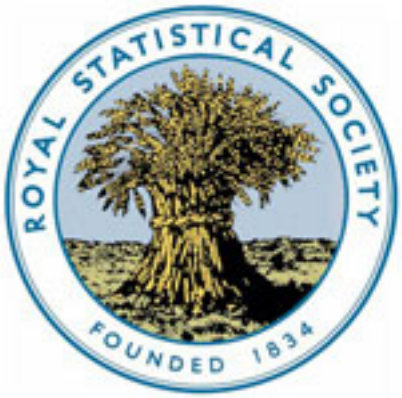

Appl. Statist. (2004)
53, *Part* 1, *pp.* 95–108

Chebyshev's inequality for nonparametric testing with small N and α in microarray research

T. Mark Beasley, Grier P. Page and Jaap P. L. Brand,

University of Alabama at Birmingham, USA

Gary L. Gadbury

University of Missouri at Rolla, USA

and John D. Mountz and David B. Allison

University of Alabama at Birmingham, USA

[Received December 2002. Final revision April 2003]

Summary. Microarrays are a powerful new technology that allow for the measurement of the expression of thousands of genes simultaneously. Owing to relatively high costs, sample sizes tend to be quite small. If investigators apply a correction for multiple testing, a very small p-value will be required to declare significance. We use modifications to Chebyshev's inequality to develop a testing procedure that is nonparametric and yields p-values on the interval [0, 1]. We evaluate its properties via simulation and show that it both holds the type I error rate below nominal levels in almost all conditions and can yield p-values denoting significance even with very small sample sizes and stringent corrections for multiple testing.

Keywords: Chebyshev's inequality; Microarrays; Multiple testing; Nonparametrics; Type I error

1. Introduction

Microarrays are a powerful new technology that allow for the measurement of the expression of thousands of genes simultaneously. Often such measurements are obtained on samples of cases from two or more populations that differ with respect to some characteristic and investigators wish to test whether gene expression levels differ across the populations for each gene. This presents challenging multiple-testing issues. As Gibson (2002) stated,

> '... it is not clear how to assess the appropriate level of significance for microarrays in which thousands of comparisons are performed...'

(page 20). Additionally, because of the relatively high costs of this type of research, the sample sizes tend to be quite small. For example, Kayo *et al.* (2001) used only $n = 3$ monkeys per group when studying aging and caloric restriction. Corominola *et al.* (2001) used only $n = 4$ humans per group when studying obesity and diabetes in humans.

To motivate this issue, suppose a hypothetical example in which a microarray experiment examining the changes in gene expression in response to the application of tumour necrosis factor (TNF-α) in normal rheumatoid arthritis synovial fibroblast (RASF) cells compared with RASF cells where Ad-IκB-DN has blocked the action of TNF-α. RASFs are abnormal cells

Address for correspondence: T. Mark Beasley, Department of Biostatistics, Ryals Public Health Building 343C, University of Alabama at Birmingham, 1665 University Boulevard, Birmingham, AL 35294, USA.
E-mail: mbeasley@uab.edu

 0035–9254/04/53095

that are found in the synovium around joints in individuals with rheumatoid arthritis (RA) and are associated with inflammation of the joints (Mountz *et al.*, 2001). TNF-α, a cytokine (a class of secreted proteins that can stimulate cells to grow or differentiate), is a key pro-inflammatory molecule that contributes to the initiation and perpetuation of RA (Mountz and Zhang, 2001) and is known to cause RASF cells to grow and divide and thus increasing the progression of RA (Miyazawa *et al.*, 1998). Drugs that inhibit TNF-α are effective treatments for RA. TNF-α binds to RASFs and causes the phosphorylation of the protein inhibitor of NF-kB (I-κB) and nuclear translocation of the nuclear factor kB (NF-κB). NF-κB then causes many genes that are involved in inflammation, cell growth and cell division to be turned on. The I-κB gene may be mutated in such a way that TNF-α can no longer cause its phosphorylation and blocks the translocation of NF-κB. This mutation is called a dominant negative mutation and it has been cloned into an adenovirus (Ad-IκB-DN), which enables high efficiency transduction of RASF and results in high levels of expression of I-κB-DN.

In this experiment, primary RASF cells are isolated from synovial fibroblasts from a single human and grown *in vitro* to retain their original *in vivo* dependence on TNF-α. Six samples are taken from these primary passage RASFs. The Ad-I-κB-DN construct is added to three of the samples. A control construct is added to the other three. After 15 h TNF-α is added to all six samples. After three more hours the ribonucleic acid is extracted from the six samples. The ribonucleic acid is then labelled and run in the Affymetrix (Santa Clara, California) U95Av2 microarray. Each chip was scanned on the same scanner using Affymetrix gene scan analysis software version 4.0. Also, suppose that these two groups of $n = 3$ samples are compared for differential expression across 12625 genes.

This ratio of a large number of variables (k) to a small number of total cases (N) renders multivariate solutions difficult and in some cases intractable. Thus, researchers often conduct multiple univariate tests. However, if investigators test each gene separately without any correction for multiple testing, then the experimentwise type I error rate will be unacceptably high (Tusher *et al.*, 2001). Therefore, with thousands of variables to test, microarray researchers may wish to apply a correction for multiple testing (Allison and Coffey, 2002; Wu, 2001).

Using standard adjustments for multiple tests yields extremely small α-values, which reduces statistical power. For example, with $\alpha = 0.05$ and $k = 12625$ variables (genes) the Bonferroni adjustment results in $\alpha_{BON} = \alpha/k = 0.00000396$; whereas the Dunn–Šidák (Dunn, 1961; Šidák, 1967) correction yields a slightly larger value, $\alpha_{DS} = 1 - (1 - \alpha)^{1/k} = 0.00000406$. These adjustments are based on conducting multiple independent tests. It is extremely likely, however, that there is some correlation (i.e. dependence) between several tests conducted on data from the same study, and thus the Bonferroni and Dunn–Šidák adjustments tend to be overcorrections for the probability of at least one type I error. Allison and Beasley (1998) proposed a Monte-Carlo-based procedure for adjusting α when multiple dependent tests are conducted. But, with the small sample sizes that are available to microarray researchers, it may not be possible to obtain stable and accurate correlation coefficients that are necessary for this procedure. Other approaches use gene-specific scatter (e.g. Tusher *et al.* (2001)) or empirical Bayes methodology (Efron and Tibshirani, 2003) and false discovery rate methods (Benjamini and Hochberg, 2000) to incorporate information from other genes and potentially to reduce the number of tests conducted. Allison *et al.* (2002) proposed a method in which the p-values from the multiple univariate tests are used in context with false discovery rate methods to reduce the number of tests in second-level analyses. Suppose that after using the procedure of Allison *et al.* (2002) the number of suspected genes has been determined to be $k = 800$. With $\alpha = 0.05$ and $k = 800$ tests (suspected genes) the Bonferroni adjustment results in $\alpha_{BON} = \alpha/k = 0.0000625$. Thus, even with such methods, a large number of tests

still results, which calls for a very small per test α-level, regardless of the adjustment procedure.

