BIOINFORMATICS

Vol. 20 no. 18 2004, pages 3553–3564
doi:10.1093/bioinformatics/bth442

The Baumgartner-Weiß-Schindler test for the detection of differentially expressed genes in replicated microarray experiments

Markus Neuhäuser[1,*] and Roswitha Senske[2]

[1]Institute for Medical Informatics, Biometry and Epidemiology, University of Duisburg-Essen, Hufelandstr. 55, D-45122 Essen, Germany and [2]Department of Mathematics and Statistics, University of Otago, P.O. Box 56, Dunedin, New Zealand

Received on December 23, 2003; revised on June 30, 2004; accepted on July 26, 2004
Advance Access publication July 29, 2004

ABSTRACT

Motivation: An important application of microarray experiments is to identify differentially expressed genes. Because microarray data are often not distributed according to a normal distribution nonparametric methods were suggested for their statistical analysis. Here, the Baumgartner-Weiß-Schindler test, a novel and powerful test based on ranks, is investigated and compared with the parametric *t*-test as well as with two other nonparametric tests (Wilcoxon rank sum test, Fisher-Pitman permutation test) recently recommended for the analysis of gene expression data.

Results: Simulation studies show that an exact permutation test based on the Baumgartner-Weiß-Schindler statistic *B* is preferable to the other three tests. It is less conservative than the Wilcoxon test and more powerful, in particular in case of asymmetric or heavily tailed distributions. When the underlying distribution is symmetric the differences in power between the tests are relatively small. Thus, the Baumgartner-Weiß-Schindler is recommended for the usual situation that the underlying distribution is a priori unknown.

Availability: SAS code available on request from the authors.

Contact: markus.neuhaeuser@medizin.uni-essen.de

INTRODUCTION

DNA microarray technologies, such as cDNA arrays and oligonucleotide arrays, can be used to measure the expression of thousands of genes simultaneously. These technologies are rapidly becoming common laboratory tools and promise to revolutionize biological research. They are used in biomedical research, but also in other areas such as ecology and evolution (Gibson, 2002). Often the question is whether gene expression is different for two (or sometimes more) groups of organisms that differ with respect to a characteristic such as exposure to some environmental stimuli, genotype or age (Gadbury *et al*., 2003). In this paper, we consider the comparison of two groups in order to detect differentially expressed genes based on replicated measurements of expression levels of each gene.

From now on, the expression levels can refer to a summary measure of relative red to green channel intensity, a radioactive intensity or a summary difference of the perfect match and mis-match scores; furthermore, the gene expression levels may have been preprocessed using dimension reduction, normalization and data transformation (Pan, 2002).

Several authors pointed out that expression data from microarrys are often not distributed according to a normal distribution, even after some preprocessing (Hunter *et al*., 2001; Thomas *et al*., 2001; Pan, 2002; Craig *et al*., 2003; Giles and Kipling, 2003; Liu *et al*., 2003; Zhao and Pan, 2003). According to Thomas *et al*. (2001) the normality assumption is certainly inappropriate for a subset of genes despite any given transformation. Therefore, nonparametric tests were recommended for the analysis of microarrays (Troyanskaya *et al*., 2002; Gadbury *et al*., 2003; Xu and Li, 2003). The advantage of nonparametric methods is that no specific distribution has to be assumed.

Giles and Kipling (2003) applied the Shapiro-Wilks test to Affymetrix microarray expression data and showed that non-normal distributions are common (up to 46% of probe sets). However, Giles and Kipling (2003) argued that the deviations from normality are often modest and, therefore, they recommend parametric tests such as the *t*-test. Indeed, the *t*-test is quite robust, but its optimal power properties apply only if the observations are drawn from a normal distribution (Blair *et al*., 1980). When the assumption of normality is violated, a nonparametric test can be more powerful (see e.g. Hunter and May, 1993). Zimmerman and Zumbo (1993) demonstrated that the Wilcoxon rank sum test (equivalent to the Mann–Whitney *U* test) is more powerful than the *t*-test when outliers (unusual extreme data values) are present. As Liu *et al*. (2003) pointed out, outliers are an accepted fact of

*To whom correspondence should be addressed.

life when dealing with microarray data. Lönnstedt and Speed (2002) also noted that outliers occur frequently in microarray experiments. Moreover, it should be noted that expression data are often non-normal even after outliers have been removed (see e.g. Magusin, 2003).

In microarrays the sample sizes, i.e. the numbers of replications, are usually very small (Gadbury *et al*., 2003; Zhao and Pan, 2003). This fact is, according to Giles and Kipling (2003), an additional argument for parametric tests. However, historically, nonparametric tests have most often been recommended as a technique for dealing with small samples (Zimmerman and Zumbo, 1993, p. 483). According to Blair *et al*. (1980) 'in the small sample situation the t test might not be as robust to population non-normality as one would wish, and in this situation the Mann–Whitney test would be especially useful in controlling the Type I error rate'. Moreover, for relatively small sample sizes, Blair and Higgins (1980) found the Wilcoxon test to be more powerful than the t-test in many cases.

In the case of small sample sizes, a nonparametric test should be carried out as a permutation test. For a permutation test all possible permutations under the null hypothesis are generated and the test statistic is calculated for each permutation. The null hypothesis can then be accepted or rejected using the permutation distribution of the test statistic, the p-value being the probability of the permutations giving a value of the test statistic as supportive or more supportive of the alternative than the observed value (Manly, 1997; Good, 2000). Thus, inference is based upon how extreme the observed test statistic is relative to other values that could have been obtained under the null hypothesis. The alternative to a permutation test is to rely on an asymptotic distribution, e.g. on the asymptotic normality of the Wilcoxon rank sum which is appropriate when the sample size exceeds eight in each group (Troyanskaya *et al*., 2002). However, in the presence of ties (observations with identical values) the appropriateness of the asymptotic approximation depends on the number and on the distribution of the ties (Brunner and Munzel, 2002, p. 68).

The Wilcoxon rank sum test was recommended for the analysis of microarray data (Wu, 2001; Troyanskaya *et al*., 2002). Some authors (Troyanskaya *et al*., 2002; Xu and Li, 2003) considered the Fisher–Pitman permutation test, also called randomization test or nonparametric t-test, for the nonparametric analysis of microarrays. Recently, it was shown that an exact test based on the Baumgartner-Weiß-Schindler statistic is preferable to the Wilcoxon test (Neuhäuser, 2000, 2003). The Baumgartner-Weiß-Schindler statistic is also based on ranks. So far, it has not been compared to the Fisher–Pitman test. It is the aim of this paper to compare these tests for the detection of differentially expressed genes in replicated microarray experiments. The parametric t-test is included in this comparison because it seems to be the most frequently used test for identifying differentially expressed genes.

