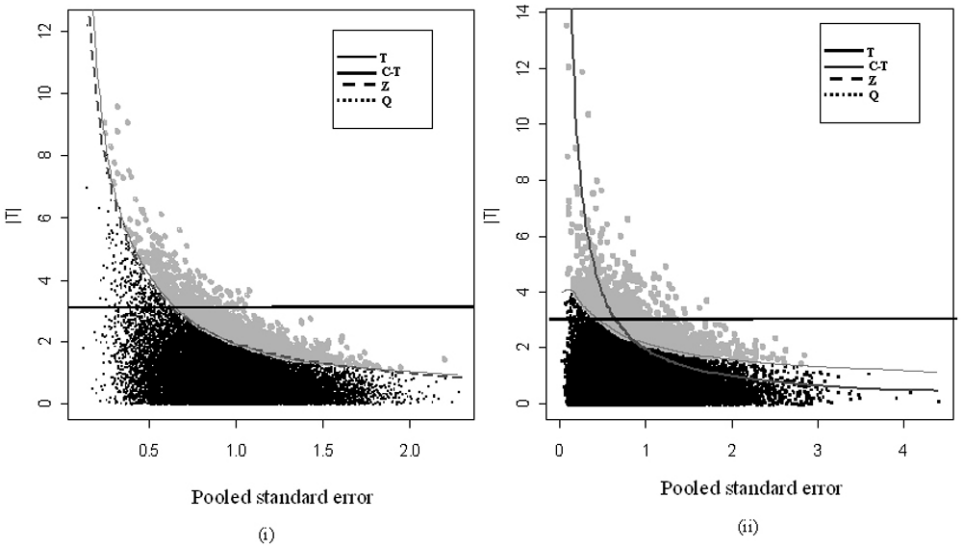

Figure 1. Scatterplot of t test statistics, $|T_g|$, versus pooled standard errors, s_g, for a simulated dataset with $n = 4$ and $F = N(0, 1)$. In (i), $\sigma_g = 1$ for all g. In (ii) $H_\sigma = \chi^2_3/3$. The gray points refer to genes that are found significant by the Ct method at $\alpha = 5\%$; the black dots are the ones that are not.

estimator $t = Q_\alpha(s)$ through the points. This clearly demonstrates that, in the null situation, $Q_\alpha(s)$ is a good estimator of $\eta_\alpha(s)$ under homoscedasticity.

In fact, the wedge shape appears not only for very small sample sizes but also for moderate sample sizes and for other error distributions and under heteroscedasticity (see Figure 2 for a variety of situations where the wedge is present).

For microarray data, the assumption of homoscedasticity may not be valid. Even though we could try to transform the data to near homoscedasticity [using, e.g., the approach of Rocke and Durbin (2003) or Cui, Kerr, and Churchill (2003)], in many cases this might not be possible. Therefore, it is imperative to also address the situation of heteroscedasticity.

To see what happens to the rejection regions of the t and z tests under heteroscedasticity we performed the same simulation as above but with $\sigma_g^2 \sim \chi^2_\nu/\nu$ for different values of ν. Figure 1(ii) shows that $\eta_\alpha(s)$ (dotted curve) and $Q_\alpha(s)$ (solid curve) no longer coincide and Table 1 indicates that the z test loses efficiency compared to the t test. We also see that a rejection region defined by a quantile curve like $Q_\alpha(s)$ (column marked Ct) has still better coverage. This is the method that we now describe.

3. THE CONDITIONAL *t* METHOD

The idea now is to construct a rejection region for the t test statistic T_g that corresponds to the z test under normality and homoscedasticity but that adapts itself when the situation is otherwise. This can be done as follows. First, establish a null distribution for T_g, which can be done via model-based resampling to avoid over-dependence on parametric assumptions. Next, establish the rejection region, which can be

