BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 10 2011, pages 1377–1383
doi:10.1093/bioinformatics/btr152

Gene expression Advance Access publication March 26, 2011

Comparison of global tests for functional gene sets in two-group designs and selection of potentially effect-causing genes

Klaus Jung, Benjamin Becker, Edgar Brunner and Tim Beißbarth*

Department of Medical Statistics, University Medical Center Göttingen, D-37099 Göttingen, Germany

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: An important object in the analysis of high-throughput genomic data is to find an association between the expression profile of functional gene sets and the different levels of a group response. Instead of multiple testing procedures which focus on single genes, global tests are usually used to detect a group effect in an entire gene set. In a simulation study, we compare the power and computation times of four different approaches for global testing. The applicability of one of these methods to gene expression data is demonstrated for the first time. In addition, we propose an algorithm for the detection of those genes which might be responsible for a group effect.

Results: We could detect that the power of three of the approaches is comparable in many settings but considerable differences were detected in the computation times. Our proposed gene selection algorithm was able to detect potentially effect-causing genes in artificial sets with high power when many genes were altered with a small effect, while classical multiple testing was more powerful when few genes were altered with a large effect.

Availability: An R-package called 'RepeatedHighDim' which implements our new global test procedures is made available from http://cran.r-project.org/.

Contact: tim.beissbarth@ams.med.uni-goettingen.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on January 21, 2011; revised on March 17, 2011; accepted on March 21, 2011

1 INTRODUCTION

A typical problem in genomic analysis is the comparison of gene expression levels recorded by DNA microarrays in two distinct groups of patients, for example therapy responders or non-responders. Usually, the comparison comprises the search for differentially expressed genes by means of multiple hypothesis testing (Dudoit *et al.*, 2003), potentially resulting in a list of significant genes. Multiple testing procedures conduct a statistical test for each single gene and adjust the resulting P-values for a desired error rate (e.g. the false discovery rate) to avoid too much false positives. The biological or medical interpretation of the resulting list of differentially expressed genes is, however, a difficult and tedious task. Furthermore, expression changes in single genes could easily be missed, for example, due to small sample size and limited power of the analysis. Therefore, it can be beneficial to search directly for significant group effects in gene sets, e.g. differential pathways or functional groups. Gene sets defined by signaling pathways (e.g. cell proliferation or apoptosis) or other certain functional groups (e.g. transcription factors, tyrosine kinase activity) are often in the special focus of disease research. This allows for a direct association between the particular gene set and the levels of the group response. This will not only make the biological interpretation of the data easier but also reduce the multiplicity problem to a more overseeable number of gene groups to test. Further, a global group effect might also become detectable via a group test when many genes within a group are just slightly differential but none of them is significant in an individual gene test by itself.

Many approaches for gene set enrichment analysis exist. In principle, they can be grouped in two different categories: (i) approaches that perform the individual significance test for each gene first and then test for enriched gene sets based on either a cutoff (Beißbarth and Speed, 2004) or on the gene order (Mootha *et al.*, 2003). (ii) Approaches that perform a global test for each gene set directly. In principle, these different categories test completely different null hypotheses and are appropriate for different biological questions. A more detailed categorization and discussions about the utility of these approaches can be found in Pavlidis *et al.* (2004), in Draghici (2007) and in Goeman and Bühlmann (2007). Here we will focus on the latter global test approaches. In particular, the purpose of this article is first to compare different global test methods for functional gene sets in two-group designs, and secondly to provide a method for detecting those genes (of a functional set) which might be responsible for a detected group effect.

A first suggestion for global testing in microarray experiments was given by Simon *et al.* (2003) who use a permutation-based approach in order to deal with the high dimensionality of gene expression data. Goeman *et al.* (2004) developed a global test ('*GlobalTest*') using a logistic regression approach, where the dichotomous response $Y \in \{1,2\}$ represents two clinical or biological groups and genes are the predictors. They study thus the hypothesis that the phenotype is independent from the gene expression X, i.e. $H_0: P[Y|X] = P[Y]$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$, $n \ll d$, is the matrix of normalized gene expression levels with d genes and n samples. The test statistic of this method employs the low-dimensional $(n \times n)$ covariance matrix between samples instead of the large $(d \times d)$ covariance matrix between the genes. Thus, the method avoids long computation times and does not require a large working space. Under the null hypothesis, the test statistic has, asymptotically, a normal distribution. However, for small sample sizes a permutation test is provided and recommended. Rocke *et al.* (2005) proposed a test that is based on the distribution of

*To whom correspondence should be addressed.