HYPOTHESIS TESTING FOR A SINGLE GENE

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ denote the independent observations regarding one gene for two groups to be compared, the sample means are $\bar{X}$ and $\bar{Y}$. Within groups, it is assumed that the observations are independent and identically distributed according to distribution functions F and G. In the location-shift model the distribution functions are the same except perhaps for a change in their locations; i.e., $F(t) = G(t - \theta)$ for every t, $-\infty < \theta < \infty$. The null hypothesis is $H_0 : \theta = 0$, whereas the alternative states $\theta \neq 0$.

Let $R_1 \leq \cdots \leq R_n$ ($H_1 \leq \cdots \leq H_m$) denote the combined-samples ranks of the X-values (Y-values) in increasing order of magnitude. The Wilcoxon statistic is defined as $W = \sum_{i=1}^{n} R_i$, i.e., W is the sum of the ranks of the observations from the first group.

The nonparametric statistic introduced by Baumgartner *et al*. (1998) is $B = \frac{1}{2} \cdot (B_X + B_Y)$, where

$$B_X = \frac{1}{n} \sum_{i=1}^{n} \frac{(R_i - [(m+n)/n] \cdot i)^2}{[i/(n+1)] \cdot (1 - i/(n+1)) \cdot [m(m+n)]/n}$$

and

$$B_Y = \frac{1}{m} \sum_{j=1}^{m} \frac{(H_j - [(m+n)/m] \cdot j)^2}{[j/(m+1)] \cdot (1 - j/(m+1)) \cdot [n(m+n)]/m}.$$

Large values of B support the alternative. This novel statistical test competes well with the Wilcoxon test and other nonparametric tests such as the Kolmogorov–Smirnov test. Baumgartner *et al*. (1998) demonstrated this using the asymptotic distribution of the test statistics. However, the asymptotic test based on B can have an inflated type I error rate in case of small sample sizes (Neuhäuser, 2000). Consequently, the exact test based on the permutation distribution of B proposed by Neuhäuser (2000) is more appropriate for the analysis of microarray data. When comparing exact tests, the one based on B is less conservative and more powerful than the Wilcoxon test for continuously distributed data (Neuhäuser, 2000) and in the presence of ties (Neuhäuser, 2003).

Troyanskaya *et al*. (2002) also mentioned the conservatism of the Wilcoxon rank sum test. The Baumgartner-Weiß-Schindler test is less conservative because the exact permutation distribution of B is less discrete than that of the rank sum W. Consider e.g. ten replications per group and no ties. Then, there are $\binom{2\times10}{10} = 184,756$ possible permutations. When calculating the statistic B and the rank sum W for all these permutations one obtains 11,833 different values for B, but only 101 different values for W. Consequently, the distribution of B has much more mass points, i.e. it is less discrete, than that of W. As a result, the Baumgartner-Weiß-Schindler test is less conservative than the Wilcoxon test and smaller p-values are possible. The latter point is particularly important for microarray analysis since the significance level may

Table 1. Type I error rates of the four tests for different sample sizes ($\alpha = 0.05$)

n	m	W test	B test	FP test					t test				
				Uniform	Normal	Cauchy	χ^2 (3 df)	Expon.	Uniform	Normal	Cauchy	χ^2 (3 df)	Expon.
5	5	0.0317	0.0476	0.049	0.048	0.045	0.048	0.051	0.056	0.050	0.019	0.042	0.043
6	6	0.0411	0.0498	0.049	0.046	0.046	0.045	0.050	0.053	0.047	0.018	0.039	0.039
7	7	0.0379	0.0490	0.051	0.052	0.049	0.050	0.051	0.054	0.052	0.019	0.045	0.042
8	8	0.0499	0.0499	0.052	0.051	0.054	0.051	0.051	0.054	0.051	0.022	0.046	0.042
9	9	0.0400	0.0499	0.047	0.051	0.049	0.050	0.050	0.048	0.050	0.019	0.044	0.042
10	10	0.0433	0.0500	0.051	0.048	0.051	0.050	0.050	0.053	0.048	0.021	0.046	0.045
8	5	0.0451	0.0497	0.054	0.050	0.049	0.043	0.053	0.056	0.050	0.023	0.038	0.043
9	7	0.0418	0.0500	0.051	0.048	0.049	0.050	0.051	0.053	0.047	0.021	0.045	0.043
10	5	0.0400	0.0500	0.049	0.052	0.045	0.049	0.050	0.052	0.051	0.025	0.044	0.043

have to be adjusted for multiple testing because many genes are considered simultaneously.

There are several equivalent test statistics for the Fisher–Pitman permutation test (Manly, 1997, pp. 15–16). One possibility is

$$FP = \left| \sum_{i=1}^{n} X_i - n \cdot \frac{n\bar{X} + m\bar{Y}}{n+m} \right|$$

as proposed for the two-sided alternative by Pitman (1937).

As the Wilcoxon test is based on ranks it does not use all available information, in contrast to the Fisher–Pitman permutation test. Nevertheless, the Wilcoxon test can be more powerful, as demonstrated by Keller-McNulty and Higgins (1987), van den Brink and van den Brink (1989) and Tanizaki (1997) for asymmetrical and heavily tailed distributions. Rasmussen (1986) showed that the Wilcoxon test outperforms the Fisher–Pitman permutation test in case of contaminated normal distributions; such a mixture of normals was also considered in the simulations of Xu and Li (2003).

The articles mentioned above (Rasmussen, 1986; Keller-McNulty and Higgins, 1987; van den Brink and van den Brink, 1989; Tanizaki, 1997) considered continuous distributions only. In this paper, we also investigate the power in the presence of ties. Furthermore, the Baumgartner-Weiß-Schindler test, a novel and powerful nonparametric test based on ranks, is included in the comparison.

SIMULATION STUDY

The different tests were compared in a Monte Carlo simulation study performed using SAS version 8.2; 10,000 simulation runs were generated for each configuration. With the exception that the parametric t test is based on the t distribution, the permutation distributions, not the asymptotic distributions, of the test statistics were used for inference. For sample sizes larger than $n = m = 10$ the number of possible permutations is very large. In this case, the permutation tests were performed based on simple random samples of 100,000 permutations.