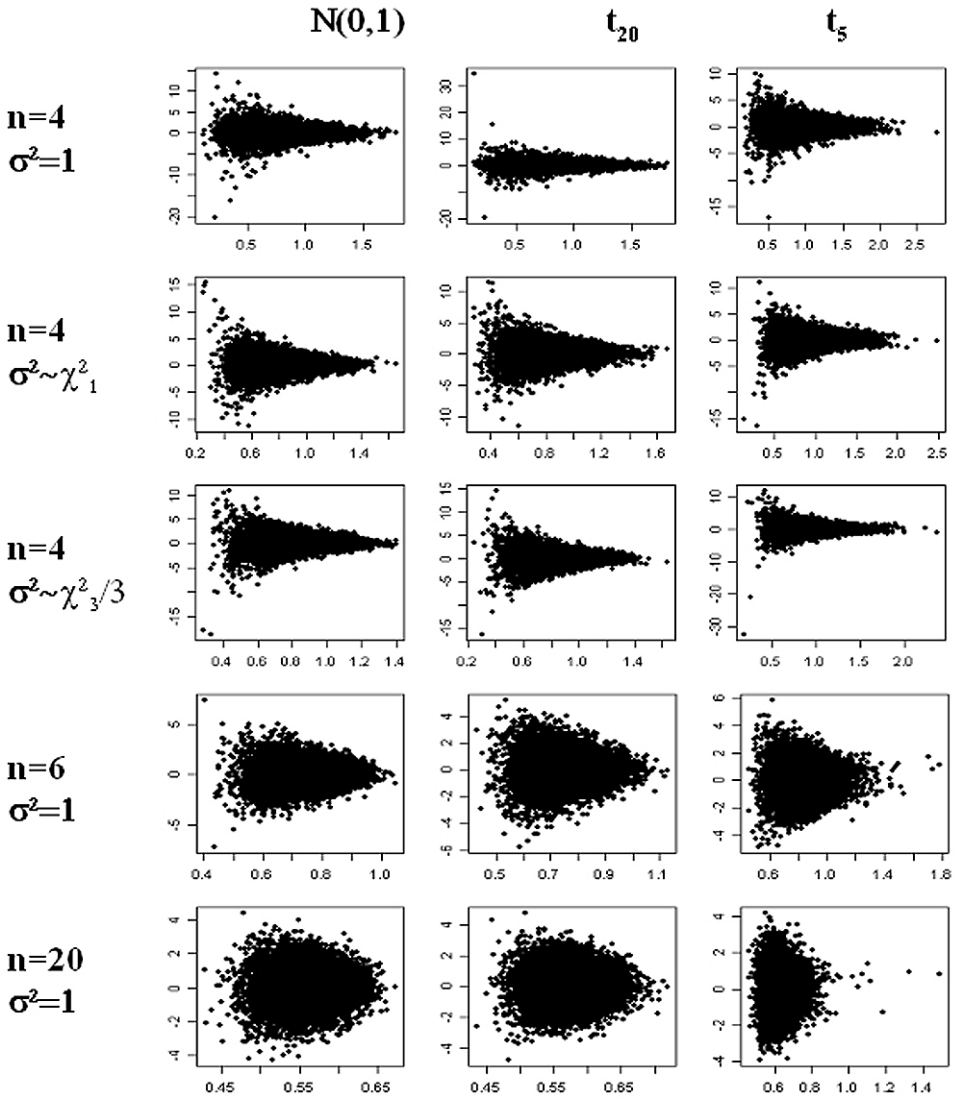

Figure 2. Wedge shapes for a spectrum of models. Each panel shows the scatterplot of t test statistics, T_g, versus pooled standard errors, s_g. The data on each panel consists of two groups of chips of size $n = 4$, 6, or 20 and of $G = 10{,}000$ genes per chip. The data were randomly generated from one of three distributions (i) $N(0, 1)$, (ii) t_{20}, (iii)t_5. In addition for some panels we set $H_\sigma = \chi^2_3/3$ or χ^2_1, otherwise $\sigma = 1$.

done by studying the relationship between null T values and their corresponding s values and running a $100(1-\alpha)\%$ quantile curve, $Q_\alpha(s)$, through this relationship. This $Q_\alpha(s)$ defines the rejection region. Details are now provided.

To borrow strength across the genes, we shall assume that F_g is the same for all the genes (i.e., $F_g = F$ for all g) and $(\mu_{g1}, \sigma_g^2) \sim H_{\mu,\sigma}$, an unspecified distribution with finite first and second moments and marginal distributions H_μ and H_σ. The unknown distributions can be estimated by their respective empirical distribution functions (edf). Thus, in particular, the edf of $\{s_g^2\}$ is an estimate, $\hat{H}_\sigma$, of H_σ and the edf of the standardized residuals $\{R_{gij} = (X_{gij} - \overline{X}_{gj})/s_g\}$ is an estimate, $\hat{F}$, of F.

The *Conditional t* (or Ct) procedure to test H_0:τ_g =0 for all g via a suite of tests proceeds as follows:

(A1) Draw a random sample, s^2, from $\hat{H}_\sigma$.

(A2) Draw a random sample (with replacement) of size N from $\hat{F}$: $r_{ij}^* \sim \hat{F}$ for $i = 1, \ldots, n_j$, $j = 1, 2$.

(A3) Combine these to form pseudo-data: $X_{ij}^* = sr_{ij}^*$.

(A4) Calculate the pooled standard error s^* and t test statistic T^* for the pseudo-data $\{X_{ij}^*\}$.

(A5) Repeat steps (A1)–(A4) a large number, B (where $B = 100{,}000$, say), of times.

(A6) Given α, estimate $t_\alpha(s_g)$ according to $P(|T^*| > t_\alpha(s_g)|s_g) = \alpha$ by computing the nonparametric regression quantile curve for the $(1-\alpha)$th quantile of the $|T^*|$ versus s^* relationship; this is the $100(1-\alpha)\%$ *Ct critical curve*.

(A7) Genes that fall outside the Ct critical curve defined by $t_\alpha(s_g)$ are said to be *Ct-significant* at level α and would be the genes that comprise Γ.

Lemma 2 *The overall Type I error rate of the Conditional t suite of tests is* α.

Ideally, F and H_σ could be estimated by the empirical distributions of their corresponding statistics as stated earlier. However, two modifications are needed particularly when the sample sizes are very small.

Modification 1: When n_1 and n_2 are very small (say no more than 3 or 4 each), the edf of $R_{gij} = (X_{gij} - \overline{X}_{gi})/s_g$ does not adequately describe the error distribution F. In such cases, it is preferable to use only the genes that are the least likely to be differentially expressed (e.g., the genes whose $|T_g| < 1$) and pool the two samples together and use as $\hat{F}$ the edf of $R'_{gij} = (X_{gij} - \overline{X}_g)/sd_g$, where $\overline{X}_g$ and sd_g are the mean and standard deviation of the N observations for the gth gene.

Lemma 3 *The empirical distribution,* $\hat{H}_\sigma$*, of* s_g*, is a biased estimator of* H_σ.

Modification 2: The bias in $\hat{H}_\sigma$ as an estimator of H_σ could be especially large for the very small sample sizes typical of many microarray experiments. Thus, we should not ignore it. The bias can be corrected using a method initially proposed by Amaratunga and Cabrera (2001) for cDNA microarrays

and that is itself a version of the "target estimation" procedure of Cabrera and Fernholz (1999). The idea is to estimate the function $h:[0:1] \to [0,1]$ defined by $h(H_\sigma(x)) = \hat{H}_\sigma(x)$. Since h is strictly monotonic, it can be inverted in order to obtain an estimate of $H_\sigma(x)$. The steps are as follows:

(B1) Assume $\hat{H}_\sigma$ is the true distribution of σ and draw a random sample, s^{*2}, from $\hat{H}_\sigma$.