gene-wise test statistics. At first, the two-sample t-test is performed per gene, and than either the one-sample t-test or the one-sample Wilcoxon test is applied on the set of gene-wise test statistics. The idea of this procedure is that the distribution of the test statistics is biased under a global group effect, and that this bias is detected by the one-sample tests. The authors called their procedure test of test-statistics ('*ToTS*'). Further, an ANCOVA model for global testing ('*GlobalAncova*') was proposed by Mansmann and Meister (2005) and was extended by Hummel *et al*. (2008). In this ANCOVA model, the null hypothesis is tested that the gene expression is independent of the group level, i.e. $H_0: P[X|Y=1] = P[X|Y=2]$. In this method, the high dimensionality is handled by assembling the residual sums of squares (RSS) of gene-wise models to the RSS of a global model. The combined RSS are used in the test statistic which has, asymptotically, a χ^2-distribution under H_0. Like *GlobelTest*, *GlobalAncova* provides a permutation test for small sample sizes. Finally, Brunner (2009) presented a test for group effects in high-dimensional repeated measures data, where the repeated measures factor is allowed to have more levels than the number of available samples. This approach allows the covariance matrices to be of an arbitrary structure. Here, we demonstrate for the first time the applicability of this approach in the setting of global tests for gene sets by regarding the genes as the levels of the repeated measures factor. Since this test is still unnamed, we call it '*RepeatedHighDim*'. The test of Brunner (2009) does not use a permutation algorithm to assess the correlation between genes. Instead, the correlation information is incorporated using the representation theorem of a quadratic form (Mathai and Provost, 1992) and Box's approximation (Box, 1954). Thus, it is likewise not necessary to estimate the large covariance matrix between genes. This new approach is further detailed in the Section 2.

We compare the attained significance levels, the power and computation times of *GlobalTest*, *ToTS*, *GlobalAncova* and *RepeatedHighDim* in simulation studies, considering different settings of covariance matrices, sample sizes and dimensions. As described by Meinshausen (2008), after global testing it can be important to step downwards to the level of individual features. Therefore, we additionally propose a procedure for detecting the relevant genes of a potential group effect. Lists of genes detected with this procedure are compared with lists resulting from of a classical procedure of multiple hypothesis testing. Furthermore, we apply the three tests to a functional gene set of 1747 genes associated with cell proliferation. The samples were taken from normal and tumor tissues of patients with colorectal cancer (Groene *et al*., 2006).

2 METHODS AND DATA

In this section, we first give an illustration of the *RepeatedHighDim* method. Next, we specify the settings of the simulation study and specify the algorithm for detecting the effect causing genes. Finally, the data example of patients with colorectal cancer is presented. All analyses were done with the free software R (version 2.8, http://www.r-project.org). The R-packages 'globaltest', 'GlobalAncova' and 'RepeatedHighDim' were employed for calculating the associated methods in the simulation study and to analyze the cancer data.

2.1 Repeated measures models

The procedure described by Brunner (2009) is based on the idea of Box (1954) who suggested a scaled χ^2-distribution to approximate numerator and denominator of the classical ANOVA statistic. In particular, Brunner (2009) provides two models. The first model is a simple repeated measures model which can be used in the case of microarray data for comparing paired observations (design A), e.g. tumor and mucosa samples from the same patients. The second model is a repeated measures model involving the comparison of two independent groups (design B), e.g. therapy responders with non-responders.

In design A, the differences $D_{kj} = X_{1kj} - X_{2kj}$ between each pair of observations are analyzed, where X_{ikj} denotes the normalized expression level of gene j in the tissue sample i of individual k ($k = 1,...,n$; $j = 1,...,d$; $i = 1,2$). The vector $\boldsymbol{D}_k = (D_{k1},...,D_{kd})'$ is supposed to have a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{S}$. In order to test that there is no group effect, the hypothesis $H_0: \boldsymbol{H\mu} = \mathbf{0}$ can be stated. In the case of design A, the contrast matrix $\boldsymbol{H}$ is given by $\boldsymbol{P}_d = \boldsymbol{I}_d - \frac{1}{d}\boldsymbol{J}_d$ with $\boldsymbol{I}_d$ being the $(d \times d)$ identity matrix and $\boldsymbol{J}_d$ being the $(d \times d)$ matrix with all elements equal to 1. This leads to the following test statistic

$$F_n = \frac{n \cdot \bar{\mathbf{Z}}_{\cdot}' \bar{\mathbf{Z}}_{\cdot}}{tr(\hat{\mathbf{\Sigma}}_n)},$$

where $\bar{\mathbf{Z}}_{\cdot} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{T}\boldsymbol{D}_k$ and $\boldsymbol{T} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{-}\boldsymbol{H}$ and where $\hat{\mathbf{\Sigma}}_n$ denotes the sample covariance matrix of $\mathbf{Z}_k = \boldsymbol{T}\boldsymbol{D}_k$. The symbol $(\cdot)^{-}$ denotes a g-inverse. Box (1954) showed that F_n is approximately F-distributed with f and $(n-1)\cdot f$ degrees of freedom. The parameter f, given by $tr(\mathbf{\Sigma})^2 / tr(\mathbf{\Sigma}^2)$, is estimated by a procedure given in Brunner (2009). With regard to computational aspects and to the high dimension of microarray data, it is fortunately not necessary to calculate the large $(d \times d)$-matrix $\hat{\mathbf{\Sigma}}_n$ in the denominator of F_n as well as in the estimator of f. Instead, one can transform the $(d \times n)$ data matrix $\boldsymbol{D}'$ to a matrix $\tilde{\boldsymbol{D}} = \boldsymbol{P}_n\boldsymbol{D}$, where $\boldsymbol{P}_n = \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{J}_n$. It follows by some routine algebra that the trace of the $(n \times n)$-matrix $\frac{1}{n-1}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{D}}'$ equals the trace of the $(d \times d)$-matrix $\hat{\mathbf{\Sigma}}_n$. Also the quantities $tr(\mathbf{\Sigma})^2$ and $tr(\mathbf{\Sigma}^2)$ are estimated from certain $(n \times n)$-matrices (Brunner, 2009).