Table 2. Simulated power of the four tests for different distributions ($n = m = 10$, $\alpha = 0.05$)

$\tilde{\theta}$[a]	Test	Uniform	Normal	Cauchy	χ^2(3 df)	Expon.
0.5	*FP*	0.15	0.19	0.13	0.17	0.15
	W	0.13	0.17	0.19	0.18	0.19
	B	0.13	0.18	0.22	0.20	0.21
	t	0.16	0.18	0.07	0.15	0.10
1.0	*FP*	0.47	0.56	0.29	0.47	0.41
	W	0.41	0.51	0.47	0.52	0.49
	B	0.42	0.52	0.55	0.58	0.56
	t	0.48	0.55	0.19	0.46	0.29
1.5	*FP*	0.83	0.89	0.43	0.76	0.67
	W	0.75	0.85	0.69	0.80	0.75
	B	0.76	0.86	0.77	0.85	0.82
	t	0.84	0.89	0.32	0.76	0.54

[a] $\theta = f \cdot \tilde{\theta}$ with $f = 4/15$ (uniform distribution), $f = 0.7$ (exponential distribution), $f = 1$ (normal distribution), and $f = 2$ (Cauchy and χ^2). The values of f were chosen on empirical grounds in order to obtain powers of comparable size.

For smaller sample sizes all permutations were considered for the rank-based tests. For the Fisher–Pitman test all permutations were considered if there were not more than 100,000 possible permutations; if there were more, an approximate Fisher–Pitman test was carried out, using 100,000 randomly selected permutations.

Distributions with different properties were used to simulate data: uniform distribution on (0,1), i.e. a symmetric distribution with short tails; standard normal distribution, i.e. a symmetric distribution with medium tails; Cauchy distribution, i.e. a symmetric distribution with heavy tails; and two asymmetric distributions: χ^2 distribution with three degrees of freedom (df), and exponential distribution with scale parameter $\lambda = 1$. These distributions, together with the investigated mixtures of normal distributions (see below), represent different types of distributions that are possible for actual gene expression data. Therefore, the power observed in the simulations is likely to transfer to real data applications.

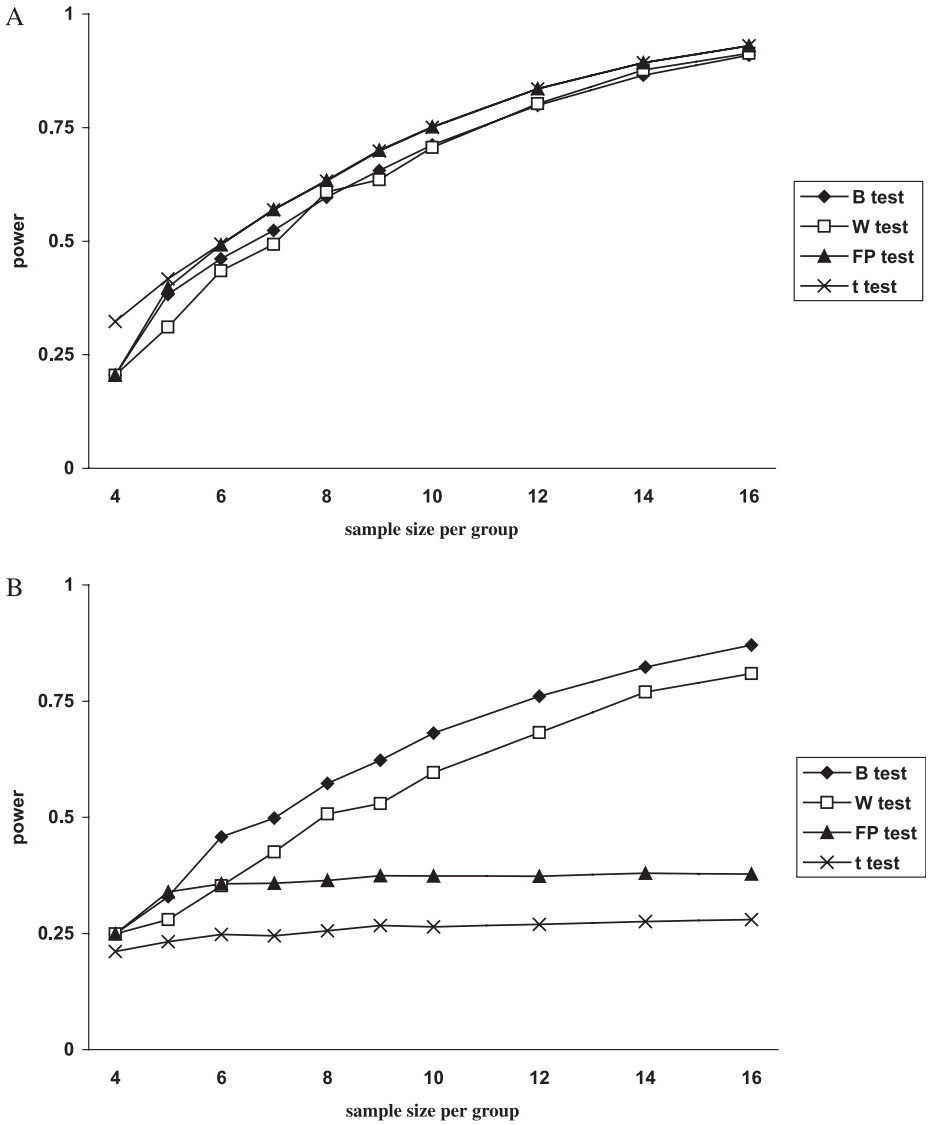

Fig. 1. Simulated power of the four tests for (**A**) the normal distribution with variance 1 (location shift: $\theta = 1.25$, $\alpha = 0.05$) and (**B**) the Cauchy distribution (location shift: $\theta = 2.5$, $\alpha = 0.05$)

Hereafter, the following abbreviations will be used: *FP* test for the Fisher–Pitman permutation test, *W* test for the Wilcoxon test, and *B* test for the test based on *B*.

Type I error rates are presented in Table 1. The type I error rate of the rank tests can be derived analytically since it depends on the ranks only. The type I error rates of the *t* and the *FP* tests were simulated. The *FP* and *B* tests have a type I error rate very close to the nominal significance level α, even for small sample sizes. The *W* test, however, is conservative. The *t* test can have a type I error rate close to α, but it can be very conservative, too, as it is in case of the Cauchy distribution. As exact permutation tests, all three nonparametric tests guarantee a type I error rate less than or equal to α. Therefore, a test statistic can be chosen purely on the basis of power (Kennedy, 1995).

The simulated power is given in Table 2. For the skewed and heavily tailed distributions the *B* test outperforms the other three tests. For the uniform and the normal distributions the *t* and the *FP* tests are more powerful than the other tests. However, in these cases the difference between the tests is much smaller than for other distributions such as Cauchy or exponential. Consequently, the simulation results indicate that the *B* test is a good choice when the underlying distribution functions are unknown as they usually are. This conclusion also holds for other sample sizes, see the results presented in Figure 1. However, for small sample sizes such as 4 or 5 per group the differences in power between the tests are negligible.