(B2) Draw a random sample (with replacement) of size N from $\hat{F}: r^*_{ij} \sim \hat{F}$ for $i = 1, \ldots, n_j$, $j = 1, 2$.

(B3) Combine these to form pseudo-data: $X^*_{ij} = s^* r^*_{ij}$.

(B4) Calculate the pooled standard error s^{**} for the pseudo-data $\{X^*_{ij}\}$.

(B5) Repeat steps (B1)–(B4) a large number (say 100,000) of times and record, for each iteration, the pair of values $\{(s^{*2}, s^{**2})\}$.

(B6) Let $\hat{H}_{\sigma^*}(x)$ be the empirical distribution of the s^{**}_g's. Then the estimator of h is obtained by mapping the empirical distribution $\hat{H}_\sigma$ into $\hat{H}_{\sigma^*}$. More precisely

$$\hat{h}(y = \hat{H}_\sigma(x)) = \hat{H}_{\sigma^*}(\hat{H}^{-1}_\sigma(y)) \quad \text{and} \quad \hat{h}^{-1}(y) = \hat{H}_\sigma(\hat{H}^{-1}_{\sigma^*}(y)).$$

Hence the bias-corrected estimator of H_σ is:

$$\tilde{H}_\sigma(x) = \hat{H}_\sigma(\hat{H}^{-1}_{\sigma^*}(\hat{H}_\sigma(x))).$$

This bias-corrected edf, $\tilde{H}_\sigma$, can be used to generate the pooled standard errors in Step (A1).

Remark: Table 1 shows that the Ct method outperforms the *t* test and performs on par with the *z* test when the data are normal and homoscedastic. In fact, Figure 1(i) shows that the Ct method produces a rejection region identical to $\eta_\alpha(s)$ (i.e., the rejection region of the *z* test) and to $Q_\alpha(s)$ in this case. Table 1 also shows that the Ct method outperforms the *t* test and the *z* test when the data are normal but not necessarily homoscedastic. In this case, Figure 1(ii) shows that its rejection region is essentially identical to $Q_\alpha(s)$. Further performance assessments are given in Section 5.

Remark: An alternative to resampling (with replacement) from the edfs is to take a permutation approach a la Tusher et al. (2001). However, in preliminary simulations, this approach turned out to be not competitive, tending to have consistently lower TDR compared to Ct. We conjecture that this could be due to differentially expressing genes contaminating the null distribution.

4. STRATEGY FOR CT-BASED GENE SELECTION

When applied to example GM-1, the *t* test and the Ct procedure identified 998 and 965 genes (respectively) as being statistically significant at the 5% level. The $|T_g|$ versus s_g scatterplot for this data (Figure 3) shows that the genes picked up by Ct are much more evenly distributed over the range of $\{s_g\}$ than the

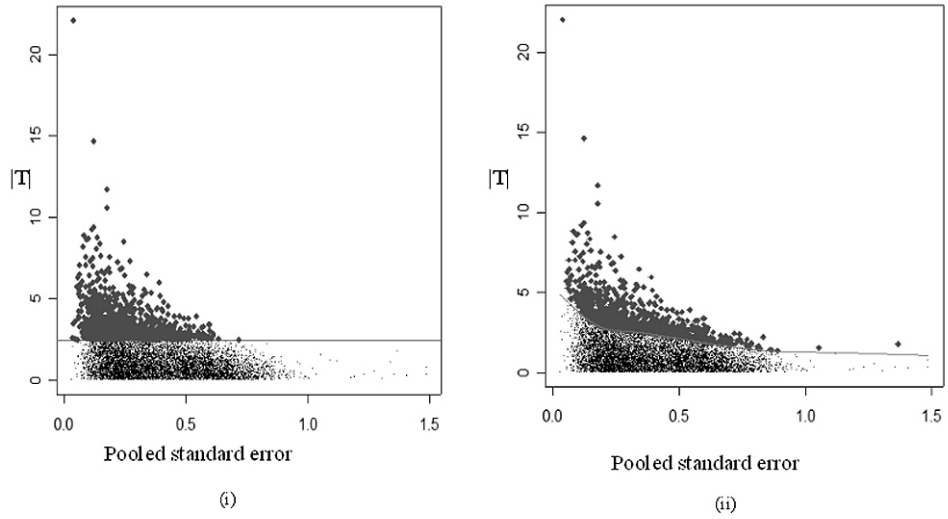

Figure 3. Scatterplot of t test statistics, $|T_g|$, versus pooled standard errors, s_g, for example GM-1. The horizontal line in the left-hand plot indicates the usual 5% critical value for the t test. The curve in the right-hand plot indicates the 5% Ct critical curve. The black filled circles refer to genes that are found significant at $\alpha = 5\%$; the black dots are the ones that are not.

genes picked up by t. It is worth noting that Ct (but not t) identified five genes with relatively high variance as being significantly differentially expressed. On closer investigation, these included progastricsin, a gastric aspartic proteinase that is synthesized in the gastric mucosa, and chymotrypsinogen, which is involved in mucosal digestion. In fact, since RG-U34A includes genes from the whole rat genome, we (and the investigator who carried out the experiment) found it reassuring that several genes being declared significant were known to be associated with the organ system under investigation.

The pFDR of the Ct procedure can be approximated as V/R, where R is the number of genes declared significant (i.e., $R = 965$) and V is the number of genes incorrectly declared significant, which can be approximated as αG (i.e., $V \cong .05 \times 8799 = 439.95$); thus pFDR $\cong$ 46%.

However, as stated in the Introduction, the objective of most microarray experiments is not so much to decide whether or not each and every gene is differentially expressed in an absolute sense, but rather to produce a short list Γ^* of genes that is likely to contain a reasonably large proportion of truly differentially expressed genes. One way to generate such a list is to rank all the genes so that a gene's rank represents its degree of differential expression relative to the other genes and then pick the highest ranking genes.