This problem of a large k/N ratio is pernicious because it presents several other complications. First, estimates of skewness and kurtosis are also unstable with small samples; therefore, applied researchers cannot test whether data are normally distributed with any reasonable degree of power (Esteban *et al.*, 2001). Also, owing to small sample sizes, parametric statistical tests of the differences between the mean levels of gene expression for each of the genes will be more sensitive to assumed distributional forms of the expression data (i.e. normality) and, therefore, the resulting p-values may not be accurate when there are departures from normality. Moreover, by applying an α-adjustment due to the large number of tests, the resulting α-value will be extremely small. It is well documented that even when parametric tests are robust at, for example, the $\alpha = 0.05$ level under violations of normality they are often far less robust when very small α-levels are used (Bradley, 1968; Hotelling, 1961). Thus, researchers cannot rely on robust asymptotic properties of parametric tests with such small sample sizes and small αs and cannot test whether their data deviate from the normality assumption.

Although a nonparametric test might be sought, conventional nonparametric tests also have severe limitations in this context. Consider the study by Lee *et al.* (2000), which used three mice per group. With $n = 3$ per group, conventional nonparametric tests for comparing two groups (e.g. the Mann–Whitney U-test; Di Bucchianico (1999)) cannot possibly yield two-tailed p-values that are less than 0.10. This is because conventional nonparametric tests are based on the number of group combinations of ranks and there are a limited number of unique combinations for finite data sets. Thus, in many situations, it is impossible to obtain '$p < 0.05$', let alone the far smaller p-values that are required if the significance level is corrected for multiple testing. Furthermore, in terms of robustness to heterogeneity of variance, rank-based tests can be nearly as sensitive as parametric procedures with small unbalanced samples (Brunner and Munzel, 2000; Zimmerman, 1996). When considering heterogeneity of variance in the present context, however, we are more concerned about robustness in terms of the type II error rate than we are about the type I error rate for reasons articulated below (see Section 5 also).

Bootstrap techniques are often suggested as an alternative (Kerr and Churchill, 2001) because they need not assume normality or homogeneity of variance (Good, 2000) and are therefore less restrictive. If we choose the bootstrap as an alternative method to produce the referent distribution and to compute p-values 'nonparametrically', however, a similar complication arises when resampling from very few cases. Specifically, the maximum number of different bootstrap samples for a two-group design with sample sizes of n_1 and n_2 is

$$W_{\max} = \frac{(2n_1 - 1)!}{n_1!\,(n_1 - 1)!} \frac{(2n_2 - 1)!}{n_2!\,(n_2 - 1)!}$$

(Efron and Tibshirani, 1993). If sample sizes are very small (e.g. $n < 5$), the p-values will be affected by the discreteness of the bootstrapped distribution and there will be a limited number of 'distinct' p-values.

Nevertheless, even with $n = 3$ per group, at times a group difference is so large that common sense suggests that the observed sample difference is significant regardless of the fact that nonparametric or bootstrap methods cannot yield significant p-values or the fact that normality assumptions are virtually impossible to assess. Real life examples of huge differences in gene expression are readily found in the literature (Table 1). One example is the effect of knocking out the interleukin-6 gene, which causes roughly a 35-fold change in expression of the IGFBP-1 gene (Li *et al.*, 2001). Such results sometimes lead biologists to state that they do not need statisticians to tell them that a difference is 'real' and, if statisticians say that the difference is not significant,

Table 1. Examples of huge differences in gene expression

Reference	*Gene*	*Fold change*	*Reported p-value*
Chen *et al.* (2002)	EST (GenBank AI840975)	442.0	None reported
	Calcineurin (GenBank J05479)	410.0	
Hernan *et al.* (2003)	OSF-2	224.6	<0.0001
Li *et al.* (2002)	Dihydrodiol dehydrogenase (GenBank U05598)	142.3	None reported
López *et al.* (2003)	Proenkephalin (GenBank S49491)	117.5	None reported
Lee *et al.* (2002)	CAB 48 (GenBank P12329)	71.1	None reported
Phadtare *et al.* (2002)	creD	47.0	None reported
Han *et al.* (2002)	NCA (GenBank AA054073)	38.3	None reported
Myers *et al.* (2002)	Laminin 37 kd	30.2	None reported
	Alpha-3 ontegrin	33.1	
Campos *et al.* (2003)	Osteopontin	11.4	0.000012

something must be wrong with statisticians or their methods. We acknowledge the sensibleness of this position, but we believe that most scientists wish to say more than 'it is *obviously* real', although we note that several references that were reviewed for this paper do not report inferential statistics (see Table 1). So, we seek a method that is nonparametric and yet theoretically capable of yielding p-values that are continuous on the interval $0 < p \leqslant 1$ with *any* $N > 4$ cases (i.e. a balanced two-group experiment with $n_1 = n_2 > 2$). To elaborate, parametric procedures test differences in location (e.g. means) because other distributional differences are assumed not to exist. Thus, by nonparametric we mean that these methods make no assumptions about the distribution of the error term (e.g. normality or constant variance). Thus, a sufficiently large test statistic indicates that the two groups significantly differ in their distribution of expression levels. Because the methods herein are based on test statistics that employ mean differences, they are most sensitive to differences in location (Bradley, 1968). However, a significant result may be attributable to a difference in location, spread and/or shape (Wilcox, 1993). Thus, we contend that, if a test statistic becomes sufficiently large to become a 'significant result' when the normality or homoscedasticity assumptions are not met, even though population means are identical, then it is still a valuable result to microarray researchers (see Cliff (1993)). This issue is elaborated in Section 5.

2. Test procedures

We use Chebyshev's classic inequality or variations thereof (DasGupta, 2000; Saw *et al.*, 1984) to construct tests which we shall call the 'Chebby checkers' (CCs). We recognize that these methods are likely to be very conservative and have very low power, but we expect them to be used only in those situations where investigators have very small samples and because of multiple variables (genes) must test at very small α-values to make confident and objective statements that their largest effects are significant. We note that other methods (e.g. Allison *et al.* (2002) and Benjamini and Hochberg (2000)) could be incorporated to reduce the number of plausible tests for which to correct.