So far, only continuous distributions were considered. But, in practice, ties occur frequently in a variety of settings

Table 3. Type I error rates of the four tests in the presence of ties ($n = m = 10, \alpha = 0.05$)

No. of ties	*W* test	*B* test	*FP* test					*t* test				
			Uniform	Normal	Cauchy	χ^2 (3 df)	Expon.	Uniform	Normal	Cauchy	χ^2 (3 df)	Expon.
1	0.0452	0.0500	0.050	0.048	0.050	0.050	0.050	0.053	0.047	0.021	0.046	0.046
2	0.0499	0.0500	0.051	0.047	0.050	0.050	0.050	0.052	0.047	0.021	0.046	0.045
3	0.0499	0.0499	0.050	0.048	0.050	0.050	0.052	0.052	0.048	0.021	0.046	0.047

Table 4. Simulated power of the four tests for different distributions in the presence of ties ($n = m = 10, \alpha = 0.05$)

No. of ties	Test	Uniform ($\theta = 6/15$)	Normal ($\theta = 1.5$)	Cauchy ($\theta = 3$)	χ^2 (3 df) ($\theta = 3$)	Expon. ($\theta = 1.05$)
1	*FP*	0.83	0.89	0.43	0.75	0.67
	W	0.75	0.86	0.69	0.80	0.75
	B	0.76	0.85	0.77	0.84	0.82
	t	0.83	0.89	0.32	0.75	0.66
2	*FP*	0.83	0.89	0.43	0.75	0.67
	W	0.77	0.87	0.71	0.81	0.77
	B	0.76	0.85	0.77	0.84	0.82
	t	0.83	0.89	0.32	0.75	0.66
3	*FP*	0.83	0.88	0.43	0.75	0.67
	W	0.76	0.87	0.71	0.81	0.77
	B	0.75	0.85	0.77	0.84	0.82
	t	0.83	0.88	0.32	0.75	0.66

including microarrays. Even when the underlying distribution is continuous rounding can lead to ties. In addition, data modifications can create ties. For example, in the microarray analysis presented by Tschentscher *et al*. (2003) expression levels below 50 were set to 50 prior to performing the *W* test. Therefore, the behavior of the tests in the presence of ties is of interest. Unfortunately, as mentioned above, non-continuous distributions were previously not considered in the comparison of *FP* versus *W* test. In the presence of ties the usual way of dealing with these values is to assign average ranks, the statistics *W* and *B* can be calculated in that way. The asymptotic BWS test can have an inflated type I error rate in this case (Neuhäuser, 2002). However, permutation tests as investigated here can be applied whether or not ties occur (Good, 2000).

Data sets with ties were generated as follows: First, data were simulated according to continuous distribution functions. In the second step, ties were created. In the case of one tie, the values corresponding to the ranks 5 and 6 were replaced by the average of these two values. To create two tied groups, the values corresponding to the ranks 10 and 11 were also replaced by their average. For three ties, the values corresponding to the ranks 15 and 16 were averaged to create a further tie.

As the results in Table 3 indicate ties affect the type I error rate of the tests only marginally. The only difference is the larger type I error rate of the *W* test in case of two and three ties. The reason is that the type I error rate of the *W* test heavily depends on the location of the few mass points of the very discrete distribution of *W*. The power displayed in Table 4 is also similar to the results for continuous distributions. The slightly increased power of the *W* test for two and three ties can be explained by the larger type I error rate.

As mentioned above, Rasmussen (1986) showed the superiority of the *W* test to the *FP* test for contaminated normal distributions. We also investigated some of these distributions that are defined as follows: The data are standard normal with probability 0.7, and with probability 0.3 they are normally distributed with mean 5 and standard deviation 0.5 (CN1) and 4 (CN2), respectively. An additional mixture was used in order to reflect the situation that one or two outliers are present. The data of this distributions (CN3) are standard normal with probability 0.9, and with probability 0.1 they are normally distributed with mean 10 and standard deviation 1. The results given in Table 5 and Figure 2 show that the *W* test is indeed preferable to the *FP* test, but the *B* test is much more powerful than the other tests. Again, ties affect the power only marginally. The results for the distribution CN3 (Figure 2B) confirm the statement mentioned in the *Introduction* that rank-based tests are especially appropriate in the presence of outliers. The *B* and the *W* test are much more powerful than the *t* and the *FP* tests, the *B* test being the most powerful.

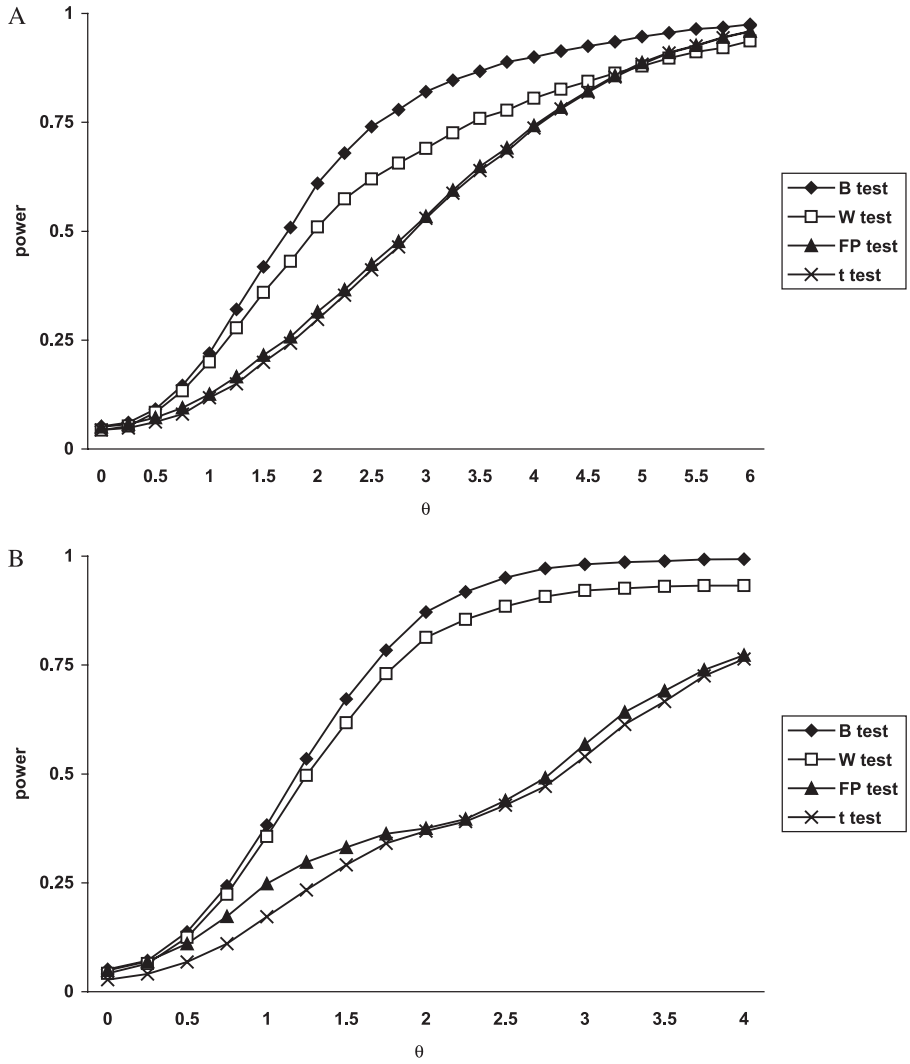

Fig. 2. Simulated power of the four tests for the contaminated normal distribution (**A**) CN2 ($n = m = 10, \alpha = 0.05$); (**B**) CN3 ($n = m = 10$, $\alpha = 0.05$)