Computing an approximate p value for each gene can produce such a ranking. Recall that the Ct procedure generates curves, $c_\alpha(s)$, such that, under the null hypothesis of no differential expression, $P(|T_g| > c_\alpha(s_g)|s_g) = \alpha$. Starting with a set of curves, $c_{\alpha_1}(s_g) < \cdots < c_{\alpha_k}(s_g)$, for a set of prespecified values, $\alpha_1 > \cdots > \alpha_k$, consider the relationship between $v_i = \log(-\log(\alpha_i))$ and $u_i = \log(c_{\alpha_i}(s_g))$ (our experience is that the relationship between v_i and u_i is quite linear and much easier to work with than the relationship between α_i and $c_{\alpha_i}(s_g)$). To assign an approximate p value to the gth gene, if $|T_g| \le c_{\alpha_k}(s_g)$, interpolate the relationship between the $\{u_i\}$ and the $\{v_i\}$. For cases when $|T_g| > c_{\alpha_k}(s_g)$, linear extrapolation may be used.

In the example, the top 20 genes by p value have a pFDR $\cong \alpha' G/20 = 23.4\%$; here $a' = 0.00053$, the largest p value among the 20. The pFDR indicates that only 4 or 5 genes (roughly) of the 20 genes selected are likely to false and, on examining the gene ontologies (Gene Ontology Consortium 2000) of the 20, were once again reassured by the fact that many of them had been found to be related to various gastric functions.

5. PERFORMANCE ASSESSMENT

To assess the performance of Ct, we continued the NID simulations referred to in Section 2. We covered a variety of cases by including several different values of n, G_{sig}, Δ, and m and various choices for F and F_σ. Table 1 shows a representative sample of results.

Tusher et al (2001)'s SAM t statistic (where SAM stands for "significant analysis of microarrays") was included in the simulation as it is a widely used scheme of dealing with the wedge effect. In SAM, the t statistic is regularized by adding a fudge factor λ to the denominator:

$$T_g^\lambda = (\overline{X}_{g2} - \overline{X}_{g1})/(s_g(1/n_1 + 1/n_2)^{1/2} + \lambda).$$

The value of λ is chosen so as to attempt to flatten the relationship between var(T_g^λ) and s_g as much as possible. There are a few different implementations of SAM; here we use the version in the R package DNAMR version 1.1 (see Section 6); this implementation follows the algorithm given by Tusher et al. (2001). Since SAM is not generally presented as a critical value-based approach, TDR (as described in Section 2) is used throughout as a measure for comparing procedures.

Each run of the simulation is represented in Table 1 by half a row. For a run, $G = 1{,}000$ genes were sampled from the corresponding model with n observations in each of the two groups. Of the $G = 1{,}000$ genes, G_{sig} genes were sampled from the alternative with a differential expression of Δ and the remaining $G - G_{\text{sig}}$ genes were sampled from the null model (i.e., with $\Delta = 0$).

For each of the five methods, we selected the m most differentially expressed genes and calculated the average proportion of true discoveries or TDR over the runs. The average TDR is reported in Table 1.

The simulation tells a very clear story about the relative operating characteristics of these methods.

(i) When the gene variances are either constant or very similar across genes (e.g., when F_σ has a χ^2 distribution with three or more degrees of freedom), the *Ct* and *Z* methods perform the best and are followed by SAM with *t* at the bottom. *Ct* and SAM are both borrowing strength, which explains why they outperform *t*.

(ii) As the variances become more and more spread out with respect to the level, the performance of the methods gets closer to one another. However, the order of performance changes only for the *Z* procedure, which in some cases is worse. The advantage of borrowing strength slowly dissipates as the spread of the variances increases.

Table 2. Simulation results for the correlated situations: The tables show the True Discovery Rate (TDR) for different values of n, G_{sig}, Δ and m and different choices for F and F_σ. TDR values that are within 0.010 of the corresponding TDR value for Z_{null} are highlighted in bold; when there are none, the TDR value closest to the TDR value for Z_{null} is highlighted in bold.

					TDR for $\Delta = 1$					TDR for $\Delta = 2$				
F	F_σ	G_{sig}	m	n	Z_{null}	Z	T	Ct	SAM	Z_{null}	Z	T	Ct	SAM
Table 2a:														
$N(0,1)$	const	100	100	4	0.427	**0.427**	0.378	**0.426**	0.411	0.667	**0.667**	0.592	**0.665**	0.655
$N(0,1)$	$\chi^2_{10}/10$	100	100	4	0.447	0.424	0.400	**0.432**	0.423	0.674	0.657	0.598	**0.661**	0.656
$N(0,1)$	$\chi^2_3/3$	100	100	4	0.499	0.414	0.453	**0.472**	0.463	0.685	0.629	0.618	**0.663**	0.653
$N(0,1)$	χ^2_1	100	100	4	0.589	0.391	0.545	**0.555**	0.551	0.714	0.617	0.662	**0.690**	0.665
Table 2b:														
$N(0,1)$	const	100	100	4	0.431	**0.431**	0.379	**0.429**	0.395	0.664	**0.664**	0.586	**0.662**	0.644
$N(0,1)$	$\chi^2_{10}/10$	100	100	4	0.439	0.424	0.392	**0.427**	0.400	0.671	0.657	0.596	**0.661**	0.650
$N(0,1)$	$\chi^2_3/3$	100	100	4	0.481	0.420	0.434	**0.458**	0.440	0.694	0.650	0.627	**0.676**	0.663
$N(0,1)$	χ^1_2	100	100	4	0.579	0.387	0.537	**0.543**	0.537	0.725	0.623	0.672	**0.703**	0.678

(iii) As the sample size gets larger, the differences in performance get smaller, but the *Ct* method continues to exhibit the best performance in general.

(iv) As the degree of differential expression increases (e.g., from 1 to 2) there is a small reduction of the differences of performance between the methods. In the case when the sample size is large ($n = 20$) and $\Delta = 2$, all the methods give essentially the same answer.