For the case of two independent groups (design B), the data can be denoted by $\boldsymbol{X}_{ik} = (X_{ik1},..,X_{ikd})' \sim N(\boldsymbol{\mu}_i, \boldsymbol{S}_i)$, $i = 1,2$; $k = 1,...,n_i$. Here, $\boldsymbol{\mu}_i = E(\boldsymbol{X}_{ik})$ and $\boldsymbol{S}_i = Cov(\boldsymbol{X}_{ik})$. Thus, unbalanced settings as well as unequal covariance matrices are allowed in this model. To compare the two groups with regard to a global group effect, one tests the hypothesis $H_0: \boldsymbol{H}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{0}$ by using the test statistic

$$F_N = \frac{(\bar{\mathbf{Z}}_{1\cdot} - \bar{\mathbf{Z}}_{2\cdot})'(\bar{\mathbf{Z}}_{1\cdot} - \bar{\mathbf{Z}}_{2\cdot})}{tr(\hat{\boldsymbol{S}}_N)},$$

where $\bar{\mathbf{Z}}_{i\cdot} = \frac{1}{n_i}\sum_{k=1}^{n_i} \boldsymbol{T}\boldsymbol{X}_{ik}$ and with $\hat{\boldsymbol{S}}_N = \frac{1}{n_1}\hat{\mathbf{\Sigma}}_{\mathbf{1}} + \frac{1}{n_2}\hat{\mathbf{\Sigma}}_{\mathbf{2}}$ being an estimate of $Cov(\bar{\mathbf{Z}}_{1\cdot} - \bar{\mathbf{Z}}_{2\cdot})$. Here, $\hat{\mathbf{\Sigma}}_i = \frac{1}{n_i - 1}\sum_{k=1}^{n_i}(\mathbf{Z}_{ik} - \bar{\mathbf{Z}}_{i\cdot})(\mathbf{Z}_{ik} - \bar{\mathbf{Z}}_{i\cdot})'$ is an estimate of the covariance matrix of $\mathbf{Z}_{ik} = \boldsymbol{T}\boldsymbol{X}_{ik}$, where $\boldsymbol{T} = \boldsymbol{H}'(\boldsymbol{H}\boldsymbol{H}')^{-}\boldsymbol{H}$. The hypothesis matrix $\boldsymbol{H}$ has the same form as in design A. Under H_0, the statistic F_N has, approximately, an F-distribution with f and f_0 degrees of freedom, where

$$f = \frac{tr(\boldsymbol{S}_N)^2}{tr(\boldsymbol{S}_N^2)}$$

and

$$f_0 = \frac{tr(\boldsymbol{S}_N)^2}{\sum_{i=1}^{2} tr(\mathbf{\Sigma}_i^2)/[n_i^2(n_i - 1)]}.$$

The technical details on the estimation of the traces in F_N, f and f_0 can be found in Brunner (2009).

Besides the above two designs, other, more general hypotheses (with different contrast matrices $\boldsymbol{H}$) can be tested with this method. In addition, this method is also valid for any arbitrary positive definite covariance matrix (Ahmad *et al*., 2008).

2.2 Simulation study

At first, we applied the four global test approaches on simulated gene expression data with d genes and sample sizes $n_1 = n_2 = 20$ for the two groups. The expression levels were drawn from multivariate normal distributions with mean vectors $\mu_1 = \mathbf{0}$ for group 1 and μ_2 for group 2, where the

entries of μ_2 decreased evenly from δ to 0 in order to represent different effects. Thus, the group differences stored in μ_2 were $\{\frac{d}{d}\delta, \frac{d-1}{d}\delta, ..., \frac{1}{d}\delta, 0\}$. For computational simplicity, the effects represented by μ_2 were generated in a decreasing order, just to simulate that there are effects of different sizes, where the order is of no further relevance. The parameter δ can be seen as the common $\log_2$ fold change ($\delta = 1$ corresponds to a 2-fold increase of expression levels). Under H_0, where $\delta = 0$, the mean vectors are the same.

Three different types of covariance matrices **S** were simulated: (i) a block structure, (ii) an autoregressive structure and (iii) an unstructured one. The block-type covariance matrices were constructed such that in each block of 20, the genes had a covariance of 0.5 and all other genes had a covariance of zero. In the autoregressive covariance matrix, the covariance for genes j and j' ($j,j' = 1,...,d; j \neq j'$) was constructed by $0.5^{\lfloor((|j-j'|-1)/4)+1\rfloor}$. To generate the elements for the unstructured covariance matrices, a sample from the standard normal distribution was drawn. The absolute values of this sample were taken as covariances. In order to account for different variances of the genes, the diagonal elements of **S** in group 1 were increased evenly from 1 to 2. In group 2, the variances increased from 1 to $2+\delta$. Thus, both groups had the same variances under H_0, while group 2 had larger variances under the alternative.