Chebyshev's inequality states that the probability that a random variable τ exceeds any real value $T > 0$ is

$$P\left(\left|\frac{\tau - \mu_\tau}{\sigma_\tau}\right| \geqslant T\right) \leqslant \frac{1}{T^2}, \tag{1}$$

where μ_τ and σ_τ are the mean and standard deviation of τ respectively. For unimodal, symmetrically distributed random variables, Gauss (1823) showed that Chebyshev's original inequality can be tightened by multiplying the right-hand side by 4/9 (see Mallows (1956)). DasGupta (2000) proved that for a normally distributed random variable this bound can be tightened further by multiplying the right-hand side by $\frac{1}{3}$. Furthermore, he showed that this improvement holds for a larger family of distributions that includes the normal distribution.

Thus, the following inequality can be used to compute two-tailed p-values for hypothesis tests when the assumptions of the usual test statistic (e.g. normality) are potentially violated but not testable:

$$P\left(\left|\frac{\tau - \mu_\tau}{\sigma_\tau}\right| \geqslant T\right) \leqslant \frac{1}{3T^2}. \tag{2}$$

Saw *et al.* (1984) proposed a variant of Chebyshev's inequality for sample data, i.e. the population mean and variance are not known but are replaced with sample estimates. Saw *et al.* (1984) showed that for a sample of fixed size N the upper bound (i.e. the largest plausible p-value) for the Chebyshev inequality approaches

$$P\left(\left|\frac{\tau - \mu_\tau}{\sigma_\tau}\right| \geqslant T\right) \leqslant \frac{1}{N+1} \tag{3}$$

as the random variable τ becomes large. We propose to combine Chebyshev's inequality (1) and the bound (3) of Saw *et al.* (1984), defining

$$P\left(\left|\frac{\tau - \mu_\tau}{\sigma_\tau}\right| \geqslant T\right) \leqslant \frac{1}{(N+1)T^2}, \tag{4}$$

which provides a way to reduce the largest plausible p-value and leads to a procedure that is more sensitive to increases in sample size.

For testing differences between two group means, we suggest the independent samples t-test:

$$t = (\bar{Y}_1 - \bar{Y}_2)/s_{(\bar{Y}_1 - \bar{Y}_2)} \tag{5}$$

as the random variable τ for these procedures. If the gene expression levels are continuously distributed, as they should be when dealing with clonal or inbred strains (Goddard, 2001), the distribution of τ will be unimodal and will be symmetric under the null hypothesis of no mean difference between the two groups, $\mu_\tau = 0$, if the group sample sizes are equal. Under normality, the t-statistic follows Student's t-distribution with $\nu = N - 2$ degrees of freedom and has a standard deviation of $\sigma_\tau = \{\nu/(\nu - 2)\}^{1/2}$. If the dependent variable Y is negatively kurtotic, however, then the conventional t-test will be conservative (Box and Watson, 1962). By contrast, if Y has positive kurtosis, then the conventional t-test will be too liberal (for some values of α), but the sampling variance of s^2 will increase at a faster rate than the sampling variance of $\bar{Y}$, which further implies that, under positive kurtosis, $\sigma_\tau < \{\nu/(\nu - 2)\}^{1/2}$. Furthermore, if Y is skewed, the t-test can be very liberal, especially with unequal sample sizes (Wilcox, 1993, 1997). We note that statistics other than the t-test (5) could be used. For example, if a researcher considers group differences in variances as a 'nuisance' outcome, rather than an interesting result, or is primarily interested in mean differences despite differences in variance, then a Satterthwaite (1949) adjustment could be applied to the t-test. In general, we would not recommend the Mann–Whitney U-test because of its discrete, truncated distribution with small N, as previously discussed.

Importantly, these variants of Chebyshev's inequality are asumed to hold for *any* distribution with finite first and second moments. Furthermore, it is not necessary for the first two moments to be known or estimated for Chebyshev's inequality to hold; they only need to exist

(DasGupta, 2000). This is particularly useful because the mean and variance of a distribution of t-statistics (5), although they may be close to their expected values, are typically unknown when underlying assumptions (e.g. normality) are severely violated. Thus, by defining τ as the t-test (5) and σ_τ as $\{\nu/(\nu-2)\}^{1/2}$, we can be reasonably certain that τ has finite first and second moments, even if Y is non-normal, i.e., when the data are non-normal, the first two moments of the distribution of t-statistics may not follow Student's t-distribution exactly, but they most assuredly exist. Therefore with small sample sizes, we can employ the t-statistic as a random variable τ and apply the modification of the version of Chebyshev's inequality (4) of Saw *et al.* (1984).

By defining τ as the t-statistic (5) and σ_τ as $\{\nu/(\nu-2)\}^{1/2}$, the p-value from inequality (2), $P_{(2)}$, is always less than the two-tailed p-value for a computed t-statistic (5), $P_{(t)}$. For small values of the t-statistic, p-values resulting from inequality (4), $P_{(4)}$, are smaller than $P_{(t)}$. However, there is a 'crossover point' where $P_{(4)} > P_{(t)}$. Thus, we propose to use

$$P_{(6)} = \max(P_{(4)} P_{(t)}). \tag{6}$$

Thus, for smaller values of t-statistic (5), $P_{(6)} = P_{(t)}$ and, for larger values of t, $P_{(6)} = P_{(4)}$. Also, for significance testing at larger αs (e.g. $\alpha = 0.05$ or $\alpha = 0.01$), where the t-test tends to be more robust, $P_{(t)}$ tends to be used. Likewise, at smaller αs, $P_{(4)}$ tends to be used. To illustrate, Table 2 shows these p-values for several values of t-statistic (5) and N.

Table 3 shows data from the hypothetical RASF microarray example for five extreme cases. Means, standard deviations, t-statistics (5), $P_{(t)}$ and p-values from the CC procedures are also displayed. As can be seen, the t-test (5) is statistically significant for all five of these genes ($P_{(t)} < 0.0000625$). However, we may question whether these data were sampled from a normal distribution. Without knowledge of the distributions from which these data were sampled, we may also question whether these results are valid or whether they are false positive results (type I errors) due to the sensitivity of the t-test to departures from normality. The CC(6) method yields statistically significant p-values for the 408-at and 35922-at genes. Even the less powerful