To study situations in which there are correlations among the genes (as is the case in reality), the simulation was continued with a correlation structure imposed upon the differentially expressing genes. These $G_{\text{sig}} = 100$ genes were divided into 10 groups of 10 each and each group was sampled using a correlation matrix with off-diagonal elements equal to ρ, where $\rho = 0.5 + 0.05(k-1)$ for the kth 10×10 block ($k = 1, \ldots, 10$). The rest of the correlations were set to zero. This generates a block diagonal correlation structure for the differentially expressing genes with correlations ranging from moderate to high. The results of this simulation are shown in Table 2a. We also performed a simulation in which this same block diagonal correlation pattern was repeated for all the remaining nine blocks of 100 genes, so that all genes, whether differentially expressing or not, have a block diagonal correlation structure. The results of this simulation are shown in Table 2b. The findings of these simulations are consistent with the findings listed earlier.

Although the NID simulation covered a lot of territory, it may not fully capture some of the characteristic features of microarray data, such as the complex correlational structure among genes, in part because of the impracticality of estimating or reproducing in a simulation the correlation structure with any reasonable sample size due to the large number of genes. *Quasi-simulation*, the use of faux data fabricated from real data, overcomes this difficulty. The idea is that quasi-simulated data would retain the error structure of an actual dataset, but would have an invented, and therefore known structure (in this case, a set of differentially expressed genes) that the procedure is expected to "find."

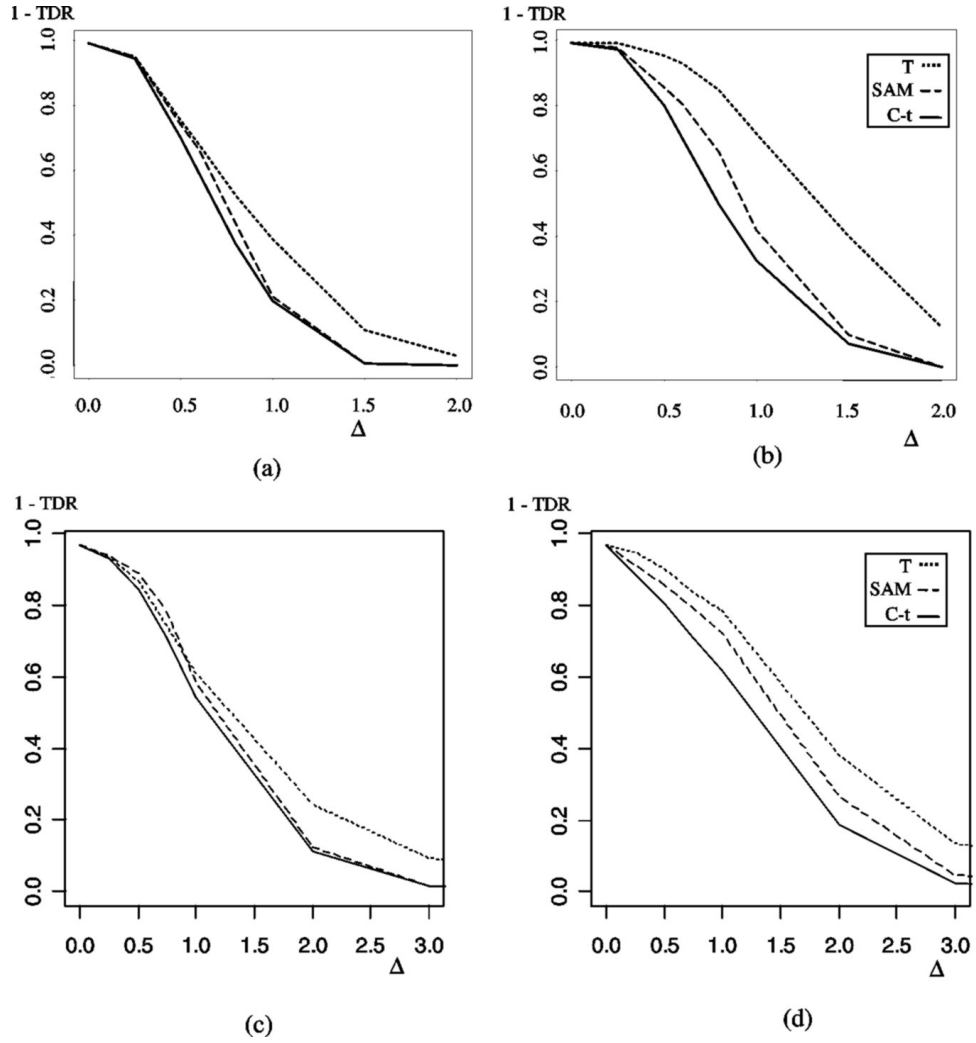

Figure 4. Quasi-simulation results. The plots show $1 - \text{TDR}$ versus Δ for the *t* test, Ct and SAM for Ewing tumor data in (i) and (ii) and rhabdomyosarcoma data in (iii) and (iv). In (i) and (iii), the "truly" differentially expressed genes are scattered throughout all the genes, whereas in (ii) and (iv) they occur only among the higher expressing genes.

We used Khan's pediatric tumor dataset (Khan et al. 2001), familiar to microarray researchers, as the basis for quasi-simulation. From this dataset, we used eight patients with Ewing's sarcoma tumors. We split these equally into a control group and a treatment group, so that $G = 2{,}308$, $n_1 = 4$, $n_2 = 4$. We added or subtracted a quantity Δ to $G_{\text{sig}} = 100$ randomly selected genes for the treatment group, analyzed the resulting data by each method and computed TDRs just as in the NID simulation with $m = 100$. We used eight values of Δ between 0 and 2 a total of 500 times. The results are shown graphically in Figure 4(i). This process was repeated with eight patients with rhabdomyosarcoma tumors from the same dataset; these results are shown in Figure 4(iii). Both figures show that Ct clearly outperforms *t* and does slightly better than SAM; for the Ewing tumor data, the maximum difference in $1 - \text{TDR}$ between Ct and SAM is 7.5% and between Ct and *t* is 14%.