Table 5. Simulated power of the four tests for the contaminated normal distribution CN1 ($\theta = 2.5, n = m = 10, \alpha = 0.05$)

No. of ties	*FP* test	*W* test	*B* test	*t* test
0	0.56	0.56	0.69	0.57
1	0.56	0.56	0.69	0.56
2	0.56	0.61	0.68	0.56
3	0.56	0.60	0.67	0.57

Application to actual data

We applied the different tests to data from microarray experiments. First, we used cDNA data from a comparison of two types of breast cancer (Hedenfalk *et al*., 2001): for 3226 genes there are seven replicates from patients with germ-line mutations of *BRCA1* and eight replicates regarding *BRCA2*. Second, the different tests were applied to the oligonucleotide microarray data from Huang *et al*. (2001). In this comparison of normal thyroid and papillary tumor tissues more than 12,000 genes were investigated. We arbitrarily selected the first-listed 2000 genes for the analyses presented here, there are eight replicates per group. In the case of this sample size ($n = m = 8$) neither test is conservative (see Table 1). For both data sets we obtained the data from http://microarray.cpmc.Columbia.edu/pavlidis/pub/gxrep (see Pavlidis *et al*., 2003). In addition, we used data from patients with uveal melanomas with and without monosomy 3 (Tschentscher *et al*., 2003; see http://www.uni-essen.de/humangenetik/download). The sample size in this microarray experiment is 10 per group. As mentioned above, expression levels below 50 were set to 50. However,

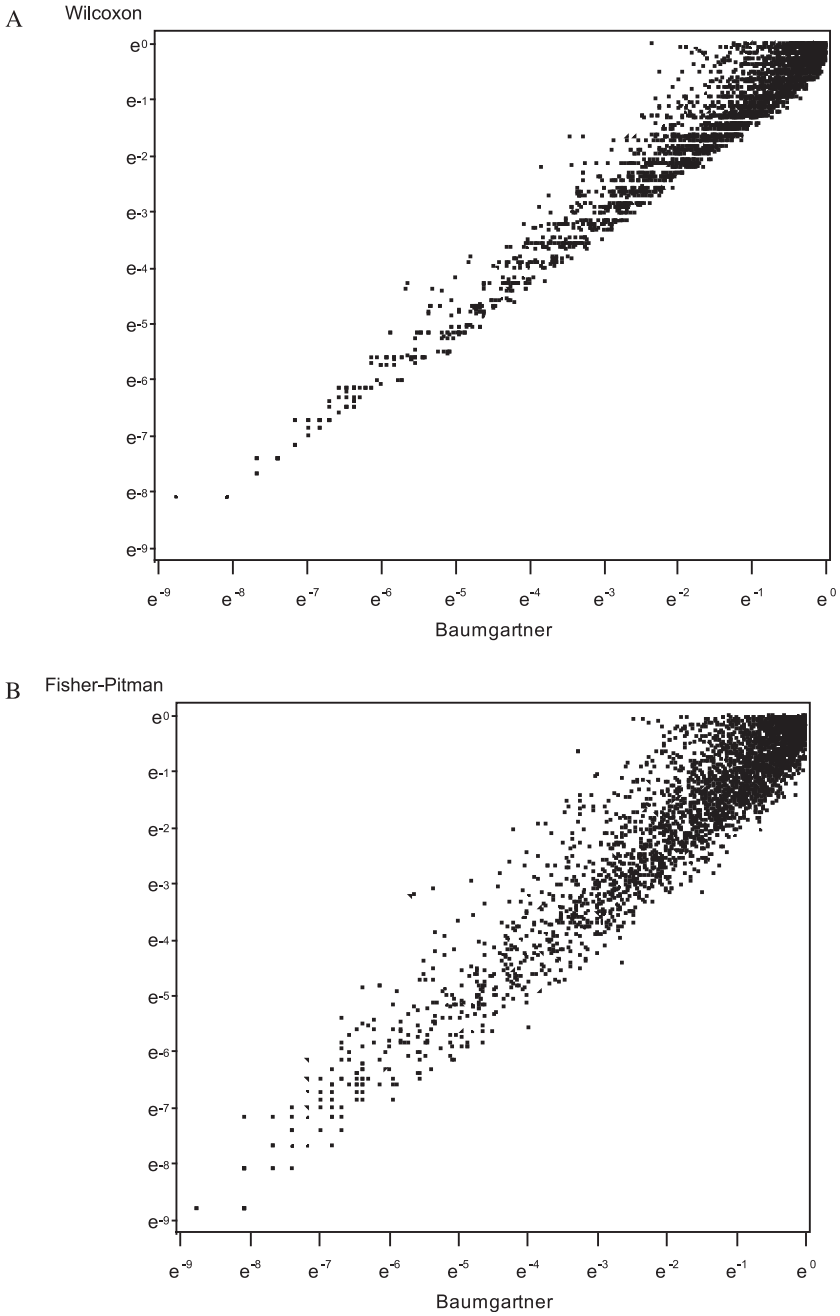

Fig. 3. *P*-values of Baumgartner-Weiß-Schindler test *B* and (**A**) the Wilcoxon test *W* and (**B**) the Fisher–Pitman permutation test *FP* for 3226 genes from the cDNA microarray of Hedenfalk *et al*. (2001).

we considered the first-listed 2000 genes for which this modification was not necessary.