Table 2. Calculated two-tailed p-values for the methods proposed†

$\tau = t$ (5)	*p-values for N = 6*					*p-values for N = 8*				
	$P_{(t)}$	$P_{(1)}$	$P_{(2)}$	$P_{(4)}$	$P_{(6)}$	$P_{(t)}$	$P_{(1)}$	$P_{(2)}$	$P_{(4)}$	$P_{(6)}$
2.0	0.11612	0.50000	0.16667	0.07143	0.11612	0.09243	0.37500	0.12500	0.04167	0.09243
2.5	0.06677	0.32000	0.10667	0.04571	0.06677	0.04653	0.24000	0.08000	0.02667	0.04653
3.0	0.03994	0.22222	0.07407	0.03175	0.03994	0.02401	0.16667	0.05556	0.01852	0.02401
3.5	0.02490	0.16327	0.05442	0.02332	0.02490	0.01283	0.12245	0.04082	0.01361	0.01361
4.0	0.01613	0.12500	0.04167	0.01786	0.01786	0.00712	0.09375	0.03125	0.01042	0.01042
4.5	0.01082	0.09877	0.03292	0.01411	0.01411	0.00410	0.07407	0.02469	0.00823	0.00823
5.0	0.00749	0.08000	0.02667	0.01143	0.01143	0.00245	0.06000	0.02000	0.00667	0.00667
5.5	0.00533	0.06612	0.02204	0.00945	0.00945	0.00151	0.04959	0.01653	0.00551	0.00551
6.0	0.00388	0.05556	0.01852	0.00794	0.00794	0.00096	0.04167	0.01389	0.00463	0.00463
7.0	0.00219	0.04082	0.01361	0.00583	0.00583	0.00042	0.03061	0.01020	0.00340	0.00340
8.0	0.00132	0.03125	0.01042	0.00446	0.00446	0.00020	0.02344	0.00781	0.00260	0.00260
$p < 0.00005$	18.53	200.01	115.48	75.59	75.59	10.26	173.21	100.00	57.74	57.74

†$\tau = t(5)$; $\mu_\tau = 0$; $\sigma_\tau = \{\nu/(\nu-2)\}^{1/2}$; $\nu = N-2$; thus $T = t\{(N-2)/(N-4)\}^{-1/2}$. The last row ($p < 0.00005$) shows the t-statistic (5) that is necessary for each method to reject the null hypothesis at $\alpha = 0.00005$.

Table 3. Two-tailed *p*-values for the methods proposed based on the computational example†

Gene	*Treated*	*Control*	*t (5)*	$P_{(t)}$	$P_{(1)}$	$P_{(2)}$	$P_{(6)}$
408-at	116.8	1540.7	−76.29	0.00000037‡	0.00017180	0.00005727‡	0.00002454‡
	60.8	1528.7					
	51.3	1506.2					
Mean	76.3	1525.2					
Standard deviation	35.4	17.5					
35992-at	302.4	1619.4	−57.22	0.00000082‡	0.00030542	0.00010181	0.00004363‡
	248.5	1615.4					
	229.0	1576.6					
Mean	260.0	1603.8					
Standard deviation	38.0	23.6					
38488-at	−0.4	75.9	−47.68	0.00000105‡	0.00043980	0.00014660	0.00006281
	2.1	76.6					
	4.5	77.9					
Mean	2.1	76.8					
Standard deviation	2.5	1.1					
37762-at	834.5	390.1	47.68	0.00000116‡	0.00043982	0.00014661	0.00006282
	851.9	407.4					
	844.4	416.4					
Mean	843.6	404.6					
Standard deviation	8.7	13.4					
37032-at	510.2	1314.4	−42.92	0.00000179‡	0.00054297	0.00018099	0.00007757
	558.4	1306.6					
	540.8	1278.3					
Mean	536.5	1299.8					
Standard deviation	24.4	19.0					

†$\tau = t(5)$; $\mu_\tau = 0$; $\sigma_\tau = \{\nu/(\nu-2)\}^{1/2}$; $\nu = N-2$; thus $T = t\{(N-2)/(N-4)\}^{-1/2}$. The Kruskal–Wallis test was $\chi^2 = 3.857$ with an exact two-tailed $p = 0.10$ for all cases.
‡Statistically significant at $\alpha_{BON} = 0.0000625$.

CC(2) procedure yields a statistically significant *p*-value for the 408-at gene. This demonstrates that the CC methods can be used to make confident and objective statements that the large effects are statistically significant.

More information about these procedures are available at `http://www.soph.uab.edu/Statgenetics/Research/chebby.htm`.

3. Simulation study of type I error rate

3.1. Method

To evaluate the performance of these testing procedures, we conducted a simulation study. We simulated a two-group design with balanced and unbalanced sample sizes of $N = 6$ and $N = 8$. Data for the two groups were sampled separately (i.e. independently) from the normal and several non-normal distributions with homogeneous variances. Therefore, the conditions of our simulations assume that the error terms for each group were IID$[0,\sigma^2]$. A more detailed elaboration of this choice is in Section 5. Consistent with the challenge that is faced by microarray researchers, we suppose a situation in which the researcher will test differences in group means across $k = 1000$ genes and thus to control for inflation of the type I error rate we shall use a very small α-value of $\alpha_{BON} = 0.0005$. Because α was set at such a small value we conducted 5 million

Table 4. Skewness, kurtosis and λ-coefficients for the generation of the generalized λ-distributions†

Distribution	λ_1	λ_2	λ_3	λ_4	*Skewness* (γ^3)	*Kurtosis* (γ^4)
Generalized λ A	−1.2251	0.1995	0.00685	0.3356	0.8	3.0
Generalized λ B	0	−0.4595	−0.17740	−0.1774	0.0	14.7
Generalized λ C	0	0.5773	1.00000	1.0000	0.0	1.8
Generalized λ D	−0.2937	−0.3658	−0.10470	−0.1833	2.0	21.9
Normal					0.0	3.0
Log-normal					6.2	113.9

†The normal and log-normal distributions were generated by using the SAS RANNOR function.

replications for the expected number of rejections to be reasonably large (i.e. 250 rejections) under the null hypothesis.

Using SAS procedure IML, we simulated normally distributed data Y by using the RANNOR function (SAS Institute, 2001). To create the non-normal distributions, we used a system of non-linear transformations from the generalized λ-distribution (Karian and Dudewicz, 2000):

$$Y = \lambda_1 + \frac{U^{\lambda_3} - (1-U)^{\lambda_4}}{\lambda_2}, \tag{7}$$

where U is a random deviate from the uniform distribution generated with the SAS RANUNI function. The resultant variable Y has a mean of 0 and unit variance with the skewness and kurtosis specified in Table 4. To simulate an extremely non-normal distribution, we generated a log-normal distribution by $Y = \exp(X)$, where X is a random deviate from the unit normal distribution generated with the RANNOR function. The resultant variable Y has a mean of $\exp(0.5\sigma^2) = 1.65$ and a variance of $\exp(\sigma^2)\{\exp(\sigma^2) - 1\} = 4.67$ with a skewness of $\{\exp(\sigma^2) + 2\}\sqrt{\{\exp(\sigma^2) - 1\}} = 6.18$ and a kurtosis of $\exp(\sigma^2)^4 + 2\exp(\sigma^2)^3 + 3\exp(\sigma^2)^2 - 3 = 113.94$. The Y-values were then standardized to have a mean 0 and unit variance. The log-normal distribution was chosen because this has been shown in past research to be especially troublesome for parametric statistics (Cressie and Whitford, 1986; Wilcox, 1997).