Finally, since Ct exploits the relationship between *t* and *s*, we quasi-simulated a situation where the

"truly" differentially expressed genes are only among those genes with large-ish expression levels (i.e., from among the 50% of genes with the highest average expression levels). The results are shown in Figure 4(ii) for the Ewing tumor patients and Figure 4(iv) for the rhabdomyosarcoma patients. Here TDR is substantially higher for Ct than for both SAM and *t* (for the Ewing tumor data, the maximum difference in TDR between Ct and SAM is 17% and between Ct and *t* is 24%), illustrating SAM's tendency to downplay the significance for high expressing genes, a feature also reported by Broberg (2003).

Overall, the performance evaluation encompassed a wide variety of situations, including iid data, data with normal and non-normal errors, non-iid data and quasi-simulated data. In all situations, Ct provided a moderate to strong improvement over SAM and was clearly superior to *t*.

6. COMPUTATIONAL ISSUES AND AVAILABLE SOFTWARE

Despite it being a computationally intensive procedure, Ct is actually not as time-consuming to run as it might appear; the analysis of a typical dataset such as GM-1 takes only a few minutes on a moderate laptop.

A few steps of the procedure do require some nontrivial computing.

Algorithm for Step A6: The quantile curve can be fitted using a kernel-type procedure. With $B = B_1 B_2$ points (where, if $B = 100{,}000$, we may set $B_1 = 100$ and $B_2 = 1{,}000$), allocate the B points into B_1 bins of B_2 points each sorted by s_g. For the jth bin (or the "super-bin" defined by bins $j - j'$ to $j + j'$ for some small j'), calculate the $(1-\alpha)$th quantile of the T_g values and call it $t_{(j)}$, then calculate the median s_g and call it $s_{(j)}$ $j = 1, \ldots, B_1$. Then estimate $t_\alpha(s_g)$ by running a smoother such as lowess to the $t_{(j)}$ versus $s_{(j)}$ relationship (actually, it is better to take the log of $t_{(j)}$ and $s_{(j)}$ before running the smoother). Koenker and Park (1994) described a more elaborate algorithm.

Algorithm for Step B6: Since $\hat{F}_{\sigma^*}(x)$ is nondecreasing and based on a very large sample we use linear interpolation to approximate the inversion. One alternative is to construct a smooth version $\hat{F}_{\sigma^*}(x)$ using cubic splines, but as most spline algorithms do not guarantee monotonicity, we use linear interpolation. The other steps of the calculation of $\tilde{F}_\sigma(x)$ are also done by using linear interpolation. The algorithm could be iterated by resampling from $\tilde{F}_\sigma(x)$ and obtaining a new $\hat{F}^{-1}_{\sigma^*}$. Then the new estimate becomes $\tilde{F}_{\sigma *}(x) = \tilde{F}_\sigma(\hat{F}^{-1}_{\sigma^*}(\hat{F}_\sigma(x)))$. This step may be repeated until convergence, but our experience is that two or three iterations often suffice.

Software: We have made our R implementation of the Ct procedure freely available in the DNAMR library available from the Web site: *http://www.rci.rutgers.edu/~cabrera/DNAMR*.

It can also be accessed from: *http://www.geocities.com/damaratung/*.

7. A FEW MISCELLANEOUS COMMENTS

We have observed that the Ct procedure is an effective way of borrowing strength across genes in the two-group small-sample case. We now note a few associated developments that are explored in greater detail elsewhere.

Mean-variance relationship: A typical characteristic of microarray data is the relationship between the means and the variances of the expression levels corresponding to individual genes. This is always strong in the raw data, and in many, but not all, cases is either eliminated or at least greatly reduced after log transformation. Otherwise, other variance stabilizing transformations could be considered as stated earlier. However, even if the mean-variance relationship is not entirely eliminated, the model as posited in Section 2 remains valid as it allows for different genes to have different variances. Alternatively, it is possible to try to increase the power of the Ct method by modifying it to account for the mean-variance relationship. This is done by estimating the joint distribution of $(\mu_{g1}, \sigma_g^2) \sim F_{\mu,\sigma}$ and applying a two-dimensional version of the algorithm in Section 3 items B1–B6. The algorithm will estimate the conditional distribution of $t_\alpha|(\mu_{g1}, \sigma_g^2)$ and calculate cutoffs $t_\alpha(\overline{x}_g, s_g)$. Doing this is computationally harder than the one-dimensional version because of the need to invert a mapping between two-dimensional functions. An easier way to compute $t_\alpha(\overline{x}_g, s_g)$ is to condition first on $\overline{x}_g$ by splitting the data into several groups by the values of $\overline{x}_g$. Contiguous groups may be chosen to overlap with each other in order to make for a more smooth transition. We apply the algorithm A1–A6 of Section 2 to each group and use it to calculate $t_\alpha(s_g)|\overline{x}_g$. This routine requires only small modifications of the algorithm described in Section 3 and preliminary simulations show that in cases where the mean-variance relationship remains after transformation there is further improvement in doing this, but the extent of the improvement is worthwhile only if the relationship is still very strong after transformation. Further details, including comparisons with other existing methods, are beyond the scope of this article and are reported elsewhere.

Conditional F: The Ct procedure can be adapted for experimental situations other than the two-sample situation. For example, when comparing across k groups (where $k > 2$), each with few replicates, the wedge effect occurs for F versus MSE; exploiting this and positing the model in much the same way as for Ct, a Conditional F suite of tests can be devised.

Software: For the convenience of users, we have developed, in addition to the R script mentioned in Section 6, a Web implementation of Ct (also available in DNAMR), where users can import their data onto a Web page and perform a Ct analysis.