The Figures 3– 5 show *p*-values. Although, for all data sets, the *p*-values of the different tests are often similar, the two tests *W* and *FP* can have much larger *p*-values than the *B* test. For the following analyses we selected, for each of the three data sets, the 100 genes with the strongest difference in expression, i.e. the 100 genes with the smallest *p*-values of the *B* test. Within these sets of genes there are 97, 93, and 92 (Hedenfalk *et al*., 2001; Huang *et al*., 2001; Tschentscher *et al*., 2003) of the genes with the 100 smallest *W* test *p*-values. Figure 6 shows that the same genes have the smallest *p*-values irrespective whether the *B* or the *W* test is applied. That means that the *B* test, in comparison to the *W* test, does not detect distinct genes as differentially expressed. However, because of its higher power (see above) the *B* test likely identifies

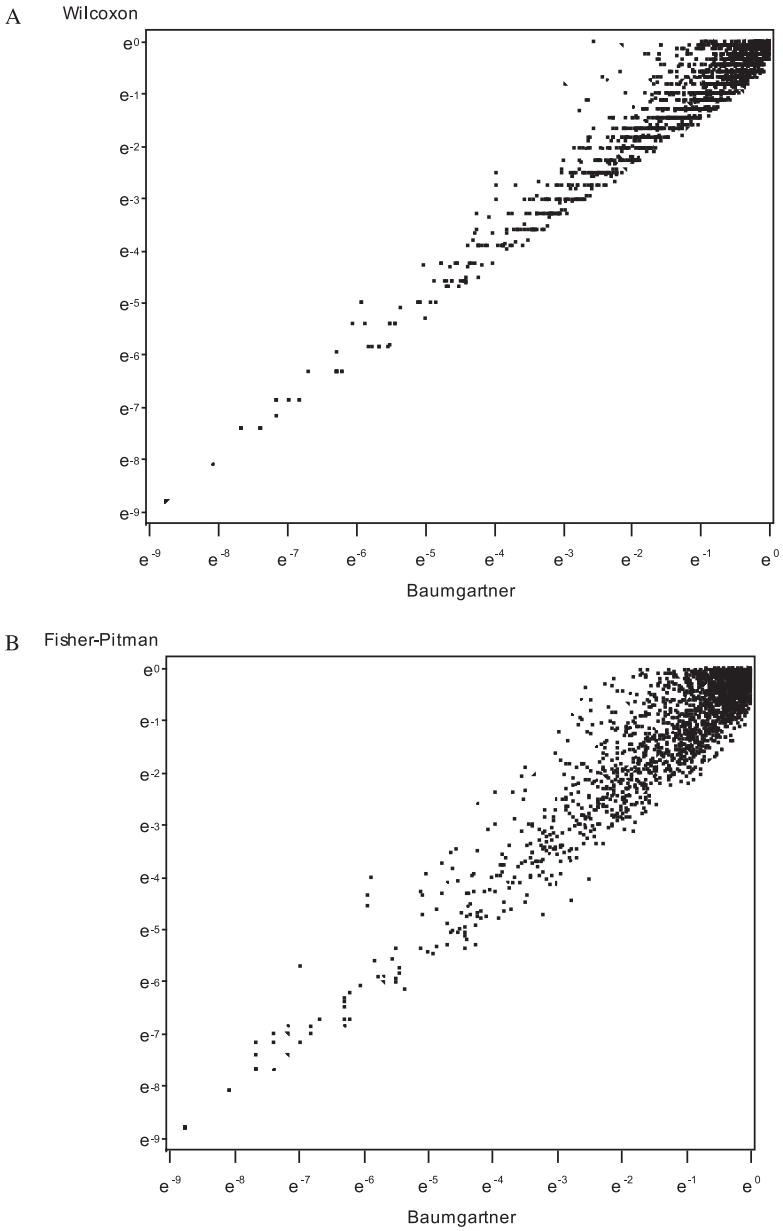

Fig. 4. *P*-values of the Baumgartner-Weiß-Schindler test *B* and (**A**) the Wilcoxon test *W* and (**B**) Fisher–Pitman permutation test *FP* for 2000 genes from the oligonucleotide microarray of Huang *et al*. (2001).

more genes as differentially expressed. Between the *FP* test (or *t*-test) and a rank test the differences are slightly larger. The numbers of genes out of those with the 100 smallest *p*-values within the set of the 100 genes with the smallest *B* test *p*-values indicate this. These numbers are 85, 88 and 91 for the *FP* test, and 74, 85 and 83 for the *t*-test (Hedenfalk *et al*., 2001; Huang *et al*., 2001; Tschentscher *et al*., 2003).

In addition, we consider the Affymetrix spike-in experiment. Because transcripts were spiked-in at known concentrations (Irizarry *et al*., 2003a), the truth is known for these data. We applied the data available at www.affymetrix.com/analysis/download_center2.affx to the robust multi-array analysis (RMA, Irizarry *et al*., 2003b) before the different tests were performed. Since a two-sample comparison is investigated here, the experiments M to T and the series 4, 6 and 8 are considered. Thus, we have two groups with 12 values each. According to Cope *et al*. (2004) we regard 16 spiked-in probe sets as differentially expressed. In total, there are 12,626 probe sets. Table 6 displays the *p*-values of the 16 transcripts with differences and the corresponding numbers of smaller or equal *p*-values within all probe sets.