The above setting was also performed with d = 200 and d = 1000 genes, respectively, to assess the impact of the dimensionality onto the power and the attained level.

As an extreme example, we further simulated groups of different samples sizes, i.e. with n_1 = 20 and n_2 = 40, or vice versa. Thus, in one case, the larger group had the smaller variances, while in the other case, the smaller group had also the smaller variances.

All simulations were performed with 1000 runs and the number of rejected null hypothesis was counted for assessing the power in each run. Null hypothesis was rejected at a significance level of $\alpha = 5\%$.

2.3 Detection of potentially effect-causing genes

When having detected a group effect, it is interesting to know which of the genes are responsible for this effect. We, therefore, propose the following algorithm for detecting these genes. Assume the columns of the data matrix X represent samples and the rows represent the genes.

(1) For $j = 1,\ldots,d$: calculate the test statistic of the desired global test procedure when gene j is omitted. (Omitting a gene with no effect would not change the value of the test statistic much, while omitting a gene with a large effect will diminish the value of the test statistic.)

(2) Order the rows of the data matrix by increasing size of the test statistic associated with each omitted row.

(3) For $j = 1,\ldots,d'$ ($d' \leq d$): calculate again the global test and remove thereafter the top row of the data matrix.

The algorithm stops in step three when the global test fails to be significant, i.e. when all effect-causing genes have been removed from the data matrix.

We evaluated the performance of this algorithm in combination with *RepeatedHighDim* again in simulation studies. In each of 100 simulation runs, the data sets consisted of expression levels of $d = 200$ genes in 20 samples per group (i.e. $n_1 = n_2 = 20$). Expression levels were drawn from multivariate normal distributions with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and the same covariance matrix $\boldsymbol{S}$ in both groups. In particular, the above detailed block-type matrix was employed here. The mean vector for group 1 was $\boldsymbol{\mu}_1 = \mathbf{0}$ in all datasets, while $\boldsymbol{\mu}_2$ represented different alterations in group 2. In particular, we altered either 2.5 or 75% randomly chosen genes by a log fold change of δ. We use the above given procedure as well as a classical gene-wise testing procedure to detect the altered genes. The classical procedure was the Empirical Bayes (EB) method (Smyth, 2004) combined with a P-value adjustment by the method of Benjamini and Yekuteli (2001) to control the false discovery rate (FDR). We evaluate the performance of the procedures by determining the portion of true positive findings (called the average power rate) as well as the portion of false positive findings (i.e. FDR).

2.4 Data example

Three of the global tests (*GlobalTest*, *GlobalAncova* and *RepeatedHighDim*) and the algorithm for the detection of the effect-causing genes were applied to a functional gene group of 1747 genes associated with cell proliferation. The selection algorithm was performed in combination with each of the three global tests. The samples were taken from normal and tumor tissues of 12 patients with colorectal cancer (Groene *et al*., 2006). Groene *et al.* extracted pathway information from the NCBI database (http://www.ncbi.nlm.nih.gov/) and the Affymetrix web site (http://www.affymetrix.com/). The dataset is available within the 'GlobalAncova' package for R. Before the actual analysis, expression levels were log2-transformed and normalized by the quantile method (Bolstad *et al*., 2003).

3 RESULTS

3.1 Simulated level

In all simulation settings, a similar scheme regarding the maintenance of the prespecified significance level could be observed (Table 1). In detail, the asymptotic results of *GlobalAncova* and *GlobalTest* were too liberal and too conservative, respectively, when block-type and autoregressive covariance matrices were simulated. In addition, the failure of these two asymptotic tests became even more pronounced when the dimension was increased, i.e. when a higher number of genes was simulated. Under the unstructured covariance matrices, the asymptotic results of the two approaches maintained the prespecified level of 5% clearly better. *GlobalAncova* and *GlobalTest* maintained the level also very well when their permutation result was used. *RepeatedHighDim* behaved very well, too, yielding simulated levels between 41% and 68%. In all simulations, the *ToTS* procedure was much to liberal yielding simulated levels about 45% for block-type and autoregressive covariance matrices and between 83.6% and 93.9% for unstructured covariance matrices. The different settings of sample sizes did not seem to have a relevant effect on the attained level. In particular, no difference could be observed between the cases of balanced and unbalanced sample sizes.

Because *ToTS* did not maintain the prespecified level, it was excluded from the further analyses.

3.2 Power assessment

Like in the level simulations, the results of the power study were quite similar across the different settings. When looking at the asymptotic results of *GlobalTest* and *GlobalAncova*, the latter one yielded the higher power. The power of *RepeatedHighDim* was always between those of the two competitors (Fig. 1). If, instead, the permutation results of *GlobalTest* and *GlobalAncova* were used, the power curves of these two approaches shifted approximately to that of *RepeatedHighDim* (Fig. 2). We observed this tendency in all simulation settings.

In all settings, the largest power was achieved when the block-type or the autoregressive matrices were simulated. The power resulting under unstructured covariance matrices was clearly smaller (Fig. 3). However, the behavior of the three global tests to each other is nearly the same as described in the first paragraph under the different covariance matrices.