The relative bias was calculated by dividing the empirical type I error rate by the nominal α. If the type I error rate is controlled then the relative bias should be 1. Using Bradley's (1978) stringent criterion we considered any relative bias that was higher that 1.1 to be an inflation of the type I error rate.

3.2. Results

Table 5 shows the relative bias for the four test procedures under normality assumptions (i.e. NID[0,σ^2]) and under five non-normal IID[0,σ^2] conditions. As can be seen, the parametric t-test (5) inflated the type I error rates with a mesokurtic (i.e. a distribution with kurtosis equal to the normal), but skewed distribution (generalized λ A) and with a light-tailed distribution (generalized λ C). It performed even worse for the extremely skewed and heavy-tailed log-normal distribution. The CC procedures held the type I error rates well below the nominal α-value and in fact rarely committed a false positive result. Consistent with previous studies (Bradley, 1968; Hotelling, 1961), Table 6 shows that the type I rate inflation that occurred at $\alpha = 0.00005$ did not occur to any great extent at $\alpha = 0.05$. This shows the importance of conducting such simulation studies with the small α-levels that may eventually be used in microarray research.

Table 5. Relative bias for $\alpha = 0.000\,05$ based on 5 million replications†

Distribution	*Test*	*Relative biases for the following values of* n_1 *and* n_2.			
		$n_1 = 3, n_2 = 3$	$n_1 = 2, n_2 = 4$	$n_1 = 4, n_2 = 4$	$n_1 = 3, n_2 = 5$
Normal, $\gamma^3 = 0, \gamma^4 = 3.0$	$t(5)$	0.964	1.004	0.928	0.940
	CC(1)	0.000	0.000	0.000	0.000
	CC(2)	0.000	0.000	0.000	0.000
	CC(6)	0.004	0.004	0.000	0.000
Generalized λ A, $\gamma^3 = 0.8, \gamma^4 = 3.0$	$t(5)$	2.124‡	2.552‡	2.224‡	2.832‡
	CC(1)	0.000	0.000	0.000	0.000
	CC(2)	0.000	0.000	0.000	0.000
	CC(6)	0.004	0.012	0.000	0.000
Generalized λ B, $\gamma^3 = 0, \gamma^4 = 14.7$	$t(5)$	0.568	0.880	0.540	0.544
	CC(1)	0.000	0.000	0.000	0.000
	CC(2)	0.000	0.000	0.000	0.000
	CC(6)	0.000	0.000	0.000	0.000
Generalized λ C, $\gamma^3 = 0, \gamma^4 = 1.8$	$t(5)$	3.640‡	3.124‡	4.624‡	4.300‡
	CC(1)	0.000	0.000	0.000	0.000
	CC(2)	0.000	0.000	0.000	0.000
	CC(6)	0.016	0.020	0.000	0.000
Generalized λ D, $\gamma^3 = 2.0, \gamma^4 = 21.9$	$t(5)$	0.600	1.064	0.500	0.692
	CC(1)	0.000	0.000	0.000	0.000
	CC(2)	0.004	0.000	0.000	0.000
	CC(6)	0.004	0.004	0.000	0.000
Log-normal, $\gamma^3 = 6, \gamma^4 = 100$	$t(5)$	1.404‡	5.300‡	0.944	2.696‡
	CC(1)	0.000	0.000	0.000	0.000
	CC(2)	0.000	0.000	0.000	0.000
	CC(6)	0.000	0.012	0.000	0.000

†γ^3 refers to the skewness and γ^4 refers to the kurtosis of the simulated distribution (see Table 4). $t(5)$ refers to the Student t-test. The number in parentheses in the CC methods refers to the equation number in the text.
‡Significantly inflated type I error rate at Bradley's stringent criterion 1.1α.

4. Simulation study of statistical power

To investigate the statistical power of these methods, a simulation study similar to the first simulation study was conducted. A constant δ was added to the dependent variable Y for the first group to create standardized group mean differences (i.e. effect sizes). Effect size constants of $\delta = 2, 5, 10, 20$ were used. Because the rejection rates were expected to be larger than $\alpha = 0.000\,05$, 1 million replications were conducted.

Tables 7 and 8 show the empirical power estimates for effect sizes of $\delta = 2$ and $\delta = 10$ respectively. Conditions in which the type I error rates exceeded the nominal α with a relative bias greater than 1.1 are not reported. For brevity, the power estimates for $\delta = 5$ and $\delta = 20$ are not reported, but these results are consistent with those for $\delta = 2$ and $\delta = 10$. A full set of tables can be viewed at `http://www.soph.uab.edu/Statgenetics/Research/chebby.htm`. As would be expected the CC tests had low statistical power, even with an extremely large effect size of $\delta = 10$. For several conditions, the t-test (5) did not inflate the type I error rate and was more powerful, and thus would be preferable. However, a researcher is not likely to know when these conditions hold in small samples. Of the remaining tests, the procedure detailed in equation (6) was slightly more powerful than either CC(1) or CC(2).

Table 6. Relative bias for $\alpha = 0.05$ based on 5 million replications†

Distribution	*Test*	*Relative biases for the following values of* n_1 *and* n_2*:*			
		$n_1 = 3, n_2 = 3$	$n_1 = 2, n_2 = 4$	$n_1 = 4, n_2 = 4$	$n_1 = 3, n_2 = 5$
Normal, $\gamma^3 = 0, \gamma^4 = 3.0$	$t(5)$	0.998964	0.998336	0.998696	0.997432
	CC(1)	0.063324	0.063860	0.030408	0.031028
	CC(2)	0.435928	0.434168	0.389708	0.390232
	CC(6)	0.998964	0.998336	0.998696	0.997432
Generalized λ A, $\gamma^3 = 0.8, \gamma^4 = 3.0$	$t(5)$	1.022748	1.034504	0.974360	0.975356
	CC(1)	0.091324	0.102608	0.044104	0.048376
	CC(2)	0.489004	0.513460	0.400964	0.410632
	CC(6)	1.022748	1.034504	0.974360	0.975356
Generalized λ B, $\gamma^3 = 0, \gamma^4 = 14.7$	$t(5)$	0.788436	0.872180	0.808344	0.845452
	CC(1)	0.041440	0.053612	0.017180	0.019548
	CC(2)	0.313712	0.367832	0.275252	0.296748
	CC(6)	0.788436	0.872180	0.808344	0.845452
Generalized λ C, $\gamma^3 = 0, \gamma^4 = 1.8$	$t(5)$	1.213360‡	1.117312‡	1.122276‡	1.093584
	CC(1)	0.136524	0.114896	0.072296	0.065404
	CC(2)	0.626492	0.557924	0.515480	0.488040
	CC(6)	1.213360‡	1.117312‡	1.122276‡	1.093584
Generalized λ D, $\gamma^3 = 2.0, \gamma^4 = 21.9$	$t(5)$	0.798420	0.889220	0.812072	0.849284
	CC(1)	0.044600	0.060020	0.018900	0.022412
	CC(2)	0.324784	0.385412	0.281020	0.304480
	CC(6)	0.798420	0.889220	0.812072	0.849284
Log-normal, $\gamma^3 = 6, \gamma^4 = 100$	$t(5)$	0.673532	0.862252	0.614032	0.689568
	CC(1)	0.058584	0.118964	0.020712	0.034876
	CC(2)	0.308380	0.458080	0.215088	0.270964
	CC(6)	0.673532	0.862252	0.614032	0.689568