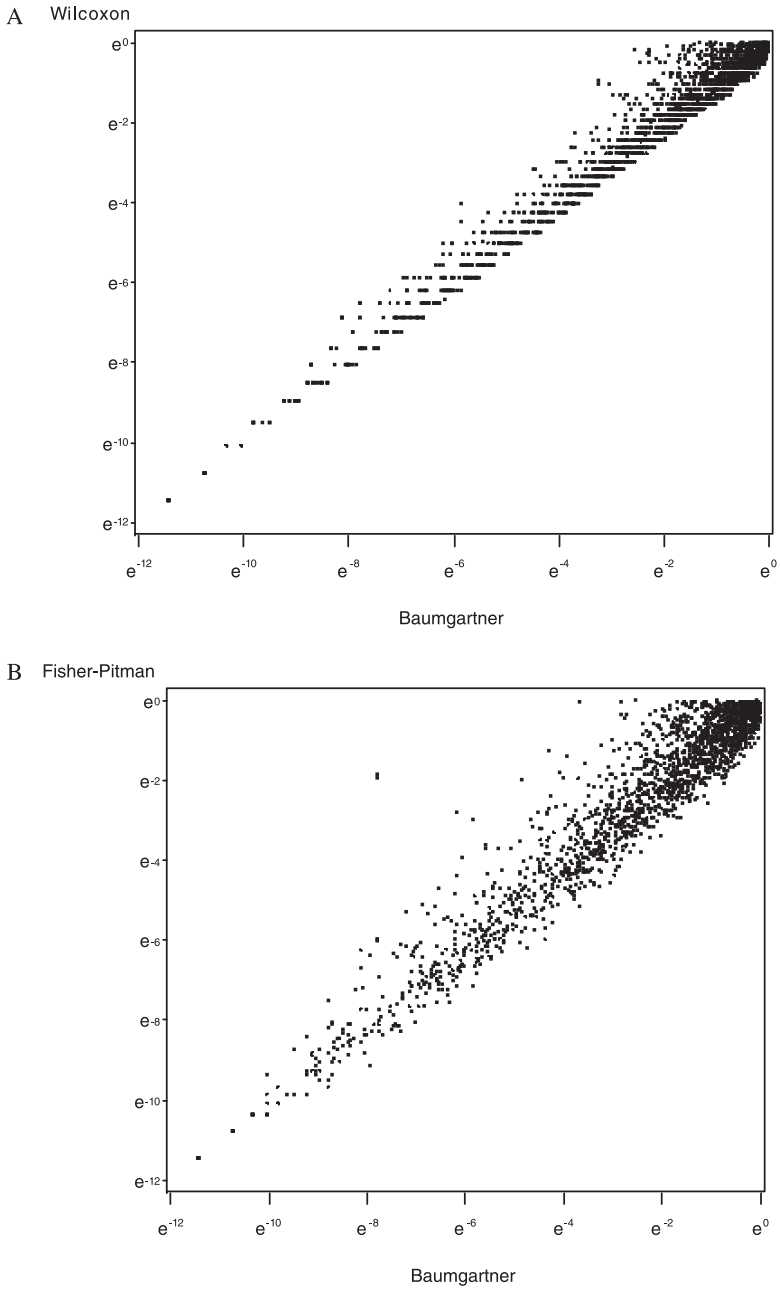

Fig. 5. *P*-values of the Baumgartner-Weiß-Schindler test *B* and (**A**) the Wilcoxon test *W*; (**B**) the Fisher–Pitman permutation test *FP* for 2000 genes from the oligonucleotide microarray of Tschentscher *et al*. (2003).

As this table shows there are very marginal differences only between the tests.

DISCUSSION

Nonparametric tests such as the Wilcoxon rank sum test were recommended for the analysis of microarray data. As mentioned above, no specific distribution has to be assumed for nonparametric methods. Disadvantages of these tests are that they can be conservative and computer-intensive. However, the presented test based on the Baumgartner-Weiß-Schindler statistic is less prone to the first problem. And the second issue is less relevant now due to faster algorithms (see e.g. Good, 2000, chap. 13) and the advent of high-speed PCs. Furthermore, one can carry out a permutation test based on a random sample out of the possible permutations.

Previous research demonstrated that the Wilcoxon rank sum test is more powerful than the *t* and the *FP* test when the data follow an asymmetric and/or a heavily tailed distribution.

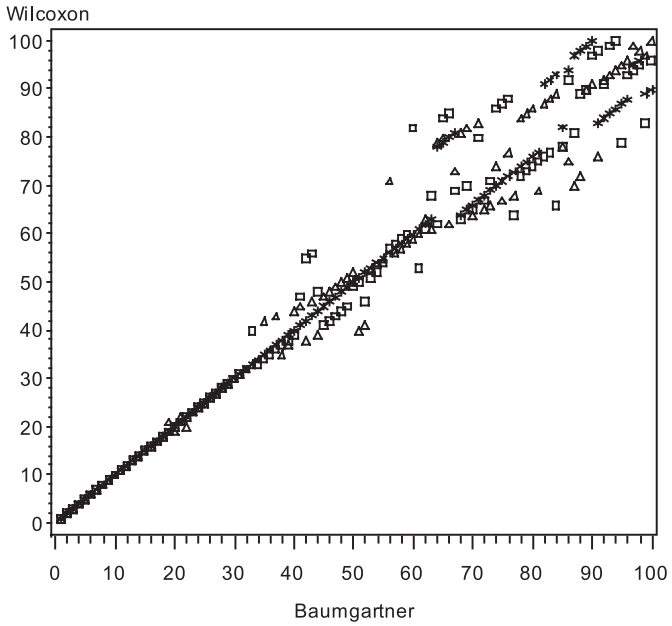

Fig. 6. The ranks of the p-values of the Baumgartner-Weiß-Schindler test B and the Wilcoxon test W test for the 100 genes with the smallest p-values of the B test, for each of the three data sets [Δ: Hedenfalk, $\Box$: Huang, $*$: Tschentscher].

Table 6. P-values of the 16 transcripts with differences in the Affymetrix spike-in experiment and the corresponding numbers of smaller or equal p-values within all probe sets

Transcripts with identical p-values	B test p-value	# of smaller or equal p-values	W test p-value	# of smaller or equal p-values	FP test p-value	# of smaller or equal p-values
33818_at, 40322_at	0.00000074	14	0.00000074	14	0.00000037	2
546_at, 684_at, 1024_at, 1091_at, 36085_at, 36202_at, 36311_at, 36889_at, 37777_at, 38734_at, 39058_at	0.00000074	14	0.00000074	14	0.00000074	13
1708_at	0.00001183	19	0.00001405	19	0.00000740	18
407_at	0.00001849	20	0.00002219	20	0.00001997	20
1597_at	0.00180759	50	0.00182978	49	0.00161899	49

Note that outliers, common in gene expression, can lead to heavy-tailed distributions (Wu, 2001; Liu *et al*., 2003). We demonstrated that the advantage of a rank test can be more pronounced when a novel statistic, the Baumgartner-Weiß-Schindler statistic B, is used instead of the rank sum. Since the Baumgartner-Weiß-Schindler test is, if at all, only marginally less powerful than the t or the FP test for symmetric distributions, this test can be recommended in case of an a priori unknown distribution, a situation quite common in practice. As the test we recommend is based on ranks, it also has the advantage that it is less sensitive to outliers.

Our approach for the identification of differentially expressed genes is to consider a univariate testing problem for each gene. A correction for the multiplicity of genes is a subsequent step, which, like the previous step of normalizing the data, outside the scope of this paper. A common approach to the multiplicity problem is to consider a procedure for testing the genes simultaneously for differential expression with the test on an individual gene being implied in the simultaneous test. For such a procedure different proposals have been made recently. Methods based on the p-values of the tests from individual genes were introduced by Zaykin *et al*. (2002), Storey and Tibshirani (2003), and Dudbridge and Koeleman (2003).

Only two-sided alternative hypotheses are considered here; one-sided alternatives can be handled in a similar manner. Due to the squares in the numerators of B_X and B_Y, the statistic B is not suitable for a one-sided test, but a modification with absolute values instead of squares has been proposed for

one-sided test problems (Neuhäuser, 2001). Using this modification, the resultant test has been compared to the one-sided W, FP and t-tests. The results of this comparison are analogous to the outcomes presented in this paper for testing the two-sided alternative.

ACKNOWLEDGEMENTS

We would like to thank three anonymous reviewers for helpful comments and suggestions.

REFERENCES

Baumgartner,W., Weiß,P. and Schindler,H. (1998) A nonparametric test for the general two-sample problem. *Biometrics*, **54**, 1129–1135.