Table 1. Simulated levels for the four test procedures in different simulation settings ($N = 1000$ simulations, samples sizes n_1 and n_2, number of genes d, prespecified significance level: 0.05)

Covariance matrix	Sample sizes	Number of genes	*GlobalAncova*		*GlobalTest*		*Repeated-HighDim*	*ToTS*
			asymp.	perm.	asymp.	perm.		
Block structure	$n_1 = n_2 = 20$	d = 200	0.131	0.042	0.015	0.047	0.053	0.445
		d = 1000	0.216	0.061	0.000	0.057	0.054	0.448
	$n_1 = 20, n_2 = 40$	d = 200	0.111	0.049	0.017	0.056	0.048	0.481
		d = 1000	0.154	0.044	0.000	0.047	0.056	0.472
	$n_1 = 40, n_2 = 20$	d = 200	0.106	0.042	0.011	0.048	0.067	0.471
		d = 1000	0.184	0.054	0.000	0.057	0.054	0.455
Autoregressive	$n_1 = n_2 = 20$	d = 200	0.117	0.041	0.007	0.043	0.058	0.445
		d = 1000	0.183	0.036	0.000	0.048	0.051	0.454
	$n_1 = 20, n_2 = 40$	d = 200	0.121	0.055	0.011	0.058	0.055	0.443
		d = 1000	0.175	0.048	0.000	0.059	0.060	0.448
	$n_1 = 40, n_2 = 20$	d = 200	0.110	0.052	0.004	0.058	0.045	0.433
		d = 1000	0.147	0.047	0.000	0.049	0.059	0.464
Unstructured	$n_1 = n_2 = 20$	d = 200	0.068	0.047	0.049	0.051	0.056	0.840
		d = 1000	0.069	0.040	0.038	0.043	0.060	0.937
	$n_1 = 20, n_2 = 40$	d = 200	0.059	0.050	0.051	0.049	0.064	0.836
		d = 1000	0.047	0.036	0.037	0.041	0.063	0.939
	$n_1 = 40, n_2 = 20$	d = 200	0.065	0.044	0.051	0.055	0.068	0.835
		d = 1000	0.061	0.049	0.053	0.056	0.062	0.937

In the case of unequal sample sizes, either the smaller or the larger group has bigger variances.

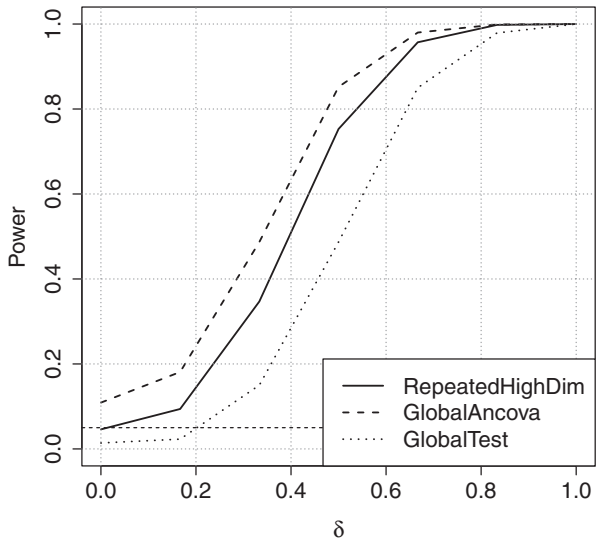

Fig. 1. Power curves for the asymptotic results of *GlobalTest* and *GlobalAncova* in comparison to the power curve of *RepeatedHighDim*. The figure shows the results of the simulation setting with $d = 200$ genes, equal sample sizes $n_1 = n_2 = 20$ and block-type covariance matrices. The largest power is achieved by *GlobalAncova* with the disadvantage of exceeding the prespecified level of 5%. The smallest power shows *GlobalTest* which is also too liberal.

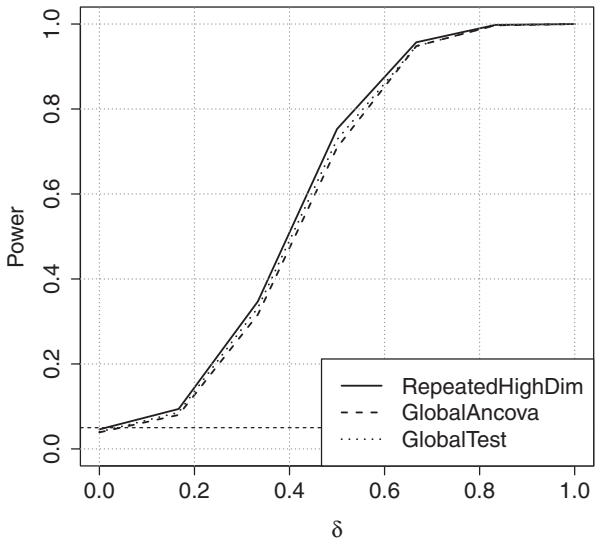

Fig. 2. Power curves for the permutation results of *GlobalTest* and *GlobalAncova* in comparison to the power curve of *RepeatedHighDim*. The figure shows the results of the simulation setting with d = 200 genes, equal sample sizes $n_1 = n_2 = 20$ and block-type covariance matrices. The permutation tests shift the power of *GlobalTest* and *GlobalAncova* nearly to that of *RepeatedHighDim*.