†γ^3 refers to the skewness and γ^4 refers to the kurtosis of the simulated distribution (see Table 4). $t(5)$ refers to the Student t-test. The number in parentheses in the CC methods refers to the equation number in the text.
‡Significantly inflated type I error rate at Bradley's stringent criterion 1.1α.

5. Discussion

It should be reiterated that we simulated data under IID[0,σ^2] assumptions. It is well known that the type I error rate of the parametric t-test (5) is affected by between-group differences in variance (e.g. Boneau (1960) and Scheffé (1959)) and skewness (Cressie and Whitford, 1986; Wilcox, 1993), especially for unequal sample sizes. Preliminary simulations studies have shown that even the most conservative CC method, CC(1), can inflate type I error rates with heterogeneous variances or with different distributions for the two groups. Satterthwaite (1949) adjustments could possibly remedy the heteroscedasticity problem (Algina *et al.*, 1994). The Cressie and Whitford (1986) correction for non-zero skewness could be used to adjust for differences in distributional shapes and spread. However, this method has not been thoroughly evaluated in simulation studies, especially with small samples sizes where the estimates of variance, skewness and kurtosis are not stable. Thus, the assumptions of independent and identically distributed (IID) data are critical to valid type I error rates for any of the procedures herein.

It may be asked why we did not use conditions that violate the assumptions of IID data. First, these assumptions are difficult to evaluate with small sample sizes; thus, for practical purposes a researcher would never know whether they hold. More importantly, we contend that

Table 7. Power for $\alpha = 0.00005$ based on 1 million replications with $\delta = 2$†

Distribution	*Test*	*Powers for the following values of* n_1 *and* n_2*:*			
		$n_1 = 3, n_2 = 3$	$n_1 = 2, n_2 = 4$	$n_1 = 4, n_2 = 4$	$n_1 = 3, n_2 = 5$
Normal, $\gamma^3 = 0, \gamma^4 = 3.0$	$t(5)$	0.001223	0.001034	0.004810	0.004118
	CC(1)	0.000000	0.000000	0.000000	0.000000
	CC(2)	0.000000	0.000000	0.000000	0.000000
	CC(6)	0.000005	0.000004	0.000000	0.000000
Generalized λ A, $\gamma^3 = 0.8, \gamma^4 = 3.0$	$t(5)$	0.001849‡	0.002177‡	0.006487‡	0.007274‡
	CC(1)	0.000000	0.000000	0.000000	0.000000
	CC(2)	0.000001	0.000002	0.000000	0.000000
	CC(6)	0.000007	0.000013	0.000000	0.000002
Generalized λ B, $\gamma^3 = 0, \gamma^4 = 14.7$	$t(5)$	0.002467	0.002089	0.012032	0.010731
	CC(1)	0.000000	0.000000	0.000000	0.000000
	CC(2)	0.000002	0.000000	0.000000	0.000000
	CC(6)	0.000007	0.000007	0.000000	0.000000
Generalized λ C, $\gamma^3 = 0, \gamma^4 = 1.8$	$t(5)$	0.001504‡	0.001191‡	0.004551‡	0.003936‡
	CC(1)	0.000000	0.000000	0.000000	0.000000
	CC(2)	0.000001	0.000003	0.000000	0.000000
	CC(6)	0.000006	0.000003	0.000001	0.000000
Generalized λ D, $\gamma^3 = 2.0, \gamma^4 = 21.9$	$t(5)$	0.002780	0.002649	0.013622	0.012882
	CC(1)	0.000000	0.000000	0.000000	0.000000
	CC(2)	0.000002	0.000001	0.000001	0.000000
	CC(6)	0.000010	0.000008	0.000002	0.000002
Log-normal, $\gamma^3 = 6, \gamma^4 = 100$	$t(5)$	0.002491‡	0.003530‡	0.008926	0.009598‡
	CC(1)	0.000000	0.000000	0.000000	0.000000
	CC(2)	0.000001	0.000003	0.000000	0.000000
	CC(6)	0.000010	0.000014	0.000002	0.000006

†γ^3 refers to the skewness and γ^4 refers to the kurtosis of the simulated distribution (see Table 4). $t(5)$ refers to the Student t-test. The number in parentheses in the CC methods refers to the equation number in the text.
‡Power estimate invalid owing to a significantly inflated type I error rate.

if a test statistic becomes sufficiently large to become a 'significant result' when the assumptions of IID data are not met, even though population means are identical, then it is still a valuable result to microarray researchers (see Cliff (1993)), i.e. differences in variance or distributional shape are *results*, not merely nuisances. Therefore, the procedures and conditions of IID data that we simulated seem reasonable, because a microarray researcher would not want to contend that there are group differences when in fact the populations are identical (in terms of location, spread and shape), but the rejection was due to the sensitivity of the t-test to non-normality. However, if the assumptions of IID data do not hold, it should be viewed as a result that can be detected with some degree of statistical power. Typically, departures from this assumption can be tested with nonparametric methods (Mann–Whitney U or Cliff's d), but as stated previously they suffer from a limited range of p-values for small samples (see Table 2).

In the majority of this paper, we refer to testing for differences between two groups, but the method generalizes to many other testing situations. For example, with multiple-group comparisons, we could use the analysis-of-variance F-statistic or some *post hoc* test statistic (e.g. Tukey's honestly significant difference or Fisher's least significant difference) as the random variable τ and their known expected values and standard deviations as μ_τ and σ_τ respectively.