Blair,R.C. and Higgins,J.J. (1980) A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* statistic under various non-normal distributions. *J. Educational Statistics*, **5**, 309–335.

Blair,R.C., Higgins,J.J. and Smitley,D.S. (1980) On the relative power of the *U* and *t* tests. *British J Mathematical Statistical Psychol.*, **33**, 114–120.

Brunner,E. and Munzel,U. (2002) *Nichtparametrische Datenanalyse*. Springer, Berlin, Germany.

Dudbridge,F. and Koeleman,B.P.C. (2003) Rank truncated product of *P*-values, with application to genomewide association scans. *Genet. Epidemiol.*, **25**, 360–366.

Cope,L.M., Irizarry,R.A., Jaffee,H.A., Wu,Z. and Speed,T.P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics*, **20**, 323–331.

Craig,B.A., Black,M.A. and Doerge,R.W. (2003) Gene expression data: the technology and statistical analysis. *J. Agric. Biol. Environ. Stat.*, **8**, 1–28.

Gadbury,G.L., Page,G.P., Heo,M., Mountz,J.D. and Allison,D.B. (2003) Randomization tests for small samples: an application for genetic expression data. *Appl. Stat.*, **52**, 365–376.

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

Giles,P.J. and Kipling.D. (2003) Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics*, **19**, 2254–2262.

Good,P.I. (2000) *Permutation Tests*, 2nd edn. Springer, New York, NY.

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

Huang,Y., Prasad,M., Lemon,W.J., Hampel,H., Wright,F.A., Kornacker,K., LiVolsi,V., Frankel,W., Kloos,R.T., Eng,C., Pellegata,N.S. and de la Chapelle,A. (2001) Gene expression in papillary throid carcinoma reveals highly consistent profiles. *Proc. Natl Acad. Sci. USA*, **98**, 15044–15049.

Hunter,L., Taylor,R.C., Leach,S.M. and Simon,R. (2001) GEST: a gene expression search tool based on a novel Bayesian similarity metric. *Bioinformatics*, **17** (Suppl. 1), S115-S122.

Hunter,M.A. and May,R.B. (1993) Some myths concerning parametric and nonparametric tests. *Can. Psychol.*, **34**, 384–389.

Irizarry,R.A., Bolstad,B.M., Collon,F., Cope,L.M., Hobbs,B. and Speed,T.P. (2003a) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Irizarry,R.A., Hobbs,F.C.B., Beaxer-Barclay,Y., Antonellis,K., Scherf,U. and Speed,T.P. (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Keller-McNulty,S. and Higgins,J.J. (1987) Effect of tail weight on power and type-I error of robust permutation tests for location. *Commun. Stat. – Simulat. Comput.*, **16**, 17–35.

Kennedy,P.E. (1995) Randomization tests in econometrics. *J. Business Economic Statistics*, **13**, 85–94.

Liu,L., Hawkins,D.M., Ghosh,S. and Young,S.S. (2003) Robust singular value decomposition analysis of microarray data. *Proc. Natl Acad. Sci. USA*, **100**, 13167–13172.

Lönnstedt,I. and Speed,T. (2002) Replicated microarray data. *Stat. Sin*, **12**, 31–46.

Magusin,A. (2003) Complementary techniques of clustering and composite pattern analysis to *Saccharomyces cerevisiae* gene expression. *Appl Bioinformatics*, **2** (3 Suppl), S37–S46.

Manly,B.F.J. (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edn. Chapman and Hall, London, U.K.

Neuhäuser,M. (2000) An exact two-sample test based on the Baumgartner-Weiß-Schindler statistic and a modification of Lepage's test. *Commun. Stat. – Theory Meth.*, **29**, 67–78.

Neuhäuser,M. (2001) One-sided two-sample and trend tests based on a modified Baumgartner-Weiß-Schindler statistic. *J. Nonparametr Stat.*, **13**, 729–739.

Neuhäuser,M. (2002) The Baumgartner-Weiß-Schindler test in the presence of ties (letter to the editor). *Biometrics*, **58**, 250.

Neuhäuser,M. (2003) A note on the exact test based on the Baumgartner-Weiß-Schindler statistic in the presence of ties. *Comput. Stat. Data Anal.*, **42**, 561–568.

Pan,W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **18**, 546–554.

Pavlidis,P., Li,Q. and Noble,W.S. (2003) The effect of replication on gene expression microarray experiments. *Bioinformatics*, **19**, 1620–1627.

Pitman,E.J.G. (1937) Significance tests which may be applied to samples from any populations. *Suppl. J Royal Stat. Soc.*, **4**, 119–130.

Rasmussen,J.L. (1986) An evaluation of parametric and nonparametric tests on modified and non-modified data. *Brit. J. Mathemat. Stat. Psychol.*, **39**, 213–220.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad Sci. USA*, **100**, 9440–9445.

Tanizaki,H. (1997) Power comparison of non-parametric tests: small-sample properties from Monte Carlo experiments. *J. Appl. Stat.*, **24**, 603–632.

Thomas,J.G., Olson,J.M., Tapscott,S.J. and Zhao,L.P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res.*, **11**, 1227–1236.

Troyanskaya,O.G., Garber,M.E., Brown,P.O., Botstein,D. and Altman,R.B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454–1461.

Tschentscher,F., Hüsing,J., Hölter,T., Kruse,E., Dresen,I.G., Jöckel,K.-H., Anastassiou,G., Schilling,H., Bornfeld,N., Horsthemke,B., Lohmann,D.R. and Zeschnigk,M. (2003) Tumor classification based on gene expression profiling shows that uveal melanomas with and without monosomy 3 represent two distinct entities. *Cancer Res.*, **63**, 2578–2584.

van den Brink,W.P. and van den Brink,S.G.J. (1989) A comparison of the power of the *t* test, Wilcoxon's test, and the approximate permutation test for the two-sample location problem. *Brit. J. Mathemat. Stat. Psychol.*, **42**, 183–189.

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

Xu,R. and Li,X. (2003) A comparison of parametric versus permutation methods with applications to general and temporal microarray gene expression data. *Bioinformatics*, **19**, 1284–1289.

Zaykin,D.V., Zhivotovsky,L.A., Westfall,P.H. and Weir,B.S. (2002) Truncated product method for combining *P*-values. *Genet. Epidemiol.*, **22**, 170–185.

Zhao,Y. and Pan,W. (2003) Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments. *Bioinformatics*, **19**, 1046–1054.

Zimmerman,D.W. and Zumbo,B.D. (1993) The relative power of parametric and nonparametric statistical methods. In Keren,G. and Lewis,C. (eds), *A handbook for data analysis in the behavioral sciences: methodological issues*. Lawrence Erlbaum Associates, Hillsdale, NJ, pp. 481–517.