The effect of the dimension d and of the samples sizes n_1 and n_2 was not considerable. In fact, the power differences between $d = 200$ and $d = 1000$ were not very large. Likewise, it did not make a large difference, whether sample sizes were balanced or unbalanced. Again, the relationship of the three global tests to each other was not considerably affected by the simulated dimension.

3.3 Evaluation of detection algorithm

We applied the algorithm for the detection of effect-causing genes in combination with *RepeatedHighDim* as well as a classical procedure of gene-wise testing on the artificial datasets described in Section 2.4. In addition, we combined our new algorithm with the conservative asymptotic version of *GlobalTest*. In the case that

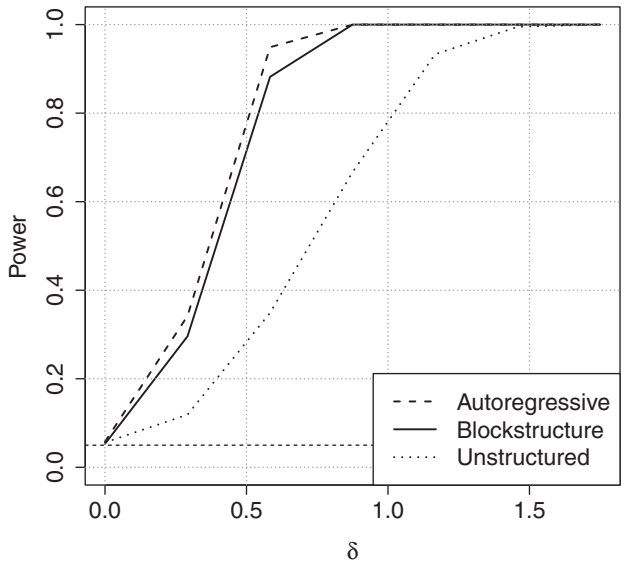

Fig. 3. Power curves for *RepeatedHighDim* under different covariance matrices. The figure shows the results of the simulation setting with d = 200 genes and equal sample sizes $n_1 = n_2 = 20$.

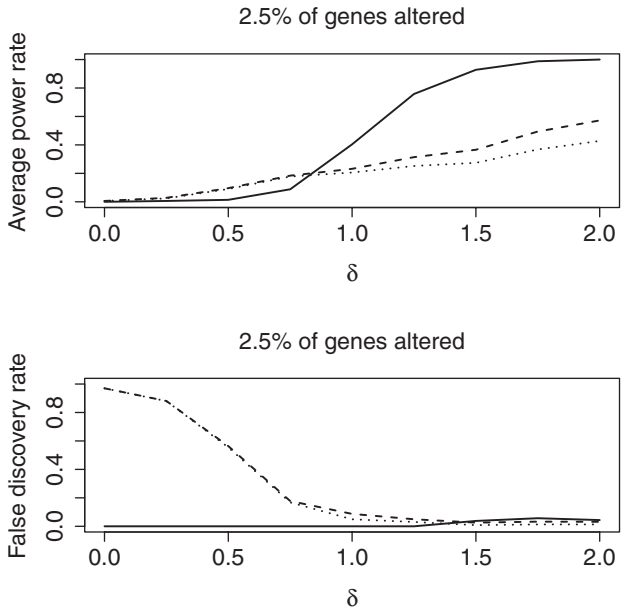

Fig. 4. Average power rate and FDR of a gene-wise testing procedure (solid line) and of our new procedure in combination with *RepeatedHighDim* (dashed lined) or the asymptotic *GlobalTest* (dotted line) for detecting genes that potentially cause a group effect.

only a small number of genes was altered between the two groups and when the effects δ were large, the classical gene-wise testing yielded a higher average power rate than our new procedure (Fig. 4). For small effects, however, the average power rate of our procedure outperformed that of the classical testing procedure. This gain in power for our procedure is, however, accompanied by a very high FDR, while the classical procedure maintains a prespecified FDR of 5%. Our algorithm in combination with *RepeatedHighDim* has a little bit larger power rate than the combination with *GlobalTest*, but with both method nearly the same FDR is observed.

When the number of altered genes was increased, our new procedure nearly approached the average power rate of the gene-wise testing procedure for larger effects δ also (Fig. 5). Again, the higher power of our procedure is connected to a high FDR. However, there is a small window between $\delta = 0.5$ and $\delta = 1.0$, where our procedure outperforms the classical gene-wise testing and maintains a prespecified FDR of 5% as well. Again, we could not observe a substantial difference in power rate and error rate between the combination with *RepeatedHighDim* and that with *GlobalTest*.