A. APPENDIX

Lemma 1: When $\sigma_g = \sigma$ for all g, $\text{Prob}_{\text{H}_0}(|T| > (\hat{\sigma}/s)z_{\alpha/2}|s) = \alpha$.

Proof: Follows trivially from the fact that $T_g = Z_g(\hat{\sigma}/s_g) \simeq Z_g(\sigma/s_g)$.

Lemma 2: The overall Type I error rate of the Conditional t suite of tests is α.

Proof: Let s and t be the random variables representing, respectively, the pooled standard error estimate and the observed t statistic for a randomly selected gene. Let $f(t,s)$ be the joint probability density function of t and s. This is a mixing distribution since s has a distribution that depends on the gene. Because the Ct procedure consists of rejecting a null hypothesis if $|t| > t_\alpha(s)$ and conditioning on s, the conditional probability of Type I error is α. The following calculation shows that the overall unconditional probability of Type I error is also α.

$$\begin{aligned}\int_0^\infty \int_{t_\alpha(s)}^\infty f(t,s)dtds &= \int_0^\infty \left(\int_{-\infty}^\infty f(t,s)dt\right) \frac{\int_{t_\alpha(s)}^\infty f(t,s)dt}{\int_{-\infty}^\infty f(t,s)dt} ds \\ &= \int_0^\infty \int_{-\infty}^\infty f(t,s)dt\alpha ds = \alpha \int_0^\infty \int_{-\infty}^\infty f(t,s)dtds = \alpha.\end{aligned}$$

Lemma 3: The empirical distribution, $\hat{H}_\sigma$, of s_g, is a biased estimator of H_σ.

Proof: For ease of exposition, we shall drop all subscripts and assume, without losing generality, that $\mu = 0$, so that we can write the model simply as just $X = \sigma\varepsilon$. By the independence of σ and ε, it follows that

$$E(X) = E(\sigma\varepsilon) = 0.$$

Now, consider a sample of n observations $\{X_1, X_2, \ldots, X_n\}$ from the above model. Let s^2 be the sample variance and let s_ε^2 be the sample variance of the corresponding sample of errors $\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\}$; $\mathrm{E}(s_\varepsilon^2) = 1$ because s_ε^2 is an unbiased estimator of $\mathrm{var}(\varepsilon) = 1$. Then, we have

$$s^2 = \sigma^2 s_\varepsilon^2,$$

$$E(s^2) = E(\sigma^2 s_\varepsilon^2) = E(\sigma^2)E(s_\varepsilon^2) = E(\sigma^2),$$

$$E(s^4) = E(\sigma^4 s_\varepsilon^4) = E(\sigma^4)\mathrm{E}(s_\varepsilon^4).$$

Since

$$\mathrm{var}(s_\varepsilon^2) = E(s_\varepsilon^4) - E(s_\varepsilon^2)^2 = E(s_\varepsilon^4) - 1,$$

and

$$\mathrm{var}(s_\varepsilon^2) > 0,$$

it follows that $E(s_\varepsilon^4) > 1$.

Hence

$$E(s^4) > E(\sigma^4)$$

whence:

$$\mathrm{var}(s^2) = E(s^4) - E(s^2)^2 > E(\sigma^4) - E(\sigma^2)^2 = \mathrm{var}(\sigma^2).$$

Thus, $\mathrm{var}(s^2) > \mathrm{var}(\sigma^2)$ and therefore $\hat{H}_\sigma$ is a biased estimator of H_σ.

The potential severity of the bias in small sample situations can be illustrated by a simulation. Suppose that $\sigma^2 \sim \chi_3^2$ (or, equivalently, $\sigma \sim \chi_3$) and that the sample size is $n = 4$ and $\varepsilon \sim N(0,1)$. A simulation