The methods that we have proposed are certainly not the sole methods that should be applied

Table 8. Power for $\alpha = 0.000005$ based on 1 million replications, with $\delta = 10$†

Distribution	*Test*	*Powers for the following values of* n_1 *and* n_2*:*			
		$n_1 = 3, n_2 = 3$	$n_1 = 2, n_2 = 4$	$n_1 = 4, n_2 = 4$	$n_1 = 3, n_2 = 5$
Normal, $\gamma^3 = 0, \gamma^4 = 3.0$	$t(5)$	0.220771	0.185707	0.913637	0.891029
	CC(1)	0.000030	0.000022	0.000000	0.000003
	CC(2)	0.000269	0.000205	0.000031	0.000040
	CC(6)	0.001433	0.001084	0.000896	0.000767
Generalized λ A, $\gamma^3 = 0.8, \gamma^4 = 3.0$	$t(5)$	0.236473‡	0.206769‡	0.908810‡	0.880787‡
	CC(1)	0.000050	0.000043	0.000004	0.000005
	CC(2)	0.000416	0.000395	0.000072	0.000071
	CC(6)	0.002107	0.001872	0.001655	0.001517
Generalized λ B, $\gamma^3 = 0, \gamma^4 = 14.7$	$t(5)$	0.358471	0.317768	0.886256	0.871281
	CC(1)	0.000087	0.000073	0.000004	0.000006
	CC(2)	0.000783	0.000631	0.000195	0.000192
	CC(6)	0.003918	0.003146	0.004110	0.003622
Generalized λ C, $\gamma^3 = 0, \gamma^4 = 1.8$	$t(5)$	0.165405‡	0.134663‡	0.964096‡	0.942765‡
	CC(1)	0.000020	0.000024	0.000001	0.000003
	CC(2)	0.000205	0.000172	0.000022	0.000018
	CC(6)	0.001055	0.000834	0.000469	0.000379
Generalized λ D, $\gamma^3 = 2.0, \gamma^4 = 21.9$	$t(5)$	0.377997	0.338651	0.889098	0.874494
	CC(1)	0.000088	0.000076	0.000007	0.000009
	CC(2)	0.000838	0.000732	0.000238	0.000192
	CC(6)	0.004278	0.003678	0.004893	0.004221
Log-normal, $\gamma^3 = 6, \gamma^4 = 100$	$t(5)$	0.212114‡	0.196178‡	0.542337	0.524592‡
	CC(1)	0.000118	0.000100	0.000010	0.000014
	CC(2)	0.000861	0.000823	0.000239	0.000226
	CC(6)	0.003994	0.003686	0.003845	0.003509

†γ^3 refers to the skewness and γ^4 refers to the kurtosis of the simulated distribution (see Table 4). $t(5)$ refers to the Student t-test. The number in parentheses in the CC methods refers to the equation number in the text.
‡Power estimate invalid owing to a significantly inflated type I error rate.

to microarray data. Other approaches use gene-specific scatter (e.g. Tusher *et al.* (2001)) or empirical Bayes methodology (Efron and Tibshirani, 2003) and false discovery rate methods to incorporate information from other genes and potentially to reduce the number of tests conducted. We note that related methods (e.g. Allison *et al.* (2002) and Benjamini and Hochberg (2000)) could be incorporated to reduce the number of plausible tests before applying the CC methods herein. We acknowledge that these methods have very low power for all except the largest effects. Nevertheless, in many cases, those are exactly the effects that are observed and of interest. As Richmond and Somerville (2000) stated,

> 'in "marker discovery" experiments, the goal is to discover a limited number of highly specific marker genes for a cell type, a developmental stage or an environmental treatment. In such experiments, the researcher is often interested only in genes that show a dramatic and selective induction or repression of expression'

(page 108).

Acknowledgements

This research was supported in part by National Institutes of Health grants R01DK56366, P30DK56336, P01AG11915, R01AG018922, P20CA093753, R01AG011653, U24DK058776

and R01ES09912, National Science Foundation grant 0090286 and a grant from the University of Alabama at Birmingham Health Services Foundation.

References

Algina, J., Oshima, T. C. and Lin, W. Y. (1994) Type I error rates for Welch's test and James's second-order test under nonnormality and inequality of variance when there are two groups. *J. Educ. Behav. Statist.*, **19**, 275–291.

Allison, D. B. and Beasley, T. M. (1998) Method and computer program for controlling the family-wise alpha rate in gene association studies involving multiple phenotypes. *Genet. Epidem.*, **15**, 87–101.

Allison, D. B. and Coffey, C. S. (2002) Two-stage testing in microarray analysis: what is gained? *J. Gerontol.* A, **57**, B189–B192.

Allison, D. B., Gadbury, G. L., Heo, M., Fernández, J. R., Lee, C.-K., Prolla, T. A. and Weindruch, R. (2002) A mixture model approach for the analysis of microarray gene expression data. *Comput. Statist. Data Anal.*, **39**, 1–20.

Benjamini, Y. and Hochberg, Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Statist.*, **25**, 60–83.

Boneau, C. A. (1960) The effects of violations of assumptions underlying the t test. *Psychol. Bull.*, **57**, 49–64.

Box, G. E. P. and Watson, G. S. (1962) Robustness to non-normality of regression tests. *Biometrika*, **49**, 93–106.

Bradley, J. V. (1968) *Distribution-free Statistical Tests*. Englewood Cliffs: Prentice Hall.

Bradley, J. V. (1978) Robustness? *Br. J. Math. Statist. Psychol.*, **31**, 144–152.

Brunner, E. and Munzel, U. (2000) The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometr. J.*, **42**, 17–25.

Campos, A. H., Zhao, Y., Pollman, M. J. and Gibbons, G. H. (2003) DNA microarray profiling to identify angiotensin-responsive genes in vascular smooth muscle cells: potential mediators of vascular disease. *Circuln Res.*, **92**, 111–118.

Chen, C.-Z., Li, M., de Graaf, D., Monti, S., Göttgens, B., Sanchez, M.-J., Lander, E. S., Golub, T. R., Green, A. R. and Lodish, H. F. (2002) Identification of endoglin as a functional marker that defines long-term repopulating hematopoietic stem cells. *Proc. Natn. Acad. Sci. USA*, **99**, 15468–15473.

Cliff, N. (1993) Dominance statistics: ordinal analyses to answer ordinal questions. *Psychol. Bull.*, **114**, 494–509.

Corominola, H., Conner, L. J., Beavers, L. S., Gadski, R. A., Johnson, D., Caro, J. F. and Rafaeloff-Phail, R. (2001) Identification of novel genes differentially expressed in omental fat of obese subjects and obese type 2 diabetic patients. *Diabetes*, **50**, 2822–2830.

Cressie, N. A. and Whitford, H. J. (1986) How to use the two sample *t*-test. *Biometr. J.*, **28**, 131–148.

DasGupta, A. (2000) Best constants in Chebychev inequalities with various applications. *Metrika*, **51**, 185–200.