3.4 Analysis of data example

As a proof of concept, we applied the different methods to a real microarray dataset including tumor and normal tissue samples. For the group tests, we used a set of 1747 genes connected to cell proliferation. Since cell proliferation genes are usually highly connected to cancer tissue in comparison to normal tissue and should therefore yield strong positive effects. Indeed, applying *GlobalTest* (permutation approach), *GlobalAncova* (permutation approach) and *RepeatedHighDim* to the colorectal data, a highly significant group effect was detected between tumour and normal samples. Each test yielded $P < 0.01$. When searching for potentially effect-causing genes, the EB method with FDR-adjusted P-values yielded 199 significant genes, whereas the new detection algorithm identified considerably more effect-causing genes. In detail, the algorithm

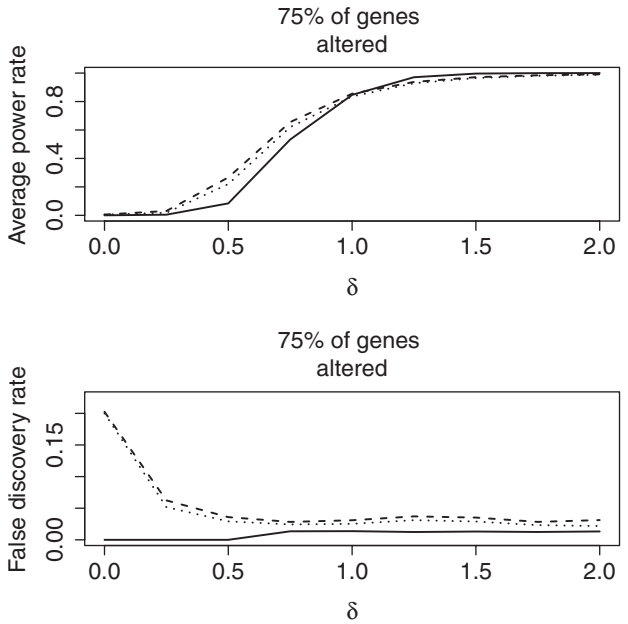

Fig. 5. Average power rate and FDR of a gene-wise testing procedure (solid line) and of our new procedure in combination with *RepeatedHighDim* (dashed lined) or the asymptotic *GlobalTest* (dotted line) for detecting genes that potentially cause a group effect.

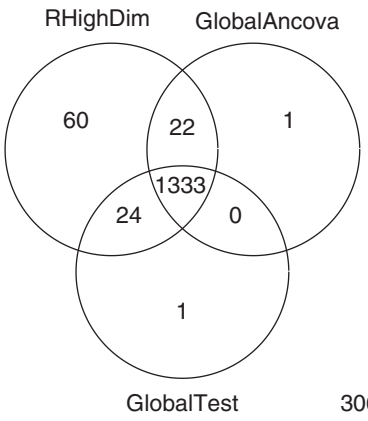

Fig. 6. Intersections of potentially effect-causing genes in a cancer dataset detected with the new selection algorithm in combination with *RepeatedHighDim*, *GlobalAncova* or *GlobalTest*. A classical multiple testing procedure detected 199 genes, nearly all of them were among these three lists.

detected 1439 genes in combination with *RepeatedHighDim*, 1356 with *GlobalAncova* and 1358 with *GlobalTest*. The intersection of these three lists were 1333 genes (Fig. 6). Nearly all (i.e. 196, 196 and 194, respectively) of the 199 genes resulted from the EB method were included in the other result lists.

3.5 Computation times

In order to compare the computation times between *GlobalTest*, *GlobalAncova* and *RepeatedHighDim*, each approach was applied 100 times on the same artificial dataset. The resulting computation times were as follows. *RepeatedHighDim*: 3 s, *GlobalAncova*: 29 s (∼42 s), *GlobalTest*: 20 s (∼7 s). *RepeatedHighDim* shows thus the fastest computation time. The calculations were done on a Intel Core2 CPU with 2.1 GHz and 0.99 GB RAM.

4 DISCUSSION

We have compared the attained level, the power and computation times of four different test procedures for global group comparisons in microarray experiments in the case of functional gene sets. Noticeable differences in the attained level and power were especially observed when the asymptotic results of *GlobalAncova* and *GlobalTest* were compared to *RepeatedHighDim*. Within this setting, *GlobalTest* becomes sometimes too conservative and *GlobalAncova* too liberal, whereas *RepeatedHighDim* maintains the preassigned significance level accurately. The permutation results of *GlobalAncova* and *GlobalTest* maintained the preassigned level very well, too. The *ToTS* approach clearly exceeded the prespecified significance level in all simulation settings. Here, we showed only the results of *ToTS* when applying the one-sample *t*-test on the set of test statistics. However, the simulated levels were very similar when using the one-sample Wilcoxon test.

We have additionally studied some situations with non-normally distributed data and found that the *RepeatedHighDim* procedure is quite robust against deviations from the normal distribution, i.e. with regard to maintenance of the preassigned significance level (see Supplementary Material).

Accordingly to the level simulations, the power was largest for the asymptotic result of *GlobalAncova* and smallest for the asymptotic result of *GlobalTest*. When using their permutation results, the power curves of these two methods nearly approached that of *RepeatedHighDim*. Differences in power were also observed for the different types of covariance matrices but neither for different numbers d of genes nor for the different numbers of samples sizes.

The analysis of computation times of the latter three methods yielded that *RepeatedHighDim* is much faster than the permutation or asymptotic tests performed with *GlobalAncova* and *GlobalTest*. This plays a particular role in the proposed detection algorithm where a global test is carried out up to $2 \times d$ times. Under this regard, and with the results of Section 3.5, one can conclude that the selection algorithm takes about 1–2 min in combination with *RepeatedHighDim* and several more minutes in combination with the two other global test approaches, depending on the size d of the gene set. The asymptotic result of *GlobalTest* is also obtained very fast but has the drawback of a rather low power. Therefore, *RepeatedHighDim* is preferable for being combined with the proposed algorithm. This speed comparison is of course only based on the available software. There may be faster implementations possible for each method.