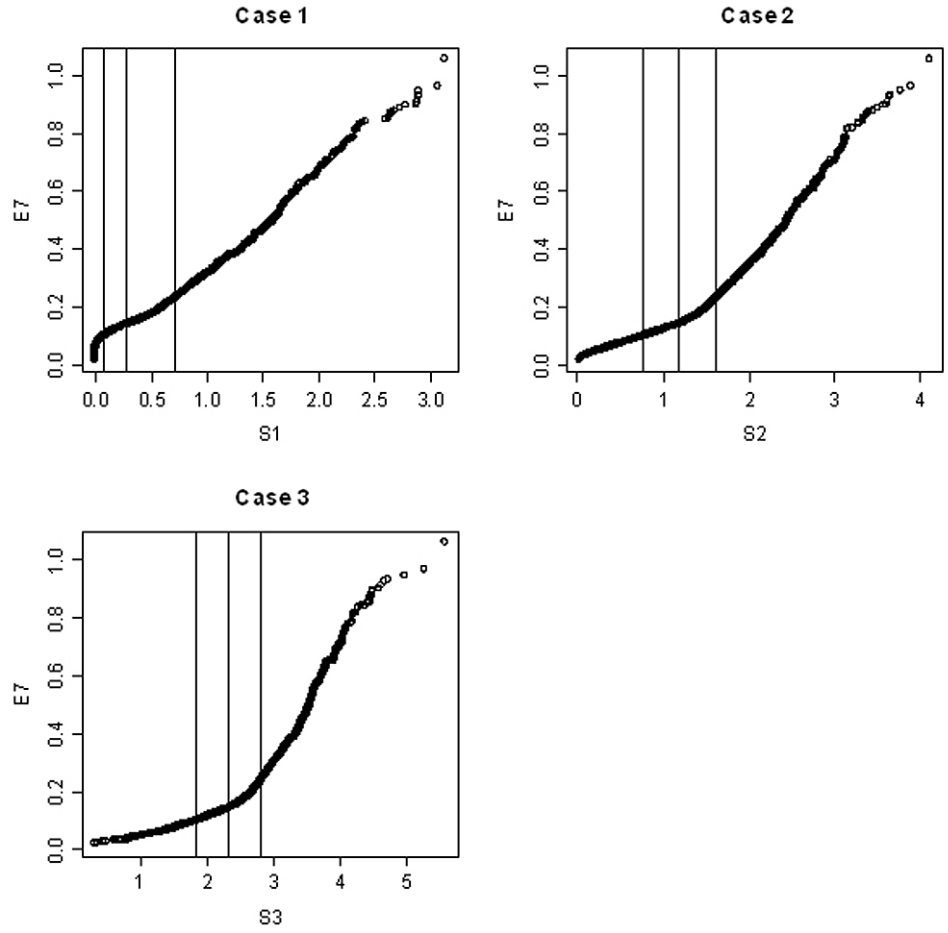

Figure A.1. Quantile-quantile plots of the standard errors, s_g from a situation in which $n = 4$, $F = N(0, 1)$, and $F_\sigma = \chi^2_3$ with three comparator distributions: (a) $\chi^2_{0.5}$ (b) χ^2_2 (c) χ^2_6; the vertical lines are at the quantiles.

from this situation shows that the marginal distribution of s has much heavier tails than a χ_3. This is illustrated in Figure A.1(i). Figure A.1(ii), in fact, implies that the distribution of s is very close to $1.5 \times \chi_{1.5}$, a distribution with much heavier tails than a χ_3. Not surprisingly, the simulation gives $\text{var}(s^2) = 15$, a value exceeding $\text{var}(\sigma^2) = 6$ substantially, confirming that, in this case, $\hat{H}_\sigma$ is a seriously biased estimator of the distribution H_σ.

ACKNOWLEDGMENTS

Javier Cabrera is funded in part by NSF Grant DBI-0629346. The authors thank Jim Colaianne (J&J PRD) for his support during this project. They also thank the editor, associate editor, and the referees for a careful reading of the manuscript and for their many constructive suggestions that greatly improved the article.

[Received September 2006. Revised June 2007.]

REFERENCES

Affymetrix (2002), Microarray Suite User's Guide, Version 5.0, *http://www.affymetrix.com/products/software/specific/mas.affx*.

Amaratunga, D., and Cabrera, J. (2001), "Statistical Analysis of Viral Microchip Data," *Journal of the American Statistical Association*, 96, 1161–1170.

——— (2004), *Exploration and Analysis of DNA Microarray and Protein Array Data*, New York: John Wiley.

Amaratunga, D., Göhlmann, H., and Peeters, P. (2007), "DNA Microarrays," in *Comprehensive Medicinal Chemistry II*, eds. D. J. Triggle and J. B Taylor, Oxford: Elsevier.

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*, 7, 509–519.

Broberg, P. (2003), "Ranking Genes with Respect to Differential Expression," *Genome Biology*, 4, R41.

Cabrera, J., and Fernholz, L. T. (1999), "Target Estimation for Bias and Mean Square Reduction," *Annals of Statistics*, 27, 1080–1104.

Cui, X., Kerr, M.K., and Churchill, G.A. (2003) "Transformations for cDNA Microarray Data," *Statistical Applications in Genetics and Molecular Biology*, 2, 1–20.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001), "Empirical Bayes Analysis of a Microarray Experiment," *Journal of the American Statistical Association*, 96, 1151–1160.

Gene Ontology Consortium (2000), "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics*, 25, 25–29.

Irizarry, R., Wu, Z., and Jaffee, H. (2006), "Comparison of Affymetrix GeneChip Expression Measures," *Bioinformatics*, 22, 789–794.

Ishwaran, H., and Rao, J. S. (2003), "Detecting Differentially Expressed Genes in Microarrays using Bayesian Model Selection," *Journal of the American Statistical Association*, 98, 438–455.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001), "Classification and Diagnostic Prediction of Cancers Using Gene Expression Profiling and Artificial Neural Networks," *Nature Medicine*, 7, 673–679.

Koenker, R., and Park, B.J. (1994), "An Interior Point Algorithm for Nonlinear Quantile Regression," *Journal of Econometrics*, 71, 265–283.

Lee, M. L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000), "Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence from Repetitive cDNA Hybridizations," *Proceedings of the National Academy of Sciences*, 97, 9834–9839.

Lönnstedt, I., and Speed, T. P. (2002), "Replicated Microarray Data," *Statistica Sinica*, 12, 31–46.

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001), "On Differential Variability of Expression Ratios: Improving Statistical Inference About Gene Expression Changes from Microarray Data," *Journal of Computational Biology*, 8, 37–52.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting Differential Gene Expression with a Semiparametric Hierarchical Mixture Method," *Biostatistics*, 5, 155–176.

Rocke, D. M., and Durbin, B. (2003), "Approximate Variance-Stabilizing Transformations for Gene Expression Microarrays," *Bioinformatics*, 19, 966–972.

Rose, A. C., Barrett, T. D., Powell, J. M., Morton, M. F., Fernandez, J. P., Zhang, Y., Cabrera, J., Amaratunga, D., and Shankley, N. P. (2003), "Changes in Rat Gastric Mucosal Gene Expression in Response to Gastrin-Mediated Cholecystokinin 2 Receptor Activation," unpublished.

Schena, M. D. (1999), *DNA Microarrays: A Practical Approach*, Oxford University Press.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995), "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," *Science*, 270, 467–470.

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, Article 3.

Storey, J. D., and Tibshirani, R. (2001), "Estimating False Discovery Rates Under Dependence, With Applications to DNA Microarrays," Technical Report 2001-18, Department of Statistics, Stanford University, Stanford.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001), "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proceedings of the National Academy of Sciences*, 98, 5116–5121.

Wright, G.W., and Simon, R. (2003), "A Random Variance Model for Detection of Differential Gene Expression in Small Microarray Experiments," *Bioinformatics* 19, 2448–1455.

Yang, Y. H., Dudoit, S., Lu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002), "Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation," *Nucleic Acids Research*, 30, e15.