Di Bucchianico, A. (1999) Combinatorics, computer algebra and the Wilcoxon-Mann-Whitney test. *J. Statist. Planng Inf.*, **79**, 349–364.

Dunn, O. J. (1961) Multiple comparison among means. *J. Am. Statist. Ass.*, **56**, 52–64.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Efron, B. and Tibshirani, R. J. (2003) Empirical Bayes methods and false discovery rates for microarrays. *Genet. Epidem.*, **23**, 70–86.

Esteban, M. D., Castellanos, M. E., Morales, D. and Vajda, I. (2001) Monte Carlo comparison of four normality tests using different entropy estimates. *Communs Statist. Simuln Computn*, **30**, 761–785.

Gauss, C. F. (1823) Theoria combinationis observationum erroribus minimus obnoxiae, pars prior. *Comm. Soc. Reg. Sci. Gotting. Rec.*, **5**.

Gibson, G. (2002) Microarrays in ecology and evolution: a preview. *Molec. Ecol.*, **11**, 17–24.

Goddard, M. E. (2001) The validity of genetic models underlying quantitative traits. *Livstck Prod. Sci.*, **72**, 117–127.

Good, P. I. (2000) *Permutation Tests: a Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd edn. London: Springer.

Han, H., Bearss, D. J., Browne, L. W., Calaluce, R., Nagle, R. B. and Von Hoff, D. D. (2002) Identification of differentially expressed genes in pancreatic cancer cells using cDNA microarray. *Cancer Res.*, **62**, 2890–2896.

Hernan, R., Fasheh, R., Calabrese, C., Frank, A. J., Maclean, K. H., Allard, D., Barraclough, R. and Gilbertson, R. J. (2003) ERBB2 up-regulates *S100A4* and several other prometastatic genes in medulloblastoma. *Cancer Res.*, **63**, 140–148.

Hotelling, H. (1961) The behavior of some standard statistical tests under non-standard conditions. In *Proc. 4th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1 (ed. J. Neyman), pp. 319–359. Berkeley: University of California Press.

Karian, Z. A. and Dudewicz, E. J. (2000) *Fitting Statistical Distributions: the Generalized Lambda Distribution and Generalized Bootstrap Methods*. New York: CRC.

Kayo, T., Allison, D. B., Weindruch, R. and Prolla, T. A. (2001) Influences of aging and caloric restriction on the transcriptional profile of skeletal muscle from rhesus monkeys. *Proc. Natn. Acad. Sci. USA*, **98**, 5093–5098.

Kerr, M. K. and Churchill, G. A. (2001) Statistical design and the analysis of gene expression microarray data. *Genet. Res.*, **77**, 123–128.

Lee, C. K., Weindruch, R. and Prolla, T. A. (2000) Gene-expression profile of the ageing brain in mice. *Nat. Genet.*, **25**, 294–297.

Lee, J.-M., Williams, M. E., Tingey, S. V. and Rafalski, J. A. (2002) DNA array profiling of gene expression changes during maize embryo development. *Funct. Integr. Genom.*, **2**, 13–27.

Li, W., Liang, X., Leu, J. I., Kovalovich, K., Ciliberto, G. and Traub, R. (2001) Global changes in interleukin-6-dependent gene expression patterns in mouse livers after partial hepatectomy. *Hepatology*, **6**, 1377–1386.

Li, Y., Li, Y., Tang, R., Xu, H., Qiu, M., Chen, Q., Chen, J., Fu, Z., Ying, K., Xie, Y. and Mao, Y. (2002) Discovery and analysis of hepatocellular carcinoma genes using cDNA microarrays. *J. Cancer Res. Clin. Oncol.*, **128**, 369–379.

López, I. P., Marti, A., Milagro, F. I., Zulet, M., Moreno-Aliaga, M. J., Martinez, J. A. and De Miguel, C. (2003) DNA microarray analysis of genes differentially expressed in diet-induced (cafeteria) obese rats. *Obes. Res.*, **11**, 188–194.

Mallows, C. L. (1956) Generalizations of Tchebycheff's inequalities (with discussion). *J. R. Statist. Soc.* B, **18**, 139–176.

Miyazawa, K., Mori, A., Yamamoto, K. and Okudaira, H. (1998) Transcriptional roles of CCAAT/enhancer binding protein-beta, nuclear factor-kappaB, and C-promoter binding factor 1 in interleukin (IL)-1beta-induced IL-6 synthesis by human rheumatoid fibroblast-like synoviocytes. *J. Biol. Chem.*, **273**, 7620–7627.

Mountz, J. D., Hsu, H. C., Matsuki, Y. and Zhang, H. G. (2001) Apoptosis and rheumatoid arthritis: past, present, and future directions. *Curr. Rheum. Rep.*, **3**, 70–78.

Mountz, J. D. and Zhang, H. G. (2001) Regulation of apoptosis of synovial fibroblasts. *Curr. Dir. Autoimmun.*, **3**, 216–239.

Myers, C., Charboneau, A., Cheung, I., Hanks, D. and Boudreau, N. (2002) Sustained expression of homeobox D10 inhibits angiogenesis. *Am. J. Pathol.*, **161**, 2009–2016.

Phadtare, S., Kato, I. and Inouye, M. (2002) DNA microarray analysis of the expression profile of *escherichia coli* in response to treatment with 4,5-Dihydroxy-2-Cyclopenten-1-One. *J. Bacteriol.*, **184**, 6725–6729.

Richmond, T. and Somerville, S. (2000) Chasing the dream: plant EST microarrays. *Curr. Opin. Plant Biol.*, **3**, 108–116.

SAS Institute (2001) *SAS/IML Users Guide (Release 8.2)*. Cary: SAS Institute.

Satterthwaite, F. E. (1949) An approximate distribution of estimates of variance components. *Biometr. Bull.*, **2**, 110–114.

Saw, J. G., Yang, M. C. K. and Mo, T. C. (1984) Chebyshev inequality with estimated mean and variance. *Am. Statistn*, **38**, 130–132.

Scheffé, H. (1959) *The Analysis of Variance*. New York: Wiley.

Šidák, Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. *J. Am. Statist. Ass.*, **62**, 626–633.

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

Wilcox, R. R. (1993) Robustness in ANOVA. In *Applied Analysis of Variance in the Behavioral Sciences* (ed. E. Edwards), pp. 345–374. New York: Dekker.

Wilcox, R. R. (1997) *Introduction to Robust Estimation and Hypothesis Testing*. San Diego: Academic Press.

Wu, T. D. (2001) Analysing gene expression data from DNA microarrays to identify candidate genes. *J. Pathol.*, **195**, 53–65.

Zimmerman, D. W. (1996) A note on homogeneity of variance of scores and ranks. *J. Exper. Educ.*, **64**, 351–362.