In principle, the introduced algorithm for the detection of effect-causing genes can be applied with all three testing approaches. Although, the average power rate of this new algorithm sometimes outperforms that of classical multiple testing, it has the drawback of too high FDR levels in these cases. We therefore intent, to work further on the improvement of this algorithm. With regard to the power rate of our selection algorithm, one can further argue that researchers perhaps tend to take adjusted *P*-values for the classical multiple testing approach from the whole dataset and not just from the functional set. In that case, our algorithm would even have a larger power rate in other situations than in those observed in our simulations. An alternative idea was given by Srivastava and Kubokawa (2008) who use Akaike's information criterion to select components of the mean vector in high-dimensional data.

In this article, we have shown the applicability of the approach proposed by Brunner (2009) to microarray data in the two-group setting. *GlobalAncova* and *GlobalTest* provide also the possibility to test other experimental factors. In general, study designs with more than one group factor on two levels can also be tested with *RepeatedHighDim* by constructing an appropriate contrast matrix $\boldsymbol{H}$.

While in this article we focused on the comparison of the new models by Brunner (2009) with three of the most popular global test methods, a certain number of other global test approaches exists, which were not included to our simulations. Dinu *et al*. (2007) for example proposed the *SAM-GS* method which was also compared to *GlobalAncova* and *GlobalTest* by Liu *et al*. (2007) with similar results for the latter two approaches than ours. Another approach, based on Hotelling's T^2-statistic, was proposed by Lu *et al*. (2005).

ACKNOWLEDGEMENTS

We thank Reinhard Meister (University of Applied Science, Berlin) for helpful remarks on the *GlobalAncova* procedure and Andreas Leha (Department of Medical Statistics, University Medical Center Göttingen) for his help on performing the simulations. We also thank a former associate editor for the reference of the *ToTS* method.

Funding: Deutsche Forschungsgemeinschaft (KFO 179).

Conflict of Interest: none declared.

REFERENCES

Ahmad,R. *et al*. (2008) Analysis of high dimensional repeated measures designs: the one sample case. *Comput. Stat. Data Anal.*, **53**, 416–427.

Beißbarth,T. and Speed,T. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

Benjamini,Y. and Yekuteli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Bolstad,B.M. *et al*. (2003) A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, **119**, 185–193.

Box,G.E.P. (1954) Some theorems on quadratic forms applied in the study of analysis if variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Ann. Math. Stat.*, **25**, 484–498.

Brunner,E. (2009) Repeated measures under non-sphericity. In *Proceedings of the 6th St. Petersburg Workshop on Simulation*, VVM.com Ltd., pp. 605–609.

Draghici,S. (2007) A systems biology approach for pathway level analysis. *Genome Res.*, **17** , 1537–1545.

Dinu,I. *et al*. (2007) Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics*, **8** , 242.

Dudoit,S. *et al*. (2003) Multiple hypothesis testing in microarray experiments. *Bioinformatics*, **18**, 71–103.

Goeman,J.J. *et al*. (2004) A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, **20**, 93–99.

Goeman,J.J. and Bühlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Groene,J. *et al*. (2006) Transcriptional census of 36 microdissected colorectal cancers yields a gene signature to distinguish UICC II and III. *Int. J. Cancer*, **119**, 1829–1836.

Hummel,M. *et al*. (2008) GlobalANCOVA: exploration and as-sessment of gene group effects. *Bioinformatics*, **24**, 78–85.

Liu,Q. *et al*. (2007) Comparative evaluation of gene-set analysis methods. *BMC Bioinformatics*, **8** , 431.

Lu,Y. *et al*. (2005) Hotelling's T^2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, **21**, 3105–3113.

Mansmann,U. and Meister,R. (2005) Testing differential gene expression in functional groups. *Methods Inf. Med.*, **44**, 449–453.

Mathai,A.M. and Provost,S.B. (1992) *Quadratic Forms in Random Variables*. Marcel Dekker Inc., New York, pp. 25–37.

Meinshausen,N. (2008) Hierachical testing of variable importance. *Biometrika*, **95**, 265–278.

Mootha,V.K. *et al*. (2003) PGC-1α responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Pavlidis,P. *et al*. (2004) Using the Gene Ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.*, **29**, 1213–1222.

Rocke,D.M. *et al*. (2005) A method for detection of differential gene expression in the presence of inter-individual variability in response. *Bioinformatics*, **21**, 3990–3992.

Simon,R.M. *et al*. (2003) Global tests of gene expression differences between classes. In Simon,R.M. *et al*. (eds) *Design and Analysis of DNA Microarray Investigations*. Springer, New York, pp. 86–88.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 3.

Srivastava,M.S. and Kubokawa,T. (2008) Akaike information criterion for selecting components of the mean vector in high dimensional data with fewer observations. *J. Japan Stat. Soc.*, **38**, 259–283.